COMPRENSIÓN DE LOS INTERVALOS DE CONFIANZA EN ESTUDIANTES DE EDUCACIÓN SUPERIOR

CARLOS MANUEL SARMIENTO SOTO
WOLFANG ALEXANDER OSMA CASTELLANOS

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
BUCARAMANGA
2010

COMPRENSIÓN DE LOS INTERVALOS DE CONFIANZA EN ESTUDIANTES DE EDUCACIÓN SUPERIOR

CARLOS MANUEL SARMIENTO SOTO
WOLFANG ALEXANDER OSMA CASTELLANOS

Trabajo de Tesis para optar el grado académico de Licenciado en Matemáticas

Director

Dr. GABRIEL YÁÑEZ CANAL

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
BUCARAMANGA
2010

DEDICATORIA

A Dios por darnos la oportunidad de ser personas de bien y permitirnos formarnos académicamente como profesionales.

A nuestros padres por su paciencia y dedicación a lo largo de nuestras vidas.

A mi hermano Giovanni Sarmiento Soto por su apoyo y solidaridad.

A mi hijo Sebastián Osma quien ha sido mi razón de vivir y mi fuente de inspiración.

AGRADECIMIENTOS

Esta tesis, si bien ha requerido de nuestro esfuerzo, estudio y dedicación, no hubiese sido posible su realización, sin la participación de personas que nos brindaron su colaboración y apoyo para que este trabajo llegara a feliz término. Para nosotros es muy importante poderles expresar nuestros sinceros agradecimientos:

En primer lugar, al Dr. Gabriel Yáñez Canal por aceptarnos para realizar esta tesis bajo su dirección. Por su paciencia, colaboración, capacidad para orientar nuestras ideas y acompañamiento constante que marcaron el rumbo del presente proyecto de investigación.

Al grupo de doctores: Gabriel Yáñez Canal, Carlos Orozco y Carlos Conde, por admitirnos participar en el seminario del modelo Rasch donde nos dedicaron sus conocimientos.

Al grupo de profesores que nos brindaron su cooperación en el desarrollo de la investigación al permitirnos aplicar el cuestionario a sus estudiantes.

A la Ingeniera Martha Patricia Prada por su colaboración y dedicación en la elaboración del informe final.

Y por supuesto, el agradecimiento más profundo y sentido para nuestras familias, sin su apoyo, colaboración e inspiración habría sido imposible alcanzar nuestras metas. A nuestros padres Antonio María Sarmiento y Alicia Soto, Bleidy

Castellanos, por su amor, ejemplo de lucha, honestidad, y sostén en el trascurrir de nuestras vidas universitaria y profesional.

Alexander:

A mi esposa Clarena, por su dedicación y trabajo, por creer en mí, brindarme todo su amor y estar a mi lado a lo largo de mis estudios y la elaboración del proyecto de grado.

Carlos:

A Giovanni Sarmiento por su apoyo incondicional y generosidad.

A Laura Barragán quien me acompañó en todo momento de mi formación como profesional, ofreciéndome su cariño y amistad.

CONTENIDO

pág.

INTRODUCCIÓN 19

1.INTERVALOS DE CONFIANZA.. 21

1.1. GENERALIDADES 21

1.2. ESTRUCTURA CONCEPTUAL ASOCIADA AL INTERVALO DE……. CONFIANZA 21

1.3. CONSTRUCCIÓN DE INTERVALOS DE CONFIANZA PARA LA MEDIA 24

1.3.1 Construcción con varianza conocida 25

1.3.2 Construcción con varianza desconocida. 28

2. CUESTIONARIO Y COMPONENTES 32

2.1. SELECCIÓN DE ÍTEMS DEL CUESTIONARIO Y ANÁLISIS DE…….. DISTRACTORES 32

2.2. ANÁLISIS DE LAS POSIBLES RESPUESTAS A LOS ITEMS 33

3. LA TEORÍA CLÁSICA DE TEST Y EL MODELO RASCH 45

3.1. TEORIA CLASICA DE TEST 46

3.1.1. Análisis de ítems 47

3.1.1.1. Índice de dificultad $(\mathbf{D})$ 48

3.1.1.2. Índice de discriminación $(\mathbf{H})$. 50

3.1.1.3. Índice de validez.$(\mathbf{V})$. 52

3.1.2. Análisis del test. 52

3.2. EL MODELO RASCH 54

3.2.1. Supuestos básicos para la construcción del modelo 59

3.2.1.1. Unidimensionalidad 59

3.2.1.2 . Independencia local 59

3.2.2. Características del modelo 59

3.2.2.1. Medición conjunta 59

3.2.2.2. Estadísticos Suficientes 60

3.2.2.3. Objetividad Específica 60

3.2.2.4. Propiedades de Intervalo 60

3.2.2.5. Especificidad del error estándar 60

3.2.3. Bases de estimación en el modelo 60

3.2.4. Curva característica de un ítem 75

3.2.5. Método de estimación de parámetros. 77

3.2.6. Estimador de máxima verosimilitud. 77

3.2.7. Ajuste de los datos al modelo. 88

3.2.7.1. Estadísticos de ajuste. 91

3.2.7.2. Criterios 92

4. ESTUDIO DE EVALUACIÓN 95

4.1. MÉTODO CLÁSICO (TCT) 96

4.1.1. ANÁLISIS DE ÍTEMS 96

4.1.1.1. Análisis de ítems de opciones múltiples 96

4.1.1.2. Análisis de ítems abiertos 111

4.2. MODELO RASCH 120

4.2.1. ANÁLISIS DE DATOS POR MEDIO DEL PROGRAMA WINSTEPS® 120

4.2.2. ANÁLISIS DE LOS RESULTADOS 122

4.2.3. ANÁLISIS DE LOS ITEMS POLITÓMICOS DEL TEST 126

4.2.3.1. Análisis del ítem 4. 126

4.2.3.2. Análisis del ítem 7. 128

4.2.4. ANÁLISIS DE LOS ÍTEMS DICOTÓMICOS DEL TEST 130

4.2.4.1. Análisis del ítem 1. 131

4.2.4.2. Análisis del ítem 2. 132

4.2.4.3. Análisis del ítem 3. 134

4.2.4.4. Análisis del ítem 5. 135

4.2.4.5. Análisis del ítem 6. 137

4.2.4.6. Análisis del ítem 8. 138

4.2.4.7. Análisis del ítem 9 140

4.2.4.8. Análisis del ítem 10. 141

4.2.5. ANÁLISIS GLOBAL Y CONJUNTO DE PERSONAS E ÍTEMS 143

4.2.6. ANÁLISIS DE LA ESTIMACIÓN DEL CONJUNTO DE ÍTEMS 148

4.2.6.1. ENTRY NUMBER 148

4.2.6.2. TOTAL SCORE 149

4.2.6.3. COUNT 149

4.2.6.4. MEASURE 149

4.2.6.5. MODEL S.E. 149

4.2.6.6. INFIT MNSQ 149

4.2.6.7. INFIT ZSTD.....151
4.2.6.8. OUTFIT MNSQ 152
4.2.6.9. OUTFIT ZSTD.....153
4.2.6.10. ITEM. 154
4.2.6.11. MEAN.....155
4.2.7. ANÁLISIS DE LA ESTIMACIÓN DEL CONJUNTO DE ESTUDIANTES....155
4.2.7.1. ENTRY NUMBER 158
4.2.7.2. RAW SCORE.....158
4.2.7.3. COUNT. 158
4.2.7.4. MEASURE 159
4.2.7.5. MODEL S.E. 159
4.2.7.6. INFIT MNSQ 160
4.2.7.7. INFIT ZSTD.....162
4.2.7.8. OUTFIT MNSQ 163
4.2.7.9. OUTFIT ZSTD.....165
4.2.7.10. PERSON.....166
4.2.7.11. MEAN.....166
4.2.8. ANÁLISIS DE LOS RESULTADOS PARA CADA UNIVERSIDAD…….. EVALUADA.....166
5. CONCLUSIONES 170
BIBLIOGRAFIA.....178
ANEXOS.....168

LISTA DE TABLAS

pág.

Tabla 1. Matriz de encuestados vs. Ítems. 48

Tabla 2. Resultados de la comparación de Carlos con Alex saltando vallas. ... 65

Tabla 3. Matriz de probabilidad de posibles resultados. 65

Tabla 4. Probabilidad de extracción de una bola blanca. 79

Tabla 5. .Frecuencias y porcentajes de las opciones del ítem 1. 97

Tabla 6. Frecuencias y porcentajes de las opciones del ítem 2. 98

Tabla 7. Frecuencias y porcentajes de las opciones del ítem 3. 100

Tabla 8. Frecuencias y porcentajes en las opciones del ítem 5. 102

Tabla 9. Frecuencias y porcentajes en las opciones del ítem 6. 104

Tabla 10. Frecuencias y porcentajes en las opciones del ítem 8. 105

Tabla 11. Frecuencias y porcentajes de las opciones del ítem 9. 107

Tabla 12. Frecuencias y porcentajes de las opciones del ítem 10. 109

Tabla 13. Valores de estimación de los ítems. 148

Tabla 14.Resultados de los estudiantes..... 158

LISTA DE CUADROS

pág.

Cuadro 1. Comparativo de los resultados en los ítems de opción múltiple..... 111

Cuadro 2. Ejemplos de respuestas en el ítem 4. 113

Cuadro 3. Frecuencias y porcentajes a las clases de respuesta del ítem 4. .. 113

Cuadro 4. Ejemplos de respuestas en el ítem 7. 115

Cuadro 5. Frecuencias y porcentajes a las clases de respuestas del ítem 7.. 115

Cuadro 6 . Comparativo de los resultados en los ítems abiertos. 117

Cuadro 7. Resultados en los índices de dificultad y discriminación de los…… ítems..... 118

LISTA DE FIGURAS

pág.

Figura 1. Términos matemáticos ligados al intervalo de confianza (Olivo, 2008) ..24

Figura 2. Distribución normal estándar. 26

Figura 3. Intervalo de confianza del 95% para una muestra. 28

Figura 4. La distribución t- student. 30

Figura 5. Interpretación del nivel de confianza en el intervalo para la media de una distribución normal. 32

Figura 6. Matriz de respuestas a un test, índice de dificultad. 49

Figura 7. Matriz de respuestas a un test, índice de discriminación. 51

Figura 8. Saltadores buscando superar las valla. 61

Figura 9. Matriz donde se observan saltadores y los resultados con cada una de las valla. 62

Figura 10. Matriz de probabilidades de los saltadores con cada una de las valla. 64

Figura 11. Curva característica de un ítem en el modelo Rasch. 75

Figura 12. Curvas características de cuatro ítems. 77

LISTA DE ANEXOS

pág.

Anexo A. Procedimiento para ejecutar el programa Winsteps®……..168

RESUMEN

TITULO: Comprensión de los intervalos de confianza en estudiantyes de educación superior.[*]

AUTORES: SARMIENTO SOTO, Carlos Manuel; OSMA CASTELLANOS, Wolfang alexander. [**]

PALABRAS CLAVES: Intervalos de confianza, Ítems, Teoría Clásica de Test, Modelo Rasch.

RESUMEN:

Se presentan en este trabajo los resultados obtenidos en la investigación sobre la comprensión de conceptos que una muestra de estudiantes de dos universidades públicas del país tienen sobre los intervalos de confianza. Para estudiar estas concepciones, se aplicó un cuestionario cuyos ítems se seleccionaron de la prueba de Olivo en su tesis doctoral y que fue respondido voluntariamente por 164 estudiantes. Las respuestas observadas se analizaron y compararon utilizando la teoría clásica del test y el modelo Rasch con el apoyo del programa Winsteps® demostrando en los dos casos que las posiciones dadas a los ítems según el grado de dificultad guardan concordancia en la mayoría de ellas y que los estudiantes a los que se les aplicó la prueba tienen un bajo dominio en el tema de los intervalos de confianza. Los ítems abiertos y los que requieren análisis de gráficos fueron los de mayor dificultad en la prueba, concluyendo que los estudiantes prefieren contestar preguntas de selección múltiple que aquellas abiertas que requieren cálculos y mayor discernimiento.

Los estudiantes presentan grandes complicaciones en el tema con algunas preguntas y pocos individuos demuestran responder con facilidad el examen. En general los resultados muestran que la prueba es difícil y exige una mayor habilidad de los participantes, pues los resultados que arrojan los modelos con los que se estudio el test concluyen con argumento lo anteriormente dicho.

[*] Trabajo de grado

[**] Facultad De Ciencias. Escuela de Matemáticas, Director de Tesis YÁÑEZ CANAL Gabriel.

ABSTRACTS

TITIE: Understanding the confidence interval of higher education students.[*]

AUTHORS: SARMIENTO SOTO, Carlos Manuel; OSMA CASTELLANOS, Wolfang alexander. [**]

KEYWORDS: Confidence intervals, Items, Classical Test Theory, Rasch Model.

ABSTRACT:

This paper presents the results obtained in research on understanding concepts, from a sample of students from two public universities in the country, about the confidence intervals. To study these concepts, a questionnaire was used whose items were selected from Olivo test in his doctoral thesis which was answered by 164 students voluntarily. The observed responses were analyzed and compared using the classical test theory and Rasch model supported by Winsteps® program. Thus, signifying in both cases that the positions given to those items, according to the degree of difficulty, keep consistent in most of them and that students, who were administered the test, have low domain in the issue of confidence intervals. The open items requiring analysis and graphics were the most difficult on the test, concluding that students prefer to answer multiple choice questions to open ones that require calculations and greater insight.

Students have major complications in the topic with some questions and few individuals have been shown to easily meet the test. In general, the results show that the test is difficult and requires greater skill from the participants, as the results show the models to the study, the test is concluded with the above argument.

[*] Degree work

[**] Faculty of science. Mathematics School, Thesis Director YÁÑEZ CANAL Gabriel.

INTRODUCCIÓN

El presente trabajo busca analizar la comprensión de los conceptos sobre Intervalos de Confianza en los estudiantes universitarios, ampliando la investigación a la aplicación de diferentes modelos de análisis de resultados, y creando una herramienta que sirva de fuente de investigación para futuros trabajos relacionados con el tema.

El inicio de esta investigación se fundamenta en el estudio de los textos de Bond & Fox (2007) y Smith & Smith (2004) usando como metodología la aplicación de los conocimientos adquiridos en el seminario sobre el Modelo Rasch, realizado por un grupo interdisciplinario de docentes de la Universidad Industrial de Santander, en el cual desarrollaron los conceptos, manejos, aplicaciones, comparaciones e importancia de este modelo. Posteriormente se aplicó el instrumento de recolección de datos en dos universidades públicas del país, la Universidad Industrial de Santander (UIS) y La Universidad Francisco de Paula Santander (UFPS) a estudiantes de diferentes programas académicos teniendo como requisito estar o haber cursado en el pénsum la materia de estadística.

Este trabajo de investigación tiene como objetivo central indagar sobre la ***"****Comprensión de los intervalos de confianza en estudiantes de educación superior"* examinando las diferencias y similitudes entre el análisis clásico de datos y el análisis usando el modelo Rasch.

Este estudio está organizado en cinco capítulos, cuyos contenidos se describen brevemente a continuación.

En el primer capítulo se presenta un resumen sobre los aspectos teóricos más relevantes de los intervalos de confianza, así como algunos elementos de la estructura conceptual y la construcción de intervalos de confianza para la media.

El segundo capítulo hace referencia al cuestionario utilizado para este estudio y sus componentes. Está dividido en dos apartados, en el primero se detalla la selección de los ítems y análisis de los distractores, y en el segundo se hace una exploración de las posibles respuestas a los ítems.

En el tercer capítulo se describen la teoría clásica del test y el modelo Rasch. También dividido en dos apartados diferentes, el primero presenta un breve resumen sobre el análisis clásico, sus componentes y modo de empleo. El segundo apartado se refiere al modelo Rasch, se hace un análisis mostrando sus componentes, fundamentos y desarrollo matemático, aplicaciones y aspectos importantes que lo convierte en una herramienta confiable para ser aplicada en diferentes áreas.

El capítulo cuarto registra los resultados obtenidos en la aplicación del cuestionario a 164 estudiantes de dos universidades públicas del país, utilizando la teoría clásica del test y la aplicación del modelo Rasch con el apoyo del programa Winsteps®.

En el quinto y último capítulo se presentan las conclusiones de esta investigación, precisando algunos resultados importantes observados en el transcurso de este estudio y que pueden generar interés en la continuación de futuras investigaciones en el tema.

Al final del trabajo aparecen las referencias citadas y los anexos que dan cuenta del estudio realizado.

1. INTERVALOS DE CONFIANZA

1.1 GENERALIDADES

Siempre que se tienen datos de una muestra que hace parte de una población y se busca estimar el valor de un parámetro de ésta, los intervalos de confianza aparecen en escena.

Para la construcción de estos intervalos de confianza se requiere, con base en los datos de la muestra, obtener un estimador puntual del parámetro a estimar y conocer su distribución de probabilidad para poder garantizar con cierta confianza que el valor del parámetro se encuentra en el intervalo hallado. Más detalles acerca de la construcción de los intervalos se describen en el próximo apartado.

1.2 ESTRUCTURA CONCEPTUAL ASOCIADA AL INTERVALO DE CONFIANZA

Utilizando el sentido común, para estimar un valor que por su naturaleza es de carácter continuo, es más útil pensar en un intervalo que en un valor puntual. Es decir, calcular un rango donde posiblemente se encontrará el valor del parámetro, acompañado de un nivel de confianza que se establece de antemano. “Llamamos intervalo de confianza al intervalo que con un cierto nivel de confianza, contiene al parámetro que se está estimando” (Olivo, 2008).

¿Cómo construir dicho intervalo valiéndose de los datos que suministra una muestra aleatoria tomada de una población?

Para estar más familiarizados con las herramientas utilizadas para el desarrollo y entendimiento del intervalo de confianza, se debe conocer el significado de los siguientes términos, que intervienen en la construcción y definición del intervalo y que se relacionan entre sí como se muestra en la Figura 1.

- *Población*: el conjunto de personas, objetos o mediciones que poseen características bien definidas y conforman el grupo a estudiar.

- *Muestra:* subconjunto o grupo de la población.

- *Tamaño de la muestra* (n): representa el número de sujetos que conforman la muestra a analizar.

- *Parámetro*: constante y desconocido, es la propiedad numérica de una población que determina la distribución de la variable.

- *Estadístico:* aleatorio y conocido, es la propiedad numérica de la distribución de frecuencias de la muestra. Es un estimador del parámetro.

- *Variable aleatoria:* es la variable que se estudia en la población y tiene una distribución de probabilidad que depende de uno o varios parámetros.

- *Variable estadística:* es el conjunto de valores en la variable aleatoria de la muestra, tiene distribución de frecuencias.

- *Distribución muestral*: es la distribución de probabilidad del estadístico considerado como variable aleatoria.

- *Valor esperado*: es el promedio de todos los valores de una variable aleatoria.

- *Dispersión:* indica en promedio la variabilidad de los datos respecto a la media.

- *Nivel de confianza*: valor de probabilidad que se fija para construir el intervalo.

- *Intervalo de confianza:* intervalo aleatorio que contiene el parámetro que se está estimando con un cierto nivel de confianza.

- *Límites de variación*: son los que determinan el error en la estimación.

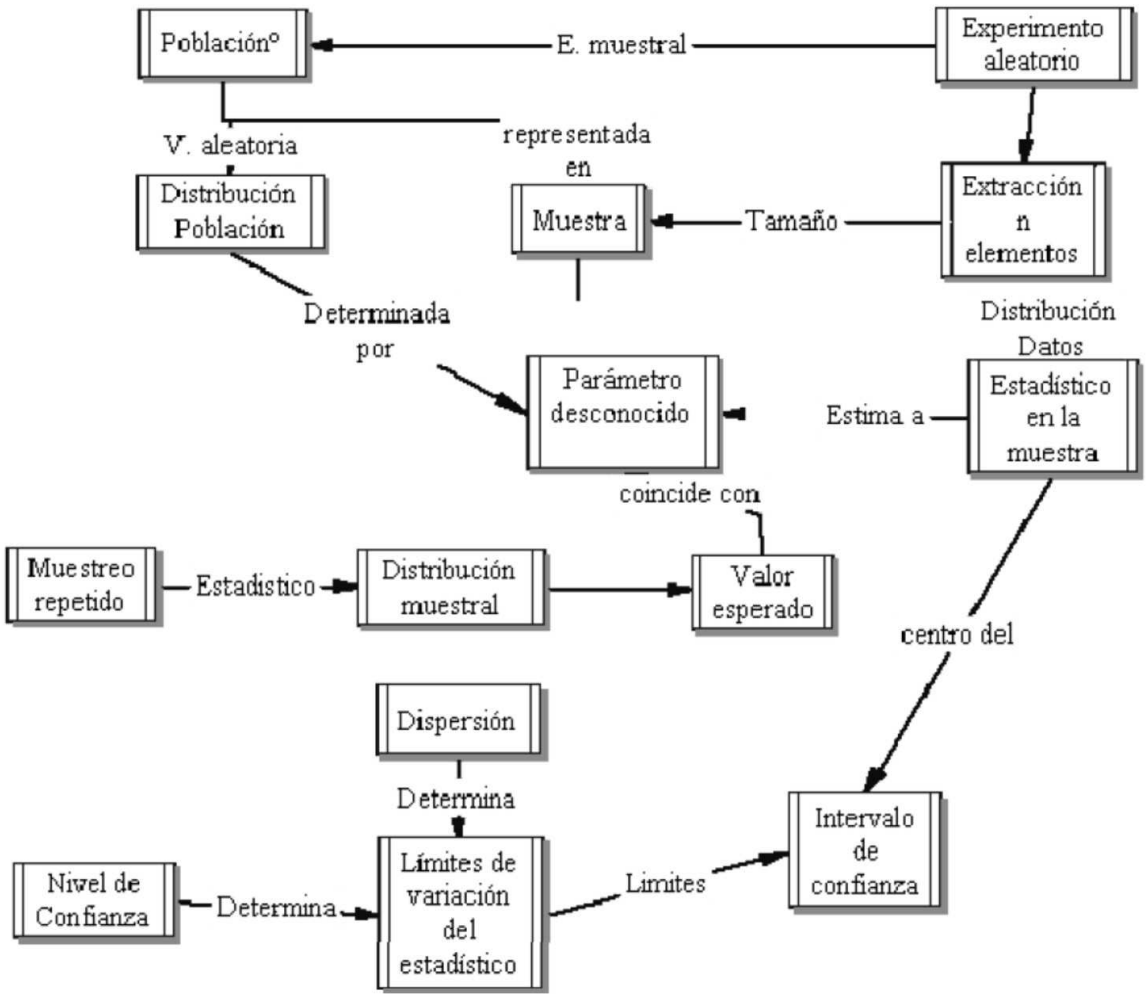


Figura 1. *Términos matemáticos ligados al intervalo de confianza* (Olivo, 2008)

1.3 CONSTRUCCIÓN DE INTERVALOS DE CONFIANZA PARA LA MEDIA

Para la construcción de los intervalos de confianza para la media se pueden presentar dos situaciones: que la varianza se conozca o que la varianza se desconozca. A continuación se describen los procedimientos para ambos casos.

1.3.1. Construcción con varianza conocida. Se tiene una población normal $N(\mu, \sigma^2)$ con varianza σ^2 conocida, y se pretende estimar μ que es la media de la población; sean $X_1, X_2, X_3, \ldots, X_n$, elementos de una muestra perteneciente a una población X entonces.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad (1)$$

$\bar{x}$ es la media de la muestra, σ es la desviación típica poblacional y n el número de elementos que tiene la muestra; la expresión (1) tiene una distribución normal $N(0,1)$. Con un nivel de confianza $1-\alpha$, se pueden seleccionar dos puntos simétricos $-Z_{\alpha/2}$ y $Z_{\alpha/2}$ (ver la Figura 2) tales que:

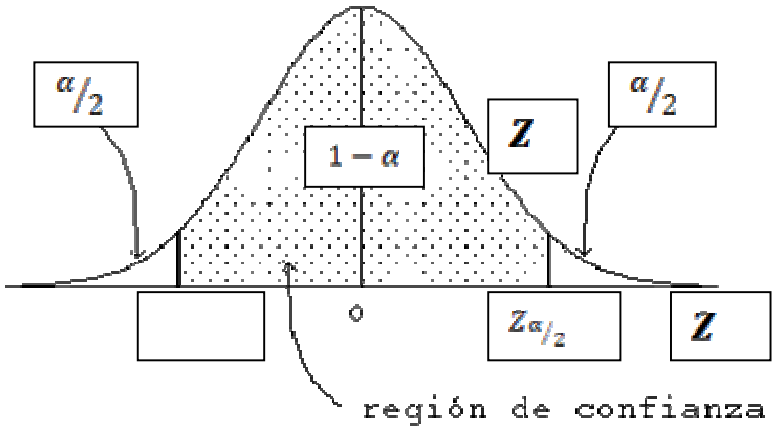


Figura 2. Distribución normal estándar.

$$P\left(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right) = 1 - \alpha$$

Siendo $Z(\alpha/2)$ el valor crítico de la distribución Z estandarizada.

Sustituyendo Z se tiene:

$$P\left(-Z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\right) = 1 - \alpha$$

O, en forma equivalente,

$$P\left(\bar{x} - Z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \bar{x} - Z_{\alpha/2}\,\sigma/\sqrt{n}\right) = 1 - \alpha$$

De donde se deduce el siguiente intervalo de confianza de nivel $1 - \alpha$ para μ

$$\left(\bar{x} - Z_{\alpha/2}\,{}^{\sigma}/_{\sqrt{n}}\ , \bar{x} + Z_{\alpha/2}\,{}^{\sigma}/_{\sqrt{n}}\right) \qquad \textbf{(2)}$$

La expresión (2), describe un intervalo aleatorio, ya que sus extremos dependen de la media muestral y ésta es una variable aleatoria.

Por lo tanto, si se quiere estimar la media de una población normal con los datos de una muestra, con un nivel de confianza del 95%, se tiene que:

$$P\left(-1.96 \leq \frac{\bar{x} - \mu}{{}^{\sigma}/_{\sqrt{n}}} \leq 1.96\right) = 95\%$$

Esto dice que la probabilidad de que el estadístico $Z = \frac{\bar{x} - \mu}{{}^{\sigma}/_{\sqrt{n}}}$ se encuentre entre -1.96 y 1.96 es de 0.95. Se tiene una confianza del 95% de esta afirmación, y 5% en contra de la hipótesis como se observa en la Figura 3.

Si se toman todas las muestras de tamaño n de la población X, y con base en cada muestra se calcula un intervalo de confianza con el mismo coeficiente de confianza, se tendría un $(1-\alpha)$ de probabilidad de encontrar a μ en estos intervalos, es decir, que el $(1-\alpha)\%$ de los intervalos de la población contiene a μ (ver Figura 3).

$$\left(\bar{x} - 1.96\,{}^{\sigma}/_{\sqrt{n}}\ , \bar{x} + 1.96\,{}^{\sigma}/_{\sqrt{n}}\right)$$

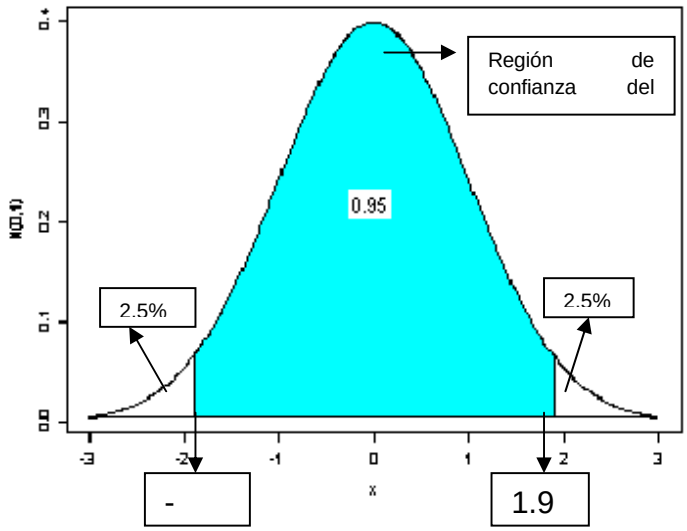


Figura 3. Intervalo de confianza del 95% para una muestra.

Este intervalo se denomina intervalo de confianza para μ, con un nivel de confianza del 95%.

1.3.2. Construcción con varianza desconocida. En este caso se va a construir un intervalo de confianza teniendo una población normal $N(\mu, \sigma^2)$ y varianza σ^2

desconocida, para estimar la media de la población μ.

Igual que en el caso anterior, se tiene una muestra $X_1, X_2, X_3, \ldots, X_n$, tomada de una distribución normal $N(\mu, \sigma^2)$ Como ahora no se conoce σ es necesario estimarla utilizando los valores de la muestra: se estima con la desviación estándar de la misma muestra y se denota *s.* De esta forma la expresión (1) toma la forma

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1} \qquad (3)$$

Donde el parámetro μ es la media de la población que se estima, $\bar{x}$ es la media de la muestra, s es su desviación típica y n el número de elementos que tiene la muestra; la anterior expresión tiene una *distribución t de student* con $(n - 1)$ grados de libertad. $1 - \alpha$. Ver Figura 4. Como se vio en la anterior interpretación, teniendo la distribución del estadístico y su nivel de confianza $1 - \alpha$, se pueden seleccionar dos puntos simétricos $-t_{n-1,\alpha/2}$ y $t_{n-1,\alpha/2}$ tales que:

$$P\left(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}\right) = 1 - \alpha$$

Sustituyendo T se tiene:

$$P\left(-t_{n-1,\alpha/2} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{n-1,\alpha/2}\right) = 1-\alpha$$

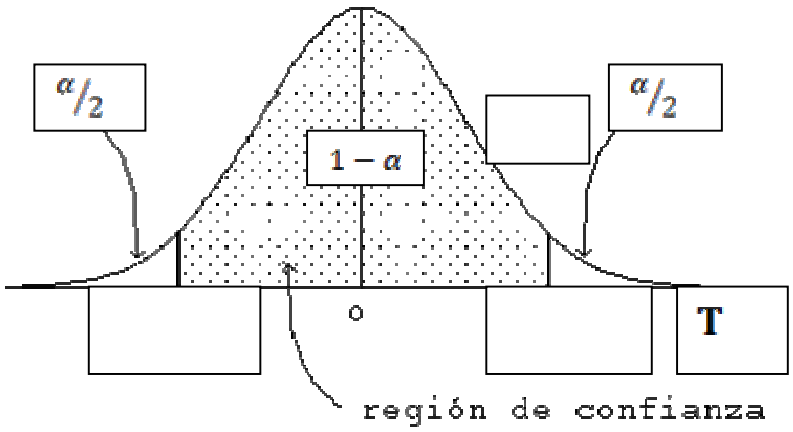


Figura 4. La distribución t- student.

Siendo $t_{\alpha/2}$ el valor crítico en la distribución t- *student*, la anterior expresión es equivalente a:

$$P\left(\bar{x} - t_{n-1,\alpha/2}\ s/\sqrt{n} < \mu < \bar{x} + t_{n-1,\alpha/2}\ s/\sqrt{n}\right) = 1-\alpha$$

De esta forma, se obtiene el intervalo que con confianza $1-\alpha$, contiene el valor de la media μ:

$$\left(\bar{x} - t_{n-1,\alpha/2}\ s/\sqrt{n}\ ,\ \bar{x} + t_{n-1,\alpha/2}\ s/\sqrt{n}\right) \qquad (4)$$

Esta expresión describe un intervalo aleatorio ya que sus extremos dependen de la media muestral y de la desviación típica muestral, que son variables aleatorias (ver Figura 4).

La explicación clásica y general del significado de intervalo de confianza es que si se toman todas las muestras de tamaño n de la población X, y a cada muestra se le haya un intervalo de confianza con el mismo coeficiente de confianza en todas las muestras, tendría un $(1-\alpha)$ de probabilidad de encontrar a μ en estos intervalos, es decir, que el $(1-\alpha)\%$ de los intervalos de la población contiene a μ. (Ver Figura 5).

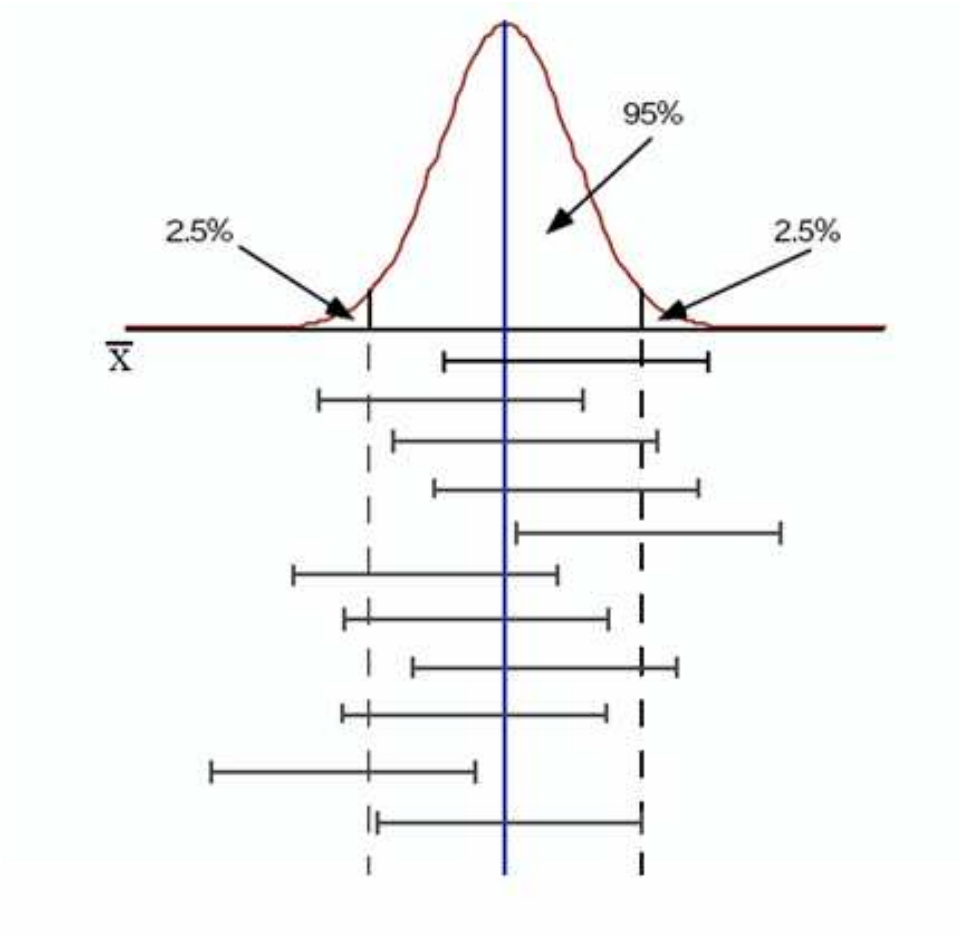


Figura 5. Interpretación del nivel de confianza en el intervalo para la media de una distribución normal.

Se puede saber que si el nivel de confianza es grande, se tendría una mayor certeza de encontrar el parámetro $\boldsymbol{\mu}$, pero al aumentar el coeficiente de confianza, el intervalo crece y se hace menos precisa la estimación de $\boldsymbol{\mu}$ ya que tendría más rango para moverse el valor que da la estimación del parámetro, aumentando con esto el margen de error. Al contrario, al disminuir el intervalo de confianza, la estimación es más precisa, pero la certeza de encontrar a $\boldsymbol{\mu}$ disminuye.

2. CUESTIONARIO Y COMPONENTES

Para tomar mejores medidas, en cualquier medio que lo amerite, es fundamental contar con una herramienta confiable. Con un metro adecuado, es posible realizar una estimación de lo que se está midiendo, en este caso, la habilidad en el tema de intervalos de confianza para la media.

A continuación se relacionan los ítems que conforman el test que se aplicó en esta investigación, se describen y justifican los argumentos que se tuvieron en cuenta para que su elección permitiera evaluar el tema de los intervalos de confianza referentes a la media, al igual que los distractores de cada pregunta.

2.1 SELECCIÓN DE ÍTEMS DEL CUESTIONARIO Y ANÁLISIS DE DISTRACTORES

Para la realización de este trabajo se escogieron 10 preguntas del test utilizado por Eusebio Olivo en su tesis doctoral (Olivo, 2008), cuyo contenido se analizará a continuación.

2.2 ANÁLISIS DE LAS POSIBLES RESPUESTAS A LOS ITEMS

Para el entendimiento del por qué de las respuestas de los estudiantes a cada una de las opciones de los ítems que componen el test, se presenta el tema a evaluar en cada pregunta y lo que el estudiante debe entender para responderlas correctamente. Las explicaciones que se dan responden a consideraciones soportadas teóricamente y con base en la experiencia en la enseñanza de estos tema o en estudios previos asociados con la comprensión de los intervalos de confianza. A continuación se presentan cada uno de los ítems:

Ítem 1: *El intervalo de confianza del 50% para la media de una población* μ *es:*

a) *El rango dentro del cual caen el 50% de los valores de la media de la muestra* $\bar{x}$.
b) *Un intervalo más ancho que el intervalo de confianza del 95%.*
c) *Un intervalo de valores calculado a partir de los datos de la muestra. En el 50% de las muestras de una población, el intervalo calculado contiene a la media de la población.*
d) *Dos veces más ancho que el intervalo de confianza del 100%.*

Lo que se está evaluando en este ítem es: *La definición de intervalo de confianza.* Se necesita saber y entender el significado de coeficiente de confianza; teniendo la idea clara de que el intervalo aumenta con la confianza, estos conceptos ayudaran a que la persona escoja la opción c) que es la correcta en este caso. Ahora, es posible que al escoger las otras opciones los evaluados tengan ideas

confusas en los conceptos que se requieren para elegir de manera acertada el ítem. A continuación se verán las respuestas que con más frecuencia eligen los estudiantes al tener ideas erróneas en los temas involucrados.

- Opción a: Se tiene una idea errónea sobre el intervalo al que se refiere la pregunta, ya que al elegir esta opción, se puede concluir que el estudiante piensa que dicho intervalo estima la media muestral y no la media poblacional, confusión entre *estadístico y parámetro*. Sin mencionar que la media muestral siempre cae dentro del intervalo de confianza calculado, de hecho, es el centro del intervalo.

- Opción b y d: Los estudiantes inclinados hacia cualquiera de estas dos alternativas, no relacionan el nivel de confianza con el ancho del intervalo. Interpretan que un intervalo con un 50% de confianza es más ancho que uno del 95%, cuando es totalmente lo contrario: a medida que aumenta la confianza, aumenta el tamaño del intervalo.

Ítem 2*: Comparado a los intervalos de confianza calculados en muestras de tamaño* $n=4$ *en una población normal, el ancho de los intervalos de confianza de la media de la población calculado en muestras de tamaño* $n=50$*:*

a) *Variará más que los anchos de los intervalos para muestras de tamaño $n=4$.*

b) *Variará, pero no tanto como lo hicieron los anchos de los intervalos para muestras de tamaño $n=4$.*

c) *Tomarán valores parecidos.*

Lo que se está evaluando en este ítem es: *El ancho de los intervalos de confianza disminuye cuando aumenta el tamaño de la muestra.* La respuesta correcta es la b). En la fórmula del intervalo de confianza se suma y se resta el producto del valor crítico por el error estándar de la media dividida por la raíz cuadrada del tamaño de la muestra, entonces, si el tamaño de la muestra es grande, ese valor que se agrega y resta a la media muestral quedará más pequeño, por ende, el intervalo disminuirá.

- Opción a: Posiblemente eligen esta opción debido a que no tienen clara la fórmula del error estándar de la media, puede ser por olvidar la fórmula o malinterpretar la misma.
- Opción c: No conoce la información para dar alguna razón de lo que puede pasar con la variabilidad del intervalo si cambia el tamaño de la muestra.

Ítem 3: Si*, manteniendo todos los demás datos fijos, el nivel de confianza se reduce (por ejemplo de 90% a 80%):*

a) El intervalo de confianza no cambia.
b) El intervalo de confianza será más ancho.
c) El intervalo de confianza será más angosto.
d) El cambio en el intervalo de confianza no es predecible.

Lo que se está evaluando en este ítem es: *El ancho de los intervalos de confianza aumenta cuando aumenta el nivel de confianza.* Este es el concepto que se debe entender para responder la opción c) que es la respuesta acertada del ítem. Esto se puede apreciar también en la fórmula para calcular el intervalo de confianza; si se busca el valor crítico para un nivel de confianza del 80% y se compara con el valor críitico para un nivel de confianza del 90%, en la tabla de distribuciones que corresponda, se podrá observar que el valor del 80% es un coeficiente más pequeño que el valor del 90%, esto hace que al multiplicarlo por el error estándar de la media, sea más pequeño y por ende, el intervalo de confianza será más angosto.

- Opción a: La falla consiste en que el individuo cree que no tiene nada que ver el nivel de confianza con la longitud del intervalo de confianza calculado, no relaciona el nivel de confianza con el intervalo.

- Opción b: Presenta problemas al pensar que si el nivel de confianza disminuye, el intervalo de confianza aumenta, siendo lo contrario y posiblemente confundiendo confianza con precisión.

- Opción d: Es claro que las personas que prefirieron esta respuesta, no pueden entender la relación entre el nivel de confianza y el ancho del

intervalo, suponen probablemente que el nivel de confianza nada tiene que ver en la fórmula para calcular el ancho del intervalo.

Ítem 4: *Explica cómo varía la anchura del intervalo de confianza si, conservando el mismo tamaño de muestra y el mismo coeficiente de confianza se toma una población con varianza cuatro veces mayor.*

Lo que se está evaluando en este ítem es: *El ancho de los intervalos de confianza aumenta cuando aumenta la varianza*. Es un hecho que la fórmula para el cálculo del intervalo de confianza debe entenderse claramente con cada uno de los conceptos que la conforman. Como se sabe, lo que se suma y resta a la media muestral, es el producto del valor crítico con la desviación típica de la población, dividida por la raíz cuadrada del tamaño de la muestra. Pero como la varianza es el cuadrado de la desviación típica, si se calcula ese intervalo con 4 varianzas, duplicaría el rango ya que se transformaría en 2 desviaciones típicas, que sería lo mismo que multiplicar por 2 lo que se suma y resta a la media muestral.

Un error común puede ser el no expresar matemáticamente en qué proporción aumenta el intervalo de confianza, cuando se decide incrementar la varianza 4 veces. También se puede confundir la varianza con la desviación típica, argumentando que el intervalo aumenta las mismas veces que la varianza, o caso contrario, que disminuye.

Ítem 5: *En un intervalo de confianza del 95% para la media:*

a) *Si se toman muchas muestras y con cada una de ellas se construye el intervalo, la media muestral* $\bar{x}$ *caerá dentro del intervalo de confianza el 95% de las veces.*

b) *La probabilidad de que $\bar{x}$ caiga dentro de un intervalo de confianza calculado de una muestra específica es 0.95*

c) *Si se toman muchas muestras de igual tamaño, el 95% de los intervalos calculados contendrá a μ.*

Lo que se está evaluando en este ítem es: *El significado del nivel de confianza (variación del intervalo en diferentes muestras).* La importancia en esta pregunta es reconocer que los intervalos calculados pretenden estimar la media poblacional, y que el nivel de confianza responde a la probabilidad del proceso generador de los intervalos y no a cada intervalo. La respuesta correcta es la c).

- Opción a: Al elegir esta opción, se puede argumentar que el estudiante cree que el intervalo del que se habla, requiere estimar la media muestral y no la poblacional, además se sabe que la media muestral siempre cae dentro del intervalo de confianza.

- Opción b: el estudiante hace una interpretación errónea del intervalo de confianza, es decir, confusión entre *estadístico y parámetro.* Además asocia el nivel de confianza con la probabilidad en un solo intervalo, concretamente en el intervalo obtenido.

Ítem 6*: La media muestral de 100 observaciones en una prueba de matemáticas es 75, encuentre el intervalo de confianza al 95% para la media de población,*

asumiendo que $\sigma=7$:

a) *(61.28, 88.72)*
b) *(73.63, 76.37)*
c) *(68, 82)*
d) *(74.3, 75.7)*

El concepto que se pretende evaluar con esta pregunta es: *Estimar el intervalo de una población normal o una muestra grande con σ conocida*. Como contenidos secundarios tiene: *definición del intervalo de confianza y elegir un modelo de distribución muestral del estadístico*. Calculando el producto del error típico con el valor critico, se le suma y se le resta a la media muestral, dando como resultado el intervalo de confianza solicitado en el ítem, de esta forma se sabrá que la repuesta es la b).

- Opción a: Este error es debido a que el estudiante calculando el intervalo, olvida o no tiene presente, dividir el producto del valor crítico con la desviación típica de la población, por el tamaño de la muestra.

- Opción c: Sencillamente calcula el intervalo de confianza sumando y restando la desviación típica poblacional a la media muestral, ignorando los demás valores como son: el hecho de que el intervalo se necesita con un nivel de confianza del 95% y que el tamaño de la muestra es 100, que se requieren para el resultado correcto, puede ser porque al calcularlo así, aparece la opción y se inclina por ella. Poca idea del concepto intervalo de confianza.

- Opción d: La falla en esta respuesta es calcular el intervalo descuidando que debe ser del 95% de confianza, es decir no tienen en cuenta el valor crítico, sumando y restando a la media muestral únicamente el error típico.

Ítem 7: *Un fabricante asegura que sus garrafones contienen un litro de cloro puro. Al tomar una muestra de 16 garrafones se determinó que en promedio contenían 0.94 litros de cloro puro, con desviación estándar de la muestra de 0.097. Construir un intervalo de confianza al 95% para el verdadero contenido promedio de litros de cloro puro. No se conoce la desviación típica de la población. (La distribución del contenido de cloro por botella puede considerarse normal).*

El concepto que se pretende evaluar con esta pregunta es: *Estimar la media de una población aproximadamente normal cuando* δ *es desconocida.* Como contenidos secundarios: *definición del intervalo de confianza y elegir un modelo de distribución muestral del estadístico.* La pregunta es abierta, donde se espera que el estudiante haga los cálculos correctos para hallar el intervalo donde se pueda estimar el verdadero contenido promedio de cloro por garrafón. Para esta solución se deben tener presentes los siguientes conceptos:

- Recordar la ecuación del cálculo del intervalo de confianza para la media poblacional, teniendo en cuenta las herramientas que ofrece la misma pregunta.

- Identificar el uso de la distribución *t*, ya que la desviación típica de la población no se conoce y solo se tiene la desviación estándar de la muestra.

Ítem 8: *Se han obtenido los siguientes datos de emisión diaria de óxidos de azufre, para una muestra de tamaño* $n=100$*, media:* $\bar{x}=18$*, y varianza muestral*

$s^2=36$. Elabore un intervalo de confianza del 95% para la verdadera emisión diaria promedio de óxidos de azufre.

a) *(17.016, 18.984).*
b) *(16.824, 19.176).*
c) *(6.24, 29.76).*
d) *(8.16, 27.84).*

Los conocimientos evaluados en esta pregunta son: *Estimar la media de una población a partir de datos experimentales, δ desconocida, muestra grande y la definición del intervalo de confianza eligiendo un modelo de distribución muestral del estadístico*. La opción correcta es la b), los que responden las demás opciones tiene los siguientes conflictos:

- Opción a: Este error se debe a que el individuo calcula el intervalo de confianza con un valor crítico de 1.64, y no con el valor 1.96 que es el correcto. Muestra mal entendimiento a la hora de buscar los valores críticos en la tabla de la distribución normal estándar.

- Opción c: El valor crítico va multiplicado por la desviación estándar muestral y dividida por la raíz de n, la falla es no dividir por este valor.

- Opción d: Se tienen los dos errores anteriormente descritos.

Ítem 9: *El nivel de confianza es de 0.95, para un intervalo de confianza para la media de la población con desviación estándar poblacional desconocida para un grupo de puntajes distribuido normalmente de tamaño $n=20$. Los valores críticos han de ser:*

a) *-1.65 y 1.65 uso de normal estándar.*
b) *-1.96 y 1.96 uso de normal estándar.*
c) *-2.093 y 2.093 uso de distribución t con 19 grados de libertad.*
d) *-2.085 y 2.085 uso de distribución t con 20 grados de libertad*.

Este ítem evalúa los conocimientos para: D*eterminar los valores críticos en la distribución del estadístico*. Utilizando la distribución *t,* que es la que se necesita por desconocerse la desviación estándar poblacional, se pueden calcular los puntos críticos del intervalo solicitado por el ítem, identificando la respuesta correcta que es la c).

- Opción a: El hecho de utilizar una distribución que se usa para distribuciones con muestras grandes y manejar otro valor de nivel de confianza hace que esta opción sea elegida.

- Opción b: El mismo error del ítem pasado, pero en este caso se usa correctamente el valor del nivel de confianza.

- Opción d: En esta opción se utiliza correctamente la distribución para muestras pequeñas pero mal el valor de los grados de libertad.

Ítem 10: *Considere el gráfico siguiente del rendimiento medio de cebada en 1980,*

1984 y 1988 junto con los respectivos intervalos de 95% de confianza.

Año	*N*	*Media*	*StDev*	----+--------+---------+---------+--
1980	*6*	*184.00*	*2.61*	*(----*---)*
1984	*5*	*212.40*	*14.36*	*(----*---)*
1988	*5*	*182.40*	*1.82*	*(-----*----)*
				----+--------+---------+---------+--

Dev. Típica conjunta = 0.19 180 195 210 225

¿Cuál de las siguientes afirmaciones es verdadera?

a) *Puesto que los intervalos de confianza para 1980 y 1988 tienen considerable solape, hay buena evidencia que las medias de las muestras difieran.*

b) *La estimación de la media de la población en 1980 es menos precisa que en 1988.*

c) *Puesto que los intervalos de confianza para 1980 y 1984 no se solapan, hay poca evidencia que las medias de las poblaciones respectivas difieran.*

d) *Puesto que los intervalos de confianza para 1980 y 1988 tienen considerable solape, hay poca evidencia que las medias de las poblaciones difieran.*

El concepto evaluado en este ítem es: *Interpretar gráficos de intervalos de confianza.* Se pretende que el estudiante relacione dos o más muestras en una misma tabla, cada una con su respectivo intervalo de confianza. Como todas la muestras pretenden dar una estimación de una población, los solapes en la gráfica son indispensables para dar una respuesta adecuada en este ítem, interpretando qué significa que haya solape entre ellas o no, la respuesta correcta es la d); otras opciones presentan los siguientes conflictos conceptuales:

- Opción a: Tener la idea contraria de lo que produce este efecto, pensando que al producirse solapes en los intervalos las medias en las poblaciones son diferentes.

- Opción b: Mal entendimiento en la variabilidad de los dos intervalos.

- Opción c: Interpretación errónea en los gráficos, ya que los intervalos en *1980 y 1988* sí se solapan. Pensar que no hay solape entre los dos intervalos.

3. LA TEORÍA CLÁSICA DE TEST Y EL MODELO RASCH

La misión principal en este capítulo es describir dos de los métodos para la calibración de ítems y evaluación de personas: la Teoría Clásica de Test (TCT) y el Modelo Rasch. Se presentan sus principales características, se resaltan sus diferencias y sus ventajas y desventajas.

3.1 TEORIA CLASICA DE TEST

Mientras los aspectos físicos (estatura, peso, volumen, entre otros) pueden ser medidos con gran precisión, debido a que son tangibles, los atributos mentales de las personas no se pueden medir directamente. Para ello, se recurre a mediciones indirectas evaluando acciones o respuestas que se considera son expresión de la característica que se quiere medir, siendo los cuestionarios el mejor instrumento para dar cuenta de este fin.

Este proceso de medición indirecta se inicia con la construcción del instrumento, teniendo claro el atributo que se va a estimar. El cuestionario está conformado por ítems que se deben responder y cuyas buenas respuestas permiten obtener una medida de la cantidad que de la característica que se pretende medir posee el sujeto que responde.

La TCT es el método comúnmente empleado en el salón de clase, donde las calificaciones de cada individuo se basan en contabilizar la cantidad o proporción de respuestas correctas sobre el total de preguntas. Cuando los ítems no tienen el mismo valor se cuenta el número total de puntos correctos sobre el total de puntos posibles.

$$\textit{calificación del estudiante} = \frac{\textit{número de puntos correctos del test}}{\textit{número total de puntos del test}}$$

Las características más significativas de este método son las siguientes:

- Los resultados de una prueba están ligados a la prueba misma. Dos exámenes diferentes que pretendan evaluar el mismo concepto pueden tener resultados distintos.
- Las propiedades de un test están en función de las personas a quienes se les aplica, es decir, la dificultad del test depende de la muestra que lo responda.
- El método acepta calificaciones perfectas y completamente nulas.

3.1.1 Análisis de ítems. Al construir un cuestionario con un grupo de preguntas se debe tener en cuenta lo que los ítems van a medir, y sobre todo que lo hagan bien, pues el objetivo es estimar el nivel en el que se ubica cada rasgo medido. Ahora bien, entre mejor medidor sea el ítem, mejores conclusiones saldrán de la investigación. Para cuantificar la eficacia en la medición de un ítem se estudian tres puntos importantes:

- Índice de dificultad
- Índice de discriminación
- Índice de validez

Para facilitar la explicación de cada uno de estos índices, se ubican los sujetos y los ítems en una matriz como se observa en la Tabla 1. En ésta se tiene un grupo

de N sujetos y X ítems; un elemento de esta matriz a_{ij} es la calificación que

obtiene la persona i en el ítem j. Al sumar las filas se tiene los puntajes de cada uno de los evaluados y al sumar las columnas se obtendrán las respuestas a los ítems.

	Ítems			
	1	2	3..... n	X
Sujeto nº 1				
Sujeto nº 2				
Sujeto nº 3				
.				
.				
.				
.				
.				
.				
Sujeto nº N				

Tabla 1. Matriz de encuestados vs. Ítems.

3.1.1.1 Índice de dificultad (**D**). Es un valor que indica el nivel de dificultad de un ítem. El ID de una pregunta j se define como el cociente entre el número de personas que respondieron acertadamente A_j y el número total de individuos que intentaron responder la pregunta N_j.

$$D_J = \frac{A_j}{N_j}$$

Cuando la persona no responde el ítem, no se contabiliza ese dato.

A continuación se muestra en la Figura 6 como calcular el ID por medio de una matriz que contiene los datos de 10 personas que respondieron a 6 ítems. Las

posibles respuestas son: si responde bien obtiene 1, si responde mal obtiene 0 y si no responde se pone un guion (-), esto ayudará a la interpretación del ID. La columna X da cuenta de los puntajes alcanzados por los sujetos.

		Ítems						
		1	2	3	4	5	6	X
Sujetos	1	0	0	0	1	1	1	3
	2	0	1	-	0	-	1	2
	3	0	0	1	-	0	1	2
	4	0	0	0	-	1	1	2
	5	0	1	0	1	-	1	3
	6	0	1	-	-	-	1	2
	7	0	0	-	1	1	1	3
	8	0	0	1	-	0	-	1
	9	0	1	0	-	0	1	2
	10	0	1	0	-	0	1	2
	A_j	0	5	2	3	3	9	
	N_j	10	10	7	4	7	9	
	D_j	0	0.5	0.29	0.75	0.43	1	

Figura 6. Matriz de respuestas a un test, índice de dificultad.

Como se ve en la matriz anterior, el ID del ítem 3 es de 0.29 que se obtiene de dividir 2 entre 7. Los valores en los otros ítems se ignoran por falta de alguna respuesta. A_j es la cantidad de respuestas correctas a cada uno de los ítems y

N_j es la cantidad de respuestas correctas e incorrectas sin contar con las respuestas nulas. Este ejemplo ayuda a evidenciar algunos aspectos:

- El máximo valor que puede tener el ID es 1 (se da cuando todos responden correctamente la pregunta) y el mínimo valor que puede tomar el ID es 0 (se da cuando todos responden mal la pregunta).
- Cuando el ID se acerca a 0, la dificultad del ítem se incrementa, cuando se acerca a 1 disminuye, y cuando se sitúa en 0.5 es media.
- El ID tiene relación con la varianza de los ítems, debido a que si el ID es 0 o 1, la varianza es 0, y aumenta cuando el ID se acerca a 0.5. La información de los ID se vuelve nula cuando son 0 o 1, porque no proporciona medida de algún rasgo.

3.1.1.2 Índice de discriminación **(H)**. También llamado índice de homogeneidad de un ítem H_j, está definido como la correlación de Pearson[1] entre los puntajes de las N personas al ítem j, y los puntajes X en el total del test, y se representa matemáticamente así:

$$H_j = r_{jx}$$

Este valor ofrece información sobre si el ítem está evaluando lo mismo que el test en general, es decir, si el ítem favorece a la homogeneidad que requiere el test internamente. Los ítems que arrojen valores cercanos a cero en el índice de discriminación reflejan que están midiendo algo diferente a la prueba, deberían eliminarse, si se pretende estimar un rasgo único.

---

[1] *El coeficiente de correlación de Pearson es un índice estadístico que mide la relación lineal entre dos variables cuantitativas y es independiente de la escala de medida de las variables.* $r = \frac{S_{xy}}{\varepsilon_x \cdot \varepsilon_y}$, $S_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$

Suponiendo que se pretenden hallar los índices de discriminación en un test que contiene 3 ítems, con un tipo de respuestas valoradas entre 1 y 5 a un grupo de 5 sujetos; la información se puede relacionar en una matriz como se muestra en la Figura 7.

		1	2	3	X
Sujetos	1	2	3	5	10
	2	3	1	0	4
	3	5	4	5	14
	4	0	1	0	1
	5	4	3	0	7

Figura 7. Matriz de respuestas a un test, índice de discriminación.

Los índices de discriminación para estos tres ítems serían:

$$H_1 = r_{1x} = 0,75$$

$$H_2 = r_{2x} = 0,94$$

$$H_3 = r_{3x} = 0,86$$

Entonces, para el ítem 1, el índice de discriminación es igual a la correlación de este ítem con el puntaje total de los individuos en la prueba, y así para los demás.

Cuando las pruebas contengan subpruebas con rasgos distintos a estimar, los H_j se deben hallar con relación a las puntuaciones de la subprueba. Cuando el índice de discriminación de un ítem H_j es negativo, se debe revisar la cantidad de puntaje que obtuvo, pues seguramente se ha cometido un error en el conteo.

3.1.1.3 Índice de validez.$(\boldsymbol{V})$. Al obtener los puntajes de un grupo de $\boldsymbol{N}$ personas a un ítem $\boldsymbol{j}$, se puede correlacionar con un criterio $\boldsymbol{Y}$ externo al test y lo que significa esta correlación; este índice se llama índice de validez del ítem $\boldsymbol{j}$ y se representa matemáticamente como:

$$\boldsymbol{V_j = r_{jy}}$$

Por ejemplo, un criterio para validar un test de intervalos de confianza, puede ser otro test que incluya cuestiones sobre el tema. Suponiendo que los datos del ejemplo anterior, se van a correlacionar con un criterio $\boldsymbol{Y}$ que tiene puntuaciones de 5 personas así:

Personas: 1 2 3 4 5

$\boldsymbol{Y}$ **:** 5 3 6 0 6

Los índices de validez para cada ítem serán:

$$\boldsymbol{V_1 = r_{1y} = 0,87}$$

$$\boldsymbol{V_2 = r_{2y} = 0,88}$$

$$\boldsymbol{V_3 = r_{3y} = 0,54}$$

Los que tengan índices de validez cercanos a cero deben eliminarse, porque no ayudan a medir la variable que se desea en la prueba. Ahora, si se quiere elegir los ítems que más colaboren con la validez del cuestionario, serian los que tengan un $\boldsymbol{V_j}$ y un bajo $\boldsymbol{H_j}$.

3.1.2 Análisis del test. Si en una prueba un estudiante obtiene una calificación $\boldsymbol{X}$, lo que debe preguntarse, es sí ese valor corresponde a la variable latente que

se está midiendo, o si existe alguna falta de información por fenómenos externos a la prueba, como problemas sicológicos, físicos, entre otros.

La TCT propone un modelo formal basado en supuestos con grandes niveles de aplicabilidad para determinar la información que ofrece un test sobre una aptitud. El modelo básico de la TCT está dado por la siguiente expresión lineal:

$$X = V + E \qquad (1)$$

Muestra que la calificación X que obtiene la persona en el test, se compone de dos valores: V que es la puntuación verdadera del individuo y E que es el error posible en la medida del atributo. Este error se considera una variable aleatoria y puede ser causado por eventos externos de la prueba que no permiten estimar el verdadero valor del atributo, el error se puede hallar restándole el valor real del rasgo al valor estimado del mismo:

$$X - V = E$$

Como en principio, los valores V y E son desconocidos, se deben ingresar otros supuestos como:

$$V = E[X] \qquad (2)$$

La expresión (2) define el valor verdadero de una puntuación como el valor esperado de las posibles puntuaciones del test, que es el promedio de las puntuaciones que obtiene la persona en varias aplicaciones del test. Del supuesto anterior se deduce que:

$$E[E] = 0$$

Asumiendo que X y E son variables aleatorias, y V es constante se puede comprobar lo anterior de la siguiente manera:

$$E[E] = E[X - V] = E[X] - E[V] = E[X] - V = V - V = 0$$

Un tercer supuesto hace referencia al hecho de la no correlación entre el verdadero valor y el error de estimación.

$$\rho_{VE} = 0 \qquad (3)$$

El tercer supuesto implica, entre otras cosas, que puntuaciones verdaderas elevadas no deben tener valores de errores elevados.

$$\rho_{E_j E_k} = 0 \qquad (4)$$

El cuarto supuesto indica que si se conocieran los errores de medida de cada persona en dos test diferentes (J y k), dado que son variables aleatorias, la correlación entre las dos seria nula.

$$\rho_{E_j V_k} = 0 \qquad (5)$$

El quinto supuesto indica que si se supieran los errores de medida E en un test j, y las puntuaciones V en un test k, la correlación entre las dos sería 0.

En la recolección de datos de un test solo se pueden conocer las puntuaciones X de los individuos, así que los supuestos anteriores por más lógicos que parezcan, no se pueden utilizar en la medida de un rasgo latente, siendo ésta, una de las principales restricciones de la TCT.

3.2 EL MODELO RASCH

Desde comienzos del siglo XX la construcción y el uso de test se ha basado principalmente en la TCT, un modelo simple que se utiliza tradicionalmente en el

aula de clases para calificar evaluaciones, práctico y muy conocido, pero que no está libre de restricciones. Por ejemplo, las calificaciones de cada individuo se basan en contabilizar la cantidad o proporción de respuestas correctas sobre el total de preguntas, que no es más que la sumatoria de respuestas individuales, o cuando los ítems no tienen el mismo valor, se cuenta el número total de puntos correctos sobre el total de puntos posibles. Así, la dificultad de cada ítem se considera igual al momento de la calificación en el primer caso, en el segundo anticipa la dificultad de cada ítem asignándole mayores puntajes a los ítems que se consideran son más difíciles, lo que no permite medir el grado de dificultad de los ítems con respecto a las respuestas de los estudiantes a los mismos, sino que establecen la dificultad antes de la aplicación del test.

Para el primer caso, en un examen de 5 preguntas, un estudiante responde correctamente 3 preguntas y las otras 2 quedan mal, su calificación es 3/5. Se puede suponer que las preguntas donde obtuvo una mala respuesta, el evaluado no tiene ningún conocimiento, y lo respondido correctamente lo entiende a la perfección. Esta suposición puede ser bastante drástica, ya que no se sabe si es discreta la variable que se mide o si las preguntas contestadas correctamente evalúan de manera completa el tema y con suficiente profundidad. Este método es de gran utilidad si las incógnitas a encontrar tienen siempre resultados discretos.

Otra limitación que se considera relevante de la TCT es lo que se conoce como efecto techo y efecto piso. Estos efectos hacen referencia al hecho de que las medidas de la cantidad de la característica latente que se quiere medir están limitadas por encima por el valor uno (efecto techo) y por debajo por el valor cero (efecto piso). Las personas que obtienen uno responden bien todos los ítems lo que hace imposible cuantificar adecuadamente su habilidad, ya que no se puede afirmar que poseen el máximo de ella. Igualmente, no tiene sentido afirmar que un individuo no posee ninguna habilidad porque no respondió bien ninguno de los

ítems. En ambos casos, la mejor respuesta es decir, simplemente, que el cuestionario no puede dar cuenta de la habilidad de estos individuos.

Los efectos techo y piso se entienden mejor si pensamos medir la estatura de un grupo de personas con una cinta métrica pegada en la pared que va de 1.90 metros en la parte más alta hasta 1.50 metros en la parte más baja. ¿Sería lógico asignarle 1.90 metros a todas las personas que tienen alturas mayores a ese punto? ¿se podría decir que los individuos con alturas menores a 1.50 metros tienen esa medida? Sin duda, es más lógico pegar una cinta métrica que vaya desde 0 metros hasta 3 metros y no acotar ni superior ni inferiormente las medidas asignando a cada uno su estatura real. Se pueden evitar estos efectos introduciendo preguntas con mayor profundidad para los más hábiles y de menor profundidad para los menos hábiles.

De otro lado, la teoría clásica de test gira en torno al test y no estudia cada ítem en particular, es decir, el valor obtenido de cada individuo en el test sólo permite observar su capacidad global sobre el examen e impide profundizar en el análisis individual de la persona con cada ítem; estas complicaciones impiden realizar predicciones sobre el comportamiento o conocimiento de los individuos ante un determinado aspecto de una prueba[2].

Uno de los principales inconvenientes de la TCT es que las características de los evaluados no pueden separarse de las características del test, es decir, la capacidad de un individuo la define el test, si la prueba es fácil o difícil las habilidades cambian; la habilidad depende exclusivamente de la dificultad del examen. De igual forma, la dificultad de un test depende de la habilidad de las personas que lo responden: si las personas son muy hábiles, el test se podría considerar fácil, y si las personas son de poca habilidad, el test podría

[2] *(www.anep.edu.uy/documentos/Matematica/CAP4.pdf).*

considerarse difícil. En resumen, no se puede en la TCT calificar a un test de difícil ni a un individuo de ser muy hábil, ambas afirmaciones son dependientes, de los individuos que lo responden la primera, y de los ítems que responde, la segunda.

Una respuesta a estas limitaciones la suministra, precisamente, el modelo Rasch. En 1960 el matemático danés George Rasch propuso un modelo de medición que ayuda a solucionar varias de las limitaciones que tiene la TCT permitiendo construir nuevas pruebas, mejor ajustadas y más eficaces. La formulación que más se conoce del modelo Rasch, se obtiene de la predicción de la probabilidad de la respuesta correcta a un ítem determinado en una prueba aplicada, a partir de la diferencia entre el nivel de habilidad que tiene la persona B y el nivel de dificultad que tiene el ítem D.

El modelo propuesto por Rasch se fundamenta en los siguientes supuestos:

- El atributo que se desea medir puede representarse en una única dimensión en la que se situarían conjuntamente las personas y los ítems.

- El nivel de la persona en el atributo y la dificultad del ítem determinan la probabilidad de que la respuesta sea correcta. Si el control de la situación es adecuado, esta expectativa es razonable y así debe representarla el modelo matemático elegido.

Son varias las métricas utilizadas para representar esta relación, la más utilizada es la escala *logit,* que se escribe como la función logística utilizada por Rasch para describir el modelo:

$$Ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i \qquad (6)$$

Esta función indica que el cociente entre la probabilidad P_{ni} de una respuesta correcta del individuo *n* al ítem *i*, y la probabilidad $1 - P_{ni}$ de una respuesta incorrecta depende de la diferencia entre la habilidad de la persona B_n y la dificultad del ítem D_i. De esta forma, cuando una persona responde a un ítem equivalente a su nivel de competencia, tendrá una probabilidad de 0.5, es decir, la misma probabilidad de respuesta correcta o incorrecta, de manera que el cociente es igual a 1 y el logaritmo natural de 1 = 0, que refleja la dificultad del ítem equivalente a la habilidad de la persona, $(B_n - D_i) = 0$.

Cuando la habilidad del individuo es mayor que la requerida por el ítem, $(B_n - D_i > 0)$ la probabilidad de una respuesta correcta será mayor. Por el contrario, si la habilidad de la persona es menor que la requerida por el ítem $(B_n - D_i < 0)$ entonces la probabilidad de respuesta correcta es menor.

Al despejar P_{ni} en la expresión (6) se obtiene la expresión (7) que permite predecir

la probabilidad de obtener una respuesta correcta un individuo de habilidad B_n

cuando se enfrenta a un ítem de dificultad D_i.

$$P_{ni} = \frac{\exp(B - D_i)}{1 + \exp(B - D_i)} \quad (7)$$

La función de probabilidad definida en (7) es continua y estrictamente creciente. Ahora bien, se sabe que este valor de probabilidad oscila entre 0 y 1, y esta

función logística tiene dos asíntotas horizontales, una en 0 y otra en 1. Se acerca a 0 la probabilidad de responder bien a un ítem cuando su nivel de dificultad es mucho mayor que el nivel de habilidad del individuo; a su vez, la probabilidad se acerca a 1 cuando es la habilidad del individuo la que es mucho mayor que la dificultad del individuo.

3.2.1 Supuestos básicos para la construcción del modelo. Igual que en todos los modelos, el modelo Rasch también tiene sus supuestos, éstos son: unidimensionalidad e independencia local.

3.2.1.1 Unidimensionalidad*.* Este supuesto especifica que solo se está midiendo una única característica en los individuos y que por lo tanto su rendimiento en cada uno de los ítems depende exclusivamente de la cantidad que de ella se posea.

3.2.1.2 Independencia local*.* Existe independencia local entre los ítems que conforman un test, si la respuesta que da una persona a uno de ellos no depende de las respuestas que da a los otros. Matemáticamente se puede escribir que la probabilidad que una persona con cierta habilidad acierte *n* número de ítems es igual al producto de la probabilidad que acierte cada uno por separado. Todas estas probabilidades están condicionadas a la habilidad que posee la persona.

3.2.2 Características del modelo. Entre las características que posee el modelo Rasch se destacan las siguientes:

3.2.2.1 Medición conjunta*.* Los parámetros de las personas y los ítems se localizan en una misma dimensión y se expresan en las mismas unidades. .

3.2.2.2 Estadísticos Suficientes*.* Para estimar los parámetros de la personas sólo se requiere el puntaje en la prueba; igualmente para estimar la dificultad de los ítems solo se requiere el número de personas que lo respondieron correctamente.

3.2.2.3 Objetividad Específica*.* La diferencia entre dos personas frente a un atributo no depende de los ítems concretos con que sean medidos, e igualmente, la diferencia entre dos ítems no depende de las personas que se utilicen para medirla. Debido a esto las comparaciones de las personas son independientes frente a los ítems suministrados, y los parámetros de los ítems son independientes de la muestra de personas que sea usada para la calibración.

3.2.2.4 Propiedades de Intervalo*.* Las diferencias entre las habilidades de los individuos así como la diferencia entre las dificultades de los ítems, sólo dependen de su magnitud y no de su posición en la escala de medida.

3.2.2.5 Especificidad del error estándar*.* El proceso de estimación, tanto de la habilidad de los individuos como de la dificultad de los ítems, suministra un error estándar de esa estimación que es propia para cada parámetro que se estima.

3.2.3. Bases de estimación en el modelo. A manera de justificación del modelo logístico propuesto por Rasch, se presenta a continuación en la Figura 8 una situación que pretende confrontar la habilidad de un par de saltadores de vallas cuando intentan saltar a varias de ellas de diferentes alturas. La idea es tomada de (Smith & Smith, 2004).

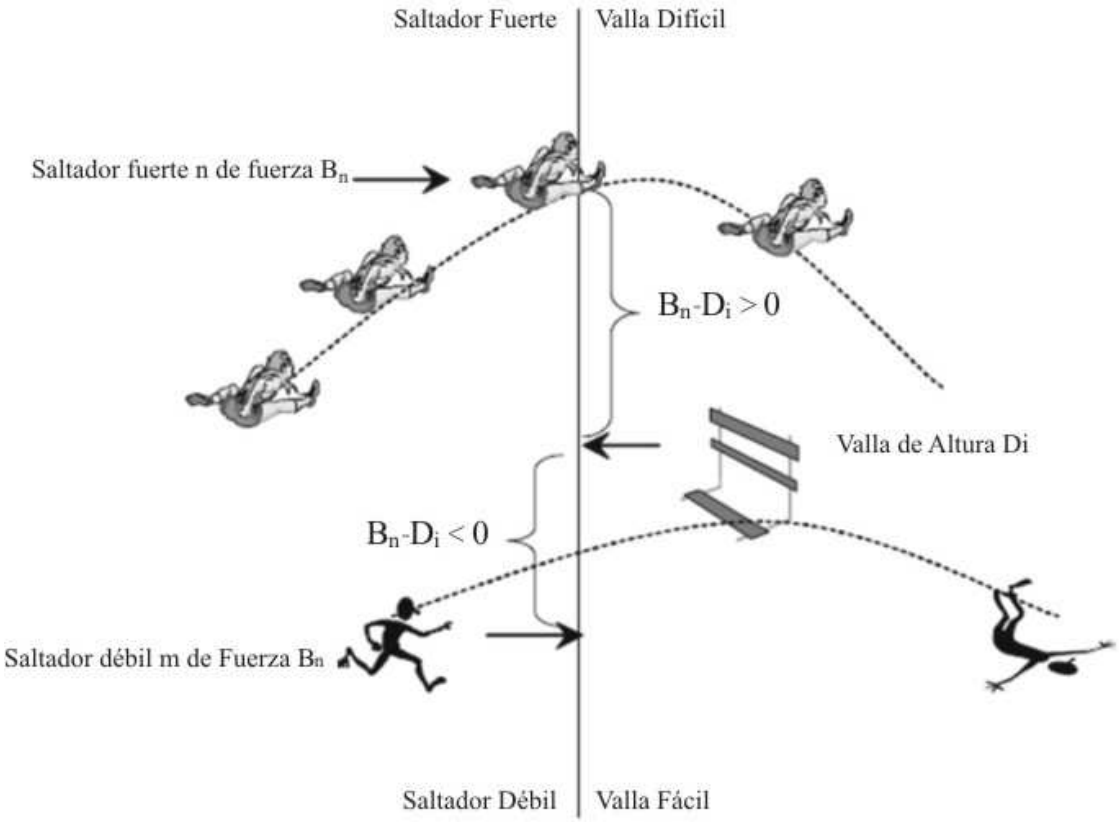


Figura 8. Saltadores buscando superar las valla.

Se presenta una situación en la que un grupo de ***N*** personas ***(n=1, 2, 3,..., N),***

intentan saltar ***L*** vallas ***(i=1, 2, 3,..., L)***, y se considera la variable ***X*** asociada con

cada saltador y con cada valla de tal forma que ***X=1*** significa que superó la valla,

en tanto que ***X=0*** da a entender que fracasó en el intento.

Los resultados se organizan en una matriz cuyas columnas se asocian con las vallas y los competidores con las filas; las celdas de la matriz responden al éxito o fracaso de los participantes al intentar saltar las vallas, como se muestra en la Figura 9.

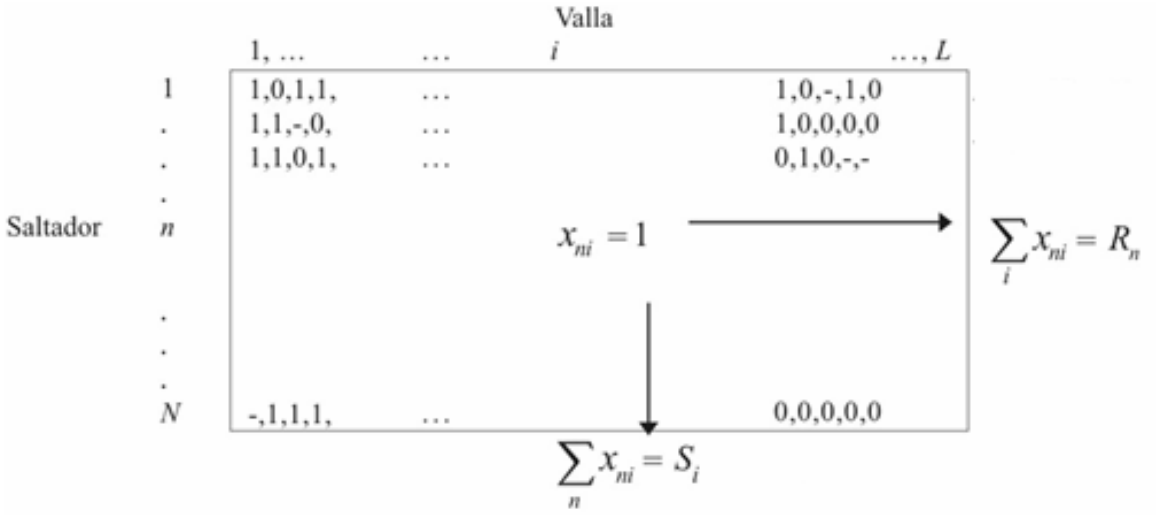


Figura 9. Matriz donde se observan saltadores y los resultados con cada una de las valla.

Las vallas se organizan de izquierda a derecha, de la menor a la más difícil de saltar, y las personas de la más hábil o fuerte, hasta la menos hábil de arriba a abajo, siendo R_n la suma de éxitos de los saltadores y S_i la suma de las veces en que la valla fue exitosamente superada.

Cuando la matriz se organiza de esta manera, se pueden observar los resultados e interpretar el nivel de fuerza que tiene cada participante, así como el nivel de dificultad de cada valla.

Una vez organizada la matriz de respuestas del test, se puede pensar en predecir la probabilidad de que un saltador n tenga éxito con una valla i; esta probabilidad estaría dada por P_{ni} (ver Figura 10). Preguntas como: ¿Cuál atleta es mejor? ¿Qué tan difícil es la valla más alta para todos los atletas? O comparar dos atletas para deducir ¿cuál de los dos es mejor? se pueden tratar de responder por medio de probabilidades. La forma de hacerlo se esquematiza a continuación.

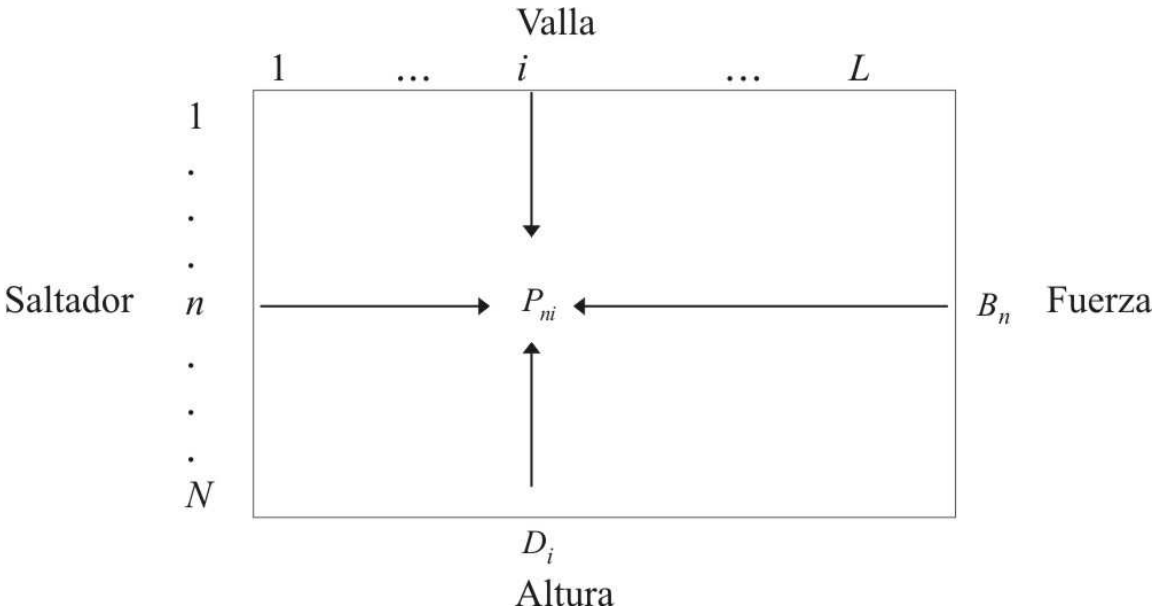


Figura 10. Matriz de probabilidades de los saltadores con cada una de las valla.

Se puede hacer un análisis con dos atletas: *Carlos y Alex*. Comparar a *Carlos* y a *Alex* para ver cuál de los dos es más hábil saltando vallas implica que *Carlos* y *Alex* realicen muchos intentos para que sea más preciso y confiable el análisis.
Si los dos saltan la misma valla, existen 4 posibilidades en el ensayo: que los dos tengan éxito; que los dos fracasen; que *Carlos* salte el obstáculo y *Alex* falle o que *Alex* salte el obstáculo y *Carlos* falle.

Los resultados después de realizar varios ensayos se resumen en la Tabla 2. Pero para poder comparar cuál de los dos es mejor, se deben descartar los resultados donde los dos fallan y los dos tienen éxito, ya que no proporcionan ninguna información sobre cuál tiene más habilidad que el otro, de manera que se dejan los intentos donde *Carlos* salta el obstáculo y *Alex* falla, y donde *Alex* salta el obstáculo y *Carlos* falla.

	Alex tiene éxito *X=1*	***Alex fracasa*** *X=0*
Carlos tiene éxito *X=1*	*Número de veces que Carlos tiene éxito y Alex también lo tiene*	*Número de veces que Carlos tiene éxito pero Alex no*
Carlos fracasa *X=0*	*Número de veces que Alex tiene éxito pero Carlos no*	*Número de veces que Carlos y Alex fracasan en el intento*

Tabla 2. Resultados de la comparación de Carlos con Alex saltando vallas.

Teniendo en cuenta que los resultados de ambos saltadores son independientes, se presentan en la Tabla 3 las probabilidades de los resultados posibles siendo P_{ni} la probabilidad de que *Carlos* tenga éxito en el salto y, por lo tanto $(1 - P_{ni})$ es la probabilidad de que fracase. Igualmente para *Alex* sería P_{mi} la probabilidad de que tenga éxito en el salto y $(1 - P_{mi})$ la probabilidad de fracaso.

	Alex tiene éxito *X=1*	***Alex fracasa*** *X=0*
Carlos tiene éxito *X=1*	$P_{ni} * P_{mi}$	$P_{mi} * (1 - P_{ni})$
Carlos fracasa *X=0*	$(1 - P_{mi}) * P_{ni}$	$(1 - P_{mi}) * (1 - P_{ni})$

Tabla 3. Matriz de probabilidad de posibles resultados.

Ahora bien, si N_{10} denota el número de veces que *Alex* salta y *Carlos* no puede y

N_{01} denota las veces que *Alex* fracasa y *Carlos* tiene éxito, la razón entre estas

frecuencias es una buena medida de la razón de habilidades entre estos dos saltadores. Si, además, se dividen estos valores entre el número total de intentos se obtiene una estimación de la razón de las probabilidades asociadas que por independencia se pueden escribir en la forma dada en la expresión (8), razón que mejor describe la superioridad de alguno de los competidores al saltar una valla determinada:

$$\frac{N_{10}}{N_{01}} = \frac{P_{mi}\,(1-P_{ni}\,)}{P_{ni}(1-P_{mi})} \quad \textbf{(8)}$$

Ahora se considera otra valla j y se repite el análisis realizado con la valla i. Al

asumir que la proporción de la habilidad de los dos saltadores es la misma en cualquier valla y no depende de su altura, se obtiene la siguiente igualdad:

$$\frac{P_{mi}(1-P_{ni})}{P_{ni}(1-P_{mi})} = \frac{P_{mj}(1-P_{nj})}{P_{nj}(1-P_{mj})}$$

Ahora para poder construir el modelo, se tiene, despejando adecuadamente que:

$$\left(\frac{P_{ni}}{1-P_{ni}}\right)=\left(\frac{P_{nj}}{1-P_{nj}}\right)\left(\frac{1-P_{mj}}{P_{mj}}\right)\left(\frac{P_{mi}}{1-P_{mi}}\right)$$

De aquí se selecciona a un competidor y a una valla tales que se tenga la misma probabilidad de éxito que de fracaso para saltar la valla. Se identifica este valla y este competidor con $j=0$ y $m=0$. Esto hace que en el lado derecho de la ecuación, el término del medio sea 1. Se garantiza la existencia de la persona 0 y la valla 0 asumiendo que tanto las medidas de habilidad de las personas como la de la dificultad de las vallas son no acotadas y que la probabilidad de éxito es creciente, continua y ocupa todo el intervalo *(0,1).* Se tiene, pues, la siguiente expresión

$$\left(\frac{P_{ni}}{1-P_{ni}}\right)=\left(\frac{P_{no}}{1-P_{no}}\right)\left(\frac{P_{0i}}{1-P_{0i}}\right)\left(\frac{1-P_{00}}{P_{00}}\right)$$

Pero como el término final de la formula da 1, queda:

$$\left(\frac{P_{ni}}{1-P_{ni}}\right)=\left(\frac{P_{no}}{1-P_{no}}\right)\left(\frac{P_{0i}}{1-P_{0i}}\right) \qquad \textbf{(9)}$$

Que escribimos en la forma:

$$\left(\frac{P_{ni}}{1-P_{ni}}\right)=f(n)*g(i) \qquad \textbf{(10)}$$

Se ve que el cociente entre la probabilidad que tiene un individuo n de responder bien a un cierto ítem i sobre la probabilidad de responderlo erradamente depende de dos funciones separadas que representan las siguientes características: $f(n)$ es una función de la habilidad del competidor n, y $g(i)$ es una función de la dificultad que representa la valla i. Sean ahora

$$f(n) = b_n \qquad y \qquad g(i) = {}^{1}/_{d_i}$$

donde se representa ${}^{1}/_{d_i}$ como una función de la dificultad del ítem y b_n como una función de la habilidad de la persona.

$$b_n = \left(\frac{P_{no}}{1-P_{no}}\right) \qquad \frac{1}{d_i} = \left(\frac{P_{0i}}{1-P_{0i}}\right)$$

Siguiendo el análisis del desarrollo del modelo, se puede ver que la habilidad de la persona n, está dada en términos de un cociente de probabilidad que depende de la probabilidad de que responda acertadamente al ítem 0 que se toma como referencia. Ahora bien, si no se tiene ninguna razón para decir que este ítem 0 no puede ser cualquiera, entonces cabe decir que la habilidad de la persona se refleja en cualquier ítem y que su probabilidad de acertar correctamente a determinado ítem, depende únicamente de su habilidad (cantidad de variable latente).

Se observa que la dificultad de la valla es independiente de la habilidad del saltador, y la habilidad del saltador es independiente de la dificultad de la valla, como se muestra en las dos funciones anteriores.

Aplicando la función logaritmo natural a ambos lados de la expresión *(9)* se obtiene la expresión

$$Ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = Ln\left[\left(\frac{P_{no}}{1-P_{no}}\right)\left(\frac{P_{0i}}{1-P_{0i}}\right)\right]$$

Ahora por propiedades de logaritmos se obtiene lo siguiente:

$$Ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = Ln\left(\frac{P_{no}}{1-P_{no}}\right) + Ln\left(\frac{P_{0i}}{1-P_{0i}}\right)$$

$$Ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = Ln\left(\frac{P_{no}}{1-P_{no}}\right) - Ln\left(\frac{1-P_{0i}}{P_{0i}}\right)$$

De manera que:

$$Ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = Ln(b_n) - Ln(d_i)$$

Sustituyendo $Ln(b_n) - Ln(d_i)$ se obtiene:

$$Ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i \qquad (11)$$

Y aplicando exponencial a ambos lados de la igualdad se tiene:

$$P_{ni} = \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)} \qquad (12)$$

Que es la probabilidad de que un saltador ***n***, tenga éxito saltando una valla ***i,*** donde se sustituyeron los valores

$$B_n = Ln\left(\frac{P_{no}}{1-P_{no}}\right) = Ln(b_n) \qquad (13)$$

Y teniendo en cuenta que:

$$\frac{1}{d_i} = \left(\frac{P_{0i}}{1-P_{0i}}\right)$$

Se obtiene que:

$$D_i = Ln\left(\frac{1-P_{0i}}{P_{0i}}\right) = Ln(d_i) \qquad (14)$$

Se identifica B_n con la habilidad del competidor n y D_i con la dificultad de la valla i; ambas medidas están dadas en lógitos, que no es otra cosa que los logaritmos de los cocientes de probabilidad dados en (13) y (14).

Para probar que esta es una escala de intervalos supongamos que existen dos personas n y m, de tal forma que la habilidad de uno es igual a la del otro más uno, así:

$$B_m = B_n + 1$$

Ahora, al analizar las diferencias en logit, se tiene:

$$Ln\left(\frac{P_{mi}}{1-P_{mi}}\right) - Ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = (B_m - D_i) - (B_n - D_i) =$$

$$B_m - B_n = (B_n + 1) - B_n = 1$$

Esto se puede escribir por propiedades de los logaritmos como:

$$Ln\left(\frac{P_{mi}/(1-P_{mi})}{P_{ni}/(1-P_{ni})}\right)=1$$

Entonces al aplicar exponencial a ambos lados de la igualdad anterior se obtiene:

$$\frac{P_{mi}/(1-P_{mi})}{P_{ni}/(1-P_{ni})}=e$$

Esto nos dice que la diferencia en una unidad en la habilidad de estos individuos (cantidad de variable latente) es una diferencia de una unidad en *log odds ratio (logaritmo de la razón de discrepancia)*, es decir, la razón de discrepancia de la habilidad de los dos individuos asociado a cualquier ítem es ***e***. Ahora, si se generaliza el evento y se supone que la diferencia entre los dos individuos es K, se puede obtener la razón de discrepancia entre los dos para una diferencia cualquiera K:

$$Ln\left(\frac{P_{mi}}{1-P_{mi}}\right)-Ln\left(\frac{P_{ni}}{1-P_{ni}}\right)=(B_m-D_i)-(B_n-D_i)=$$

$$B_m-B_n=(B_n+K)-B_n=K$$

Se obtiene la forma general:

$$Ln\left(\frac{P_{mi}/(1-P_{mi})}{P_{ni}/(1-P_{ni})}\right)=K$$

Y al aplicar exponencial a ambos lados de la igualdad se tiene:

$$\frac{P_{mi}/(1-P_{mi})}{P_{ni}/(1-P_{ni})}=e^{K}$$

El modelo Rasch permite comparar no solo dos personas n y m con una valla i, sino también comparar dos vallas i y j con una persona n y obtener el mismo resultado para la probabilidad de éxito de una persona n, al saltar una valla determinada i. Se puede ver, que el resultado es exactamente el mismo, ya que

la diferencia de habilidad entre dos personas en el atributo no depende de los ítems con que sean medidos, e igualmente, la diferencia entre dos ítems no depende de las personas que se hayan utilizado para cuantificarlos. En consecuencia, si los datos están ajustados al modelo, las comparaciones de las personas son independientes de los ítems administrados y las estimaciones de los parámetros de los ítems no estarán influenciadas por la distribución de la muestra utilizada para la calibración.

Ya que se pueden predecir los resultados acertados de un individuo ante un ítem valiéndose de probabilidades, es lógico pensar en la respuesta incorrecta, es decir, predecir la probabilidad que tiene dicha persona de responder incorrectamente determinado ítem.

Basta con calcular:

$$1 - P_{ni} = 1 - \left(\frac{\exp\,(B_n - D_i)}{1 + \exp\,(B_n - D_i)}\right)$$

Al efectuar operaciones se obtiene:

$$q_{ni} = \frac{1}{1 + \exp\,(B_n - D_i)}$$

Donde $q_{ni} = 1 - P_{ni}$ es la probabilidad de que la persona n responda

incorrectamente el ítem i.

3.2.4. Curva característica de un ítem. La expresión (12) cuando se asume que la dificultad D_i del ítem es constante, se conoce como la curva característica del ítem. Un ejemplo de curva característica se presenta en la Figura 11.

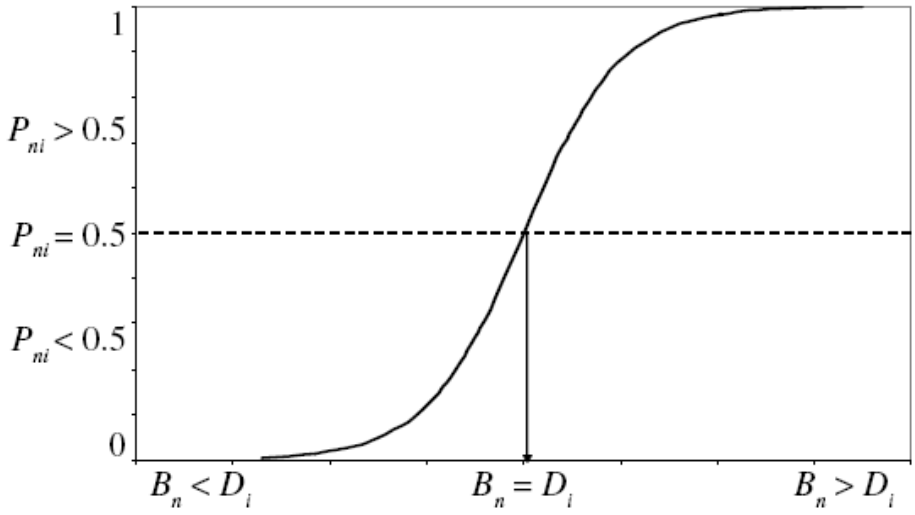


Figura 11. Curva característica de un ítem en el modelo Rasch.

Esta curva da la probabilidad que una persona de cierta habilidad responda correctamente al ítem i y depende de la diferencia entre la habilidad del individuo y la dificultad del ítem. Existen tres casos claramente diferenciados:

$(B_n - D_i)$ >0 arroja el valor de **P_{ni} > 0.5**

$(B_n - D_i)$ = 0 arroja el valor de **P_{ni} = 0.5**

$(B_n - D_i)$ <0 arroja el valor de **P_{ni} < 0.5**

La función que define la curva está dada por la siguiente expresión:

$$P_{ni} = \frac{\exp(B - D_i)}{1 + \exp(B - D_i)}$$

El nivel de dificultad del ítem está relacionado con la habilidad que se necesita para que la probabilidad de éxito de los participantes sea $1/2$, es decir, la intersección de la CCI con la recta horizontal en $P = 1/2$ identifica el punto del nivel de dificultad en el eje X, así que cuanto más grande es el valor de ese punto de intersección que se refleja en el eje X, mayor es la habilidad requerida para contestar bien dicho ítem; estos valores suelen situarse entre -2 y 2.

En la Figura 12 se presenta un gráfico con 4 ítems, en los que el nivel de dificultad de cada uno de los ítems son: b = -1, b = 0, b = 1, y b = 2, y donde $\boldsymbol{b}$ representa el

nivel de dificultad. Las curvas características de los ítems en un modelo Rasch son paralelas entre sí.

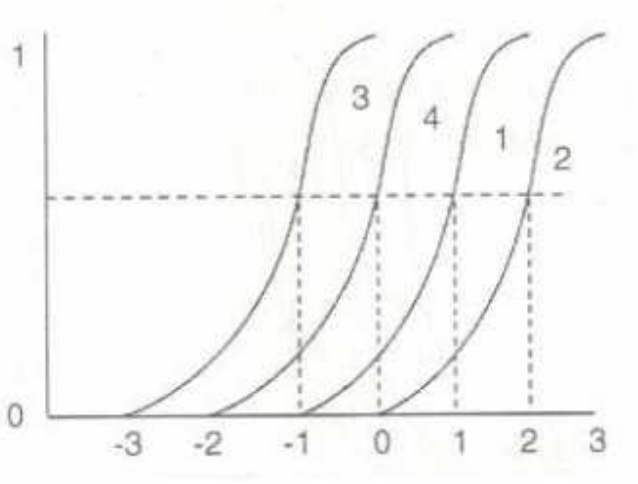


Figura 12. Curvas características de cuatro ítems.

3.2.5. Método de estimación de parámetros. Como este modelo se aplica a partir de datos observados, y no es aplicable para estimaciones probabilísticas antes de la evaluación del test, el método apropiado es el estimador de máxima verosimilitud siendo el procedimiento más justificable desde el punto de vista matemático "Este es un método a posteriori que da por resultado estimaciones de los parámetros que con mayor probabilidad habrían producido los patrones de respuesta observados en los datos" (Montesinos, 2007). La estimación de dificultad de ítems y habilidad de los individuos, se denomina calibración de ítems y habilidad de individuos por el modelo Rasch.

3.2.6. Estimador de máxima verosimilitud. Técnicamente el objetivo primario de administrar un test consiste en estimar los parámetros de las personas y de los ítems en una variable de interés. Muchas veces se tiene alguno de estos conjuntos de parámetros y usualmente se estiman los parámetros de la habilidad de las personas, conociendo los parámetros de los ítems que son obtenidos en

pruebas aplicadas anteriormente. A este procedimiento se le da el nombre de estimación condicional, pero cuando se desconocen los parámetros de las personas y los ítems, este procedimiento se llama estimación conjunta.

La lógica más frecuente es denominada de máxima verosimilitud y consiste en determinar qué parámetros hacen más probables las respuestas observadas.

En este trabajo el procedimiento usado es la estimación condicional de los parámetros de las personas, conocidos los parámetros de los ítems, y consiste en un proceso de búsqueda que calcula la probabilidad conjunta de las respuestas observadas a los ítems para cada puntuación. Ahora se asigna un valor θ más probable para el patrón de respuesta a cada persona, y este valor es denominado estimador de máxima verosimilitud.

Para entrar en la formalidad matemática del proceso de máxima verosimilitud, se empezará con un ejemplo de un evento usando la función de la probabilidad binomial.

Ejemplo: Suponiendo que una urna contiene bolas blancas y negras, y existe una relación de 3 a 1 entre los dos colores, pero se desconoce cuál de los dos colores tiene mayor número de bolas. Se necesita saber cuál es la probabilidad de extraer una bola blanca, si se tienen dos opciones: ¼ o ¾ ¿Cuál de las dos probabilidades es más posible?

Se realizan tres extracciones en la urna y se relacionan los resultados obtenidos como se ilustra en la Tabla 4 y con la función que representa este evento que es la binomial, para tomar así, una decisión entre las dos probabilidades y definir cuál de las dos concede más verosimilitud a los resultados obtenidos.

Número de blancas	**0**	**1**	**2**	**3**
Probabilidad= 3/4	1/64	9/64	27/64	27/64
Probabilidad= 1/4	27/64	27/64	9/64	1/64

Tabla 4. Probabilidad de extracción de una bola blanca.

$$f(x;n) = \binom{x}{n} p^x (1-p)^{n-x} \qquad para\ x = 1, 2, 3, \dots, n$$

Se observa que los resultados que son más probables se dan cuando:

$$p = p(x) = \begin{cases} 1/4 & para\ x = 0, 1 \\ 3/4 & para\ x = 2, 3 \end{cases}$$

Es decir, se obtiene que si se obtuvieron una o ninguna bola blanca, el valor de probabilidad que le concede mayor probabilidad a estos sucesos en ¼; si, en cambio, el número de blancas fueron 2 o 3, la mejor probabilidad de obtener estos sucesos es ¾.

Para deducir estos resultados en forma analítica, tomamos la función binomial en función de p:

$$f(x;n) = \binom{x}{n} p^x (1-p)^{n-x}$$

Como se trata de hallar el valor de *p* que genere el mayor valor de esta función, se deriva y se iguala a 0, para hallar los valores críticos de la función con respecto a p:

$$\frac{df}{dp} = \binom{n}{x} x p^{x-1} (1-p)^{n-x} - \binom{n}{x} p^x (n-x)(1-p)^{n-x-1}$$

Entonces:

$$\frac{df}{dp} = p^{x-1}(1-p)^{n-x-1}[x(1-p) - p(n-x)]$$

Igualando a 0 se tiene:

$$p^{x-1}(1-p)^{n-x-1} = 0 \quad Y \quad [x(1-p) - p(n-x)] = 0$$

Se evalúa el término de la derecha:

$$x - xp - pn + px = 0$$

$$x = pn$$

$$\widehat{p} = \frac{x}{n}$$

De esta forma se obtiene que el estimador de máxima verosimilitud de p es la

media muestral de las bolas blancas obtenidas, es decir, el promedio entre el número de éxitos x y el número total de extracciones n.

Pero para generalizar este argumento, se toma cualquier muestra aleatoria $X_1, X_2,$

$..., X_n$ con densidad $f(X;\theta)$. La función de verosimilitud de X es la densidad

conjunta de la muestra, es decir:

$$L(\theta;\ X1, X2, \dots, Xn) = L(\theta;\ x_1)L(\theta; x_2)L(\theta; x_3) \dots L(\theta; x_n)$$

Ahora, el estimador que maximiza la función de verosimilitud se encuentra derivando la función e igualándola a cero para hallar sus puntos críticos:

$$\frac{dL}{d\theta} = 0$$

Para una cantidad de parámetros k se hacen las derivadas parciales con respecto a cada uno de los parámetros, se igualan a cero cada una de las ecuaciones y se resuelve el sistema de ecuaciones resultante, para hallar el valor de cada parámetro θ.

.

Ahora se consideraran los estimadores en las respuestas de un grupo de individuos a un conjunto de ítems con el modelo Rasch. Teniendo los datos de los ítems con un grupo de personas se tienen diferentes casos hasta llegar al general.

Para una persona de una habilidad θ y un ítem i se tendría:

$$P(x_i|\theta) = P(x_i = 1|\theta)^{x_i} P(x_i = 0|\theta)^{1-x_i}$$

Para una persona de una habilidad θ y n items:

$$L(x_1 + x_2 + x + \cdots + x_n|\theta) = \prod_{i=1}^{n} P(x_i = 1|\theta)^{x_i} P(x_i = 0|\theta)^{1-x_i}$$

Para N personas y n ítems:

$$L(X|\theta) = \prod_{j=1}^{N} \prod_{i=1}^{n} P(x_{ij} = 1|\theta)^{x_{ij}} P(x_{ij} = 0|\theta)^{1-x_{ij}}$$

Aplicando logaritmo a ambos lados, y sabiendo que el logaritmo de un producto es la suma de los logaritmos, se tiene que:

$$\Delta(\theta) = lnL(X|\theta) = \sum_{j=1}^{N} \sum_{i=1}^{n} [x_{ij} lnP_{ij}(\theta) + (1 - x_{ij}) ln(1 - P_{ij}(\theta))]$$

Donde:

$$P_{ij}(\theta) = P(x_i|\theta_j) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)}$$

Se aplica el logaritmo para minimizar la dificultad del las operaciones ya que se necesita derivar la función $\Delta(\theta)$ para hallar los estimadores de máxima verosimilitud.

Se deriva la función de verosimilitud con respecto a θ_j:

$$\Delta'(\theta) = \frac{\partial ln[L(x_{ij}|\theta_j)]}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial lnL}{\partial \theta_j} * \frac{\partial P_{ij}}{\partial P_{ij}} = \sum_{i=1}^{n} \frac{\partial lnL}{\partial P_{ij}} * \frac{\partial P_{ij}}{\partial \theta_j} = 0$$

Se multiplica por $\partial P_{ij} / \partial P_{ij}$ para poder obtener la derivada con respecto a P_{ij} y llegar al objetivo:

$$\Delta'(\theta) = \sum_{i=1}^{n} \left[\frac{x_{ij}}{P_{ij}} - \frac{1 - x_{ij}}{1 - P_{ij}} \right] \frac{\partial P_{ij}}{\partial \theta_j} = \sum_{i=1}^{n} \left[\frac{x_{ij} - P_{ij}}{P_{ij}(1 - P_{ij})} \right] \frac{\partial P_{ij}}{\partial \theta_j}$$

Se observa que:

$$\mathrm{P_{ij}}(\theta) = \mathrm{P}(\mathrm{x_i}|\theta_\mathrm{j}) = \frac{\exp\left(\theta_\mathrm{j} - \beta_\mathrm{i}\right)}{1 + \exp\left(\theta_\mathrm{j} - \beta_\mathrm{i}\right)} = \exp(\theta_\mathrm{j} - \beta_\mathrm{i})[1 + \exp\left(\theta_\mathrm{j} - \beta_\mathrm{i}\right)]^{-1}$$

$$\frac{\partial P_{ij}}{\partial \theta_j} = \exp(\theta_\mathrm{j} - \beta_\mathrm{i})[1 + \exp\left(\theta_\mathrm{j} - \beta_\mathrm{i}\right)]^{-1} +$$

$$(-1)[1 + \exp\left(\theta_\mathrm{j} - \beta_\mathrm{i}\right)]^{-2} \exp(\theta_\mathrm{j} - \beta_\mathrm{i}) \exp(\theta_\mathrm{j} - \beta_\mathrm{i})$$

Como fracciones queda:

$$\frac{\partial P_{ij}}{\partial \theta_j} = \frac{exp(\theta_j - \beta_i)}{1 + exp\left(\theta_j - \beta_i\right)} - \left(\frac{exp(\theta_j - \beta_i)}{1 + exp\left(\theta_j - \beta_i\right)} \right)^2$$

Factorizando la expresión se tiene:

$$\frac{\partial P_{ij}}{\partial \theta_j} = \left[\frac{exp(\theta_j - \beta_i)}{1 + exp\left(\theta_j - \beta_i\right)} \right] \left[1 - \frac{exp(\theta_j - \beta_i)}{1 + exp\left(\theta_j - \beta_i\right)} \right]$$

Reemplazando en la ecuación anterior se obtiene:

$$\frac{\partial P_{ij}}{\partial \theta_j} = P_{ij}(1 - P_{ij})$$

Y se puede seguir con el resultado de la derivada, de manera que:

$$\Delta'(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[\frac{x_{ij} - P_{ij}}{P_{ij}(1 - P_{ij})}\right]P_{ij}(1 - P_{ij})$$

$$\Delta'(\boldsymbol{\theta}) = \sum_{i=1}^{n}[x_{ij} - P_{ij}] = 0$$

Igualando a cero la derivada, se consiguen los valores de máxima verosimilitud de la función, y estos resultados se encuentran despejando para cada parámetro:

$$\sum_{i=1}^{n} x_{ij} = \sum_{i=1}^{n} P(\boldsymbol{\theta})_{ij} = \sum_{i=1}^{n} E_{ij}$$

El valor del puntaje esperado de una persona de cierta habilidad debe ser equivalente al obtenido en la muestra (curva característica del ítem), esto permite que los estimadores de verosimilitud sean máximos, y que se logre una cercanía entre lo observado y lo esperado.

A la hora de obtener los valores de las estimaciones de la dificultad de los ítems se debe tener en cuenta que estos dependen de la cantidad de personas que respondieron bien los ítems y de la habilidad de las personas que son evaluadas, pero también, la habilidad de las personas depende de los ítems bien respondidos y de todas las dificultades de los ítems; por lo tanto, si se excluye algún ítem en el análisis, la estimación de las habilidades cambia, y eliminar cualquier evaluado, cambiaria la estimación de la dificultad de los ítems.

Ahora, para hacer la segunda derivada de la función de verosimilitud, se utiliza el siguiente resultado para minimizar los cálculos:

$$r_j = \sum_{i=1}^{n} x_{ij}$$

Se escribe la función de máxima verosimilitud:

$$\Delta'(\theta) = \sum_{i=1}^{n} [x_{ij} - P_{ij}] = 0$$

Al reemplazar se tiene:

$$\Delta'(\theta) = r_j - \sum_{i=1}^{n} P_{ij} = 0$$

Derivando $r_j = 0$ ya que es constante, queda la derivada de la sumatoria:

$$\Delta''(\theta) = -\sum_{i=1}^{n} P_{ij}(1 - P_{ij}) = 0$$

Este resultado ayuda en la estimación de las habilidades de las personas, ya que gracias a Newton-Raphson se tiene una fórmula que facilita esta estimación:

$$\theta_{j+1} = \theta_j - \frac{\Delta'(\theta)}{\Delta''(\theta)}$$

El proceso de estimación de habilidad es sumamente largo y por eso se utilizan programas de ordenador para hacer los cálculos; en este trabajo se utilizó el programa WINSTEPS®[3].

Con el siguiente algoritmo que proporciona Newton-Raphson se puede estimar la habilidad del individuo evaluado:

[3] *Rasch software más altamente usado y diseñado para medición de las personas y objetos. MCQ, escalas de valoración, parcial de crédito. El tamaño de los datos: de 30.000 artículos por 10.000.000 personas.*

$$\theta_g^{(t+1)} = \theta_g^{t} + \frac{r_g - \sum_{i=1}^{n} \hat{P}_{ig}}{\sum_{i=1}^{n} \hat{P}_{ig}\left(1 - \hat{P}_{ig}\right)} \quad para\ g = 1, 2, \ldots, n-1$$

Ahora es bueno recordar que si un individuo responde correctamente a la totalidad de los ítems, no se puede estimar nada de su habilidad ya que eso significa que el test es de bajo nivel para esta persona, y si un individuo no responde ningún ítem correctamente, tampoco se puede estimar su habilidad, ya que no se tiene ninguna información sobre su capacidad; por tanto, los resultados de un test con una cantidad *n* de ítems, debe suministrar información cuando los puntajes están entre *1* y *n-1* respuestas correctas, por esto g en el subíndice, va desde *1* hasta *n-1.*

De forma similar a la anterior, la estimación de la dificultad de los ítems se puede obtener con:

$$\beta_i^{(k+1)} = \beta_i^{k} + \frac{s_i - \sum_{g=1}^{n-1} f_g \hat{P}_{ig}}{\sum_{g=1}^{n-1} f_g \hat{P}_{ig}\left(1 - \hat{P}_{ig}\right)} \qquad para\ i = 1, 2, \ldots, n$$

Este método estima los valores de los parámetros de habilidad y dificultad juntos, pero no simultáneamente, y en cada etapa el método utiliza un procedimiento

iterativo llamado Newton-Raphson[4], y empieza usando un método de máxima verosimilitud para estimar los parámetros de las personas conocidos los parámetros de los ítems, y estima los parámetros de los ítems conocidos los parámetros de los sujetos.

Donde los valores iniciales de los parámetros β_i y θ_g para la implementación del modelo son:

$$\beta_i^0 = log\left[\frac{N - s_i}{s_i}\right] \quad y \quad \theta_g^0 = log\left[\frac{r_g}{n - r_g}\right]$$

Estas estimaciones se estandarizan y se aplica Newton-Raphson para hallar nuevos estimadores de dificultad de cada ítem, por lo menos 10 iteraciones o hasta que la diferencia entre ellas sea menos de 0.05. Todos los estimadores se reubican teniendo en cuenta su media de tal manera que el centro de los mismos vuelva a ser 0. Luego se aplica de nuevo Newton-Raphson pero esta vez a las habilidades efectuando 5 iteraciones o hasta que la diferencia entre ellas sea menor que 0.05.

El procedimiento descrito en el párrafo anterior se repite hasta que la media de las diferencias entre los estimadores de dificultad sea menor que 0.025. Los estimadores de dificultad se multiplican por (n-1)/n y los estimadores de habilidad por (n-2)/(n-1) para corregir sus sesgos.

Las desviaciones estándar de los estimadores que arroja el mismo procedimiento de máxima verosimilitud están dadas por:

[4] *El método de Newton-Raphson es un algoritmo eficiente para encontrar aproximaciones de los ceros o raíces de una función real. También puede ser usado para encontrar el máximo o mínimo de una función, encontrando los ceros de su primera derivada.*

$$DE(\beta_i) = \frac{1}{\sqrt{\sum_{g=1}^{n-1} f_g P_{ig} Q_{ig}}}$$

$$DE(\theta_g) = \frac{1}{\sqrt{\sum_{i=1}^{n} P_{ig} Q_{ig}}}$$

El valor dentro de la raíz se llama función de información del test. Debido a que el error típico de la medida es una función inversa de la de la información del test, esta particularidad es de significado parecido al de la fiabilidad en la teoría clásica.

3.2.7. Ajuste de los datos al modelo. Los procedimientos de análisis que permiten detectar los ítems y las personas que no se ajustan al modelo en este trabajo están basados en los residuos (diferencias entre las respuestas observadas y las esperadas). La fórmula del residuo es:

$$Y_{ni} = X_{ni} - P_{ni}$$

donde X_{ni} es la respuesta observada del participante n al ítem i, sea $X_{ni} = 0$ o $X_{ni} = 1$, ya que el estudio es dicotómico; P_{ni} es la respuesta esperada del participante n al ítem i, que se obtiene en valores continuos de 0 a 1 porque es una probabilidad.

Al dividir por su desviación estándar se obtiene el residuo estandarizado para cada observación X_{ni}:

$$Z_{ni} = \frac{X_{ni} - P_{ni}}{\sqrt{P_{ni}(1 - P_{ni})}}$$

Este residuo es una medida de las irregularidades observadas en cada pregunta o persona sometida al test.

Ahora bien, si los datos se ajustan al modelo, estos residuales adquieren una distribución normal con una media de 0 y una varianza de 1:

$$Z_{ni} \approx N(0, 1)$$

Se analizará qué pasa con los residuos estandarizados en ítems con respuesta dicotómica.

Los residuales son una medida de las anomalías observadas en cada ítem o persona en el análisis, y se convierten en una medida de bondad de ajuste de los datos al modelo.

A continuación se analizará cómo se comportan estos residuos estandarizados en los tipos de respuesta dicotómica.

a) Cuando la respuesta es errada, es decir cuando $X = 0$, este residuo se convierte en:

$$Z_0 = \frac{-P}{\sqrt{(P(1-P))}} = -\left[\frac{P}{1-P}\right]^{1/2}$$

Como se observa en la expresión anterior, el residuo se hace mayor cuanto más cercano a 1 se encuentra el valor esperado de respuesta, es decir, se obtienen

residuos cada vez más altos entre mayor sea la diferencia entre la habilidad del individuo y la dificultar del ítem siendo mayor la habilidad del individuo. El residuo pone en evidencia la situación extraña de un individuo muy capaz respondiendo erradamente un ítem fácil.

b) Cuando la respuesta es correcta, es decir cuando $\boldsymbol{X = 1}$ este residuo se convierte en:

$$\boldsymbol{Z_1 = \frac{1 - P}{\sqrt{(P(1 - P))}} = \left[\frac{1 - P}{P}\right]^{1/2}}$$

Esta expresión se crece cuando la probabilidad de respuesta es muy cercana a 0, es decir, cuando la habilidad del individuo es mucho menor que la dificultad del ítem. En este caso, se detectan las anomalías representadas en individuos poco capaces respondiendo acertadamente ítems difíciles.

Con el análisis anterior, se tiene una certeza de la manera como el modelo construye valores que identifican anomalías en las respuestas de los estudiantes, pues para cada predicción del modelo se espera cierta respuesta que esté cerca del pronóstico.

Para evitar la cancelación de términos al sumar los residuos para los ítems o los individuos, se elevan los residuos al cuadrado y se obtienen los residuos estandarizados al cuadrado que ahora siguen una distribución chi cuadrado.

$$\boldsymbol{Z_{ni}{}^2 = \frac{(X_{ni} - P_{ni})^2}{P_{ni}(1 - P_{ni})}}$$

3.2.7.1. *Estadísticos de ajuste*. Existen dos tipos de estadísticos utilizados para medir el ajuste de los datos al modelo, éstos son, el outfit y el infit.

- El **OUTFIT.** Es el promedio de los residuales estandarizados. Permite detectar respuestas irregulares en la parte externa de la distribución, es decir, cuando existan anomalías lejos del nivel de habilidad de las personas o lejos del nivel de dificultad de los ítems, tal como se comentó que hacen los residuos estandarizados. Su expresión algebraica para un ítem *i* es la siguiente:

$$OUTFIT = \sum Z_{ni}^{2} / N$$

donde la sumatoria se realiza sobre los individuos *n* y N es el total de individuos analizados. Una expresión semejante se obtiene para un individuo *n* cualquiera realizando la suma sobre los ítems aplicados y dividiendo por el número de ellos.

- El **INFIT.** Es la media cuadrática de residuales ponderada, es vulnerable a respuestas inesperadas según el modelo cerca del nivel de habilidad de las personas y del nivel de dificultad de los ítems, es decir, captura anomalías internas en las distribuciones tanto de personas como de ítems.

$$INFIT = \sum Z_{ni}^{2} W_{ni} / N$$

Donde N es el número de individuos observados sumados y W_{ni} las varianzas individuales que evitan la intervención de los comportamientos lejos del nivel de habilidad o de dificultad, es decir, disminuyen su alteración por respuestas externas.

3.2.7.2. *Criterios.* En la práctica existen dos criterios que se deben tener en cuenta para el ajuste de los datos al modelo, tanto para el INFIT como para el OUTFIT, y son el MNSQ (estadístico de media cuadrática) y el ZSTD (estadístico de media cuadrática estandarizado).

- El **INFIT MNSQ** tiene un valor ideal de 1, que es lo sugerido por el modelo para un ajuste perfecto de los datos, la tolerancia que se maneja en este estadístico depende de la cantidad de casos estudiados. Si la muestra tiene menos de 500 participantes, el valor máximo para un excelente ajuste deberá ser de 1.3, si la muestra está entre 500 y 1000 casos, el valor ideal máximo deberá ser de 1.2 y para más de 1000, deberá ser de 1.1; pues cuanto más grande es la muestra, la predicción del modelo se agudiza, esto pasa igual con los valores del OUTFIT MNSQ.

  Valores mayores a 1.3 indican aleatoriedad en los datos, es decir, falta de ajuste, representando irregularidades en las respuestas cerca del nivel de habilidad o dificultad; cuanto más se aleje el estadístico de 1.3, el desajuste crece. Valores menores que 1, muestran buen ajuste hasta 0.8, pero a medida que disminuye este valor, el ajuste empieza a verse más perfecto y a mostrar dependencia ya que el patrón de respuestas es muy predecible.

- El **INFIT ZSTD** es el estadístico de media cuadrática estandarizado con distribución normal estándar, que percibe desajuste cuando estos valores se salen del intervalo entre -2 y 2 medido en lógitos; el valor ideal es 0 que revela un ajuste estandarizado perfecto. Al estandarizar estos valores se obtiene ganancia en el análisis del ajuste, pues se pueden estudiar las características de la distribución normal para el ítem o la persona. (González, 2008)

Para tomar decisiones en la exclusión de datos en los análisis, se debe tener en cuenta que cuando los MNSQ presentan desajustes menores que 1, se recomienda no sacar estos datos a menos que la intención sea minimizar la cantidad de preguntas en el examen, o disminuir la cantidad de evaluados en el proceso. Los ZSTD se deben tener en cuenta cuando se quiere salvar un dato con un MNSQ mayor que 1.5, cuando la cantidad de preguntas o de personas que contengan este dato sea muy corta y sea esencial contar con dicho dato, en el caso de INFIT MNSQ. Por ejemplo, si se tiene un examen con pocas preguntas y la exclusión de alguna de ellas deja incompleto el tema a evaluar, en el caso que dicha pregunta tenga un INFIT MNSQ mayor a 1.5, se mira el valor del ZSTD con el fin de salvar este dato, ya que es necesario en la medición del parámetro que se desea estimar.

- El **OUTFIT MNSQ** también tiene un valor ideal de 1, y maneja de igual forma el intervalo entre 0.8 y 1.3, los valores muy superiores a 1.3 indican desajuste en los datos analizados. Valores menores que 0.8 indican más ajuste perfecto de los datos, esto es, un patrón de respuestas predecible y aunque parece aportar en la medida, no lo hace.

- El **OUTFIT ZSTD** indica un ajuste razonable cuando obtiene valores que estén dentro del intervalo entre -2 y 2 medido en lógitos; fuera de este intervalo, representa desajuste tanto de ítems como de individuos.

Los ZSTD se deben tener en cuenta cuando se quiere salvar un dato con un MNSQ mayor que 1.5, cuando la cantidad de preguntas o de personas que contengan este dato sea muy corta y sea esencial contar con dicho dato, en el caso de OUTFIT. Por ejemplo, si se tiene un examen con pocas preguntas y la exclusión de alguna de ellas deja incompleto el tema a evaluar, en el caso que dicha pregunta tenga un OUTFIT MNSQ mayor a 1.5, se mira el valor del ZSTD

con el fin de salvar este dato, ya que es necesario en la medición del parámetro que se desea estimar.

El significado de la variación que tienen los datos con respecto a lo que predice el modelo se interpreta de la siguiente manera para los MNSQ:

Con un valor MNSQ de 1+ B, se tiene un 100% multiplicado por B más de variación entre los datos observados y lo predicho por el modelo, es decir, si se tiene un MNSQ de 1.4 esto significa un 40% más de variación entre el patrón de datos observado y el patrón de respuestas predicho por el modelo. De igual forma pasa si los valores son menores que 1, ya que se interpretaría como menos variación. Con un valor MNSQ de 1- B, se tiene un 100% multiplicado por B menos de variación, por ejemplo, un MNSQ de 0.70 quiere decir, un 30% menos de variación que la elaborada por el modelo, esto tiende a ajustarse bien. (González, 2008).

Un punto importante es que los estadísticos INFIT y OUTFIT son independientes porque evalúan el ajuste en partes distintas de las distribuciones de personas o de ítems. Puede pasar que se obtenga un valor INFIT que represente ajuste y un OUTFIT que no lo haga en lo absoluto, y viceversa.

Con los criterios anteriores se puede analizar un conjunto de datos con el modelo Rasch y obtener resultados satisfactorios.

4. ESTUDIO DE EVALUACIÓN

Uno de los objetivos específicos de esta investigación es hacer un paralelo entre la Teoría Clásica de Test (TCT) y el Modelo Rasch, realizando un análisis comparativo con los resultados obtenidos en las dos teorías, para ofrecer una mayor comprensión de las diferencias y similitudes que éstos presentan.

En este capítulo se considerará inicialmente la teoría clásica de Test (TCT). Para ello se observará cada uno de los ítems de selección múltiple desde diferentes puntos de vista: el conteo de respuestas, los porcentajes de cada una de las opciones, el análisis conceptual de cada una de las respuestas, los diferentes índices de dificultad y discriminación (sin tener en cuenta el índice de validez, ya que no se posee una muestra anterior que haya realizado la prueba) para que facilite detectar las posibles deficiencias o razones en las respuestas de los estudiantes al test.

Los ítems de respuesta abierta se analizarán para completar el estudio, aprovechando las frecuencias de cada una de las calificaciones y utilizando estos resultados para ingresarlos en el análisis.

4.1. MÉTODO CLÁSICO (TCT)

En este estudio se tienen en cuenta todas las respuestas dadas por los participantes, con el ánimo de poder identificar las concepciones de los estudiantes respecto a los aspectos conceptuales que encierra cada ítem.

4.1.1. ANÁLISIS DE ÍTEMS

Se presenta a continuación el análisis clásico de las respuestas dadas por los estudiantes a cada ítem del cuestionario, teniendo en cuenta la cantidad de respuestas en cada una de las opciones y sus porcentajes, tanto en los ítems dicotómicos como en los politómicos.

4.1.1.1. *Análisis de ítems de opciones múltiples*

Ítem 1: *El intervalo de confianza del 50% para media de una población μ es:*

a) *El rango dentro del cual caen el 50% de los valores de la media de la muestra* $\bar{x}$.
b) *Un intervalo más ancho que el intervalo de confianza del 95%.*
c) *Un intervalo de valores calculado a partir de los datos de la muestra. En el 50% de las muestras de una población, el intervalo calculado contiene a la media de la población.*
d) *Dos veces más ancho que el intervalo de confianza del 100%.*

Opciones	*Frecuencias*	*%*
a	*66*	*40.3*

b	*8*	*4.9*
c	***88***	***53.6***
d	*2*	*1.2*
No responde	*0*	*0*
Total	*164*	*100*

Tabla 5. .Frecuencias y porcentajes de las opciones del ítem 1.

Índice de dificultad: 0,54 Índice de discriminación: 0,33

Este ítem busca evaluar la comprensión que tiene el estudiante de la definición de intervalo de confianza y su variación al calcular diferentes intervalos tomando muestras de la misma población. Los estudiantes respondieron en su mayoría la opción correcta que es la c): el 53.6% de ellos, correspondiente a 88 de los 164 evaluados. El 40.3% de los estudiantes, equivalente a 66, respondieron la opción a), este porcentaje de individuos puede que confundan el estadístico con el parámetro, al pensar que el intervalo se calcula para estimar la media muestral y no la poblacional. El 6.1% restante, 10 estudiantes, al responder la opción b) o d), creen que el ancho del intervalo aumenta mientras disminuye su coeficiente de confianza, como se observa en la Tabla 5; no tiene claro el efecto del nivel de confianza sobre la anchura del intervalo. Todos los estudiantes respondieron este ítem.

Este ítem presenta un índice de dificultad de 0,54 mostrando un grado de dificultad medio, ya que el 53,6% de respuestas correctas, señalan que gran parte del grupo domina el tema evaluado, y un índice discriminador de 0,33 el cual brinda confiabilidad en la evaluación del tema.

Al observar los resultados obtenidos en Olivo se encuentran algunas similitudes: el índice de dificultad es de 0.55, la opción más seleccionada es la c), que es la respuesta correcta con un porcentaje de 55.6%, la opción d) fue muy poco seleccionada y el número de respuestas en blanco es cero. En cuanto a la selección de las otras respuestas, la opción a) con 26% y la opción b) el 16%, contra un 40.3% y 5% respectivamente que señala este estudio. (Ver Cuadro 1).

El índice discriminador en la prueba piloto de Olivo es de 0.44[5] porcentaje con poca diferencia al 0.33 que devuelve este estudio.

Ítem 2: *Comparado con los intervalos de confianza calculados en muestras de tamaño n=4 en una población normal, el ancho de los intervalos de confianza de la media de la población calculado en muestras de tamaño n=50:*

a) *Variará más que los anchos de los intervalos para muestras de tamaño n=4.*
b) *Variará, pero no tanto como lo hicieron los anchos de los intervalos para muestras de tamaño n=4.*
c) *Tomarán valores parecidos.*

Opciones	*Frecuencias*	*%*
a	*57*	*34.7*
b	***80***	***48.8***
c	*27*	*16.5*
No responde	*0*	*0*
Total	*164*	*100*

Tabla 6. Frecuencias y porcentajes de las opciones del ítem 2.

Índice de dificultad: 0,49 Índice de discriminación: 0,44

[5] *Valor tomado de la prueba piloto de Olivo, ya que los índices discriminadores de la prueba final no los muestra al igual que los resultados de cada examen para poderlos hallar*.

Este ítem permite evaluar el conocimiento de los estudiantes frente al concepto: el ancho *de los intervalos de confianza disminuye cuando aumenta el tamaño de la muestra con desviación típica conocida o desconocida.* La opción que más respondieron los estudiantes fue la b), opción correcta, con un porcentaje del 48.8%. El 34.7% de los estudiantes respondieron la opción a) posiblemente creen que si se aumenta el tamaño de la muestra la variación en el ancho de los intervalos de confianza aumenta, pero es al contrario, la mayor variación se da cuando la muestra es más pequeña. El 16.5% de los estudiantes que respondieron a la opción c) pueden pensar que el ancho no se verá afectado, dando a entender que no relacionan el tamaño de la muestra con el ancho del intervalo de confianza (Ver Tabla 6).

El ítem 2 tiene un índice de dificultad de 0.49 mostrando un grado de dificultad medio y un índice discriminador de 0,44 que brinda confiabilidad en la evaluación del tema.

Al Igual que en Olivo, la opción más seleccionada es la b) que es la acertada, con un porcentaje de respuesta correcta de 55.2% y un índice de dificultad de 0.55 valores muy cercanos al 48.8% y 0.49 comparativamente con los que arroja este estudio. También se observa similitud con los resultados de Olivo en la selección de las otras alternativas de respuesta al escoger gran número de estudiantes la respuesta incorrecta a) con porcentajes de 33% contra un 34.7% , la opción c) con 11% frente a un 16.5% y un número de respuestas en blanco de 1 a 0 que se obtienen respectivamente en esta investigación como se aprecia en el Cuadro1. Estos bajos porcentajes, indican que son muy pocos los estudiantes que no tienen bases para dar alguna razón de lo que puede pasar con la variabilidad del intervalo si cambia el tamaño de la muestra.

El índice discriminador en la prueba piloto de Olivo es de 0.57 porcentaje éste muy poco diferente al 0.44 de esta prueba.

Ítem 3: *Si, manteniendo todos los demás datos fijos, el nivel de confianza se reduce (por ejemplo de 90% a 80%):*

a) *El intervalo de confianza no cambia.*
b) *El intervalo de confianza será más ancho.*
c) *El intervalo de confianza será más angosto.*
d) *El cambio en el intervalo de confianza no es predecible.*

Opciones	*Frecuencias*	*%*
a	*8*	*4.9*
b	*48*	*29.3*
c	***92***	***56.1***
d	*16*	*9.7*
No responde	*0*	*0*
Total	*164*	*100*

Tabla 7. Frecuencias y porcentajes de las opciones del ítem 3.

Índice de dificultad: 0,56 Índice de discriminación: 0,29

Esta pregunta busca conocer qué entendimiento poseen los estudiantes con el concepto: *el ancho de los intervalos de confianza aumenta cuando el nivel de confianza aumenta*. Es decir, se trata de examinar si los estudiantes relacionan el significado del nivel de confianza con el cálculo del valor crítico, y como éste afecta al intervalo obtenido. La opción que más respondieron los estudiantes fue la opción correcta c) con un porcentaje de 56.1%. El 29.3% de los evaluados creen

que si el nivel de confianza disminuye el intervalo de confianza aumenta escogiendo la opción b). El 9.7% de los estudiantes se inclinaron por la opción d) dando a entender que creen que el ancho del intervalo de confianza no se afecta con el cambio del nivel de confianza. El 4.9% de este grupo, a diferencia de los que eligieron la opción d), creen que el intervalo no cambia al variar el nivel de confianza, eligiendo la opción a) como se aprecia en la Tabla 7.

Comparativamente la tesis de Olivo presenta un porcentaje de respuesta correcta del 65.1% con la opción c), de igual forma se puede ver que los estudiantes tienden a seleccionar la opción b) con un 26% y las opciones a) y d) con porcentajes más bajos y muy similares a los mostrados en este estudio.

Este ítem presenta un índice de dificultad de 0,56 mostrando un grado de complejidad medio, frente a un 0.65 de las pruebas de Olivo y un índice discriminador de 0,29 presentando poca confiabilidad en la evaluación del tema. Por el contrario la prueba piloto de Olivo tiene un índice discriminador de 0.88 que marca una gran diferencia con el obtenido en este análisis. Al comparar con este ítem, se observa al igual que los dos anteriores, que los resultados en los porcentajes de respuesta y en los índices de dificultad no mostraron mayor disparidad con los encontrados en este estudio, pero el desconocer el índice discriminador de la prueba final de Olivo y los resultados de los exámenes no permite hacer un análisis amplio de este ítem.

Ítem 5: *En un intervalo de confianza del 95% para la media:*

a) *Si se toman muchas muestras y con cada una de ellas se construye el intervalo, la media muestral* $\bar{x}$ *caerá dentro del intervalo de confianza del 95% de las veces.*

b) *La probabilidad de que $\bar{x}$ caiga dentro de un intervalo de confianza calculado de una muestra específica es 0.95.*

c) *Si se toman muchas muestras de igual tamaño, el 95% de los intervalos calculados contendrá a μ.*

Opciones	*Frecuencias*	*%*
a	*35*	*21.4*
b	*62*	*37.8*
c	***65***	***39.6***
No responde	*2*	*1.2*
Total	*164*	*100*

Tabla 8. Frecuencias y porcentajes en las opciones del ítem 5.

Índice de dificultad: 0,40 Índice de discriminación: 0,44

El ítem 5 busca conocer qué tanto entienden los estudiantes el significado del nivel de confianza: variación del intervalo en diferentes muestras. La opción que más respondieron los estudiantes fue la c), que es la opción correcta, con un porcentaje del 39.6%. El 21.4% de los estudiantes se inclinaron por la opción a) creyendo que el intervalo calculado trata de estimar la media muestral y no la poblacional. El 37.8% de los estudiantes al parecer hacen una interpretación bayesiana[6] del intervalo de confianza, error bastante frecuente en la interpretación de contraste de hipótesis (Olivo, 2008). El 1.2% no respondió el ítem (Ver Tabla 8).

[6] *La interpretación Bayesiana es subjetiva y requiere de la asignación de probabilidades a priori.*

El índice de dificultad es de 0,40 un grado de dificultad medio alto dando a entender la poca comprensión que los estudiantes poseen respecto al significado del nivel de confianza asociado a un intervalo de confianza. El índice discriminador de este ítem es 0,44 lo que da a entender que posee una buena confiabilidad en la evaluación del tema deseado.

Los resultados en la prueba de Olivo presentan un porcentaje de respuesta correcta del 36.5%, 3.1 puntos porcentuales por debajo del alcanzado en este estudio y un índice de dificultad de 0.37. Se observa igualmente, un alto número de respuesta en las opciones a) y c) con porcentajes de 29.0% y 36.5%. (Ver Cuadro 1). Este ítem presenta un índice discriminador de 0.44 alcanzando una diferencia notable con el obtenido por Olivo en la prueba piloto de 0.63.

Ítem 6: *La media muestral de 100 observaciones en una prueba de matemáticas es 75, encuentre el intervalo de confianza al 95% para la media de población, asumiendo que* $\delta = 7$:

a) *(61.28, 88.72)*
b) *(73.63, 76.37)*
c) *(68, 82)*
d) *(74.3, 75.7)*

Opciones	*Frecuencias*	*%*
A	*26*	*15.8*
B	***54***	***33***
C	*43*	*26.2*
d	*31*	*18.9*
No responde	*10*	*6.1*
Total	*164*	*100*

Tabla 9. Frecuencias y porcentajes en las opciones del ítem 6.

Índice de dificultad:0,35 Índice de discriminación: 0,39

El ítem 6 pretende evaluar cómo los estudiantes *estiman la media de una población normal en una muestra grande con* σ *conocida.* La opción que más respondieron los estudiantes fue la b) que es la opción correcta con un porcentaje de 33%. El 18.9% de los estudiantes respondió la opción d), olvidando multiplicar el valor crítico de la distribución normal estándar por el error estándar. La opción a) la eligió el 15.8% de los estudiantes que no dividió la desviación típica poblacional por la raíz cuadrada del tamaño de la muestra. El 26.2% del grupo cometió los dos errores anteriores, olvidó el valor crítico de la distribución normal estándar y la raíz cuadrada de la muestra que divide a la desviación típica poblacional. El 6.1% de los evaluados no respondió (Ver Tabla 9).

Presenta un índice de dificultad de 0,35 señalando un grado de dificultad medio alto, el cual, al igual que el ítem anterior, mostrando que los estudiantes no conocen a cabalidad la expresión algebraica que permite calcular los intervalos de confianza. Presenta un índice discriminador de 0,39 mostrando una baja confiabilidad en la evaluación del tema.

A diferencia de los ítems anteriores, en este ítem el índice de dificultad presenta una variación negativa bastante notable con respecto a los resultados de la prueba de Olivo de 0.79 (más precisamente 0.44 puntos porcentuales por debajo) y un porcentaje de respuesta correcta del 79%, mostrando mayor habilidad los estudiantes que respondieron el test de Olivo (Ver Cuadro 1). Esta diferencia también se puede apreciar en el índice de discriminación con 0.88 frente a 0.39 que señala el de este estudio.

Ítem 8: *Se han obtenido los siguientes datos de emisión diaria de óxidos de azufre, para una muestra de tamaño n=100, media:* $\bar{x}$*=18, y varianza muestral* s^2*=36. Elabore un intervalo de confianza del 95% para la verdadera emisión diaria promedio de óxidos de azufre.*

a) *(17.016, 18.984).*
b) *(16.824, 19.176).*
c) *(6.24, 29.76).*
d) *(8.16, 27.84).*

Opciones	*Frecuencias*	*%*
a	*41*	*25*
b	***54***	***33***
c	*23*	*14*
d	*21*	*12.8*
No responde	*25*	*15.2*
Total	*164*	*100*

Tabla 10. Frecuencias y porcentajes en las opciones del ítem 8.

Índice de dificultad: 0,39 Índice de discriminación: 0,40

Con este ítem se quiere conocer qué tanto los estudiantes manejan el tema: *estimar la media de una población a partir de datos experimentales, con* σ *desconocida y muestra grande;* lo que hace que el grado de dificultad sea mayor al de los ítems anteriores. La opción que más respondieron los estudiantes fue la b) que es la respuesta correcta con un porcentaje de 33%. El 25% obtuvieron un valor errado del intervalo de confianza pedido, utilizando como valor crítico para el 95% de confianza 1.64 en lugar de 1.96; esto se debe a que en la tabla de la normal estándar buscan un área a la derecha de la curva de 5%, debiendo ser del

2.5% porque se debe usar el área a la izquierda también ya que sumadas dan el 5% por ser un intervalo bilateral. El 14% de los estudiantes responden la opción c), no dividen por la raíz del tamaño de la muestra al calcular el intervalo, confunden la desviación típica de la población con la desviación típica del estadístico. El 12.8% comete los dos errores anteriores, utilizan el 16.4 en el valor critico de la normal y no dividen por el tamaño de la muestra. El 15.2% no respondió (Ver Tabla 10).

El índice de dificultad es de 0,39 un grado de dificultad medio alto y un índice discriminador de 0,40 mostrando una buena confiabilidad en la evaluación del tema deseado.

Este ítem, al igual que el ítem 6, al examinarlo con la prueba de Olivo, observa una notable diferencia en todos los resultados: el porcentaje de respuesta correcta del 73% y un índice de dificultad de 0.73 contra un 0.39 que muestra este estudio, los porcentajes en las otras opciones de respuestas, el número de respuestas en blanco y el índice de discriminación de la prueba piloto determinan mayor habilidad de los estudiantes que contestaron el test de Olivo.

Ítem 9: *El nivel de confianza es de 0.95, para un intervalo de confianza para la media de la población con desviación estándar poblacional desconocida para un grupo de puntajes distribuido normalmente de tamaño n=20. Los valores críticos han de ser:*

a) *-1.65 y 1.65 uso de normal estándar.*
b) *-1.96 y 1.96 uso de normal estándar.*
c) *-2.093 y 2.093 uso de distribución t con 19 grados de libertad.*
d) *-2.085 y 2.085 uso de distribución t con 20 grados de libertad.*

Opciones	*Frecuencias*	*%*

a	*22*	*13.4*
b	*48*	*29.3*
c	***37***	***22.5***
d	*13*	*7.9*
No responde	*44*	*26.9*
Total	*164*	*100*

Tabla 11. Frecuencias y porcentajes de las opciones del ítem 9.

Índice de dificultad: 0,31 Índice de discriminación: 0,28

El ítem 9 estudia la habilidad que poseen los estudiantes al *determinar valores críticos en la distribución del estadístico*. La respuesta más común de los estudiantes fue la b) que es una opción incorrecta con un 29.3% y una frecuencia de 48 estudiantes, no manejan bien el cálculo de valores críticos en las distribuciones, buscan el valor crítico de manera incorrecta utilizando el de la normal estándar, sin reconocer los casos en los cuales se debe utilizar la distribución t. El 22.5% respondieron correctamente y entienden el procedimiento. El 7.9% de los estudiantes usó grados de libertad incorrectos en la distribución *t,* eligiendo la opción d). El 13.4% restante se equivocó al calcular el valor crítico correspondiente a $\alpha/2$ en las tablas de la normal estándar. Una cantidad de estudiantes equivalente al 26.9% no respondieron la pregunta (Ver Tabla 11).

El índice de dificultad es de 0,31 un grado de dificultad medio alto, y un índice discriminador de 0,28 que es señal de una confiabilidad baja.

En este ítem la diferencia con respecto a los resultados de Olivo también es notable. Resalta el gran número de respuestas en blanco, 44 en este estudio contra 5 de Olivo, los porcentajes de respuesta y el índice discriminador en la

prueba piloto de Olivo de 1.0. También se puede observar que los estudiantes en este ítem, al igual que en el ítem 10, seleccionaron en mayor número una respuesta incorrecta, lo que no registro la prueba de Olivo (Ver Cuadro 1).

Ítem 10: *Considere el grafico siguiente del rendimiento medio de cebada en 1980, 1984 y 1988 junto con un intervalo de 95% de confianza respectivos*

Año	*N*	*Media*	*StDev*	----+--------+---------+---------+--
1980	*6*	*184.00*	*2.61*	*(----*---)*
1984	*5*	*212.40*	*14.36*	*(----*---)*
1988	*5*	*182.40*	*1.82*	*(-----*----)*
				----+--------+---------+---------+--
Dev. Típica conjunta=0.19				*180 195 210 225*

¿Cuál de las siguientes afirmaciones es verdadera?

a) *Puesto que los intervalos de confianza para 1980 y 1988 tienen considerable solape, hay buena evidencia que las medias de las muestras difieran.*
b) *La estimación de la media de la población en 1980 es menos precisa que en 1988.*
c) *Puesto que los intervalos de confianza para 1980 y 1984 no se solapan, hay poca evidencia que las medias de las poblaciones respectivas difieran.*
d) *Puesto que los intervalos de confianza para 1980 y 1988 tienen considerable solape, hay poca evidencia que las medias de las poblaciones difieran.*

Opciones	*Frecuencias*	*%*
a	*26*	*15.9*

b	*44*	*26.8*
c	*34*	*20.7*
d	***20***	***12.2***
No responde	*40*	*24.4*
Total	*164*	*100*

Tabla 12. Frecuencias y porcentajes de las opciones del ítem 10.

Índice de dificultad. 0,16 Índice de discriminación: 0,43

El ítem 10 evalúa *la interpretación de gráficos sobre los intervalos de confianza*, la opción que más respondieron los estudiantes fue la b) que es incorrecta con un porcentaje del 26,8%, al parecer no asocian la precisión de la estimación con el ancho del intervalo, siendo más precisa la estimación cuando el intervalo es menor. El 20,7% de los estudiantes respondió incorrectamente eligiendo la opción c) asumiendo, posiblemente, que el no solape es un indicio de la igualdad de las medias, interpretación que no es ni intuitiva ni lógicamente explicable. Apenas un 12,2% de los estudiantes respondieron correctamente escogiendo la opción d), reconociendo e interpretando acertadamente esta pregunta. El 15,9% de los estudiantes respondieron a) y piensan que los argumentos del solape son que las medias difieren, con el mismo pensamiento de los que respondieron la opción c). El 24,4% no respondió (Ver Tabla 12).

El índice de dificultad es de 0,16 un grado de dificultad alto que lo hace el ítem más difícil de la prueba, lo que puede poner en evidencia el poco trabajo gráfico que realizaron al estudiar este tema de los intervalos de confianza, y un índice discriminador de 0,43 mostrando una buena confiabilidad en la evaluación del tema deseado.

En conclusión, se puede decir que estos ítems de selección múltiple con diferentes grados de dificultad, permiten evaluar a fondo las concepciones que sobre los

intervalos de confianza y sus elementos constituyentes poseen los distintos grupos analizados. Es pertinente resaltar el hecho de que fueron precisamente los estudiantes de la licenciatura de matemáticas de la UIS los que presentaron mejores resultados en el test.

En cuanto al índice discriminador, los ítems muestran una confiabilidad media situada en el rango de 0,28 a 0,44. Se esperaba que fuera mayor ya que estos ítems se seleccionaron de una tesis doctoral (Olivo, 2008) en la que se estudiaron y modificaron con el fin de ser una buena prueba a la hora de evaluar los intervalos de confianza. Al analizar los diferentes resultados arrojados por la prueba, se observó falta de interés por parte de varios estudiantes al no contestar algunos ítems o desarrollar los cálculos necesarios para hacerlo, como lo demostraron los ítems 3 y 10 que presentaron los índices de discriminación más bajos.

Como se mostró en el análisis de cada uno de los ítems politómicos y dicotómicos, todos los índices de discriminación en la prueba aplicada por Olivo en su tesis doctoral fueron más altos. Es importante recordar que los estudiantes a los que les aplicó la prueba final cursaron los temas necesarios para responderla y en el transcurso del semestre presentaron pruebas similares que les permitió un mejor desempeño en el desarrollo del test.

A continuación se puede observar en el Cuadro 1, un comparativo de las frecuencias en las respuestas de cada opción en cada ítem de selección múltiple, el porcentaje de respuestas correctas entre los resultados obtenidos en el estudio de Olivo y los obtenidos en la aplicación del test para el presente estudio.

ÍTEMS	Frecuencias resultados Olivo	Frecuencias resultados estudio

	n = 252							n = 164						
	Opciones de Respuesta					*Correcto*	*Índice de dificultad*	*Opciones de respuesta*					*Correcto*	*Índice de dificultad*
	a	***b***	***c***	***d***	***NR***	***%***	***Id***	***a***	***b***	***c***	***d***	***NR***	***%***	***Id***
1	*65*	*40*	***140***	*7*	*0*	***55.6***	*0.55*	*66*	*8*	***88***	*2*	*0*	***53.6***	*0.54*
2	*84*	***139***	*28*	*0*	*1*	***55.2***	*0.55*	*57*	***80***	*27*	*0*	*0*	***48.8***	*0.49*
3	*7*	*66*	***164***	*15*	*0*	***65.1***	*0.65*	*8*	*48*	***92***	*16*	*0*	***56.1***	*0.56*
5	*73*	*86*	***92***	*0*	*1*	***36.5***	*0.37*	*35*	*62*	***65***	*0*	*2*	***39.6***	*0.40*
6	*17*	***199***	*11*	*21*	*4*	***79.0***	*0.79*	*26*	***54***	*43*	*31*	*10*	***33***	*0.35*
8	*29*	***184***	*15*	*13*	*11*	***73.0***	*0.73*	*41*	***54***	*23*	*21*	*25*	***33***	*0.39*
9	*12*	*104*	***118***	*13*	*5*	***46.8***	*0.47*	*22*	*48*	***37***	*13*	*44*	***22.5***	*0.31*
10	*23*	*17*	*32*	***175***	*5*	***69.4***	*0.69*	*26*	*44*	*34*	***20***	*40*	***12.2***	*0.16*
***NR** = No responden*								***%** = Porcentaje de respuestas correctas*						

Cuadro 1. Comparativo de los resultados en los ítems de opción múltiple

Aunque los ítems 1, 2, 3 y 5 tienen porcentajes de respuesta parecidos, en los ítems 6, 8, 9 y 10 estos valores difieren bastante, mostrando al parecer que el orden de las preguntas afectó el rendimiento de los evaluados en este test, pues los primeros ítems se respondieron con mayor afluencia que los últimos, dejando la incertidumbre si fue por el poco tiempo otorgado para responder el mismo.

4.1.1.2. *Análisis de ítems abiertos*

Ítem 4*: Explique cómo varia la anchura del intervalo de confianza, si conservando el mismo tamaño de muestra y el mismo coeficiente de confianza se tomara una población con varianza cuatro veces mayor.*

En este ítem se desea evaluar la capacidad de los estudiantes para reconocer el efecto de la varianza sobre la anchura del intervalo. Los participantes deben darse cuenta que el ancho del intervalo de confianza se duplica.

El Cuadro 2 presenta algunas de las soluciones dadas por los estudiantes a este ítem. Se observa la complejidad de las respuestas, conceptos, propiedades y argumentos necesarios para responderlo correctamente y la puntuación dada a cada argumento.

Puntuación	**Ejemplo de respuesta**	**Concepto**
2	Como la varianza es la raíz cuadrada de la desviación típica, entonces se pone la desviación multiplicada por 2 en la fórmula del intervalo así: $\bar{x} \pm Z_{\alpha/2} {2\sigma}/{\sqrt{n}}$ Entonces el intervalo crece el doble.	Indica y argumenta de manera correcta la proporción en la que aumenta el intervalo, mostrando la fórmula y multiplicando por 2 la desviación típica.
2	La anchura del intervalo de confianza crece el doble con: $\bar{x} - Z_{\alpha/2} {(4\sigma^2)}/{\sqrt{n}}$, $\bar{x} \pm Z_{\alpha/2} {(4\sigma^2)}/{\sqrt{n}}$	Argumenta mediante la varianza en la fórmula, que el intervalo aumenta el doble.
1	El rango del intervalo de confianza aumenta, porque la varianza aumenta, entonces la desviación aumenta.	Indica que el intervalo aumenta pero no argumenta mediante la fórmula con que proporción lo hace.
0	El intervalo se hace más pequeño.	Responde relacionando incorrectamente la varianza con el intervalo de confianza.
0	No responde o da respuestas que no tienen relación con la pregunta.	Poca o nula claridad sobre el tema.

Cuadro 2. Ejemplos de respuestas en el ítem 4.

A continuación en el Cuadro 3 se presentan las frecuencias y porcentajes correspondientes a cada clase de respuesta dada en este ítem.

Clases de respuestas	Frecuencias	Porcentajes
Respuestas correctas	3	1.9
Respuestas parcialmente correctas	33	20.1
Respuestas incorrectas	47	28.6
No responde	81	49.4
Total	164	100

Cuadro 3. Frecuencias y porcentajes a las clases de respuesta del ítem 4.

El ítem 4 evalúa el efecto de la varianza sobre el ancho del intervalo de confianza, es una pregunta con respuesta abierta, y requiere de conocer a fondo la interpretación de la fórmula para hallar el intervalo de confianza, al igual que la relación entre la varianza con la deviación típica y la manera como afectan estos dos conceptos el ancho del intervalo. El 1,9% respondió correctamente este ítem, pues solo 3 personas de las 164 evaluadas entienden bien la relación entre la varianza y el ancho del intervalo. El 20,1% contestó parcialmente correcto este ítem, pues reconocen el aumento de tamaño del intervalo pero no argumentan en qué proporción. El 28,6% respondió incorrectamente creyendo que el intervalo disminuye o simplemente dieron respuestas que no tienen relación con la pregunta. El 49,4% no respondió el ítem.

En este ítem se puede observar un índice dificultad de 0,24, el cual refleja el hecho de ser un ítem abierto que requiere de habilidad y confianza en el manejo

de conceptos para crear una respuesta correcta o cercana a ella. El índice discriminador es de 0,66, que lo hace el ítem más confiable del test.

Ítem 7: *Un fabricante asegura que sus garrafones contienen un litro de cloro puro. Al tomar una muestra de 16 garrafones se determinó que en promedio contenían 0.94 litros de cloro puro, con desviación estándar de la muestra de 0.097. Construir un intervalo de confianza al 95% para el verdadero contenido promedio de litros de cloro puro. No se conoce la desviación típica de la población. (La distribución del contenido de cloro por botella puede considerarse normal).*

En este ítem se desea evaluar si los estudiantes saben estimar la media de una población cuando σ es desconocida, utilizando la distribución t con 15 grados de libertad, y conociendo el tamaño de la muestra que es 16.

En el Cuadro 4 se muestran algunas de las respuestas dadas por los estudiantes a este ítem, los argumentos y los puntajes dados a cada una.

Puntaje	Ejemplos de respuesta	Conceptos
2	Se usa la distribución t, con 15 grados de libertad por no conocerse la desviación típica de la población, con la siguiente fórmula se construye el intervalo al 95% $\bar{x} \pm t_{\alpha/2} {}^{s}/_{\sqrt{n}}$ Dando como resultado después de varias operaciones: $0.8884 < \mu < 0.9916$	Aunque las operaciones fueron mal realizadas, se toma como correcto.
1	$0.94 \pm (2.11)\frac{0.097}{\sqrt{16}}$	Toma los grados de libertad incorrectos, pero las demás operaciones están bien.

0	$0.94 \pm (1.75)\frac{0.097}{\sqrt{16}}$	Mal uso en la tabla t y los grados de libertad, por esto el intervalo es incorrecto.
0	No responde o da una respuesta sin relación con el ítem.	Poca o nula claridad sobre el tema.

Cuadro 4. Ejemplos de respuestas en el ítem 7.

En el Cuadro 5 se muestran las frecuencias, porcentajes y clases de respuestas que se obtuvieron en esta pregunta.

Clases de respuestas	Frecuencias	Porcentajes
Respuestas correctas	2	1.2
Respuestas parcialmente correctas	2	1.2
Respuestas incorrectas	49	30
No responde	111	67.6
Total	164	100

Cuadro 5. Frecuencias y porcentajes a las clases de respuestas del ítem 7.

El ítem 7 pretende evaluar si los estudiantes saben *cómo estimar la media de una población aproximadamente normal cuando* σ *es desconocida.* Es un ítem con respuesta abierta, exige el conocimiento y dominio de las tablas de distribución y los tamaños de la muestra en el cálculo del intervalo. El 1,2% respondió correctamente este ítem, es decir, solamente 2 personas de las 164 evaluadas. El 1,2% respondieron parcialmente bien, pues construyen bien la formula pero tienen

errores en el procedimiento de obtención del intervalo. El 30% responde incorrectamente mostrando errores a la hora de buscar en la tabla de distribución el valor correcto y utilizando mal los grados de libertad, entre otros errores de cálculo. El 67.6% no respondió el ítem.

Los resultados de este ítem señalan un índice de dificultad de 0,06 mostrando un grado de dificultad muy alto, con una mezcla de manejo de conceptos y desarrollo de cálculos que permitan dar o aproximarse a una respuesta correcta. En este ítem también se observa una buena confiabilidad en la evaluación del tema con un índice discriminador de 0,57.

En estos dos ítems abiertos, se observó un gran número de estudiantes que no respondieron. El hecho de ser preguntas abiertas requiere que los estudiantes generen en forma solitaria las respuestas sin tener opciones que confrontar, lo que hace que el grado de dificultad aumente.

En el Cuadro 6 que se muestra a continuación, se presenta un comparativo de las frecuencias en las respuestas de las distintas opciones en cada ítem abierto, el porcentaje de respuestas correctas, el índice de dificultad y el índice de discriminación entre los resultados obtenidos en el estudio de Olivo y los obtenidos en la aplicación del test para el presente estudio.

ITEMS	Frecuencias resultados Olivo n = 252						Frecuencias resultados estudio n = 164					
	Opciones de respuesta				*Correcto*	*Índice de dificultad*	*Opciones de respuesta*				*Correcto*	*Índice de dificultad*
	c	***pc***	***i***	***NR***	***%***	***Id***	***c***	***pc***	***i***	***NR***	***%***	***Id***
4	***108***	*66*	*68*	*10*	***42.8***	*0.42*	***3***	*33*	*47*	*81*	***1.9***	*0.24*

7	***100***	*87*	*49*	*16*	***39.7***	*0.40*	***2***	*2*	*49*	*111*	***1.2***	*0.06*
% *=Porcentaje de respuestas correctas* ***c*** *= respuestas correctas* ***pc*** *= respuestas parcialmente correctas*							***i*** *= respuestas incorrectas* ***NR*** *= No responden*					

Cuadro 6 . Comparativo de los resultados en los ítems abiertos.

Los ítems 4 y 7 están entre los más difíciles de los escogidos en la tesis de Olivo, juzgando por los porcentajes de respuesta correcta que alcanzan un 40% aproximadamente. En los resultados de esta investigación, estos porcentajes son los más bajos que muestra la prueba, con un 1.9% y 1.2% para estos ítems respectivamente. Sumado al poco tiempo dado por los profesores para contestar el test, estos ítems son de respuesta abierta y la inclinación a no responder ítems de este tipo se nota en este trabajo, pues los evaluados tienen que realizar cálculos para poder dar una respuesta y necesitan más habilidad y tiempo para ello.

El Cuadro 7 ofrece una relación entre los resultados obtenidos en el estudio de Olivo y los arrojados en el presente análisis para cada ítem del test en el índice de dificultad y de discriminación.

Clase	**ÍTEMS**	**Contenido del ítem**	**Resultados Olivo**		**Resultados estudio**	
			Índice de Dificultad	Índice de Discriminación*	Índice Dificultad	Índice de Discriminación
D*	**1**	Definición	0.55	0.44	0.54	0.33
D*	**2**	Variación con tamaño de muestra	0.55	0.57	0.49	0.44
D*	**3**	Relación con nivel de confianza	0.65	0.88	0.56	0.29

P*	**4**	Comparar varianzas sobre amplitud	0.42	0.88	0.24	0.66
D*	**5**	Variación en diferentes muestras	0.37	0.63	0.40	0.44
D*	**6**	Estimar media, σ conocida y población normal	0.79	0.88	0.35	0.39
P*	**7**	Estimar media, σ desconocida y distribución t	0.40	0.82	0.06	0.57
D*	**8**	Estimar media, σ desconocida y muestra grande	0.73	0.63	0.39	0.40
D*	**9**	Determinar valor crítico	0.47	1	0.31	0.28
D*	**10**	Interpretar gráficos	0.69	0.75	0.16	0.43
D* = *Ítem dicotómico* **P*** = *Ítem politómico*						

**Resultados de la prueba piloto efectuada por Olivo en su tesis doctoral*

Cuadro 7. Resultados en los índices de dificultad y discriminación de los ítems.

Los índices de habilidad que se comparan en el Cuadro 7 son los obtenidos por Olivo en la prueba Vs los obtenidos en este estudio. Los índices de dificultad de la prueba realizada por Olivo, muestran valores más altos en todos los ítems ¿Es acaso éste un motivo para pensar que los estudiantes que enfrentaron esa prueba tienen mayor habilidad en el tema que los estudiantes que participaron en este estudio? No necesariamente, ya que la prueba de Olivo se aplicó en un momento donde los estudiantes terminaban de ver el tema y la colaboración de los profesores para la misma fue total. Contrario en esta investigación donde la prueba se realizó en época de previos, razón por la cual, algunos profesores no otorgaron suficiente tiempo para responder con total comodidad el test.

La mayor dificultad de los ítems en el ensayo de olivo se presentó en las preguntas 4 y 7, con índices de dificultad de 0.42 y 0.40 respectivamente, mostrando que las preguntas abiertas exigen más de los alumnos, al combinar los conceptos teóricos con la práctica. Igualmente, en este estudio, los índices de

dificultad más bajos se encuentran en los ítems 4, 7 y 10, con valores de 0.24, 0.04 y 0.16 respectivamente. El ítem 10, siendo una pregunta de opción múltiple sobre el análisis de gráficos referentes al intervalo de confianza, logró que sólo muy pocos estudiantes respondieran correctamente, posiblemente por ser el último ítem, el poco tiempo para responder la prueba y la falta de interés de algunos en completarla.

Los índices de discriminación con lo que se examinó este estudio, son los obtenidos en la prueba piloto realizada por Olivo, ya que no aparecen en la tesis los índices de discriminación del último examen, afirmando:

> Los coeficientes de fiabilidad y generalizabilidad que intentan dar una medida objetiva de la estabilidad de las puntuaciones obtenidas frente a variaciones aleatorias, nos permite concluir, según los índices obtenidos, que nuestro cuestionario lo podríamos calificar como moderadamente fiable, en cuanto a generalizabilidad a otros ítems; sin embargo es altamente generalizable a otros alumnos[7]

Los índices de discriminación en la prueba piloto de Olivo con valores mayores a 0.60 se presentaron en los ítems 3, 6, 8, y 10. Sólo el ítem 5 obtuvo un valor menor a 0.40. Estos datos revelan el alto nivel de competencia que demostraron los estudiantes a los cuales se les aplicó la prueba. Por el contrario en este estudio, ítems con valores mayores a 0.60 no existen, en cambio si se tienen ítems con valores muy inferiores a 0.40 como son los ítems 4, 7 y 10, que están incluso por debajo de 0.30. Sólo los ítems 1, 2, 3 y 5 muestran valores entre 0.40 y 0.60 para éste índice

[7] Olivo, E.. *Significado de los intervalos de confianza para los estudiantes de ingeniería en México.* [Tesis Doctoral] España. Universidad de Granada. Departamento de Didáctica de las matemáticas; 2008, p.208.

La variación del intervalo en diferentes muestras, tema que se evaluó con la pregunta 5, presentó la mayor dificultad para los estudiantes de Olivo en los ítems dicotómicos. En este estudio, el ítem 10, que evalúa el análisis de gráficos de los intervalos de confianza fue el más difícil de estos ítems y del test. Opuestamente, en relación a la pregunta menos difícil, en la prueba de Olivo pretendía evaluar como estimar la media poblacional en una muestra grande con desviación conocida y corresponde al ítem 6 y para este estudio el ítem 3 que evaluaba como aumenta el intervalo de confianza cuando el nivel de confianza aumenta (Ver Cuadro 7).

Un índice discriminador de 1 y 0.88 frente a 0.28 y 0.29 en este estudio, representa la mayor diferencia con los resultados de la prueba piloto en el ítem 9 y 3 respectivamente. Estos dos ítems tienen la mejor confiabilidad a la hora de evaluar a los estudiantes en la prueba de Olivo, pero, por el contrario, la menor confiabilidad en esta investigación, al evaluar que tanto sabe el estudiante para hallar los valores críticos en la construcción del intervalo de confianza y como varia el ancho del intervalo al aumentar el nivel de confianza.

Los índices de discriminación en general, presentaron valores más altos en el trabajo piloto realizado por Olivo; los ítems 3, 4, 6, 7 y 9 tienen valores superiores a 0.8, que dan una buena discriminación de esas preguntas a la hora de aplicar la prueba. El ID de 0.44 fue el menor del cuestionario, obtenido por el ítem 1, y muestra la seguridad que representan estos valores para el test en general.

4.2. MODELO RASCH

4.2.1. ANÁLISIS DE DATOS POR MEDIO DEL PROGRAMA WINSTEPS®

En los años sesenta el matemático danés George Rasch ya había diseñado este modelo para la calibración de ítems y evaluación de personas, pero debido a la cantidad de cálculos necesarios para su ejecución, su uso no era sencillo, ni rápido y menos aún práctico, por integrar un análisis profundo de ítems y personas involucradas en un ensayo. Pero gracias a los adelantos de la tecnología, hoy se encuentran softwares que permiten realizar estos cálculos de manera eficiente y acuciosa. En este capítulo se presentará el análisis de los datos recolectados, aplicando el modelo Rasch mediante el uso del programa Winsteps®, (Linacre, J.M. 2006), que se puede encontrar y obtener en la página http://www.winsteps.com por un bajo costo, aunque la página ofrece una versión gratis para prácticas con una reducida cantidad de ítems y personas, ya que sólo admite 75 personas y 25 ítems, llamada Ministep en http://www.winsteps.com/ministep.htm.

Este programa se ejecuta a través de archivos de control, donde se tienen unas variables que permiten analizar los datos como se precisen dependiendo de las especificaciones introducidas.
Los datos obtenidos del test aplicado, se tabularon en un programa para el manejo de bases de datos (Excel), aunque para trabajar este software es recomendable guardarlos en un formato de texto, con extensión .txt o .dat, ya que el programa procesa los datos en cualquiera de estos formatos. También se necesita tener sólo los datos de respuesta a los ítems y la variable que representa las personas, puesto que WINSTEPS pretende que estos archivos estén libres de información que no concierna a datos de preguntas y personas.

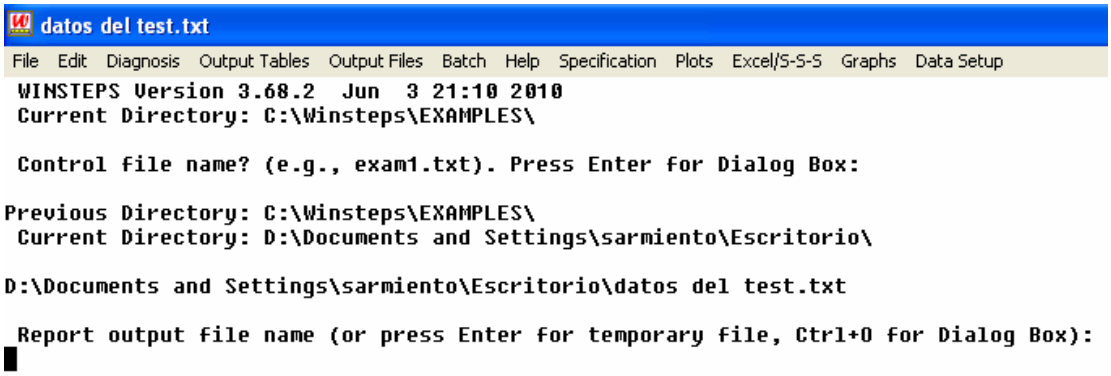
En el Anexo I se detalla el procedimiento para organizar los datos de forma que se pueda ejecutar el programa Winsteps®.

4.2.2. ANÁLISIS DE LOS RESULTADOS

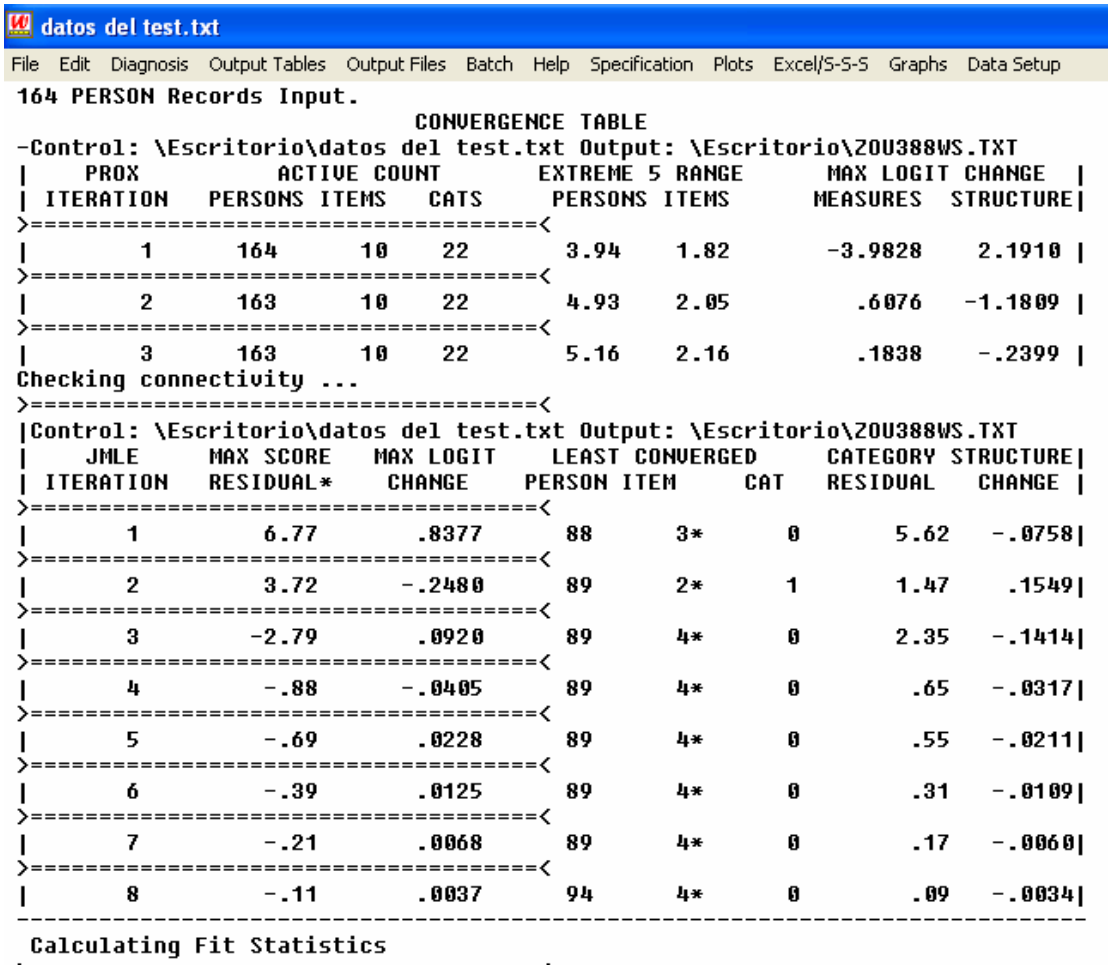
Los datos de respuesta de 164 estudiantes al test de 10 ítems fueron tabulados en Excel, donde se evaluaron ocho ítems dicotómicos, es decir, con dos clases de calificación, en este caso (0, 1) si responde mal 0 y si la respuesta es correcta 1; y dos ítems politómicos con tres opciones de calificación: 0, 1, 2; 0 si la respuesta es errada, 1 si respondió parcialmente bien, y 2 si la respuesta es correcta.

Al hacer doble click en el icono de WINSTEPS, la información inicial muestra algunas opciones, pero se puede abrir la base de datos desde el programa directamente ya que el archivo se convirtió a formato .txt. Como se desea trabajar con el archivo, se escoge no y enseguida intro, así se abre la ventana *control file,* que se muestra a continuación, para buscar el archivo de control que se guardó previamente en el escritorio con el nombre *Datos del test.*

Al aceptar el comando ofrecido y señalar el archivo de interés, las líneas que se aprecian en la siguiente ventana *datos del test .txt* aparecen en el software, se da doble click o enter para que el programa inicie. Al ejecutar esta acción, de inmediato WINSTEPS analiza todos los datos de respuesta con el modelo Rasch.



Como se muestra en la ventana que aparece a continuación, lo primero que brinda WINSTEPS en este análisis es la calibración de las personas y los ítems mediante el procedimiento iterativo PROX, que consiste en aproximarse al patrón de datos observados teniendo el parámetro de los ítems estimado (estimación condicional), en este caso fue suficiente hacer tres iteraciones inicialmente; luego el software utiliza el proceso iterativo JMLE que es más preciso en la calibración de estos parámetros, realizando ocho iteraciones sin conocer alguno de los dos parámetros (estimación conjunta).



datos del test.txt

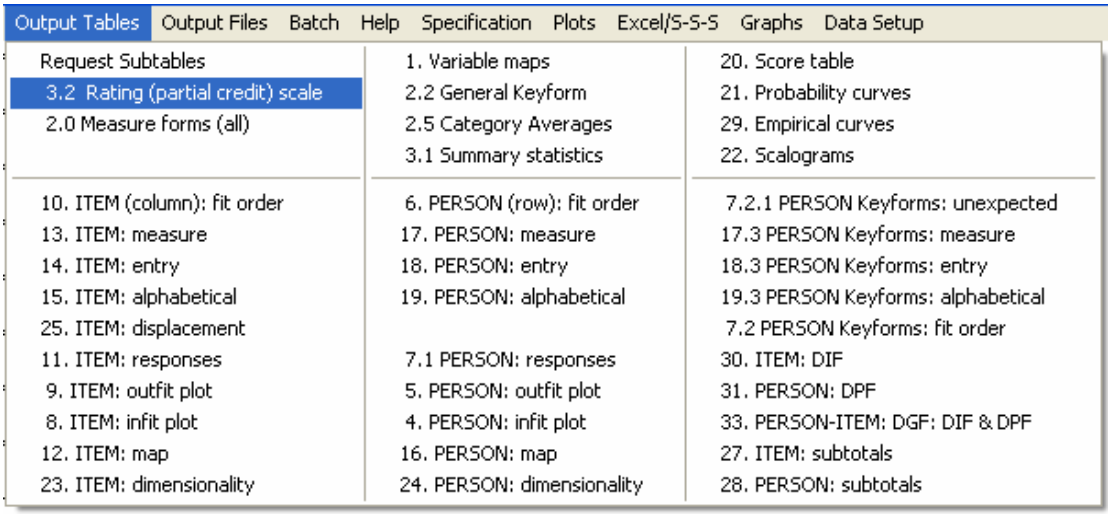
File Edit Diagnosis Output Tables Output Files Batch Help Specification Plots Excel/S-S-S Graphs Data Setup

```
164 PERSON Records Input.
                              CONVERGENCE TABLE
-Control: \Escritorio\datos del test.txt Output: \Escritorio\ZOU388WS.TXT
|    PROX          ACTIVE COUNT       EXTREME 5 RANGE      MAX LOGIT CHANGE   |
| ITERATION   PERSONS ITEMS   CATS     PERSONS ITEMS      MEASURES  STRUCTURE|
>=====================================<
|        1      164      10    22       3.94    1.82       -3.9828    2.1910 |
>=====================================<
|        2      163      10    22       4.93    2.05          .6076   -1.1809 |
>=====================================<
|        3      163      10    22       5.16    2.16          .1838    -.2399 |
Checking connectivity ...
>=====================================<
|Control: \Escritorio\datos del test.txt Output: \Escritorio\ZOU388WS.TXT
|    JMLE      MAX SCORE   MAX LOGIT    LEAST CONVERGED     CATEGORY STRUCTURE|
| ITERATION   RESIDUAL*    CHANGE    PERSON ITEM     CAT   RESIDUAL   CHANGE |
>=====================================<
|       1         6.77       .8377      88      3*      0       5.62   -.0758|
>=====================================<
|       2         3.72      -.2480      89      2*      1       1.47    .1549|
>=====================================<
|       3        -2.79       .0920      89      4*      0       2.35   -.1414|
>=====================================<
|       4         -.88      -.0405      89      4*      0        .65   -.0317|
>=====================================<
|       5         -.69       .0228      89      4*      0        .55   -.0211|
>=====================================<
|       6         -.39       .0125      89      4*      0        .31   -.0109|
>=====================================<
|       7         -.21       .0068      89      4*      0        .17   -.0060|
>=====================================<
|       8         -.11       .0037      94      4*      0        .09   -.0034|
-------------------------------------------------------------------------------
 Calculating Fit Statistics
```

Después de las iteraciones, aparecen los datos de las calibraciones de estos parámetros por separado como: media del puntaje de las personas, desviación estándar, media de la habilidad de las personas fijada en lógitos con su respectivo error, las medidas del infit, el outfit, la confiabilidad, la separación tanto de las personas como de los ítems y otros datos claves para obtener un resultado satisfactorio del análisis de respuesta y dificultad de los ítems de dicho test como se puede observar en la pantalla siguiente:

```
Standardized Residuals N(0,1)  Mean: -.02 S.D.: .94
D:\Documents and Settings\sarmiento\Escritorio\Datos del tes
-------------------------------------------------------------------------------
| PERSONS     164 INPUT     164 MEASURED              INFIT         OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN       3.3      10.0       -1.21     .84       .99     .1    .89     .3|
| S.D.       1.8        .0        1.10     .13       .26     .8    .46     .4|
| REAL RMSE    .85  ADJ.SD     .69  SEPARATION   .81  PERSON RELIABILITY  .40|
|-----------------------------------------------------------------------------|
| ITEMS      10 INPUT      10 MEASURED               INFIT         OUTFIT    |
| MEAN      53.5     164.0         .00     .22       .95     .1    .89    -.2|
| S.D.      27.1        .0        1.39     .09       .22    1.1    .34    1.2|
| REAL RMSE    .24  ADJ.SD    1.37  SEPARATION  5.75  ITEM   RELIABILITY  .97|
-------------------------------------------------------------------------------
 Output written to D:\Documents and Settings\sarmiento\Escritorio\ZOU388WS.TXT
 CODES= 012
IVALUEA= 01*
IVALUEB= 012
IREFER= AAABAABAAA
GROUPS= 0
Measures constructed: use "Diagnosis" and "Output Tables" menus
```

Al final de la pantalla aparece una sugerencia sobre alternativas para escoger y obtener información más detallada del análisis. Primero se seleccionará la opción *Output Tables* que se encuentra en la parte superior de la ventana, haciendo click en *3.2 Rating (partial credit) scale* como se ilustra a continuación:



Aquí se puede apreciar la distribución de las respuestas a cada uno de los ítems, primero aparecen los datos de los dos ítems politómicos que se encuentran en el test dado que las respuestas posibles son tres (0, 1, 2). En los resultados de los ítems 4 y 7 aparecen los números correspondientes a cada una de las categorías de respuesta y otros valores críticos en el análisis de esta clase de ítems. A continuación se verá cómo WINSTEPS presenta estas gráficas.

4.2.3. ANÁLISIS DE LOS ITEMS POLITÓMICOS DEL TEST

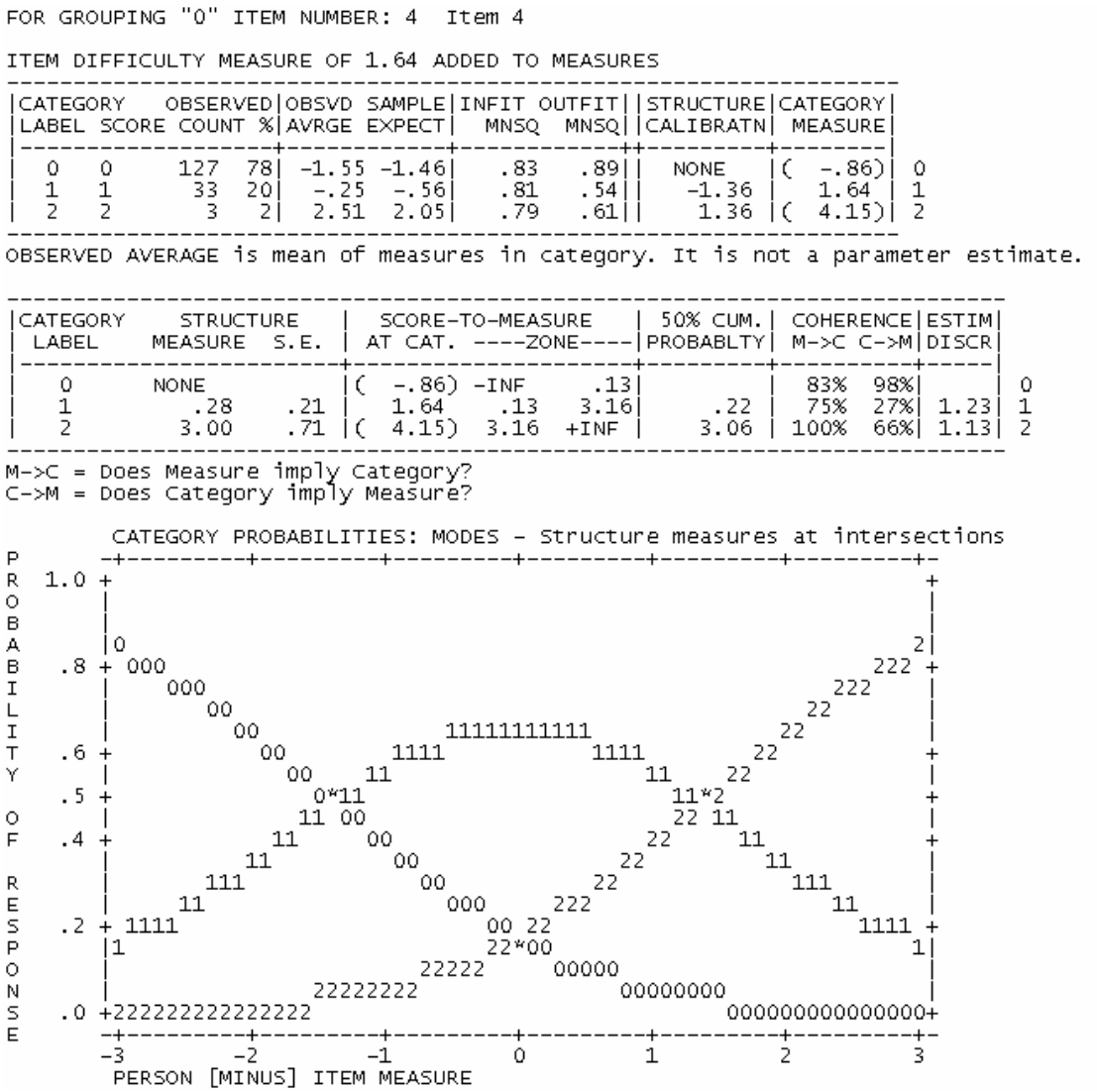
Como el programa muestra primero el análisis de los ítems politómicos, se seguirá en el estudio el mismo orden. Los gráficos son diferentes por tener tres clases de calificación, aparecen dos tablas con información decisiva, y valores de ajuste del ítem, cada uno de estos ítems es descrito y analizado en detalle con ayuda del gráfico obtenido.

4.2.3.1. *Análisis del ítem 4.* La distribución de los tres códigos posibles de respuesta a esta pregunta, tiene que ver con la probabilidad de obtener cada una de las respuestas con respecto a la habilidad de los estudiantes, en este caso, el análisis de la pregunta 4 se hace a nivel grupal. Al observar la distribución del código 2 (respuesta correcta), se desprende que la probabilidad de responder bien la pregunta con una baja habilidad del individuo cercana a -3 (en logitos), está próxima a cero y va creciendo a medida que la habilidad va en aumento; como sólo tres personas respondieron bien este ítem, la curva no crece rápidamente a medida que aumenta la habilidad, sino que sube muy despacio a medida que la habilidad crece. En la distribución del 0 (respuesta incorrecta), se ve que la curva decrece rápidamente, indicando que con mayor habilidad, menos probabilidad se tiene de obtenerla. Pero al analizar la distribución de la respuesta 1(parcialmente correcta), se observa que si la habilidad es baja, la probabilidad de obtener el 1 también lo es, pero ésta va creciendo a medida que la habilidad aumenta hasta su

media, y empieza a disminuir a medida que la habilidad crece a su máximo nivel, como se observa en la Gráfica 1. Esta distribución de campana se debe a que las personas con un alto nivel de habilidad, son más propensas a responder correctamente la pregunta, y las que tienen poca habilidad tienden más a responder de manera errada, es por esto, que su mayor probabilidad se ve próxima a la media de la habilidad de las personas. Como cada valor posible es el dominante en algún intervalo, tal como lo muestra la Gráfica 1, se puede concluir que las categorías creadas (0, 1 y 2) están bien establecidas.

Lo más importante es ver los números que representan la dificultad de este ítem, 1.64 es la dificultad, estando este valor por encima de la media del test que es 0, esto implica que el ítem fue difícil de responder siendo uno de los que menos contestaron los estudiantes. Los valores que representan el ajuste INFIT MNSQ y OUTFIT MNSQ son para 0 *(0.83 y 0.89),* para 1 *(0.81 y 0.54),* y para 2 *(0.79 y 0.61)* respectivamente, que reflejan en alguna forma que las respuestas a este ítem se presentaron como se habrían podido suponer de antemano, es decir, respuestas mayoritariamente erradas.

Los resultados obtenidos en este ítem, permiten deducir que los estudiantes no tienen una idea clara sobre lo que implica la varianza en el ancho del intervalo, ya que sólo tres respondieron correctamente este ítem, 33 respondieron parcialmente bien la pregunta y 127 o no respondieron, o respondieron mal.

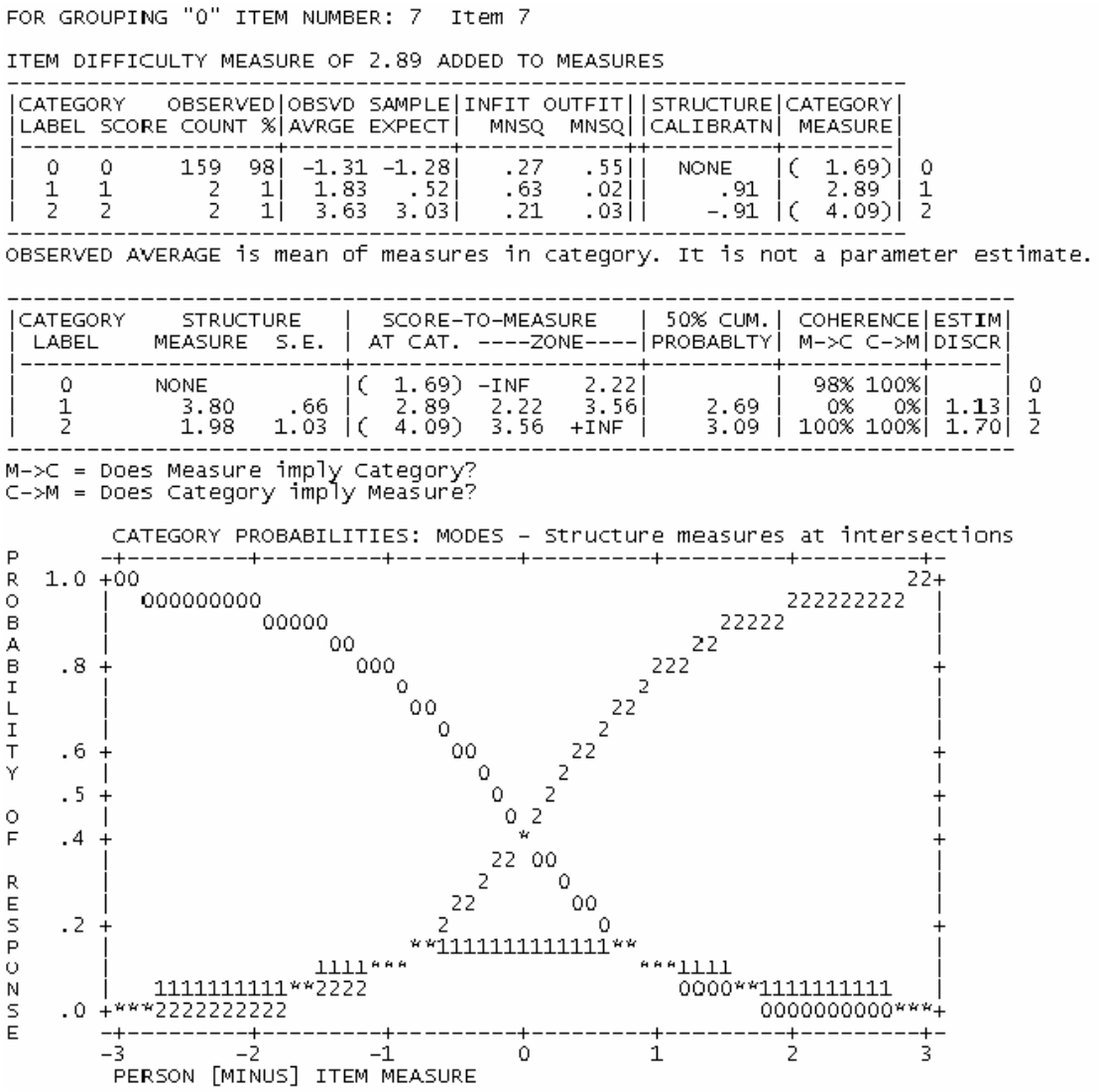


Gráfica 1. Distribución de las respuestas del ítem 4.

4.2.3.2. *Análisis del ítem 7*. Como se vio en el ítem 4, la distribución de las respuestas en este ítem se da de manera parecida, pero en este caso, el código 1 se distribuye en forma de campana muy abierta, a manera casi de una línea horizontal como se aprecia en la Gráfica 2 y sin ser la dominante en ningún intervalo de habilidad de los individuos. Este hecho gráfico está asociado con las pocas personas que lograron esta categoría en sus respuestas. Sólo dos

individuos lograron esta calificación, el 98% del grupo respondió de manera errónea este ítem.

La estimación de la dificultad que ofrece el modelo en este ítem es la mayor en el test, con un valor de 2.89 lo que se traduce en la enorme dificultad que representa para los evaluados hallar intervalos de confianza cuando la desviación poblacional no se conoce. Los valores que representan el ajuste INFIT MNSQ y OUTFIT MNSQ son para 0 *(0.27 y 0.55)*, para 1 *(0.63 y 0.2),* y para 2 *(0.21 y 0.3)* respectivamente; los valores en rojo como era de esperarse, muestran demasiado ajuste al modelo que se explica por la enorme dificultad que le representó a los estudiantes. Por los valores que arroja el software, la expectativa del modelo indica que las respuestas de este ítem ajustan demasiado bien, es decir, la dificultad sobresale y fue respondido por sólo dos personas de las 164 evaluadas, en cambio 111 personas se abstuvieron de responder la pregunta y 49 respondieron de manera errada el ítem.



Gráfica 2. Distribución de las respuestas del ítem 7.

4.2.4. ANÁLISIS DE LOS ÍTEMS DICOTÓMICOS DEL TEST

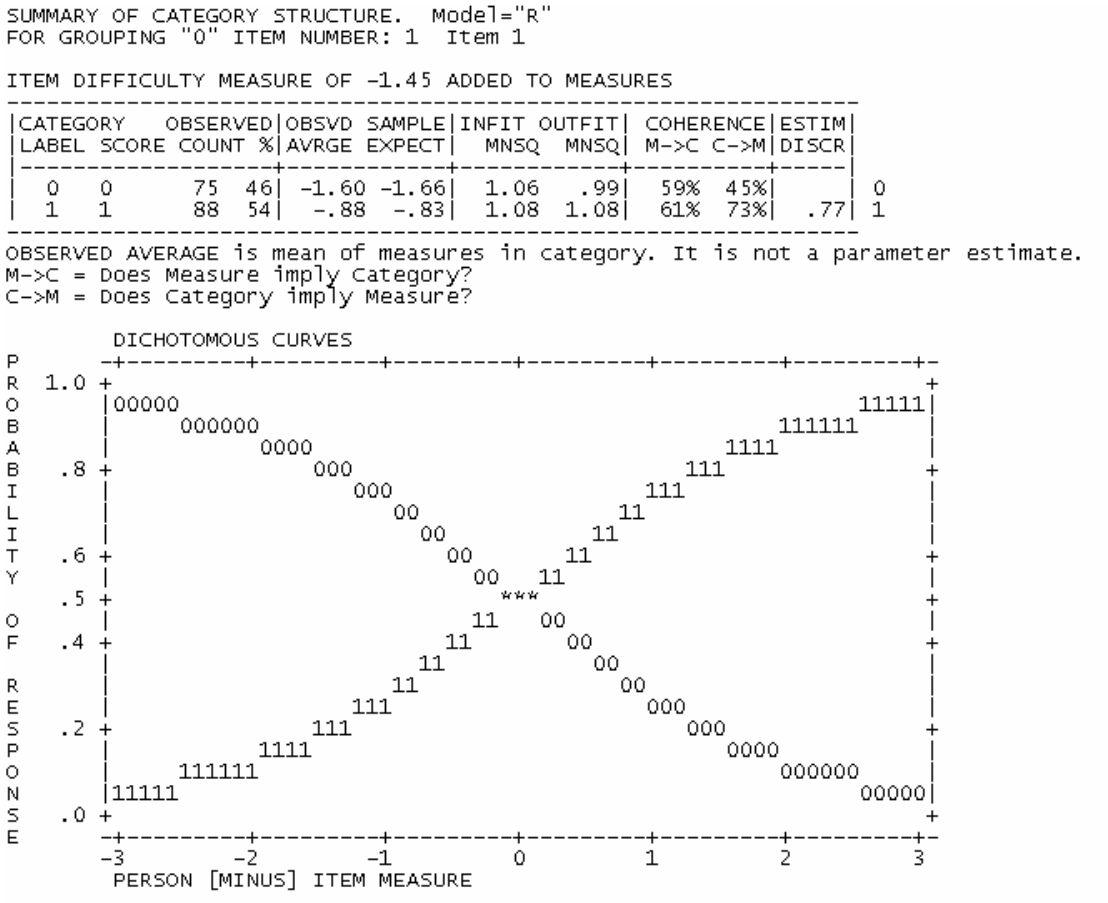
La distribución de las respuestas es igual en todos los ítems dicotómicos cuando los datos se ajustan al modelo. Como el tipo de respuesta es binaria, 0 para respuesta errada y 1 para respuesta correcta, la gráfica muestra que la probabilidad de obtener 0 es mayor cuando la diferencia entre la habilidad del

individuo y la dificultad del ítem es cada vez más pequeña (valor negativo). De manera similar, la respuesta correcta es más probable cuando la diferencia está a favor del individuo y se hace cada vez más grande.
Se señala también la cantidad de personas que respondieron correctamente el ítem, con el respectivo porcentaje para cada grupo de respuestas, el INFIT MNSQ y el OUTFIT MNSQ para 0 y 1, y la dificultad media del ítem que aparece en la parte superior de la tabla.

4.2.4.1. *Análisis del ítem 1*. Los datos que presenta la Gráfica 3 son los siguientes: 88 personas respondieron acertadamente el ítem, es decir un 54% de los evaluados y 75 lo hicieron de forma errónea correspondiente al 46% de los 164 sometidos a la prueba. La dificultad del ítem -1.45, arrojando esta pregunta un bajo nivel de exigencia ya que este valor está muy alejado del cero, que es la media de la dificultad de los ítems del test. Los valores del OUTFIT MNSQ y el INFIT MNSQ de cada una de las posibilidades de respuesta no se salen de los parámetros aceptados para ajustarse al modelo al obtener para el 0 *(1.06 y 0.99)* y para el 1 *(1.08 y 1.08)* respectivamente.

Este ítem se puede considerar fue un ítem fácil para los estudiantes evaluados aunque, aproximadamente, sólo la mitad de los evaluados lo respondieron bien. Esto hace pensar que la habilidad del grupo está por debajo de la exigencia del test, pues cuando la mitad de las personas responden correctamente una pregunta, lo que se espera es que la dificultad media de la misma este próxima a cero para que haya un equilibrio entre la habilidad del grupo y la dificultad del test.

La definición del intervalo de confianza y su variación cuando se calculan diferentes intervalos tomando muestras de la misma población, es un tema que los estudiantes entienden y se refleja en los datos arrojados por el modelo, al ser un ítem con un nivel de dificultad bajo con respecto a la media de dificultad que es cero.



Gráfica 3. Distribución de las respuestas del ítem 1.

4.2.4.2. *Análisis del ítem 2*. La dificultad de este ítem es de -1.22, presenta un poco más de exigencia que el anterior. Los estadísticos de ajuste cumplen con lo requerido permitiendo contar con este ítem en el test, ya que los valores para el INFIT y OUTFIT son para el 0 *(0.95 y 1.00)* y para el 1 *(0.92 y 0.84)* respectivamente para cada MNSQ como lo presenta la Gráfica 4.

En este ítem las respuestas de las personas se dieron casi equitativamente con un 49% para las correctas y un 51% para las incorrectas, y confirmando lo dicho en el análisis del ítem anterior. Aún más, se puede suponer que la habilidad media del

grupo debe estar cerca del valor de dificultad de este ítem, por contar con casi un 50% en las dos opciones. La siguiente Gráfica respalda la afirmación anterior.

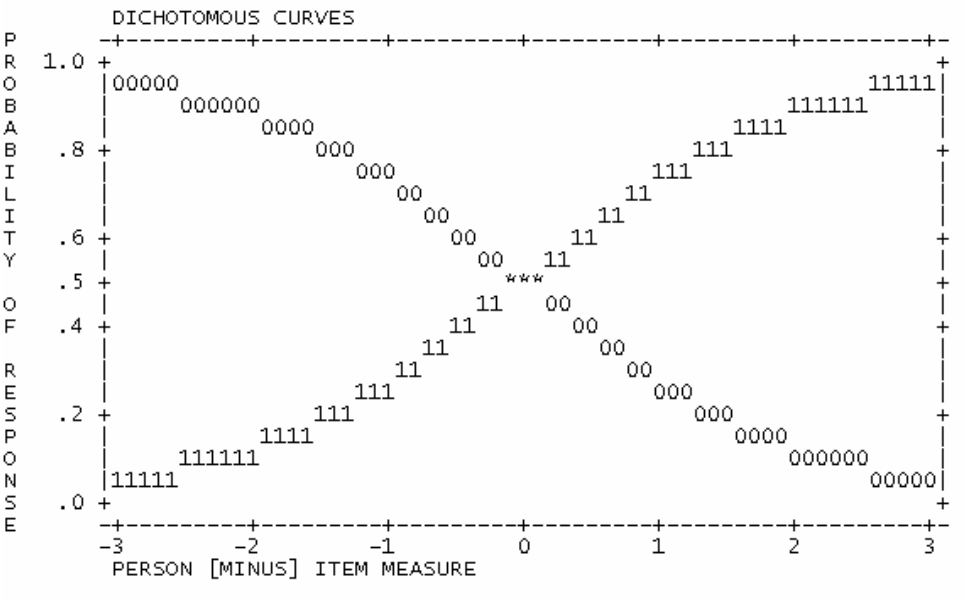
La mitad del grupo entiende que cuando el ancho de los intervalos de confianza disminuye, aumenta el tamaño de la muestra, con variación conocida o desconocida. Aunque la mitad respondió bien esta pregunta, el ítem tiene una baja dificultad con respecto a la media del test. El 31.7% creen que si aumenta el tamaño de la muestra la variación en el ancho del intervalo aumenta, los demás presentan errores de otros tipos.

SUMMARY OF CATEGORY STRUCTURE. Model="R"
FOR GROUPING "0" ITEM NUMBER: 2 Item 2

ITEM DIFFICULTY MEASURE OF -1.22 ADDED TO MEASURES

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	COHERENCE M->C	COHERENCE C->M	ESTIM DISCR	
0	0	83	51	-1.69	-1.63	.95	1.00	62%	75%		0
1	1	80	49	-.71	-.78	.92	.84	67%	52%	1.26	1

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?



Gráfica 4. Distribución de las respuestas del ítem 2.

4.2.4.3. *Análisis del ítem 3.* La media de la dificultad de este ítem es la menor de todas las preguntas, es decir, la pregunta más fácil para el grupo en general, con un porcentaje de respuestas correctas de 56%, y una media de dificultad estimada en lógitos de -1.57. Presenta unos valores de ajuste que aparte de estar dentro de los permitidos para el ajuste, son valores muy buenos con un INFIT y OUTFIT para el 0 *(1.09 y 1.08)* y para el 1 *(1.12 y 1.18)* respectivamente, como se aprecia en la Gráfica 5.

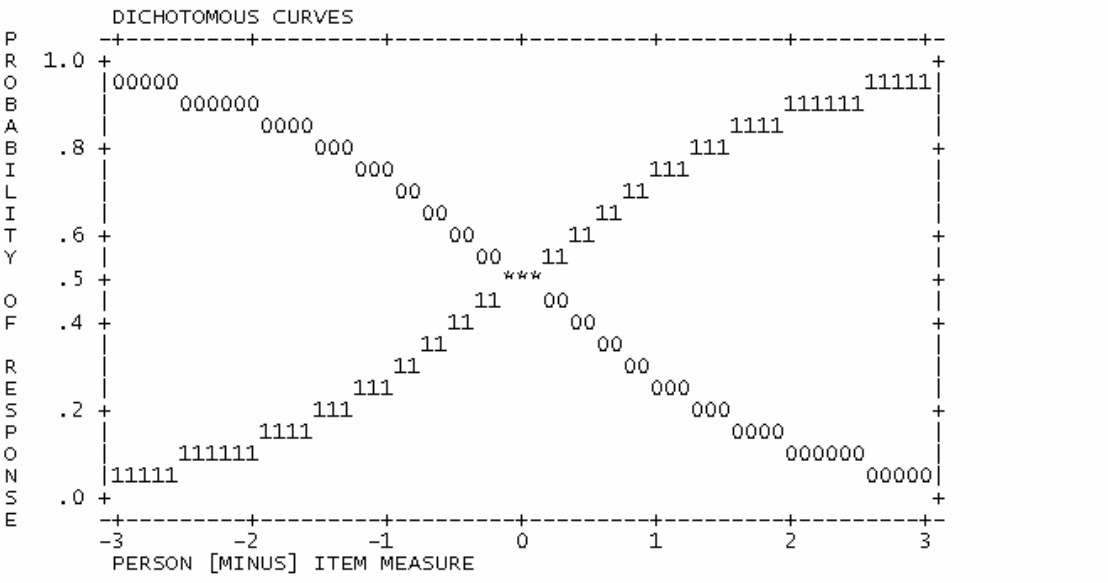
Este ítem da el mínimo valor de dificultad al que fueron enfrentados los estudiantes, y deduce que la mayoría de ellos entienden la relación que hay entre el nivel de confianza y la construcción del intervalo de confianza. Un número significativo de los que eligieron distractores, se inclinaron a pensar, que si el nivel de confianza disminuye, el intervalo de confianza aumenta, siendo la confusión más propensa en este ítem con un 28.3% de los evaluados

SUMMARY OF CATEGORY STRUCTURE. Model="R"
FOR GROUPING "0" ITEM NUMBER: 3 Item 3

ITEM DIFFICULTY MEASURE OF -1.57 ADDED TO MEASURES

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	COHERENCE M->C	C->M	ESTIM DISCR	
0	0	71	44	-1.56	-1.67	1.09	1.08	64%	52%		0
1	1	92	56	-.94	-.85	1.12	1.18	67%	78%	.60	1

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

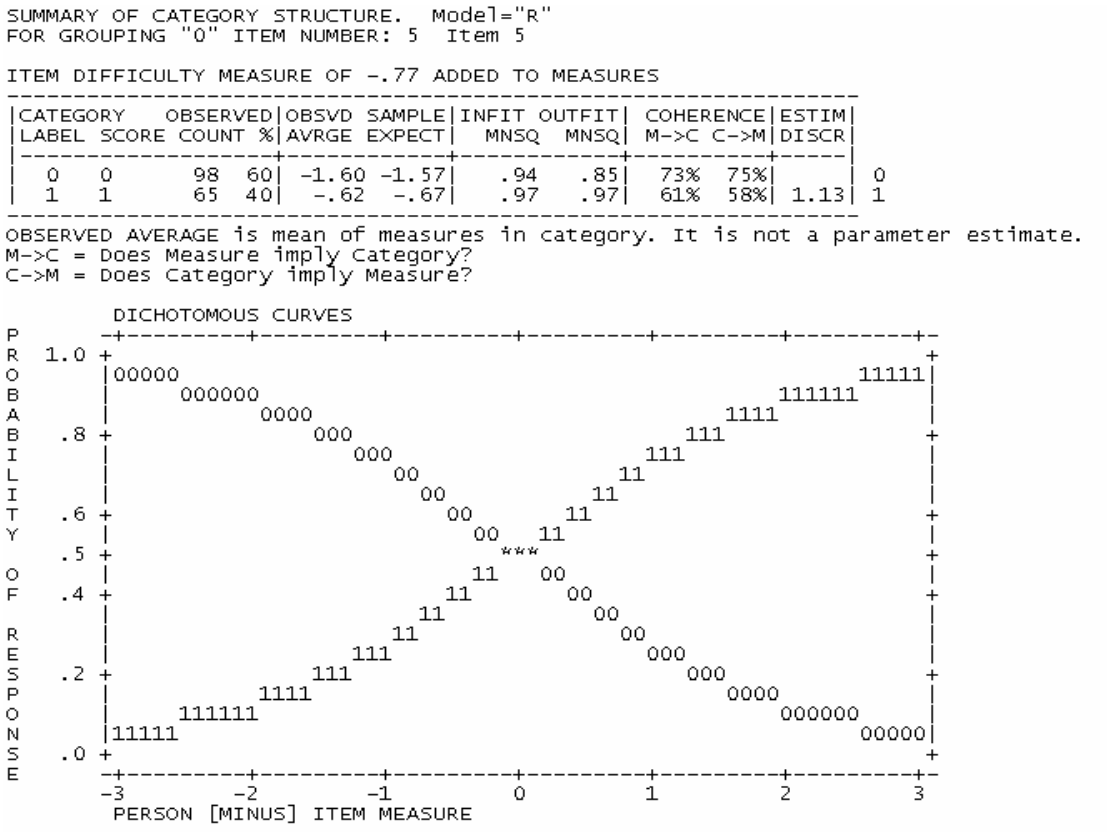


Gráfica 5. Distribución de las respuestas del ítem 3.

4.2.4.4. *Análisis del ítem 5.* La media de dificultad de esta pregunta fue de -0.77, teniendo un 40% de respuestas correctas. En general, el grupo tiene diferentes problemas con el significado del nivel de confianza, el 60% de los evaluados respondió alguno de los distractores del ítem. Los valores de ajuste aparte de estar dentro de los permitidos para el arreglo, son valores muy buenos con un INFIT y OUTFIT para el 0 *(0.94 y 0.85)* y para el 1 *(0.97 y 0.97)* respectivamente para cada MNSQ como lo muestra la Gráfica 6.

Se percibe que este ítem tiene un valor de dificultad que está por encima de la media de habilidad del grupo, puesto que el porcentaje de personas que respondieron correctamente no llegó a la mitad.

Aunque el valor de dificultad del ítem es negativo aún, el porcentaje de respuestas correctas es menos de la mitad del total; este valor de -0.77 refleja un problema con la mayoría de los estudiantes con el significado del nivel de confianza (variación del intervalo en diferentes muestras). Los demás evaluados, tienen conflictos con diferentes conceptos acerca de la construcción del intervalo de confianza.



Gráfica 6. Distribución de las respuestas del ítem 5.

4.2.4.5. *Análisis del ítem 6.* Este ítem reporta datos de mayor grado de exigencia respecto a los anteriores determinado por su medida de dificultad: -0.42, con un 33% de evaluados que respondieron correctamente frente a un 67% que respondieron erróneamente. Esto permite apreciar el desconocimiento que tiene el grupo respecto al procedimiento para estimar la media de una población normal con σ conocida. Los valores de ajuste están dentro de los permitidos para el arreglo, son valores muy buenos con un INFIT y OUTFIT para el 0 *(1.03 y 0.94)* y para el 1 *(1.01 y 0.96)* respectivamente para cada MNSQ. Ver Gráfica 7.

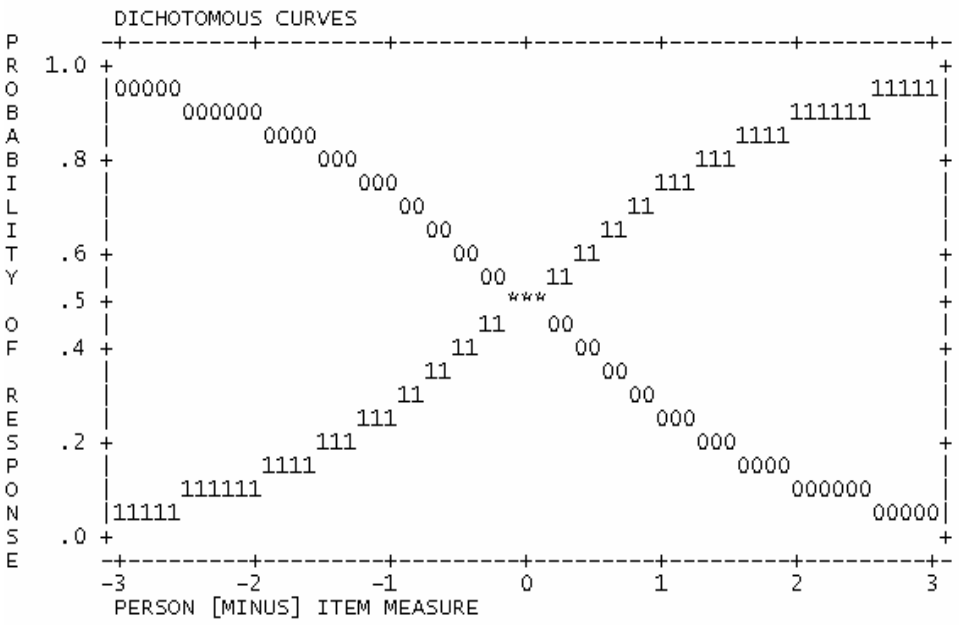
Los estudiantes sometidos a esta prueba, muestran en su mayoría en este ítem, un conflicto sobre cómo se debe estimar la media de una población normal en una muestra grande con desviación conocida. Los errores frecuentemente cometidos por los estudiantes consisten en olvidar multiplicar el valor crítico de la distribución normal estándar por el error estándar, o dividir por la desviación típica poblacional.

SUMMARY OF CATEGORY STRUCTURE. Model="R"
FOR GROUPING "0" ITEM NUMBER: 6 Item 6

ITEM DIFFICULTY MEASURE OF -.42 ADDED TO MEASURES

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	COHERENCE M->C	C->M	ESTIM DISCR	
0	0	109	67	-1.52	-1.52	1.03	.94	71%	89%		0
1	1	54	33	-.58	-.58	1.01	.96	57%	27%	.99	1

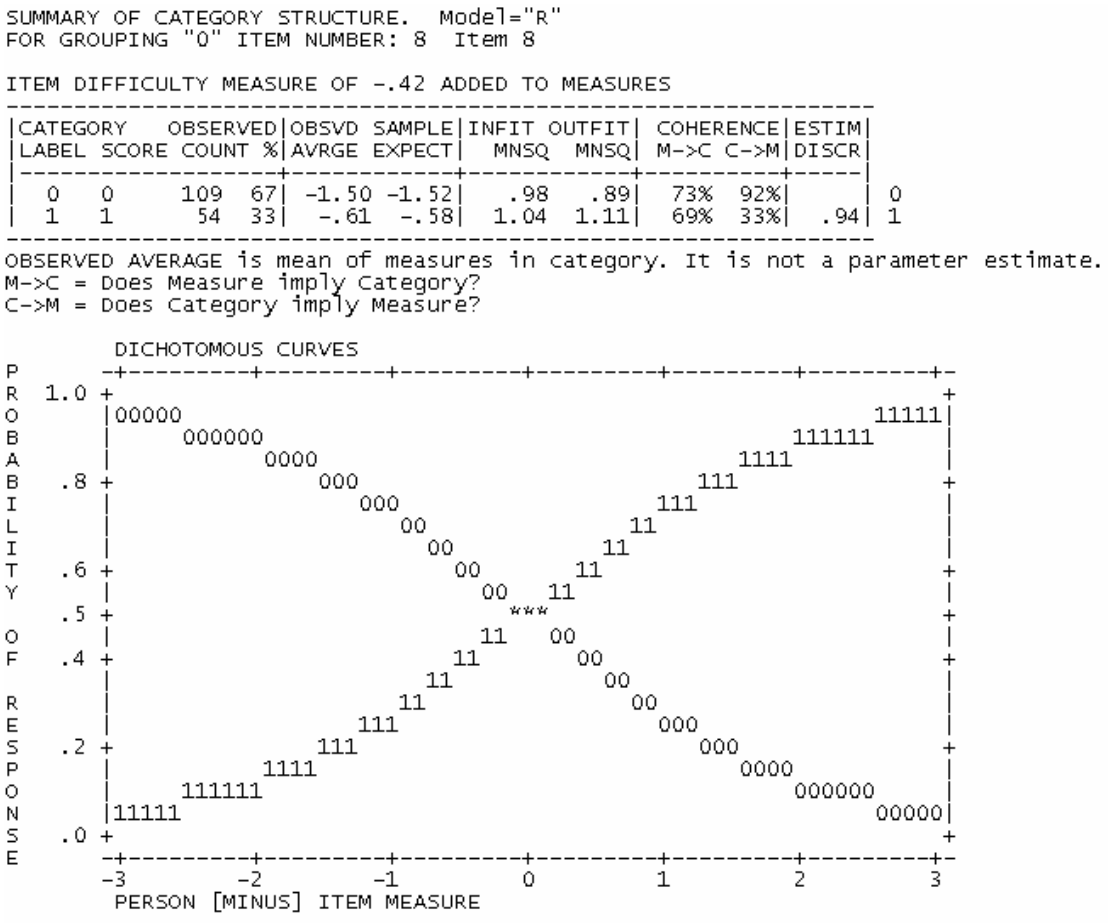
OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?



Gráfica 7. Distribución de las respuestas del ítem 6.

4.2.4.6. *Análisis del ítem 8*. La dificultad de este ítem estimada en lógitos es de -0.42, con un 33% de evaluados que respondieron correctamente esta pregunta y el 67% restante que se inclinaron por alguna de las otras opciones del ítem. Estos números señalan que el grupo tiene problemas para estimar la media de una población a partir de datos experimentales, σ desconocida y muestra grande. Los valores del INFIT y el OUTFIT para el 0 *(0.98 y 0.89)* y para el 1 *(1.04 y 1.11)* respectivamente para cada MNSQ, mostrando valores aceptados por el modelo para un buen ajuste del ítem como lo muestra la Gráfica 8 a continuación.

Los evaluados poseen problemas al estimar la media de una población a partir de datos experimentales con desviación desconocida y muestra grande. Nótese que este ítem presenta el mismo índice de dificultad que el ítem 6, con el mismo porcentaje de respuestas correctas y respuestas incorrectas. El distractor que prevalece en este tema, es el mal manejo de la tabla de la normal estándar donde se obtienen los valores críticos y se utiliza 1.64 para un 95% de confianza en vez de 1.96, al tomar cada lado de la curva con un 5% en lugar de 2.5%.



Gráfica 8. Distribución de las respuestas del ítem 8.

4.2.4.7. *Análisis del ítem 9***.** Con una medida de dificultad de 0.21 este ítem está por encima de la media de dificultad del test, obteniendo un 23% de respuestas correctas contra un 77% de respuestas erradas. Los valores del INFIT y el OUTFIT para el 0 *(1.36 y 2.62)* y para el 1 *(1.11 y 1.06)* respectivamente para cada MNSQ, mostrando el valor que está en rojo ruido y falta de ajuste, debido a que sobrepasan el intervalo exigido por el modelo para un buen análisis. Esto se debe a que personas con poca habilidad estuvieron respondiendo correctamente a este ítem y aunque en un principio se intentó disminuir este valor eliminando algunos participantes del análisis, los resultados no fueron relevantes y se continuó con todos los estudiantes en el proceso.

El valor OUTFIT de 2.62 indica que existen respuestas inesperadas o anomalías lejos del nivel de habilidad de la persona para esta calificación. Al tratar de mejorar el ajuste, este ítem no fue excluido porque la cantidad de preguntas que componen el test son necesarias para cubrir la totalidad del tema a evaluar y los otros estadígrafos de ajuste proporcionan buenas medidas (Ver Gráfica 9).

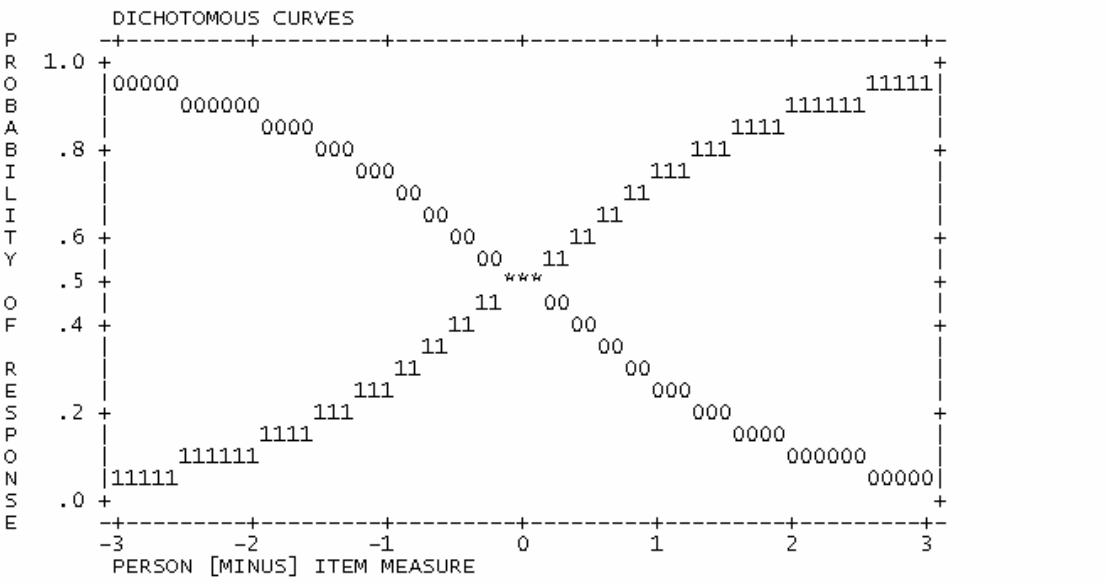
El nivel de habilidad de esta pregunta está por encima de la media, presentando gran dificultad los alumnos a la hora de determinar valores críticos en la distribución del estadígrafo, ya que la mayoría de ellos o no reconoce en que caso se usa la distribución t en lugar de la normal estándar, o utilizan los grados de libertad incorrectos.

SUMMARY OF CATEGORY STRUCTURE. Model="R"
FOR GROUPING "0" ITEM NUMBER: 9 Item 9

ITEM DIFFICULTY MEASURE OF .21 ADDED TO MEASURES

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	COHERENCE M->C	C->M	ESTIM DISCR	
0	0	126	77	-1.37	-1.46	1.36	2.62	77%	92%		0
1	1	37	23	-.68	-.37	1.11	1.06	28%	10%	.77	1

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?



Gráfica 9. Distribución de las respuestas del ítem 9.

4.2.4.8. *Análisis del ítem 10*. La medida de dificultad de esta pregunta es de 1.13 siendo el ítem más difícil de los dicotómicos con un 12% de respuestas correctas contra un 88% de respuestas incorrectas, mostrando que los evaluados poseen una gran falla a la hora de interpretar gráficos sobre los intervalos de confianza.

Los estadísticos de ajuste están en el rango aceptado de ajuste: el INFIT y el OUTFIT son para el 0 *(1.04 y 0.97)* y para el 1 *(1.01 y 0.85)* respectivamente para cada MNSQ como lo presenta la Gráfica 10.

Aunque esta pregunta no está muy lejos de la media de dificultad de los ítems, solo 20 personas la respondieron correctamente, lo que permite deducir que el nivel de habilidad del grupo evaluado es inferior al nivel de dificultar de los ítems.

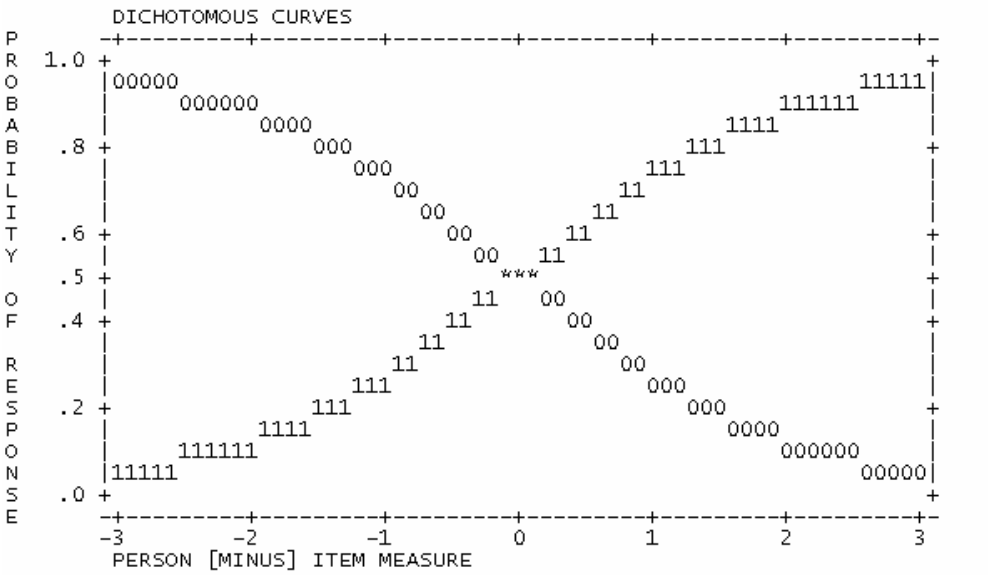
La interpretación de gráficos en los intervalos de confianza, son un déficit casi absoluto para este grupo, pues el 88% de los estudiantes presentan entre otras, confusión en conceptos sobre las diferencias de las medias poblacionales cuando no se solapan los intervalos de confianza.

SUMMARY OF CATEGORY STRUCTURE. Model="R"
FOR GROUPING "0" ITEM NUMBER: 10 Item 10

ITEM DIFFICULTY MEASURE OF 1.13 ADDED TO MEASURES

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	COHERENCE M->C	COHERENCE C->M	ESTIM DISCR	
0	0	143	88	-1.38	-1.38	1.04	.97	90%	98%		0
1	1	20	12	.01	.00	1.01	.85	71%	25%	1.00	1

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?



Gráfica 10. Distribución de las respuestas del ítem 10.

4.2.5. ANÁLISIS GLOBAL Y CONJUNTO DE PERSONAS E ÍTEMS

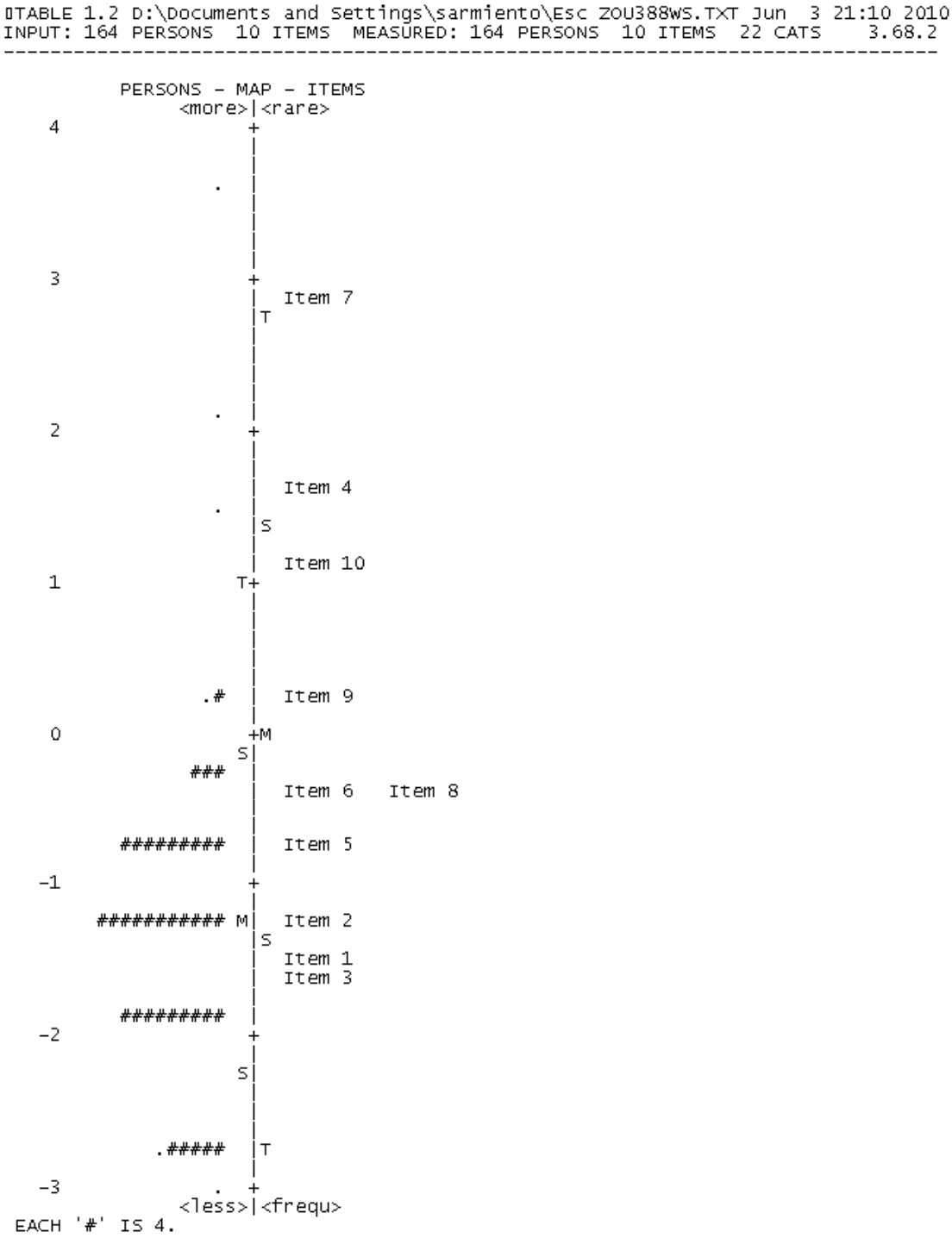
Las gráficas anteriores han mostrado la información de cada ítem por separado con respecto a cada posibilidad de respuesta (correcta e incorrecta) y sus estadísticos de ajuste; pero no se tiene claro cómo se comporta cada uno de los ítems con respecto a la distribución de la población que los responde, o cómo se comporta la habilidad de cada uno de los individuos con respecto a cada ítem y con respecto al test, pues hasta ahora no se ha presentado una herramienta que enfrente estas dos variables. Para realizar esta confrontación WINSTEPS sitúa en una misma escala las puntuaciones de las personas y de los ítems estimados en lógitos, ofreciendo una mejor manera de interpretar la información. Estos datos se ilustran en la Gráfica 11 y se explican de manera más profunda a continuación.

En la parte izquierda de la gráfica se tiene la distribución de los estudiantes representando con un signo numeral a 4 personas y con un punto a 2, las variables ubicadas a la derecha son los ítems cada uno con su respectivo nombre, y la columna de números en la parte izquierda, es la escala que se usa para las dos variables a medir que son la habilidad de los sujetos y la dificultad de los ítems.

Lo primero que se debe saber es que la medida de la habilidad de los sujetos y de la dificultad de los ítems se encuentran en la parte izquierda de tal forma que los valores bajos están en la parte inferior de la gráfica y los altos en la parte superior, por ejemplo, el ítem 3 es el más fácil y el ítem 7 el más difícil; de igual manera, la persona representada por un punto ubicado en -3 es el sujeto menos hábil y el individuo que está situado por encima del 3 es al más hábil.

La media de habilidad de las personas esta en *-1.21*, la media de dificultad del test está en *0,* el programa señala estos dos índices con la letra M que se ve en cada

lado, representando una desviación estándar con la letra S y dos desviaciones con la letra T.



Gráfica 11. Mapa de las habilidades de las personas y las dificultades de los ítems.

Ahora, si se enfrentan las medias de cada una de las variables se observa que evidentemente la habilidad de las personas está por debajo de la dificultad del test, o dicho de otra forma, el test resultó difícil para el grupo de estudiantes evaluados.

Al examinar ítems como el 7 y el 4, que tienen un grado de dificultad bastante elevado para el grupo de estudiantes, los puntos que representan las personas que tienen una probabilidad mayor a 0.5 de responder al menos uno de estos ítems correctamente, son sólo cuatro. Ambos ítems requieren conocer claramente la definición y la forma de construir intervalos de confianza, es así que el ítem 4 pregunta cómo varía el ancho del intervalo al tomar una población con varianza 4 veces mayor manteniendo los demás datos igual, y el ítem 7, pide determinar un intervalo de confianza teniendo una desviación poblacional desconocida.

Al hacer un barrido por las preguntas del test para encontrar una relación con la dificultad de los mismos, se puede confirmar lo dicho anteriormente sobre las preguntas conceptuales, pues los ítems 3, 1, 2, y 5 en su orden de dificultad, del menos difícil al más difícil, tienen un tipo de estructura interrogativa. Los temas del test que mejor entiende el grupo son los conceptos sobre definición del intervalo de confianza, la relación del tamaño de la muestra con el ancho del intervalo, la relación entre el nivel de confianza y el ancho del intervalo y el significado del nivel de confianza. El ítem 10 presenta una estructura conceptual-visual al interrogar sobre el análisis de un gráfico y señala que los estudiantes están presentando dificultades en el análisis de gráficos referentes a los intervalos de confianza.

Las preguntas que necesitan cálculos para dar una respuesta, revelan mayor dificultad que las preguntas conceptuales. La ubicación de la dificultad de los ítems 6, 8, y 9 que aparece en el gráfico y los cálculos en estas preguntas, que encierran los conceptos de deducción del intervalo de confianza con desviación poblacional conocida, efecto de la varianza sobre el cálculo de intervalos y cálculo

del intervalo con desviación poblacional desconocida, evalúan conceptos distintos aunque algunos de ellos tienen la misma dificultad.

Los espacios entre cada uno de los ítems se interpreta como falta de datos para la estimación de las habilidades de las personas, es decir, entre más densa está la distribución de ítems, más precisa es la estimación de habilidades. En forma complementaria, la inexistencia de estudiantes en frente de los ítems 9, 10, 4 y 7 impiden que la estimación de la dificultad de estos ítems se realice con mayor exactitud. Las mejores estimaciones de ítems se dan en la parte central de la distribución de habilidad de las personas, debido a que en los extremos existen dos fenómenos comunes en la aplicación de un examen, el primero, es que los ítems fáciles van a ser respondidos por una gran mayoría de personas, y en el segundo ocurre todo lo contrario, la gran mayoría no responde las preguntas difíciles, esto hace que la información para la estimación se reduzca a medida que los ítems se alejan de la media de dificultad del test, aumentando así el error de estimación en los extremos. Como se observa en la Tabla 1 los ítems fáciles como el ítem 3, 1, 2, y 5, tienen una mejor estimación que los demás por estar cerca del centro de la distribución de habilidad.

La distribución de personas es aproximadamente normal frente a la distribución de ítems que está más comprimida en la parte baja del gráfico y se va dispersando a medida que la dificultad aumenta.

En cuanto a la estimación de las personas, se observa un grupo con un nivel de habilidad que se ubica por debajo del ítem 3, que es la pregunta que requiere menos habilidad del test. Tanto estas personas, como las que se ubican en la parte superior del grupo, poseen una estimación de habilidad menos precisa que los que se ubican en la parte central de la distribución, por la escasez de ítems que los estiman. Los estudiantes que se ubican frente al ítem 2 tienen una probabilidad de responder esa pregunta de 0.5 por estar en el mismo nivel, igual

ocurre con los que están frente al ítem 5 y al ítem 9. El grupo de estudiantes que está frente al ítem 2, tiene una probabilidad de responder el ítem 3 mayor a 0.5 y una probabilidad de responder al ítem 4 menor a 0.5 por su ubicación.

4.2.6. ANÁLISIS DE LA ESTIMACIÓN DEL CONJUNTO DE ÍTEMS

Como la Gráfica 11 no ofrece con exactitud el valor de estimación de los ítems ni de las personas, WINSTEPS proporciona estos valores. La Tabla 13 contiene los valores de estimación de los ítems con sus estadísticos de ajuste y otros datos interesantes para el análisis.

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	ITEM	G
7	6	164	2.89	.46	.36	-1.4	.03	-1.6	Item 7	0
4	39	164	1.64	.20	.81	-1.4	.61	-2.3	Item 4	0
10	20	164	1.13	.27	1.02	.2	.86	-.4	Item 10	0
9	37	164	.21	.21	1.17	1.4	1.42	2.1	Item 9	0
6	54	164	-.42	.18	1.02	.3	.95	-.3	Item 6	0
8	54	164	-.42	.18	1.02	.3	1.04	.4	Item 8	0
5	65	164	-.77	.18	.96	-.6	.92	-.6	Item 5	0
2	80	164	-1.22	.17	.93	-1.2	.92	-.6	Item 2	0
1	88	164	-1.45	.17	1.07	1.2	1.03	.3	Item 1	0
3	92	164	-1.57	.17	1.11	1.8	1.12	.8	Item 3	0
MEAN	53.5	164.0	.00	.22	.95	.1	.89	-.2		
S.D.	27.1	.0	1.39	.09	.22	1.1	.34	1.2		

Tabla 13. Valores de estimación de los ítems.

La tabla anterior muestra información de los ítems del test que se describen a continuación analizando cada una de las columnas y los datos que contiene.

4.2.6.1. *ENTRY NUMBER***.** Es el número de cada ítem y aparecen ordenados de mayor a menor dificultad.

4.2.6.2. *TOTAL SCORE***.** Es el puntaje logrado para cada ítem, es decir, la cantidad de estudiantes que lo respondieron bien. El ítem con menos puntaje fue el 7 con 6 buenas respuestas, y el de mayor puntaje fue el ítem 3 que obtuvo 92 respuestas acertadas.

4.2.6.3. *COUNT***.** El número de personas que respondieron a los ítems.

4.2.6.4. *MEASURE.* En esta columna aparecen los valores de dificultad de los ítems estimado en lógitos, organizados de menor a mayor dificultad, que van desde -1.57 hasta 2.89.

Los dos ítems politómicos reflejaron mayor dificultad para los estudiantes, el ítem 4 que requiere un concepto sobre el efecto de la varianza en el intervalo de confianza y el ítem 7, una pregunta sobre un problema aplicado que pide hallar un intervalo de confianza. Confirmando que en este examen a los alumnos les fue más difícil responder este tipo de preguntas abiertas. Los ítems donde los individuos demostraron menor dificultad al responder, fueron los ítems 3, 1, 2 y 5 dicotómicos de selección múltiples y conceptuales, los demás ítems dicotómicos requirieron de mayor destreza, al parecer debido a los cálculos necesarios para resolverlos, estos fueron los ítems 8, 6 y 9.

4.2.6.5. *MODEL S.E*. Es el error de estimación de dificultad de los ítems, corroborando que efectivamente los ítems difíciles tienen un error mayor al de los ítems fáciles en esta aplicación, por tener pocas personas estimando estas preguntas. En este estudio las mejores estimaciones se encontraron en los ítems de baja dificultad por tener gran cantidad de personas estimándolos.

4.2.6.6. *INFIT MNSQ***.** Es el estadístico que señala anomalías dadas por respuestas inesperadas de estudiantes, cerca del nivel de dificultad del ítem, es decir, arroja valores alejados del intervalo entre 0.8 y 1.3, cuando los estudiantes

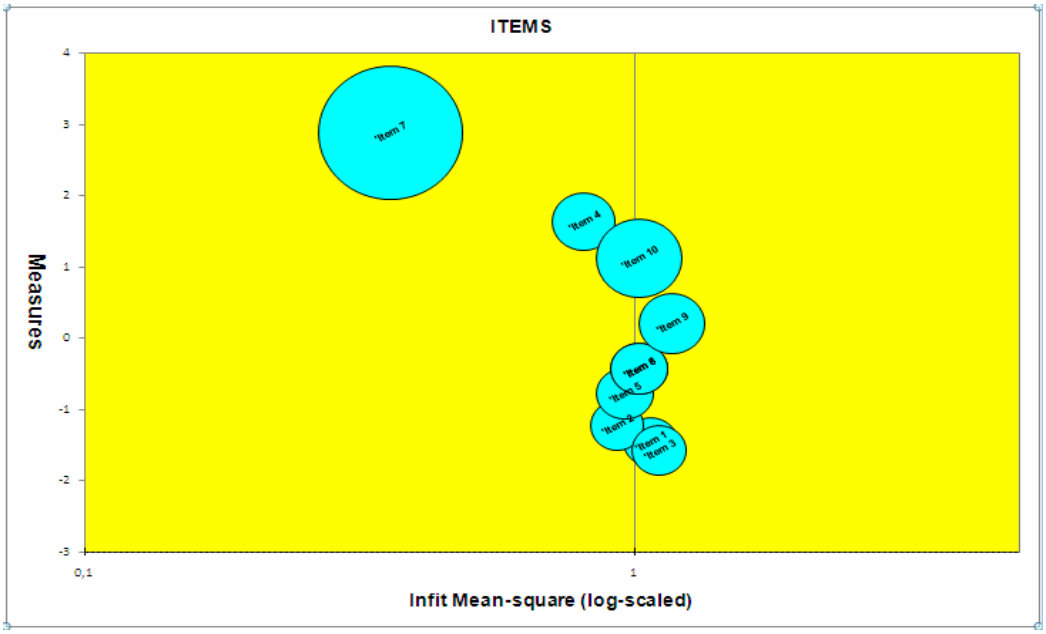
que tienen una habilidad equivalente a la de un ítem, responden de manera inesperada esta pregunta.

El ítem 7 es el único en presentar irregularidades por tener 0.36 de INFIT MNSQ. Fue respondido correctamente por solo 4 estudiantes, por esto el determinismo, o sea, el patrón de respuestas al ítem es muy predecible: casi nadie la responde.

Aunque los demás ítems ajustan bien al modelo, sería de gran utilidad ver un gráfico que mostrara estos datos de manera más clara que en la tabla. WINSTEPS presenta cada estadístico con respecto a la medida de dificultad de los ítems incluyendo el error de estimación de cada uno.

- **INFIT MNSQ Vs. Dificultad de ítems.** En la Gráfica 12 se observa el INFIT MNSQ en el eje horizontal y la dificultad de los ítems estimada en lógitos en el eje vertical. Cada una de las circunferencias representa un ítem y su diámetro el error de estimación de dificultad; cuánto más grande es el diámetro, más grande es el error de estimación. Se puede apreciar que todos los ítems, con excepción del 7, se mantienen alrededor de la vertical que tiene valor de 1; aunque este ítem está alejado de los demás con un valor de 0.36, se demuestra mejor en el gráfico que con los datos numéricos. La decisión de no excluirlo del test se tomó porque los estadísticos de ajuste ZSTD, como se verá más adelante, no exceden los límites además de que los ítems con los que cuenta son demasiado pocos.

  Los demás ítems tienen valores de este estadístico que están entre 0.81 y 1.17, ofreciendo datos muy buenos para el análisis.



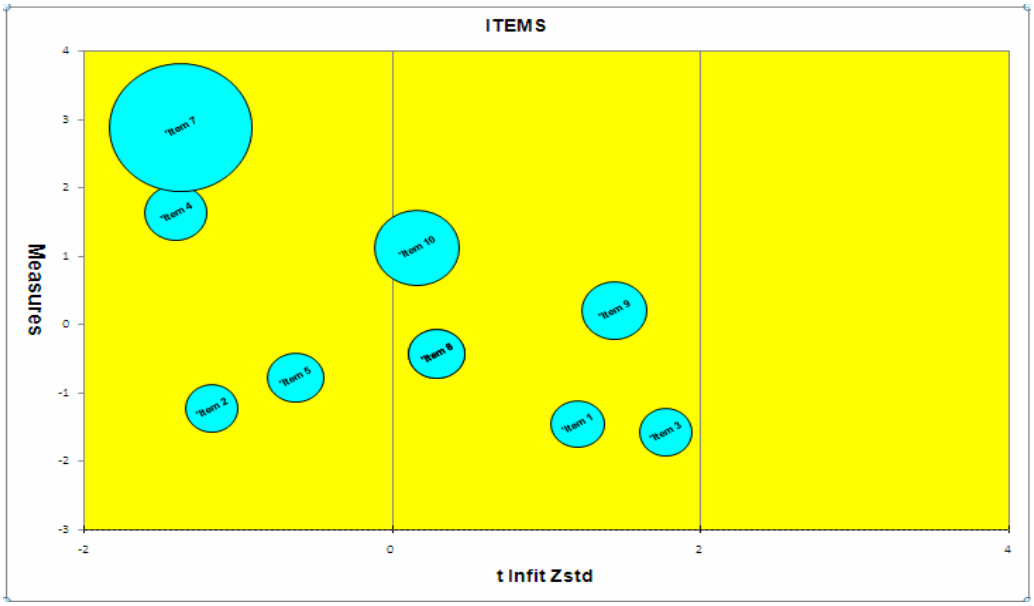
Gráfica 12. INFIT MNSQ Vs. Dificultad de los ítems.

El error de estimación del ítem 7 es mayor que el de los demás con 0.46, resultado que se explica por la poca cantidad de personas que lo estiman; el ítem 10 tiene un error de estimación de 0.27 superior al de los demás ítems, debido a que también su nivel de dificultad es mayor que el de los otros y por lo tanto contar con pocos estudiantes para estimarlo. Los demás ítems tienen errores entre 0.17 y 0.21, valores explicables por contar con una gran cantidad de personas estimándolos.

4.2.6.7. *INFIT ZSTD.* Es el mismo estadístico de medida cuadrática ponderada anterior, pero transformado de tal forma que el estadístico resultante posee distribución normal, lo que sugiere un intervalo entre -2 y 2 para un buen ajuste de los ítems al modelo; cuanto más se acerquen los valores de estos ítems a cero, el ajuste se vuelve más perfecto.

Se obtuvieron valores entre -1.4 y 1.8 en este estadístico, lo que permite mantener a todos los ítems considerados, incluso el ítem 7.

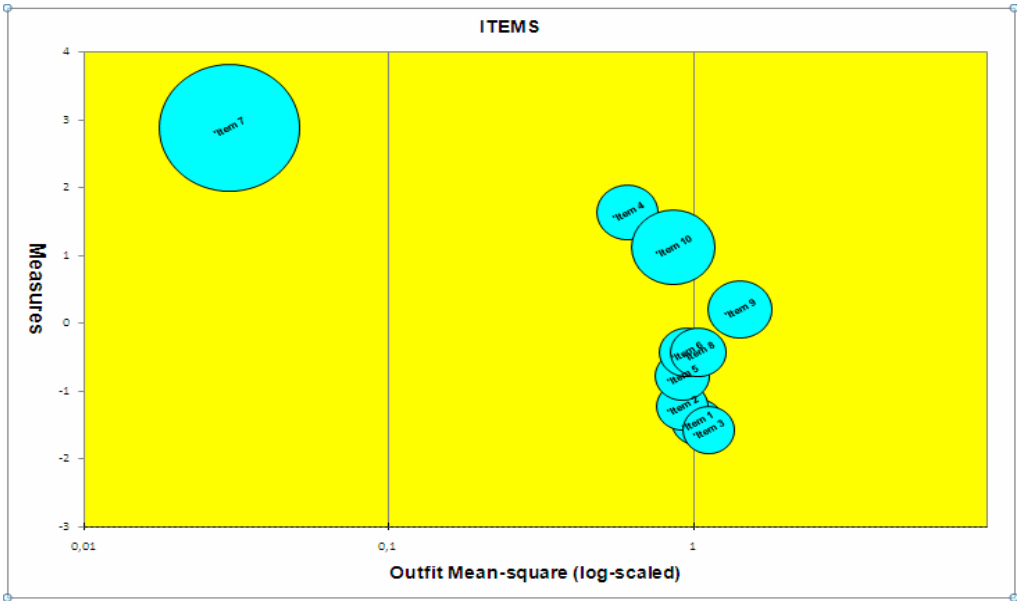
En la Gráfica 13 se encuentra el INFIT ZSTD en el eje horizontal y la dificultad de los ítems estimada en lógitos en el eje vertical. Cada una de las circunferencias representa lo mismo que en la Gráfica 12 analizada anteriormente. Aquí se puede apreciar que la totalidad de los ítems está dentro del intervalo sugerido por el modelo, obteniendo valores para este estadístico entre -1.4 y 1.8.



Gráfica 13.INFIT ZSTD vs. Dificultad de los ítems.

4.2.6.8. *OUTFIT MNSQ.* Es el estadístico de ajuste externo definido como el promedio de los residuales no ponderados, sensible a respuestas inesperadas lejos del nivel de dificultad de las preguntas, es decir, refleja las irregularidades en patrones de respuesta alejados del nivel de habilidad de los evaluados, con un intervalo entre 0.8 y 1.3 considerado como bueno para el proceso de medida de estas dificultades.

Se presentaron valores entre 0.03 y 1.42 llamando la atención el ítem 7 por su alto nivel de predicción en el patrón de respuestas adquirido, pero como se dijo anteriormente, no se elimina del test porque la idea no es disminuir la cantidad de ítems, debido a que solamente hay 10, y todos son necesarios en la evaluación. Las demás preguntas cuentan con valores que representan buen ajuste en los ítems. En la Gráfica 14 se aprecian lo valores Outfit que son aceptables para todos los ítems con la excepción del ítem 7 que posee un comportamiento muy predecible.

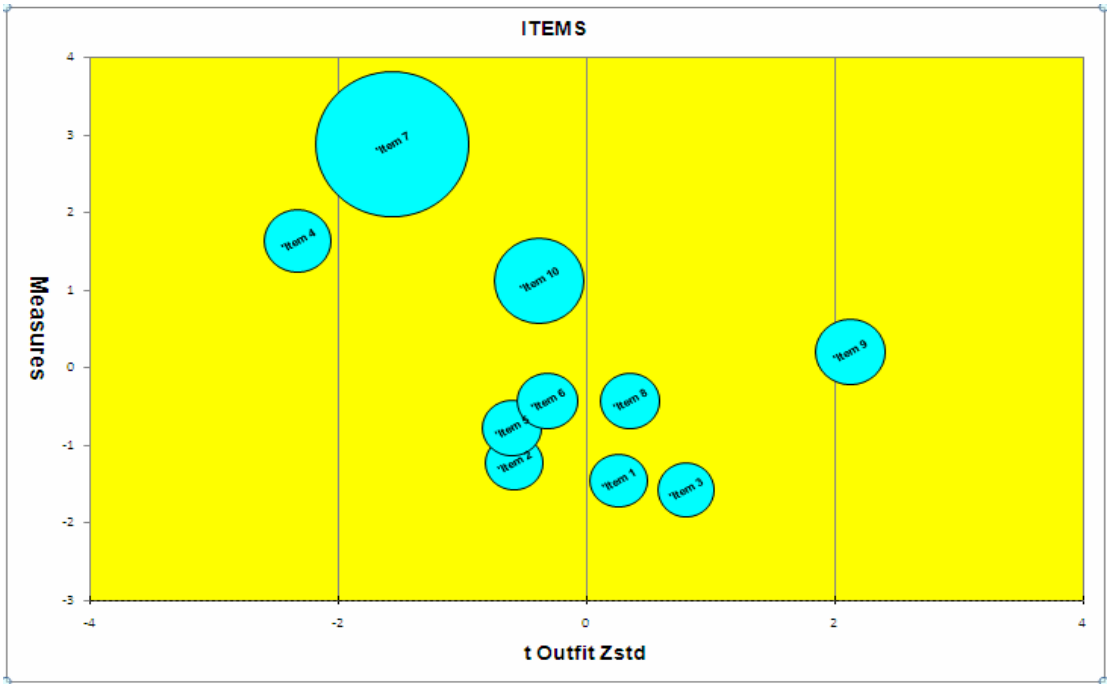


Gráfica 14. OUTFIT MNSQ vs. Dificultad de los ítems.

4.2.6.9. *OUTFIT ZSTD***.** Es el mismo estadístico de ajuste externo pero ahora transformado para que posea una distribución normal con un intervalo ideal de -2 a 2 para un buen ajuste.

Cuando un estudiante tiene una baja habilidad y termina respondiendo el ítem más difícil, o ítems que son fáciles y personas con alta habilidad terminan sin

responderlos, los valores del OUTFIT están lejos de ofrecer buenos ajustes. Los valores de este criterio oscilan entre -2.3 hasta 2.1, los ítems 4 y 9 presentaron levemente estas características, pero en los demás valores de ajuste fueron buenos, así que se determinó revisar las calificaciones en estos ítems y aunque se presentaron algunas anomalías en las respuestas, no tuvieron la suficiente fuerza para desistir de estas preguntas en el proceso. La Gráfica 15 representa claramente este criterio.



Gráfica 15. OUTFIT ZSTD vs. Dificultad de los ítems.

4.2.6.10. *ITEM.* Esta columna muestra el nombre que se le asigna a cada pregunta.

En la parte inferior de la Tabla 13 aparecen dos datos generales bastante importantes: la media de todos los temas antes analizados (MEAN) y su respectiva desviación estándar (S.D).

4.2.6.11. *MEAN.* Esta media es útil porque da información del test en general, y ofrece una idea rápida de lo que sucede en la aplicación.

La media del puntaje total en el test fue de 53.5 puntos de 166 posibles, sólo un 32% del total. Aumentan los argumentos para decir que el nivel de habilidad no fue suficiente, o que el nivel de dificultad del test es alto para este grupo de personas, es decir, los estudiantes presentan problemas al responder la mayoría de temas incorporados en el mismo. La media de dificultad de los ítems es cero y la media de los errores de estimación de éstas es de 0.22.

Con un INFIT MNSQ de 0.95 y un OUTFIT MNSQ de 0.89, las medias de estos estadísticos indican que el test ajusta bien al modelo, complementado con un INFIT ZSTD de 0.1 y un OUTFIT ZSTD de -0.2 bastante bueno por estar cerca del 0.

4.2.7. ANÁLISIS DE LA ESTIMACIÓN DEL CONJUNTO DE ESTUDIANTES

Así como se hizo el análisis de dificultad de los ítems, WINSTEPS calcula y presenta la habilidad de las personas de la misma forma, mostrando el valor de habilidad de cada uno, los errores de estimación y los estadísticos de ajuste de cada persona al modelo.

En la Tabla 14 que se muestra a continuación, aparecen los estudiantes ordenados en forma descendente respecto a su medida de habilidad; la identificación de cada individuo se encuentra localizada en la columna del extremo derecho, en este caso, se codificaron los nombres con números. La columna de la izquierda representa la localización de cada alumno en el archivo que se importó de Excel, así cuando se registren datos irregulares en los criterios de bondad del modelo, se puede revisar exactamente de que se trata la anomalía resaltada por el programa.

PERSON STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
94	11	10	3.63	1.00	1.30	.6	3.16	1.5	.05	.35	90.0	91.7	407
95	11	10	3.63	1.00	1.30	.6	3.16	1.5	.05	.35	90.0	91.7	408
88	9	10	2.18	.82	.60	-.4	.39	-.2	.73	.60	80.0	86.7	401
96	9	10	2.18	.82	.24	-1.3	.13	-.8	.83	.60	100.0	86.7	409
89	8	10	1.49	.83	.80	.0	1.66	.9	.44	.62	90.0	83.8	402
90	8	10	1.49	.83	.82	.0	.43	-.5	.81	.62	90.0	83.8	403
92	8	10	1.49	.83	1.14	.4	.81	.1	.65	.62	70.0	83.8	405
3	6	10	.27	.74	.89	-.1	1.03	.3	.53	.56	80.0	72.7	103
30	6	10	.27	.74	.88	-.2	.95	.2	.55	.56	80.0	72.7	203
43	6	10	.27	.74	.68	-.8	.56	-.4	.70	.56	80.0	72.7	216
48	6	10	.27	.74	.68	-.8	.56	-.4	.70	.56	80.0	72.7	221
50	6	10	.27	.74	1.29	.8	1.37	.7	.29	.56	60.0	72.7	223
97	6	10	.27	.74	1.89	2.0	1.33	.6	.49	.56	60.0	72.7	410
154	6	10	.27	.74	.55	-1.3	.44	-.7	.86	.56	80.0	72.7	915
2	5	10	-.25	.71	1.00	.1	.93	.2	.46	.52	70.0	70.1	102
10	5	10	-.25	.71	.97	.0	.89	.2	.48	.52	70.0	70.1	110
12	5	10	-.25	.71	.84	-.4	.82	.1	.62	.52	90.0	70.1	112
24	5	10	-.25	.71	1.18	.6	1.02	.3	.37	.52	50.0	70.1	124
36	5	10	-.25	.71	.75	-.8	.63	-.2	.69	.52	70.0	70.1	209
39	5	10	-.25	.71	.84	-.5	.71	.0	.64	.52	70.0	70.1	212
49	5	10	-.25	.71	.74	-.8	.62	-.2	.64	.52	70.0	70.1	222
60	5	10	-.25	.71	.98	.0	.90	.2	.54	.52	70.0	70.1	309
63	5	10	-.25	.71	1.35	1.1	1.35	.7	.30	.52	50.0	70.1	312
75	5	10	-.25	.71	1.75	2.1	1.69	.9	-.02	.52	30.0	70.1	324
93	5	10	-.25	.71	1.00	.1	.93	.2	.46	.52	70.0	70.1	406
155	5	10	-.25	.71	1.44	1.4	1.43	.7	.25	.52	50.0	70.1	916
5	4	10	-.76	.72	.66	-1.3	.56	-.1	.68	.46	80.0	69.6	105
7	4	10	-.76	.72	.66	-1.3	.56	-.1	.68	.46	80.0	69.6	107
13	4	10	-.76	.72	1.27	1.0	1.09	.5	.25	.46	40.0	69.6	113
14	4	10	-.76	.72	1.35	1.2	1.18	.5	.20	.46	40.0	69.6	114
15	4	10	-.76	.72	1.27	1.0	1.09	.5	.25	.46	40.0	69.6	115
16	4	10	-.76	.72	1.27	1.0	1.09	.5	.25	.46	40.0	69.6	116
19	4	10	-.76	.72	.57	-1.7	.49	-.2	.74	.46	100.0	69.6	119
20	4	10	-.76	.72	1.01	.1	.87	.3	.42	.46	60.0	69.6	120
23	4	10	-.76	.72	1.01	.1	.87	.3	.42	.46	60.0	69.6	123
25	4	10	-.76	.72	1.01	.1	.87	.3	.42	.46	60.0	69.6	125
29	4	10	-.76	.72	1.21	.8	1.05	.4	.29	.46	60.0	69.6	202
33	4	10	-.76	.72	1.35	1.2	1.18	.5	.20	.46	40.0	69.6	206
35	4	10	-.76	.72	.96	-.1	1.13	.5	.46	.46	80.0	69.6	208
37	4	10	-.76	.72	.92	-.2	.81	.2	.53	.46	80.0	69.6	210
38	4	10	-.76	.72	1.18	.7	1.03	.4	.38	.46	60.0	69.6	211
40	4	10	-.76	.72	1.16	.6	1.03	.4	.32	.46	60.0	69.6	213
42	4	10	-.76	.72	1.01	.1	.89	.3	.42	.46	80.0	69.6	215
52	4	10	-.76	.72	.93	-.2	.80	.2	.48	.46	80.0	69.6	301
62	4	10	-.76	.72	1.50	1.7	1.63	.9	.13	.46	60.0	69.6	311
67	4	10	-.76	.72	1.71	2.2	1.82	1.0	-.09	.46	40.0	69.6	316
68	4	10	-.76	.72	.83	-.5	.71	.1	.58	.46	80.0	69.6	317
74	4	10	-.76	.72	.93	-.2	.80	.2	.48	.46	80.0	69.6	323
77	4	10	-.76	.72	1.79	2.4	1.91	1.0	-.15	.46	40.0	69.6	326
80	4	10	-.76	.72	.86	-.4	.74	.1	.57	.46	80.0	69.6	329
83	4	10	-.76	.72	.92	-.2	.81	.2	.53	.46	80.0	69.6	332
100	4	10	-.76	.72	1.07	.3	.94	.3	.44	.46	60.0	69.6	703
102	4	10	-.76	.72	1.27	1.0	1.10	.5	.32	.46	40.0	69.6	705
109	4	10	-.76	.72	.66	-1.3	.56	-.1	.68	.46	80.0	69.6	712
118	4	10	-.76	.72	1.15	.6	1.31	.6	.34	.46	80.0	69.6	804
125	4	10	-.76	.72	.57	-1.7	.49	-.2	.74	.46	100.0	69.6	811
130	4	10	-.76	.72	1.35	1.2	1.19	.5	.27	.46	40.0	69.6	816
137	4	10	-.76	.72	.83	-.5	.71	.1	.58	.46	80.0	69.6	823
144	4	10	-.76	.72	1.29	1.1	1.13	.5	.31	.46	40.0	69.6	905
150	4	10	-.76	.72	.98	.0	.86	.3	.49	.46	80.0	69.6	911
152	4	10	-.76	.72	1.41	1.4	1.53	.8	.19	.46	60.0	69.6	913
158	4	10	-.76	.72	.66	-1.3	.56	-.1	.68	.46	80.0	69.6	919
4	3	10	-1.29	.75	.71	-.9	.56	.1	.60	.40	90.0	72.2	104
6	3	10	-1.29	.75	.59	-1.5	.46	.0	.67	.40	90.0	72.2	106

11	3	10	-1.29	.75	.81	-.5	.63	.2	.54	.40	70.0	72.2	111
26	3	10	-1.29	.75	.80	-.6	.65	.2	.54	.40	90.0	72.2	126
32	3	10	-1.29	.75	1.30	1.0	1.04	.5	.25	.40	50.0	72.2	205
34	3	10	-1.29	.75	.71	-.9	.56	.1	.60	.40	90.0	72.2	207
44	3	10	-1.29	.75	.87	-.3	.69	.2	.50	.40	70.0	72.2	217
45	3	10	-1.29	.75	.81	-.5	.63	.2	.54	.40	70.0	72.2	218
46	3	10	-1.29	.75	.87	-.3	.69	.2	.50	.40	70.0	72.2	219
47	3	10	-1.29	.75	.59	-1.5	.46	.0	.67	.40	90.0	72.2	220
54	3	10	-1.29	.75	1.04	.2	1.56	.8	.30	.40	90.0	72.2	303
56	3	10	-1.29	.75	1.44	1.4	1.41	.7	.03	.40	50.0	72.2	305
59	3	10	-1.29	.75	.87	-.3	.69	.2	.50	.40	70.0	72.2	308
61	3	10	-1.29	.75	.93	-.1	.88	.4	.44	.40	90.0	72.2	310
65	3	10	-1.29	.75	1.00	.1	.92	.4	.40	.40	70.0	72.2	314
70	3	10	-1.29	.75	1.08	.4	.88	.4	.38	.40	70.0	72.2	319
72	3	10	-1.29	.75	1.18	.7	.95	.4	.32	.40	50.0	72.2	321
78	3	10	-1.29	.75	1.00	.1	.92	.4	.40	.40	70.0	72.2	327
79	3	10	-1.29	.75	1.21	.8	1.11	.6	.21	.40	70.0	72.2	328
84	3	10	-1.29	.75	.71	-.9	.56	.1	.60	.40	90.0	72.2	333
85	3	10	-1.29	.75	.71	-.9	.56	.1	.60	.40	90.0	72.2	334
86	3	10	-1.29	.75	.71	-.9	.56	.1	.60	.40	90.0	72.2	335
87	3	10	-1.29	.75	1.04	.2	.95	.4	.33	.40	70.0	72.2	336
91	3	10	-1.29	.75	1.30	1.0	1.18	.6	.22	.40	50.0	72.2	404
101	3	10	-1.29	.75	1.12	.5	1.02	.5	.33	.40	70.0	72.2	704
106	3	10	-1.29	.75	.59	-1.5	.46	.0	.67	.40	90.0	72.2	709
110	3	10	-1.29	.75	1.45	1.4	2.02	1.0	.04	.40	70.0	72.2	713
112	3	10	-1.29	.75	.99	.1	.79	.3	.43	.40	70.0	72.2	715
119	3	10	-1.29	.75	.87	-.3	.69	.2	.50	.40	70.0	72.2	805
120	3	10	-1.29	.75	1.26	.9	1.73	.9	.17	.40	70.0	72.2	806
121	3	10	-1.29	.75	1.30	1.0	1.18	.6	.22	.40	50.0	72.2	807
122	3	10	-1.29	.75	.59	-1.5	.46	.0	.67	.40	90.0	72.2	808
124	3	10	-1.29	.75	.81	-.5	.63	.2	.54	.40	70.0	72.2	810
126	3	10	-1.29	.75	1.11	.4	1.61	.8	.26	.40	70.0	72.2	812
128	3	10	-1.29	.75	.90	-.2	.72	.3	.49	.40	70.0	72.2	814
135	3	10	-1.29	.75	.90	-.2	.72	.3	.49	.40	70.0	72.2	821
136	3	10	-1.29	.75	.90	-.2	.72	.3	.49	.40	70.0	72.2	822
139	3	10	-1.29	.75	1.03	.2	.95	.4	.39	.40	70.0	72.2	825
142	3	10	-1.29	.75	.78	-.7	.61	.1	.56	.40	70.0	72.2	903
143	3	10	-1.29	.75	.87	-.3	.69	.2	.50	.40	70.0	72.2	904
153	3	10	-1.29	.75	.99	.1	.79	.3	.43	.40	70.0	72.2	914
156	3	10	-1.29	.75	.87	-.3	.69	.2	.50	.40	70.0	72.2	917
159	3	10	-1.29	.75	1.52	1.6	1.36	.7	.10	.40	50.0	72.2	920
160	3	10	-1.29	.75	1.02	.2	.81	.3	.42	.40	70.0	72.2	921
8	2	10	-1.92	.84	.69	-.7	.45	.0	.54	.33	80.0	80.0	108
21	2	10	-1.92	.84	1.02	.2	.81	.3	.34	.33	80.0	80.0	121
22	2	10	-1.92	.84	1.02	.2	.81	.3	.34	.33	80.0	80.0	122
27	2	10	-1.92	.84	.69	-.7	.45	.0	.54	.33	80.0	80.0	127
28	2	10	-1.92	.84	1.02	.2	.81	.3	.34	.33	80.0	80.0	201
31	2	10	-1.92	.84	.69	-.7	.45	.0	.54	.33	80.0	80.0	204
51	2	10	-1.92	.84	.69	-.7	.45	.0	.54	.33	80.0	80.0	224
55	2	10	-1.92	.84	1.09	.4	.86	.4	.30	.33	80.0	80.0	304
57	2	10	-1.92	.84	1.30	.8	1.13	.6	.15	.33	80.0	80.0	306
58	2	10	-1.92	.84	1.01	.2	.72	.3	.36	.33	80.0	80.0	307
64	2	10	-1.92	.84	1.22	.6	.99	.5	.21	.33	80.0	80.0	313
66	2	10	-1.92	.84	1.22	.6	.99	.5	.21	.33	80.0	80.0	315
69	2	10	-1.92	.84	.77	-.4	.51	.0	.50	.33	80.0	80.0	318
71	2	10	-1.92	.84	.89	-.1	.64	.2	.42	.33	80.0	80.0	320
81	2	10	-1.92	.84	.93	.0	.67	.2	.40	.33	80.0	80.0	330
82	2	10	-1.92	.84	1.02	.2	.81	.3	.34	.33	80.0	80.0	331
98	2	10	-1.92	.84	.77	-.4	.51	.0	.50	.33	80.0	80.0	701
103	2	10	-1.92	.84	.69	-.7	.45	.0	.54	.33	80.0	80.0	706
104	2	10	-1.92	.84	1.02	.2	.81	.3	.34	.33	80.0	80.0	707
105	2	10	-1.92	.84	1.09	.4	.86	.4	.30	.33	80.0	80.0	708
108	2	10	-1.92	.84	1.09	.4	.86	.4	.30	.33	80.0	80.0	711
111	2	10	-1.92	.84	.93	.0	.67	.2	.40	.33	80.0	80.0	714
115	2	10	-1.92	.84	1.09	.4	.86	.4	.30	.33	80.0	80.0	801
117	2	10	-1.92	.84	1.41	1.0	2.66	1.3	-.06	.33	80.0	80.0	803
123	2	10	-1.92	.84	.81	-.3	.54	.1	.48	.33	80.0	80.0	809
129	2	10	-1.92	.84	1.12	.4	1.21	.6	.22	.33	80.0	80.0	815
134	2	10	-1.92	.84	.93	.0	.67	.2	.40	.33	80.0	80.0	820
140	2	10	-1.92	.84	1.41	1.0	1.54	.8	.04	.33	80.0	80.0	901
141	2	10	-1.92	.84	1.02	.2	.81	.3	.34	.33	80.0	80.0	902
146	2	10	-1.92	.84	.81	-.3	.54	.1	.48	.33	80.0	80.0	907

147	2	10	-1.92	.84	1.09	.4	.86	.4	.30	.33	80.0	80.0	908
148	2	10	-1.92	.84	1.20	.6	1.27	.6	.18	.33	80.0	80.0	909
151	2	10	-1.92	.84	.89	-.1	.64	.2	.42	.33	80.0	80.0	912
162	2	10	-1.92	.84	1.02	.2	.81	.3	.34	.33	80.0	80.0	923
163	2	10	-1.92	.84	.98	.1	.78	.3	.36	.33	80.0	80.0	924
164	2	10	-1.92	.84	1.01	.2	.72	.3	.36	.33	80.0	80.0	925
1	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	101
9	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	109
17	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	117
18	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	118
41	1	10	-2.80	1.09	1.09	.4	.87	.4	.19	.23	90.0	90.0	214
53	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	302
73	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	322
76	1	10	-2.80	1.09	.88	.1	.48	.0	.35	.23	90.0	90.0	325
99	1	10	-2.80	1.09	.88	.1	.48	.0	.35	.23	90.0	90.0	702
107	1	10	-2.80	1.09	.88	.1	.48	.0	.35	.23	90.0	90.0	710
113	1	10	-2.80	1.09	1.26	.6	2.15	1.1	-.05	.23	90.0	90.0	716
114	1	10	-2.80	1.09	1.17	.5	1.19	.6	.10	.23	90.0	90.0	717
116	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	802
127	1	10	-2.80	1.09	1.17	.5	1.19	.6	.10	.23	90.0	90.0	813
131	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	817
132	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	818
133	1	10	-2.80	1.09	.88	.1	.48	.0	.35	.23	90.0	90.0	819
138	1	10	-2.80	1.09	.83	.0	.43	.0	.38	.23	90.0	90.0	824
149	1	10	-2.80	1.09	1.09	.4	.87	.4	.19	.23	90.0	90.0	910
157	1	10	-2.80	1.09	1.17	.5	1.19	.6	.10	.23	90.0	90.0	918
161	1	10	-2.80	1.09	.88	.1	.48	.0	.35	.23	90.0	90.0	922
145	0	10	-4.13	1.87	MINIMUM MEASURE				.00	.00	100.0	100.0	906
MEAN	3.3	10.0	-1.23	.81	.99	.1	.89	.3			74.8	76.1	
S.D.	1.8	.0	1.12	.15	.26	.8	.46	.4			14.7	7.1	

Tabla 14. Resultados de los estudiantes.

Seguidamente se explicará y analizará cada una de las columnas de la tabla anterior que brindan información para el análisis.

4.2.7.1. *ENTRY NUMBER***.** Es simplemente el número que ocupa la persona en el archivo de Excel que se importó, Cuando se encuentran irregularidades en las respuestas de los estudiantes, lo más sencillo es ir al archivo de Excel y ver exactamente cuál fue la anomalía, y dar una conclusión de lo ocurrido.

4.2.7.2. *RAW SCORE.* Es el puntaje obtenido por los estudiantes en el test y permite tener una idea de sus habilidades.

4.2.7.3. *COUNT.* Hace referencia al número de ítems que fueron respondidos por los estudiantes; si algún ítem hubiera sido respondido por todos o si algún ítem no lo hubiera respondido nadie, el modelo lo excluye del estudio, pero en este caso no se excluyó ninguna pregunta. Como se ve en la Tabla 14, los dos primeros

estudiantes obtuvieron un score de 11 con 10 ítems, esto sucede porque existen dos ítems politómicos con máximo valor de 2 en cada uno de ellos y el mayor puntaje logrado por un estudiante sería 12 al tener 10 ítems el test, aunque en este estudio sólo obtuvieron un puntaje de 11.

4.2.7.4. *MEASURE.* Indica la medida de habilidad de los individuos expresada en lógitos. En la columna están organizados los estudiantes de mayor a menor habilidad con valores que van desde 3.63 hasta -2.80, con una persona excluida del análisis por no responder ninguna pregunta del test, pues no ofrece información para estimar su habilidad. Se encuentran 14 personas por encima del nivel medio de dificultad del test y 150 personas con habilidad de signo negativo, esto justifica la media de habilidad del grupo estimada por el modelo en -1.21, ya que apenas un 8.5% del grupo cuenta con una probabilidad mayor a 0.5 de responder más de la mitad del test correctamente, y el 91.5% tiene una probabilidad menor a 0.5 de responder correctamente más de la mitad del examen.

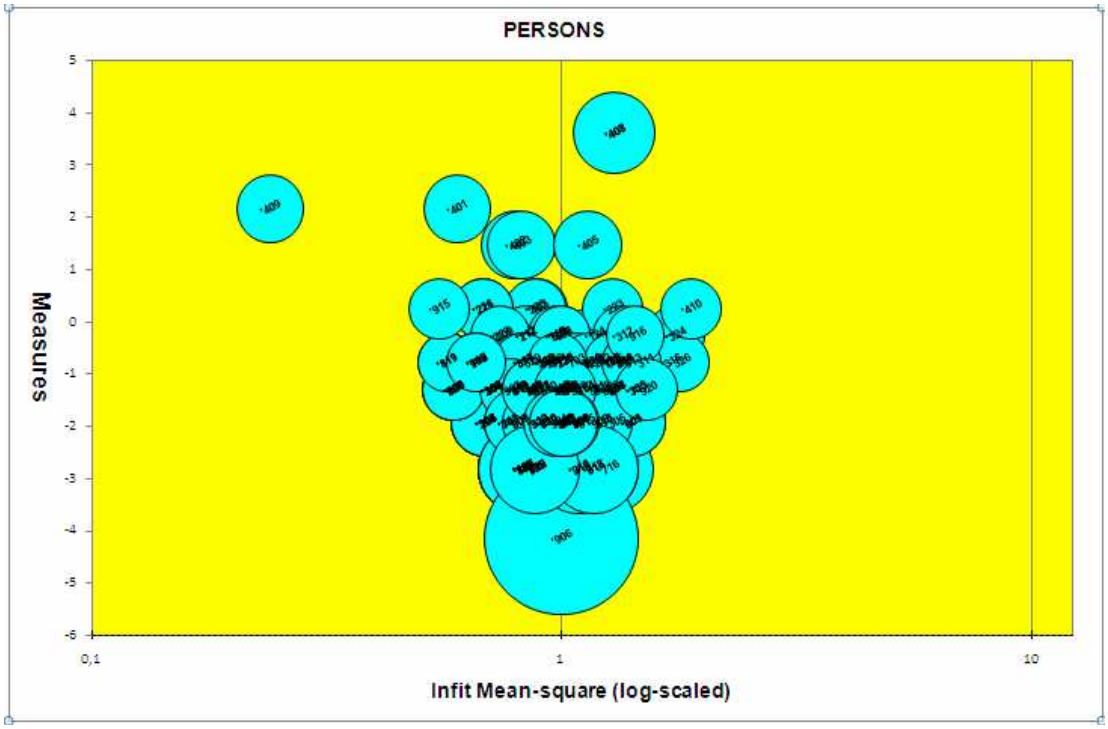
4.2.7.5. *MODEL S.E.* Indica el error de estimación de la habilidad de los estudiantes, siendo mayor en los extremos como se observa en la Tabla 2. A mayor o menor habilidad del estudiante, mayor es el error de estimación, y se reduce en la parte central de la distribución de los individuos, con errores de estimación de 1.0 en los dos estudiantes con mayor habilidad y de 1.09 para el que tiene menor habilidad, los demás tienen errores menores a estos valores.

Nótese que el estudiante que no respondió pregunta alguna en el test, tiene un error de 1.87 debido a la falta de información, pues es casi el doble de los errores más altos de estimación de habilidad que pertenecen al análisis, por esto WINSTEPS excluye a este estudiante del examen, pues su falta de información afecta las demás estimaciones, pero de igual modo le asigna a esta persona una medida de habilidad, pues no es lógico pensar en menos infinito para ésta,

encontrando el modelo un valor para este individuo (-4.13), así no tenga respuestas acertadas en el examen.

4.2.7.6. *INFIT MNSQ***.** Aunque el intervalo de 0.8 a 1.3 en este estadístico es de gran importancia, los datos pueden variar entre 0.5 y 1.5 debido a que el test fue aplicado a solo 164 personas, así que, valores dentro de este rango proporcionan un contexto válido y no degrada el proceso de la medición. Con valores entre 0.24 y 1.89 se encuentran sólo cinco estudiantes que rebasan el rango óptimo para una buena medición.

La gran mayoría del grupo se concentra entre 0.8 y 1.3 haciendo más cómoda la medición. Las personas 409, 410, 324, 316 y 326 con infit 0.24, 0.89, 1.75, 1.71 y 1.79 respectivamente, no aportan al proceso de medida, pero tampoco lo destruye por la poca cantidad de personas que presentan esta característica, el estudiante 409 tiene un INFIT MNSQ de 0.24, esto representa determinismo en los datos, ajusta demasiado bien su conjunto de respuestas con el patrón observado en el test, los demás individuos presentan aleatoriedad en los datos, sus respuestas son irregulares para sus niveles de habilidad.



Gráfica 16. INFIT MNSQ vs. Habilidad de las personas.

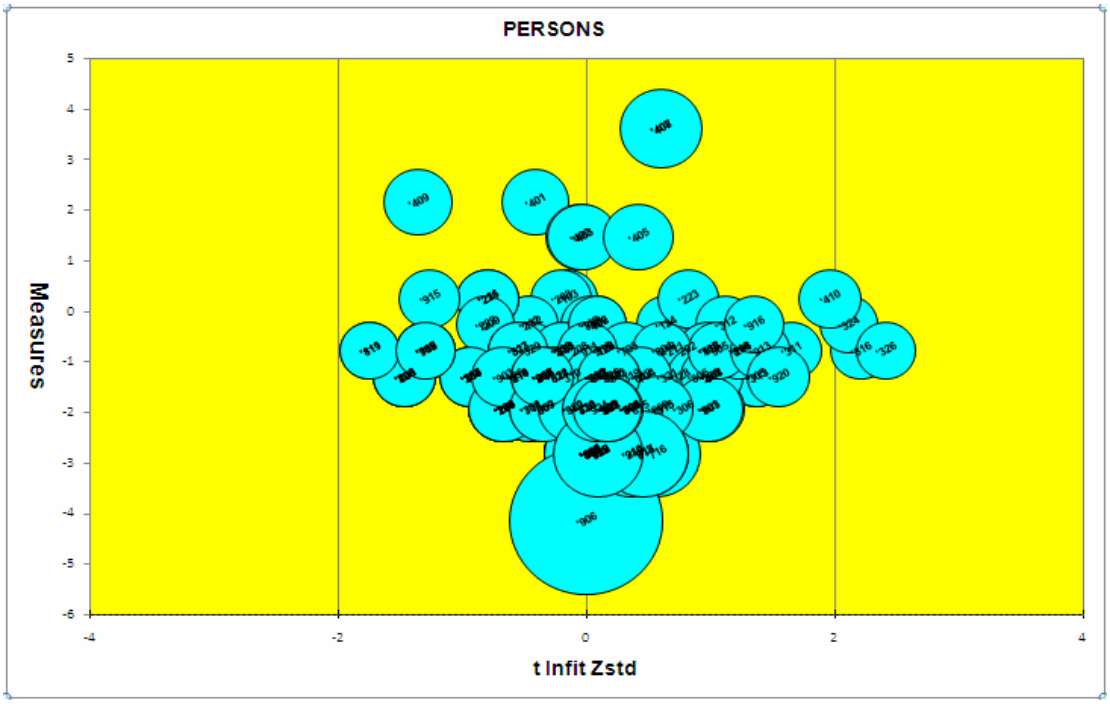
La Gráfica 16 que ofrece WINSTEPS permite relacionar este criterio con la habilidad de los estudiantes y su error de estimación, ya que los datos crudos no facilitan la identificación de las personas que presentan anomalías debido a que la cantidad de datos representados no hacen sencillo dicho proceso.

Se encuentra el INFIT MNSQ en el eje horizontal y la habilidad de los estudiantes estimada en lógitos en el eje vertical. El seguimiento visual resalta rápidamente al estudiante 409 a la izquierda de la figura, así como al 410, 326, 316 y 324, datos que exigen estudio y revisión.

Se destaca también el alumno 906 en la parte inferior, por el tamaño del error de estimación de habilidad, pues no respondió ninguna pregunta, afectando la precisión en su estimación como la del grupo, por esto, WINSTEPS lo estima pero

no lo tiene en cuenta en el proceso de medición. Se observa también que los estudiantes que se hallan en los extremos de habilidad tienen errores más grandes que los que están en el centro, así como los datos que demuestran demasiado ajuste se encuentran en la parte izquierda y en la parte derecha los que presentan ruido o aleatoriedad en los mismos.

4.2.7.7. *INFIT ZSTD***.** Es el mismo estadístico de medida cuadrática ponderada anterior, pero transformado de tal forma que el estadístico resultante posee distribución normal. El rango ideal de valores que debe tomar está entre -2 y 2 para un buen ajuste al modelo.



Gráfica 17. INFIT ZSTD vs. Habilidad de las personas.

Sólo las cuatro personas que presentaron aleatoriedad en sus respuestas para el INFIT MNSQ, obtuvieron valores fuera del rango para el INFIT ZSTD entre 2.0 y

2.4, como estos valores no están lejos del umbral sugerido por el modelo, no representan riesgo en el proceso de medición y se mantienen en el grupo.

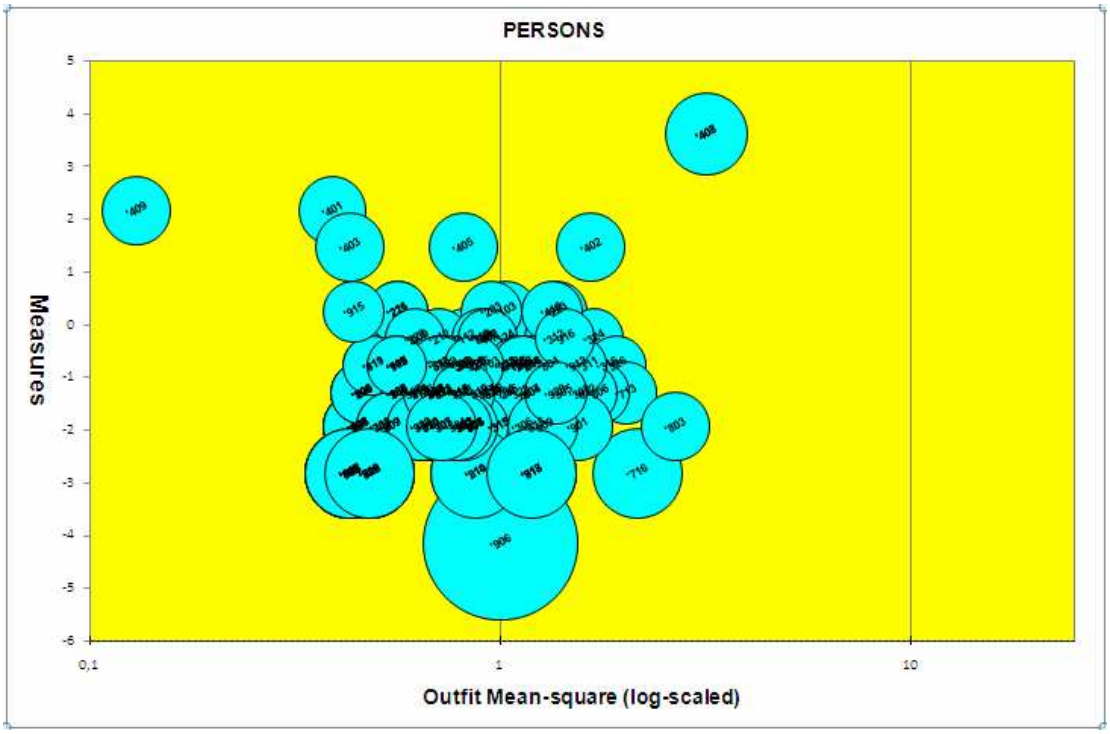
Al observar la Gráfica 17, los individuos 410, 324, 326 y 316 fácilmente se pueden identificar, mientras en la Tabla 14 se debe hacer un mayor trabajo de observación y búsqueda. Las características de los errores de estimación se mantienen, mostrando las circunferencias más grandes en los extremos de habilidad de los evaluados, mientras los errores más pequeños se ven en las personas que están a la altura de la habilidad media del grupo.

Se puede decir que todos los datos se ajustan al modelo, ofreciendo ventajas para el estudio de calibración de ítems y evaluación de personas, a excepción de la persona 906 que aparece en la parte inferior de la figura con un error de estimación de 1.87 por no responder pregunta alguna del test.

4.2.7.8. *OUTFIT MNSQ***.** Es el estadístico de ajuste externo, del que se habló anteriormente, sensible a irregularidades lejos del nivel de habilidad e indica si la persona se ajusta al modelo. El ideal es que los valores estén entre 0.8 y 1.3, intervalo considerado razonable para la cantidad de personas involucradas.

Entre 0.13 y 3.16 se encuentran los valores del OUTFIT MNSQ del grupo, con dos personas destacadas por la alta aleatoriedad en sus respuestas, el 407 y el 408 y con un valor de 3.16, son los dos estudiantes con mayor destrezas del grupo, que con una alta habilidad, dejaron de responder un ítem con una dificultad lejana a su nivel de destreza. El estudiante 326 con una baja habilidad respondió un ítem difícil, lejos de su nivel de destreza. Estos fenómenos permiten que los estadísticos arrojen valores que se salen de lo requerido por el modelo para un buen análisis. El alumno 409 respondió al test de forma demasiado predecible en su patrón de respuestas, por eso el ZSTD de 0.13. Como los dos estudiantes que presentaron irregularidad en este dato tienen valores aceptables en los demás

criterios de ajuste, se permiten en el proceso. Los demás participantes tienen valores menores a 2 con el resto de criterios aceptables.



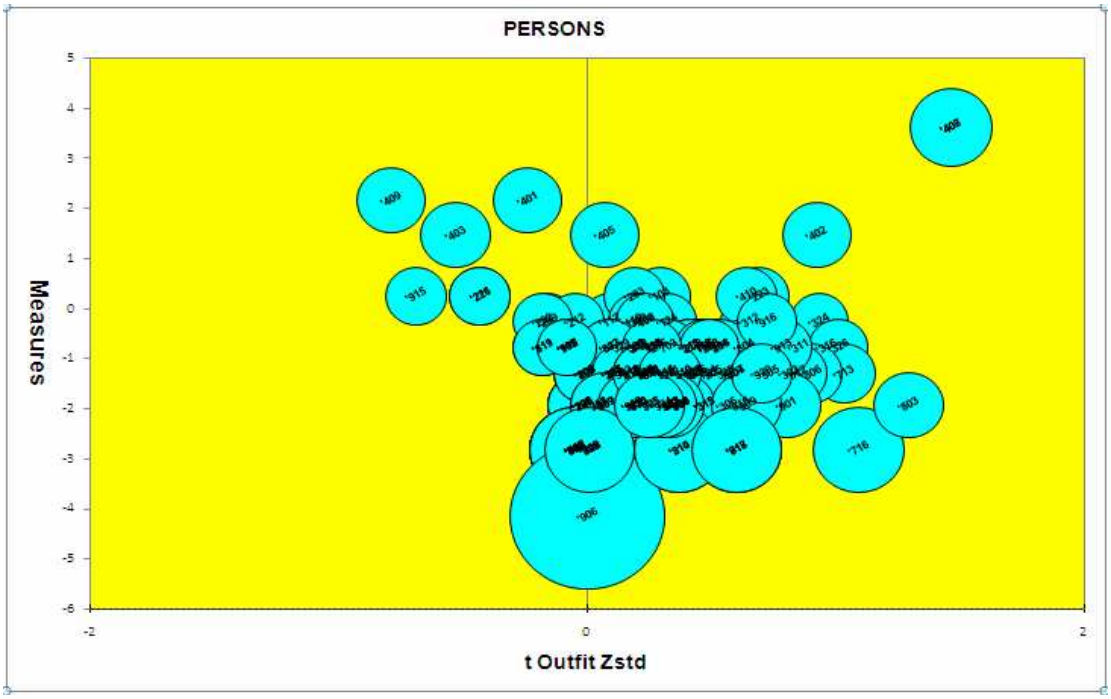
Gráfica 18. OUTFIT MNSQ vs. Habilidad de las personas.

Aunque es preciso hablar de valores, se identificará en la Gráfica 18 lo descrito en el párrafo anterior. En la parte superior derecha se encuentran los dos estudiantes de alto nivel en el test como era de esperarse, pues la anomalía en la respuesta lejos de sus niveles de habilidad, ubica a estas personas en puntos mayores a 1.5, y las sitúan en la parte superior por su nivel de habilidad con un error de estimación mayor a los estudiantes que tienen habilidad media. Ahora, el alumno 409 obtiene valores menores que 0.5 por responder el test de manera como el modelo considera un patrón de respuestas predecible, es decir, que ajusta demasiado bien para su gusto; este estudiante no se elimina a menos que se

quiera reducir el grupo de personas en el análisis, y como no se pretende hacerlo, se mantiene en el estudio.

4.2.7.9. *OUTFIT ZSTD***.** Es el mismo estadígrafo de ajuste externo interpretado como el promedio de los residuales estandarizados y transformado para que posea una distribución normal. El rango ideal para un buen ajuste que debe tomar este criterio está entre -2 y 2.

El grupo de estudio presenta valores entre -0.8 y 1.5 validando así, a todos los estudiantes para el proceso de medición.



Gráfica 19.OUTFIT ZSTD Vs. Habilidad de las personas.

En la Gráfica 19 se observa como todos los datos están dentro del rango ideal para el buen ajuste al modelo, con las mismas características del error estándar que en los anteriores análisis.

4.2.7.10. *PERSON.* Esta referencia que se presenta en el extremo derecho de la Tabla 2, corresponde al nombre con el que se identifican los estudiantes en el registro que se importa de Excel y que aparece en el archivo de control.

En la parte inferior de la tabla se encuentran dos datos generales de importancia en el análisis: la media de todos los temas antes analizados (MEAN) y su respectiva desviación estándar (S.D).

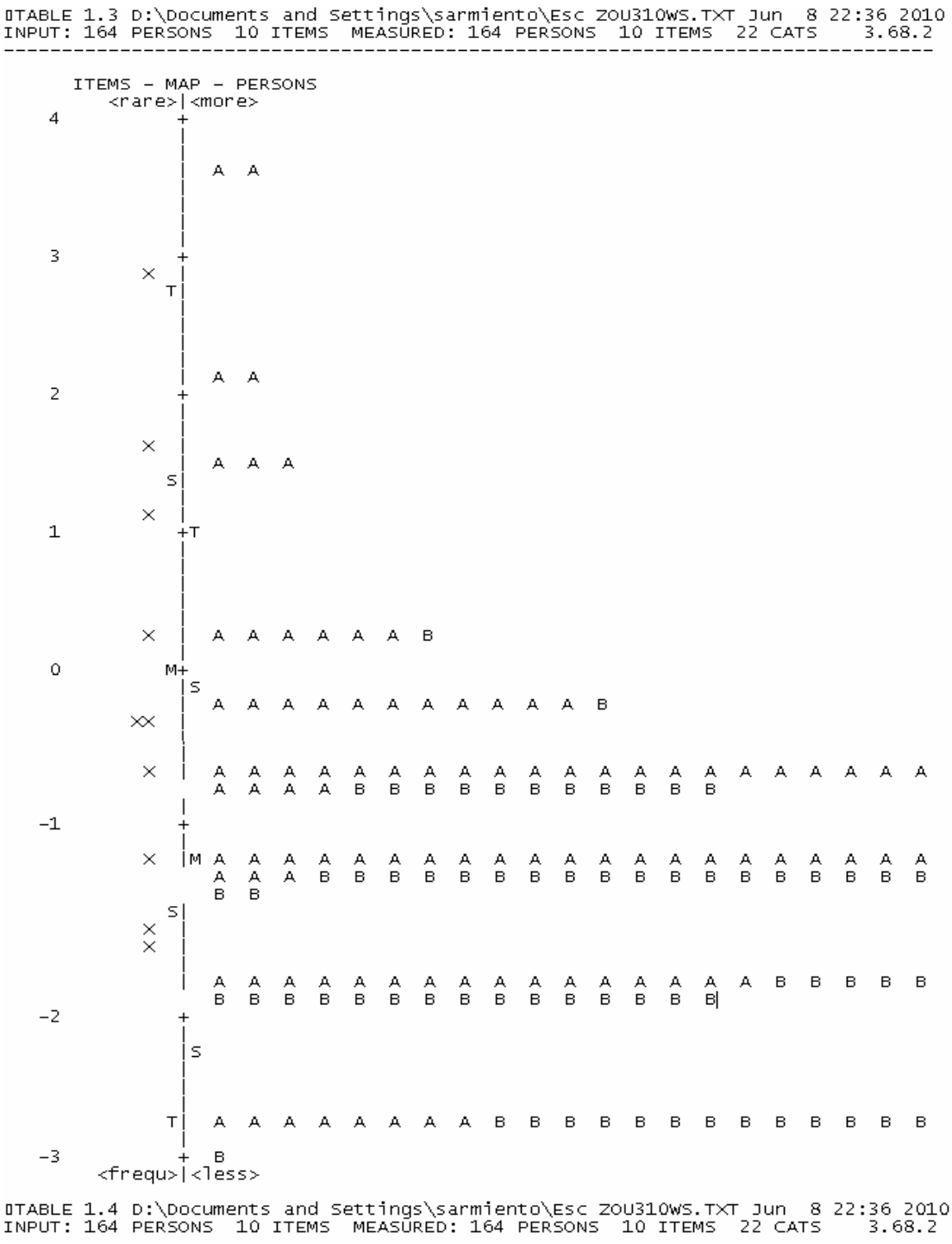
4.2.7.11. *MEAN.* Esta media es útil porque ofrece información del grupo de estudiantes en general y permite inferir sobre algunos aspectos, como por ejemplo, en la media del puntaje que obtuvieron los alumnos que fue de 3.3 puntos de 12 posibles. Bastante baja la destreza general del grupo que se refleja en la media de las habilidades que fue -1.23; al excluir al individuo 906, que no respondió pregunta alguna en el test, la media aumenta a -1.21, valor presentado en el primer pantallazo de la corrida de los datos con WINSTEPS (Ver Tabla 14). La media de los errores fue de 0.81. . Estos datos confirman que el grupo tiene una baja habilidad para responder este test.

Con un INFIT MNSQ de 0.99 y un OUTFIT MNSQ de 0.89, las medias de estos estadísticos establecen que el grupo de estudiantes se ajusta bien al modelo, complementado con un INFIT ZSTD de 0.1 y un OUTFIT ZSTD de 0.3 bastante buenos por estar tan cerca del 0.

4.2.8. ANÁLISIS DE LOS RESULTADOS PARA CADA UNIVERSIDAD EVALUADA

Como el test fue aplicado en dos universidades del país, se verá a continuación cómo fueron los resultados de los estudiantes en cada una de estas instituciones. Para tal fin, se codificaron los resultados de los estudiantes de la Universidad Industrial de Santander (UIS) con la letra (A) y los de la Universidad Francisco de

Paula Santander (UFPS) con la letra (B), para poder visualizar más claramente los datos que arroja la Gráfica 20 y sacar algunas conclusiones respecto de las habilidades de los estudiantes de cada una de estas universidades.



Gráfica 20 . Resultados para las dos universidades donde estudiaban los estudiantes evaluados.

La Gráfica 20 construida por WINSTEPS presenta los ítems en la parte izquierda cada uno con la letra x en el mismo orden de dificultad de la gráfica anterior, y en la parte derecha las personas representadas cada una con una letra.

Aunque la UIS tiene un grupo mayor de estudiantes evaluados que la UFPS en este test, las personas localizadas en la línea inferior de habilidad, son en mayor número de la UFPS con 14 personas y 8 de la UIS, todos están en el nivel más bajo de destreza en las respuestas del test, con una habilidad de -2.80. En la siguiente línea donde la medida de habilidad está en -1.92, hay 21 estudiantes de la UFPS y 15 de la UIS, manteniendo aún el mayor número de estudiantes la UFPS en este nivel. En la línea de habilidad con un valor de -1.29 se encuentran 24 estudiantes de la UIS y 20 de la UFPS, en este punto se invierten los resultados y la cantidad de estudiantes de la UIS es mayor, alimentando la conjetura de que a medida que aumenta la habilidad se incrementa la cantidad de estudiantes de la UIS con respecto a los de la UFPS. La línea que sigue a continuación señala la habilidad con un valor de -0.76, cuenta con 25 estudiantes de la UIS y 11 de la UFPS. Le sigue la línea de habilidad con valor de -0.25 y concentra 11 estudiantes de la UIS y solo 1 de la UFPS. La línea de habilidad con un valor de 0.27 agrupa 6 estudiantes de la UIS y sólo 1 de la UFPS, siendo los últimos estudiantes de la UFPS en aparecer en la escala de habilidad, pues los 7 estudiantes que se presentan con habilidades entre 1.49 y 3.63 son de la UIS.

El conteo anterior es solo una formalidad, pues claramente al observar la gráfica se puede identificar qué grupo respondió mejor. En la parte baja prevaleció la cantidad de estudiantes de la UFPS y a medida que se observaban los resultados en niveles de habilidad más altos, la proporción de estudiantes de la UIS fue creciendo con respecto a los de la UFPS. Esto obliga a pensar que en la distribución de los estudiantes de la UIS se tiene una habilidad media mayor que la de los estudiantes de la UFPS, pues la mayoría de este último grupo se congrega en la parte inferior de la gráfica que señala la habilidad más baja.

5. CONCLUSIONES

En este capítulo se presentan las principales conclusiones obtenidas en el desarrollo de este estudio:

En general, el análisis clásico aplicado a los resultados del cuestionario, señaló un margen en el índice de dificultad entre 0,06 y 0,56 con una media de 0.35 determinando una prueba difícil para los estudiantes, como se observó igualmente en la tesis doctoral de Olivo. La habilidad de los estudiantes para contestar el test, el análisis de Rasch la ubicó en -1.21 lógits y siendo el valor de la dificultad de los ítems 0.00 el que señala el modelo, reafirma la complejidad de la prueba y la mayor exigencia de habilidad en los participantes.

La mayor dificultad de la prueba se observó en los ítems abiertos 4 y 7 y en el ítem 10 que requiere análisis de gráficos, concluyendo, que los estudiantes prefieren contestar ítems de selección múltiple, que preguntas abiertas donde necesitan realizar cálculos. Los ítems que presentan menor grado de dificultad son los dicotómicos de tipo conceptual como los ítems 1, 2, 3 y 5 que a diferencia de los restantes no exigen un mayor discernimiento.

Otro aspecto destacable, es la falta de respuesta a los ítems, ya que sólo tres, fueron contestados por todos los estudiantes, siendo éstos los ítems 1, 2 y 3; cuyos Id los identifican como menos difíciles. Los ítems donde se observó falta de respuestas se relacionan con el grado de dificultad, por ejemplo, el ítem 6 con un grado de dificultad de 0,35 cuenta con el 6,1% (10 estudiantes) de respuestas en blanco, y el ítem 10, con un Id de 0.16 tiene un 24,4% (40 estudiantes) de respuestas en blanco.

En el índice de discriminación, los resultados obtenidos están comprendidos en el intervalo de 0,28 a 0,66 señalando validez con lo que se quiere evaluar. Es pertinente recordar que los ítems se seleccionaron de un cuestionario creado y analizado para evaluar el tema de los intervalos de confianza, pero al igual que Olivo en su tesis, las pruebas fueron aplicadas a muestras pequeñas.

Con los resultados obtenidos en el ítem 4, el modelo Rasch lo sitúa por encima de la media de dificultad de los ítems con 1.64 lógits, y de la media de habilidad al que fueron enfrentados los estudiantes. Sólo cuatro personas tienen una probabilidad mayor a 0.5 de responder esta pregunta bien y sus estadígrafos de ajuste aprueban el ítem con un Infit y Outfit de 0.81 y 0.61 respectivamente. En la teoría clásica el índice de dificultad de 0.24 establece que los estudiantes no tienen una idea clara sobre el efecto de la varianza en el ancho del intervalo de confianza, ya que sólo tres respondieron correctamente este ítem, 33 respondieron parcialmente bien la pregunta y 127 no respondieron, o respondieron mal. Su índice de discriminación es de 0.66, coincidiendo con la aprobación que le da el modelo.

Se deduce con los resultados del ítem 7 en el análisis de Rasch, que la dificultad que ofrece este ítem es la mayor en el test con un valor de 2.89 lógits, que se refleja en la gran complejidad que representó para los evaluados hallar intervalos de confianza cuando la desviación poblacional es desconocida. Sus estadígrafos de ajuste no aprueban el ítem, con un Infit y Outfit de 0.36 y 0.03 respectivamente, considerando un ajuste perfecto. Pero al consultar los valores del Infit ZSTD y Outfit ZSTD que están dentro del rango propuesto por el modelo y que corresponden a -1.4 y -1.6 respectivamente se decidió salvarlo para no disminuir el número de ítems. En la teoría clásica, con un índice de dificultad de 0.06, sólo 2 estudiantes respondieron correctamente este ítem, 2 respondieron parcialmente bien y 160 no respondieron, o respondieron mal. Su índice de discriminación es

de 0.57, considerando este ítem confiable en la evaluación del tema contrario a lo estipulado por el modelo.

Respecto a los ítems dicotómicos, con los resultados obtenidos en el ítem 1, el análisis de Rasch lo sitúa por debajo de la media de dificultad de los ítems con -1.45 lógits y de la habilidad media de los estudiantes; aproximadamente 104 alumnos tienen una probabilidad mayor a 0.5 de responder esta pregunta bien y sus estadígrafos de ajuste aprueban el ítem con un Infit y Outfit de 1.07 y 1.03 respectivamente La teoría clásica con un índice de dificultad de 0.54, permite considerarlo de los menos difíciles para los estudiantes aunque aproximadamente, sólo la mitad de los evaluados lo respondieron bien, osea 88 estudiantes, quienes manejan la definición de intervalo de confianza. Esto demuestra que la habilidad del grupo está por debajo de la exigencia del test, pues cuando la mitad de las personas responden correctamente una pregunta, lo que se espera es que la dificultad media de la misma este próxima a cero para que haya un equilibrio entre la habilidad del grupo y la dificultad del test. Su índice de discriminación es de 0.33, relativamente bajo a diferencia de la aprobación que le da el modelo.

El análisis de Rasch con los resultados obtenidos en el ítem 2, lo sitúa por debajo de la media de la dificultad de los ítems en -1.22 lógits, prácticamente igual a la habilidad media de los evaluados; aproximadamente 104 estudiantes tienen una probabilidad mayor o igual a 0.5 de responder esta pregunta bien y sus estadígrafos de ajuste aprueban el ítem con un Infit y Outfit de 0.93 y 0.92 respectivamente. La teoría clásica con un índice de dificultad de 0.49 considera este ítem uno de los menos difíciles para los estudiantes evaluados, el 48.8% (80 estudiantes) lo contestaron bien, comprendiendo que el ancho de los intervalos de confianza disminuye cuando aumenta el tamaño de la muestra. Su índice de discriminación es de 0.44, relativamente bajo a diferencia de la aprobación que le da el modelo.

Al ítem 3, el modelo Rasch lo ubica por debajo de la media de la dificultad de los ítems con -1.57 lógits, dándole el mínimo grado de dificultad al que fueron enfrentados los estudiantes; se deduce que la mayoría de ellos entienden la relación que hay entre el nivel de confianza y la construcción del intervalo. Con 104 estudiantes con probabilidad mayor a 0.5 de responder esta pregunta correctamente y estadígrafos de ajuste infit de 1.11 y Outfit de 1.12, aprueban este ítem. En la teoría clásica el índice de dificultad de 0.56, lo considera el menos difícil para los evaluados, un 56.1% (92 estudiantes) lo contestaron correctamente. Su índice de discriminación es de 0.29, bajo con relación a la aprobación que le da el modelo Rasch.

Con los resultados que muestra el ítem 5, el análisis de Rasch lo sitúa por debajo de la media de la dificultad de los ítems con -0.77 lógits, y por encima de la media de habilidad al que fueron enfrentados los evaluados; 60 estudiantes tienen una probabilidad mayor o igual a 0.5 de responder esta pregunta bien y sus estadígrafos de ajuste aprueban el ítem con un Infit de 0.96 y un Outfit de 0.92.
Con respecto a la teoría clásica, el índice de dificultad de 0.40 lo considera difícil para las personas evaluadas, un 39.6% (65 estudiantes) lo contestaron correctamente, concluyendo que menos de la mitad de los participantes entiende el significado del nivel de confianza (variación del intervalo en diferentes muestras) y un 59.1% (97 estudiantes) cree que el intervalo se construye para hallar la media muestral y no la poblacional o confunden estadístico con parámetro. Su índice de discriminación es de 0.44, relativamente bajo a diferencia de la aprobación que le da el modelo.

El ítem 6 el análisis de Rasch lo ubica por debajo de la media de la dificultad de los ítems con -0.42 lógits, y por encima de la media de habilidad al que fueron enfrentados los evaluados; sólo 24 estudiantes tienen una probabilidad mayor a 0.5 de responder esta pregunta bien y sus estadígrafos de ajuste aprueban el ítem

con un Infit y Outfit de 1.02 y 0.95 respectivamente. La teoría clásica con un índice de dificultad de 0.35 lo considera un ítem difícil; un 33.0% (54 estudiantes) lo contestaron correctamente, concluyendo que aproximadamente la tercera parte de los estudiantes evaluados estiman el intervalo de una población normal o una muestra grande con desviación conocida, un 61.0% (100 estudiantes) olvidan dividir por el tamaño de la muestra u olvidan el hecho de que el intervalo se necesita con un nivel de confianza de 95% y que el tamaño de la muestra es 100, y un 6.1% (10 estudiantes) no respondieron o no saben. Su índice de discriminación es de 0.39, bajo a diferencia de la aprobación que le da el modelo Rasch.

Con los resultados que se obtienen en el ítem 8, el análisis de Rasch lo sitúa por debajo de la media de dificultad de los ítems con -0.42 lógits (valor igual al ítem 6) y por encima de la media de habilidad al que fueron enfrentados los participantes; 24 estudiantes tienen una probabilidad mayor a 0.5 de responder esta pregunta bien y sus estadígrafos de ajuste aprueban el ítem con un Infit de 1.02 y un Outfit de 1.04. La teoría clásica con un índice de dificultad de 0.39, lo considera un ítem difícil para los evaluados; sólo un 33.0% (54 estudiantes) lo contestaron correctamente y se deduce que aproximadamente la tercera parte de los participantes estiman la media de una población a partir de datos experimentales con desviación desconocida y muestra grande, un 51.8% (85 estudiantes) muestran error a la hora de buscar los valores críticos en la tabla de distribución normal estándar o no dividen por el tamaño de la muestra, y un 15.2% (25 estudiantes) no responden o no saben la pregunta. Su índice de discriminación es de 0.40, bajo en relación a la aprobación que le da el modelo Rasch.

Los ítems 6 y 8 se relacionan en cuanto a su dificultad, de la misma forma en los dos análisis realizados.

El ítem 9 el análisis de Rasch lo sitúa por encima de la media de la dificultad de los ítems con 0.21 lógits y de la media de habilidad al que fueron enfrentados los participantes; sólo 12 estudiantes tienen una probabilidad mayor o igual a 0.5 de responder esta pregunta bien y sus estadígrafos de ajuste aprueban el ítem con un Infit y Outfit de 1.17 y 1.42 respectivamente. La teoría clásica señala un índice de dificultad de 0.39, considerándolo uno de los más difíciles para los evaluados; un 22.5% (37 estudiantes) lo contestaron correctamente concluyendo que menos de la cuarta parte de los estudiantes evaluados determinan los valores críticos en la distribución del estadístico, un 50.6% (83 estudiantes) utilizan una distribución para muestras grandes o utilizan mal el valor de los grados de libertad y un 26.8% (44 estudiantes) no responden o no saben. Su índice de discriminación es 0.28, valor bajo al compararlo con la aprobación que le da el modelo.

Los resultados obtenidos en el ítem 10 permiten que el análisis de Rasch lo sitúe por encima de la media de la dificultad de los ítems con 1.13 lógits, y más por encima de la media de habilidad al que fueron enfrentados los evaluados representando una gran exigencia para este grupo; sólo 6 estudiantes tienen una probabilidad mayor o igual a 0.5 de responder esta pregunta correctamente y sus estadígrafos de ajuste aprueban el ítem con un Infit y Outfit de 1.02 y 0.86 respectivamente. La teoría clásica con un índice de dificultad de 0.16, lo considera el más difícil de los ítems dicotómicos para los evaluados, un 12.2% (20 estudiantes) lo contestaron correctamente y deduce que sólo un grupo muy pequeño de participantes interpretan gráficos de intervalos de confianza, un 63.4% (104 estudiantes) tienen una idea errónea, creen que al producirse solapes en los intervalos, las medias o las poblaciones son diferentes o piensan que no hay solape en los intervalos que representan el año 1980 y 1982 (interpretación errada de los gráficos) o tienen un mal entendimiento en la variabilidad de los intervalos y un 24.4%(40 estudiantes) no responden o no saben. Su índice de discriminación es de 0.43, bajo a diferencia de la aprobación que le da el modelo Rasch.

Al comparar las posiciones dadas a los ítems según el grado de dificultad por la teoría clásica y el modelo Rasch, se aprecia concordancia en la mayoría de ellas, a excepción de los ítem 4 y 10, donde Rasch sitúa como más difícil el ítem 4 que el 10 y la teoría clásica considera al ítem 10 más difícil que el 4. En resumen los estudiantes en las dos universidades a los que se les aplicó la prueba, tienen un bajo dominio en el tema de los intervalos de confianza.

Entre las deficiencias encontradas en el desarrollo de la investigación, se observó la falta de tiempo otorgada por algunos profesores para la aplicación del test, los estudiantes evaluados contestaron el cuestionario en tiempos menores a 30 minutos.

En esta investigación, se seleccionaron libremente grupos para que respondieran la prueba a voluntad propia y algunos estudiantes que vieron el tema en semestres anteriores sin manejo frecuente del mismo, no tenían frescos dichos conceptos. También se reconoce, que por la época en que se aplicó el test, un incentivo en la materia para que las respuestas fueran más consientes, posiblemente reduciría la abstención de muchos estudiantes en algunas preguntas. Adicional a lo anterior, en esta tesis, el gran número de ítems sin respuesta pudo interferir en los resultados haciendo aún más pequeña la muestra, ya que los ítems que no mostraban respuesta alguna no se contaron para el resultado final.

Se deja la inquietud para futuras investigaciones sobre el tema, realizar la prueba en un mayor número de universidades (una muestra más grande), con un tiempo adecuado para responderla, osea procurando aplicar el test en un tiempo prudentemente y cercano al estudio que amerita el tema, para poder así, conseguir una muestra más significativa. Respecto al número de ítems del test, es recomendable buscar o crear otros ítems relacionados con el tema a evaluar, que permitan la exclusión de aquellos, cuyos índices de discriminación o estadígrafos

de ajuste no sean confiables o estén por fuera de los parámetros recomendados por cada modelo para así poder afinar la medida.

BIBLIOGRAFIA

Abad, F. J., Garrido, J., Olea, J., & Ponsoda, V. *Introducción a la psicometría. Teoría clásica de los test y teoría de la respuesta al ítem.* Madrid.2006.

González, M. M. *El análisis de reactivos en el modelo Rasch.* México: Universidad de Sonora. 2008. 100 p.

Iraurgi, O., Lozano, F., González-Saiz, & Trujols, J. *Valoración psicométrica de la escala de severidad de la dependencia a partir de 2 modelos de análisis: la teoría clásica de los test y la teoría de respuesta al ítem.* Botetín de psicología No. 93 . 2008.

Olivo, E.. *Significado de los intervalos de confianza para los estudiantes de ingeniería en México.* [Tesis Doctoral] España. Universidad de Granada. Departamento de Didáctica de las matemáticas; 2008, 327 p.

Prieto Adánez, G., & Delgado, A. R. *Análisis de un test mediante el modelo de Rasch.* Psicothema. 2003. vol 15, nº1. 94-100 p.

Prieto Adánez, G., & Dias Velasco, A. *Uso del modelo de Rasch para poner en la misma escala las puntuaciones de distintos tests.* Actual Psicologia. 2003. Vol.19,nº 106. 5-23 p.

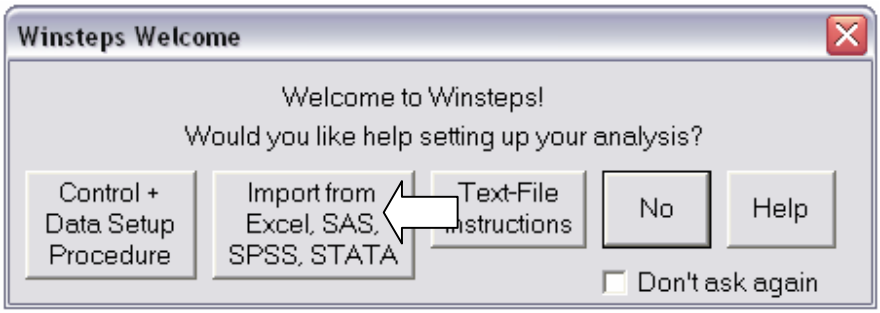
Smith, J. E., & Smith, R. M. *Introduction to Rasch measurement: theory, models, and applications.* Maple Grove MN: JAM Press. 2004. 979 p.

ANEXO A. PROCEDIMIENTO PARA EJECUTAR WINSTEPS®

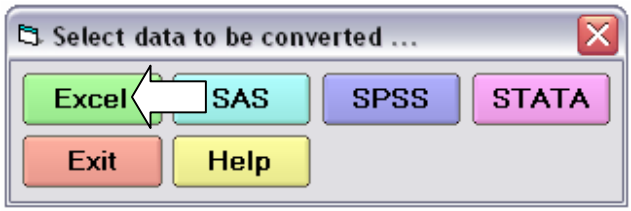
PROCEDIMIENTO PARA EJECUTAR WINSTEPS®

A continuación se detallará cómo importar la base de datos desde Excel para convertirlo a formato .txt y así el software lo pueda ejecutar.

Se abre el programa haciendo doble click sobre él, aparece la ventana de *Winsteps Welcome* que presenta un grupo de opciones como se muestra a continuación, entre ellas se puede ver la que dice *Import from Excel, SAS, SPSS, STATA*, que es la selección a elegir para poder importar la base de datos que se tiene en Excel.

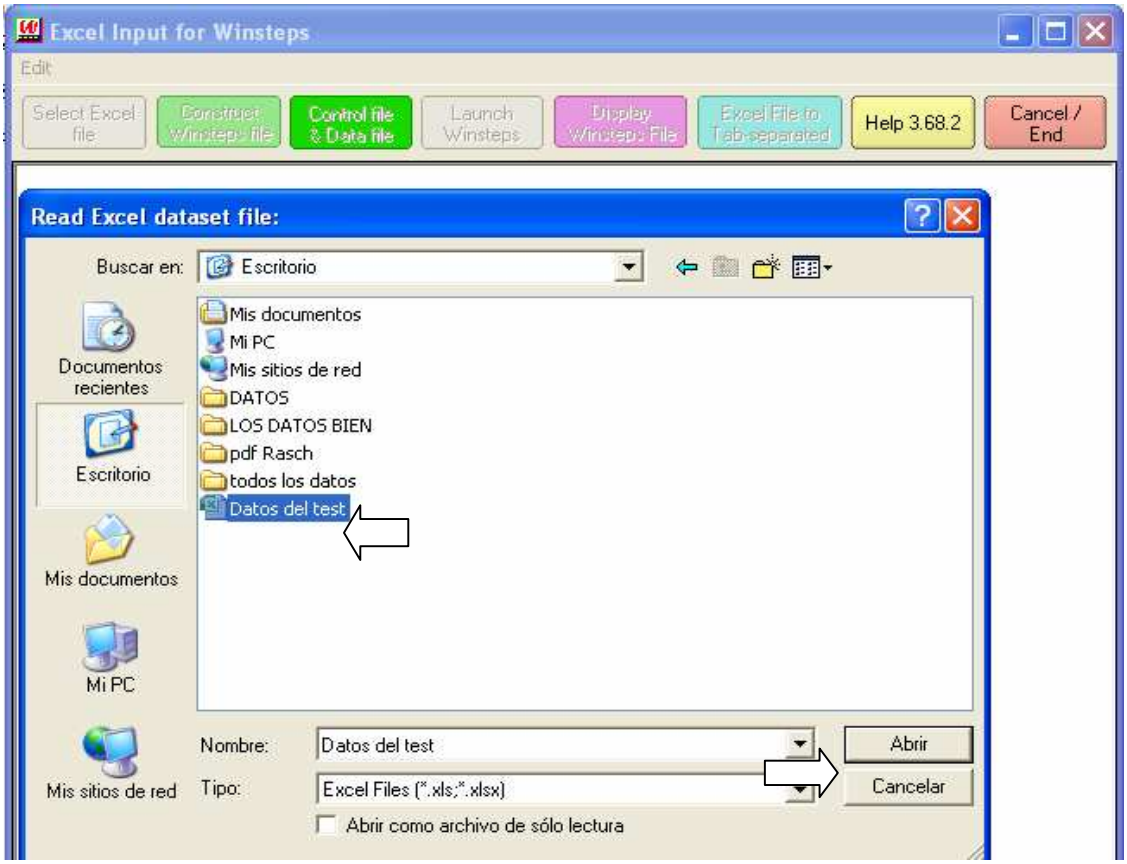


Se abre una nueva ventana *Select data to be converted…* que preguntará de dónde se desea importar la base de datos y ofrece diferentes alternativas, eligiendo en este caso la de Excel.

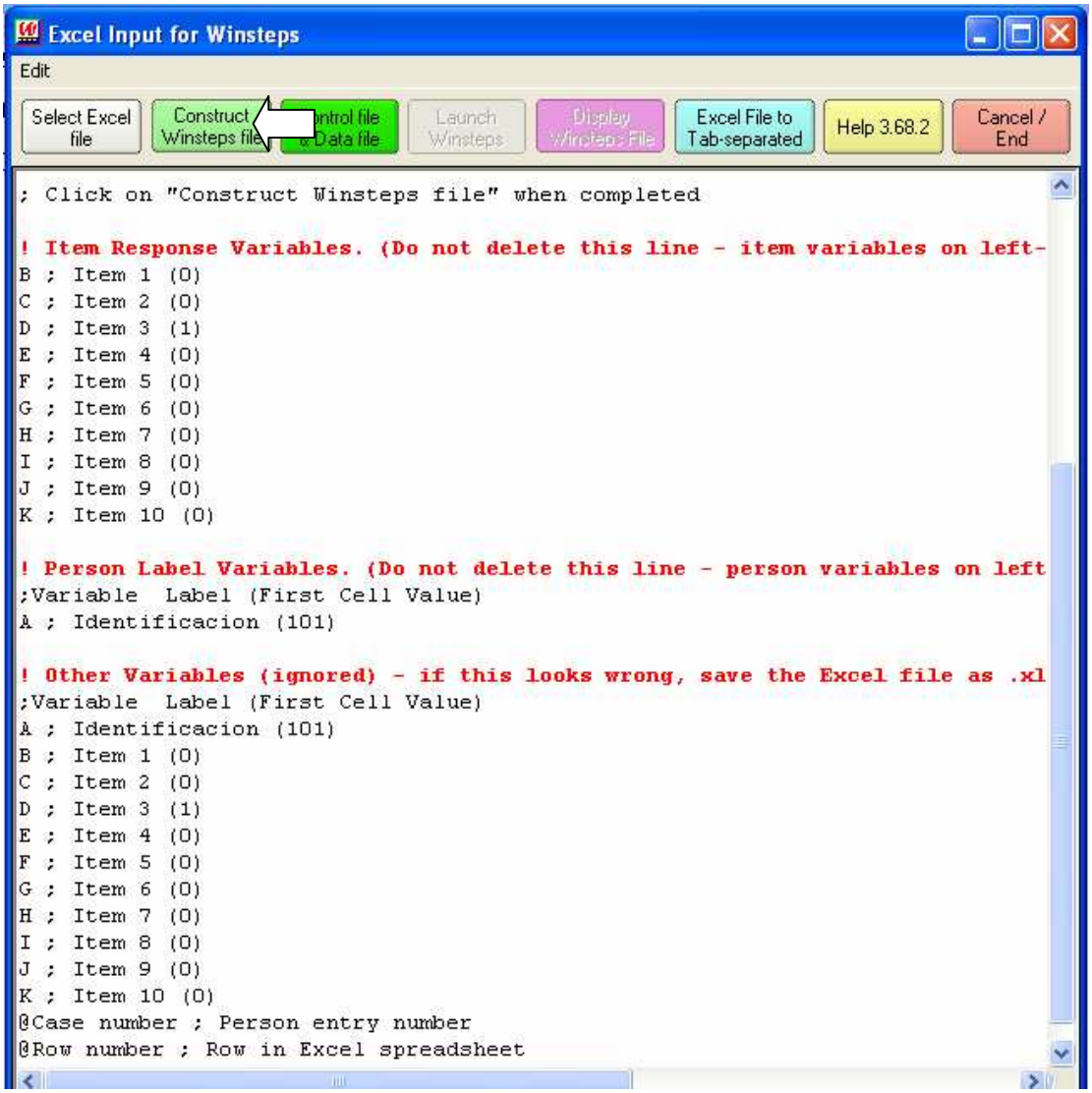


La siguiente imagen corresponde a la ventana que el programa abre para buscar en el equipo el archivo de Excel donde se tiene guardada la base de datos

Este archivo se encuentra guardado en el escritorio con el nombre *Tesis,* se selecciona como se muestra en la siguiente ventana *Excel Input for Winsteps*, de esta forma, el programa empieza a reconocer la base de datos que se señaló para iniciar la construcción de las variables de control básicas de WINSTEPS; esto hará que los datos cambien del formato original al formato .txt que requiere.

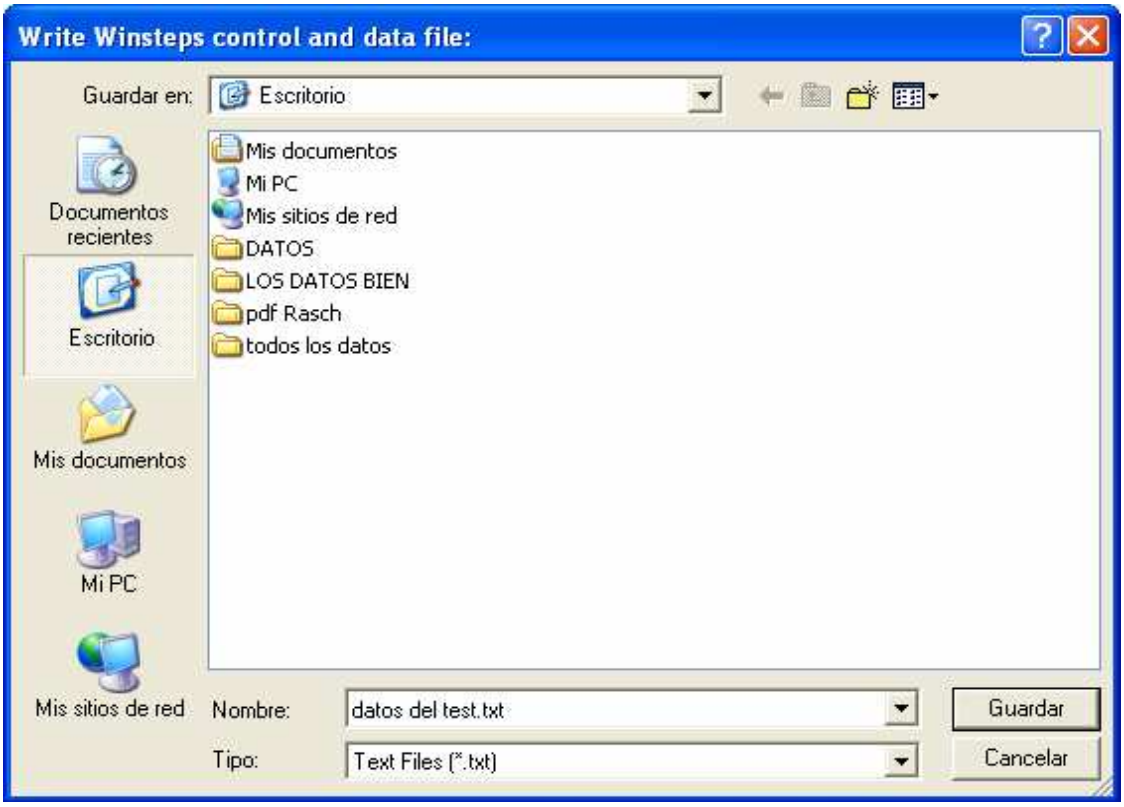


La ventana que aparece al aceptar el archivo tiene algunas indicaciones como la ubicación del archivo donde se encuentran los datos a importar, el número de casos estudiados, es decir, las personas evaluadas que son 164, las variables a analizar que son 11, contando los 10 ítems del test y la casilla de personas. En los párrafos que siguen, las instrucciones señalan que se deben copiar y pegar la cantidad de ítems debajo de la línea roja que dice: *ítem response variables*, esto hará que en el archivo de control aparezca la calificación de todos los ítems. Pero faltaría identificar a cada persona con cada línea de calificaciones, para esto se copian y pegan las instrucciones de las dos primeras líneas donde dice: identificación (1). El procedimiento se ilustra en la siguiente ventana:

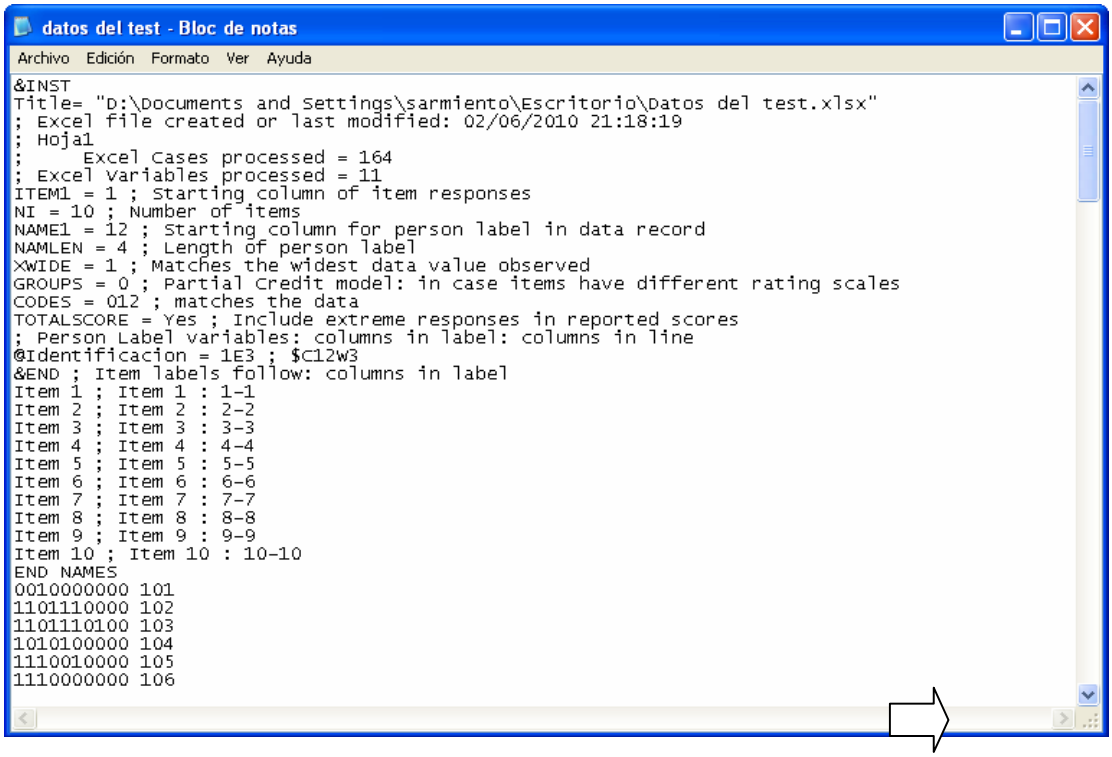


Ahora se le indica al programa que haga la estructura del archivo en la opción indicada por la flecha *Construct Winsteps file* (ver ventana anterior).

Por último aparece la opción para nombrar al archivo .txt y se procede al inicio del análisis del archivo con WINSTEPS.



Este archivo queda guardado en formato .txt donde originalmente se tenía, en este caso en el escritorio. El archivo de control queda como se muestra a continuación.

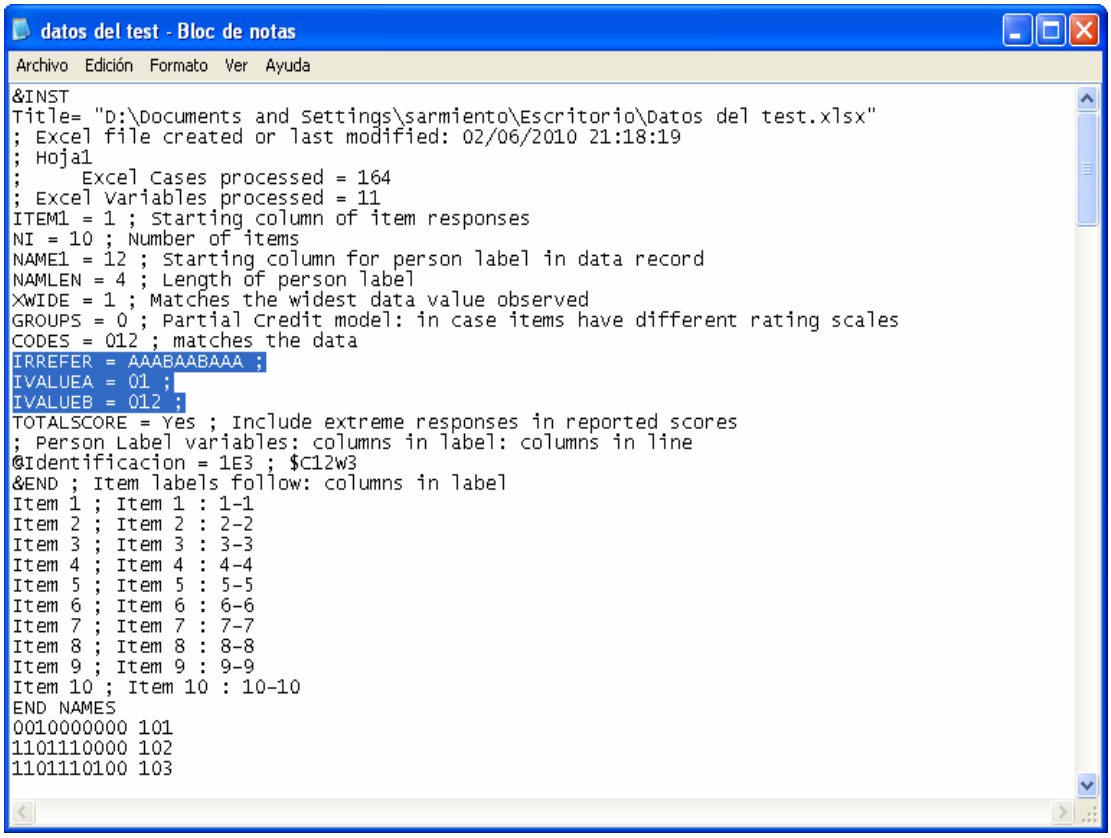


Como las calificaciones de los ítems no son todas dicotómicas, se recordará que se tienen dos ítems politómicos y ocho dicotómicos en el test, esto obliga a implantar unas variables de control extras, pues de lo contrario el programa asumirá que todos los ítems son dicotómicos. Por eso es necesario introducir las siguientes variables de control debajo de la variable *CODES* como se aprecia en la ventana *datos del test – Bloc de notas*:

IRREFER = AAABAABAAA ;

IVALUEA = 01;

IVALUEB = 012;



Estas variables de control hacen que las preguntas de categoría A, tengan una calificación de 0 ó 1 y las de categoría B tengan una calificación de 0, 1, ó 2. Ahora se puede correr este archivo en WINSTEPS de manera correcta, contiene todas las especificaciones de la cantidad de ítems, personas, códigos posibles de calificación, ubicación del archivo y listado con las calificaciones que obtuvieron los estudiantes, que empieza en la parte inferior izquierda y donde se identifica el renglón que dice "*END NAMES*" que señala los resultados de cada pregunta, seguido por el código de cada estudiante que responde la prueba y que para este estudio inicia con el número 101, continua con el 102 y así sucesivamente. Deslizando la barra se puede observar el total de estudiantes que no se aprecia en esta imagen.

IMPLEMENTACION COMPUTACIONAL

Para empezar la implementación computacional se debe tener en cuenta que el algoritmo PROX descrito en al capítulo anterior, es apropiado para la estimación de parámetros por medio de aproximaciones de cálculo manual. El procedimiento de estimación de los parámetros del modelo es el de Estimación Conjunta de Máxima Verosimilitud que en sus siglas en inglés se conoce como JMLE (Joint Máximum Likelihood Estimation).

WINSTEPS® utiliza un proceso iterativo que consiste en dos etapas: una primera con el algoritmo PROX y luego, con los valores obtenidos en la primera etapa da lugar a la segunda etapa que es el procedimiento JMLE que termina dando la estimación de los parámetros de los individuos e ítems junto con sus errores estándar. Adicionalmente, calcula los estadísticos de ajuste y realiza un gran número de tablas y gráficas que son muy útiles a la hora de interpretar la calidad de ajuste de los datos al modelo.