3.- FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS Y DEL TEST

3.1.- INTRODUCCIÓN

Durante las últimas décadas, el Funcionamiento Diferencial del Ítem (*Differential Item Functioning,* DIF) ha sido una de las áreas psicométricas de más auge. El inicio de la polémica comienza en los años cincuenta, a partir de los estudios realizados en la Universidad de Chicago donde EELLS y col. en el año 1951 pusieron de manifiesto las variaciones de los ítems en aspectos tales como contenido y formato, que reducían o exageraban las diferencias entre los grupos de sujetos comparados.

En un principio se pensó que este fenómeno se debía a que los grupos podían diferir en función de características como nivel cultural, clase social, raza u otras, que llevaba a un comportamiento diferente ante las tareas reflejadas en el test. Pero es en la década de los sesenta, coincidiendo con el movimiento de los derechos civiles de EE.UU., cuando la posibilidad de sesgo en los tests recibió amplia atención. Así cualquier test en el que se encontrasen diferencias entre grupos étnicos, culturales y socioeconómicos, se consideraba injusto y sesgado.

Para JESEN (1980), el término "sesgo" puede tener dos connotaciones: una estadística-matemática y otra psicológica: *en estadística matemática el "sesgo" se refiere a una sobre o infraestimación sistemática de un parámetro poblacional para un estadístico basado en muestras sacadas de la población. En Psicometría, el "sesgo" se refiere a errores sistemáticos en la validez predictiva o en la validez del constructo de las puntuaciones en el test de individuos que pertenecen a grupos diferentes.*

Hasta los años setenta la investigación del sesgo era realizada por sociólogos, antropólogos y educadores, fue hacia los ochenta cuando los investigadores en psicometría abordan este tema, y así ANGOFF (1982) propone sustituir el término de sesgo de los ítems, en situaciones en las que sólo se intenta dilucidar las propiedades estadísticas de los ítems en distintos grupos sin entrar en juicios de valor, por el más neutro de discrepancias de ítems. Después de más de veinte años, fueron HOLLAND y THAYLER (1988) quienes con *Funcionamiento Diferencial de los Ítems (Differential Item Functioning, DIF)* acuñaron la expresión que finalmente desplazaría al término sesgo de los ítems.

En la actualidad, el problema es suficientemente importante ya que la existencia de un posible funcionamiento diferencial de uno o más ítems en un test es una clara amenaza a la validez del propio test (ACKERMAN, 1992).

Las investigaciones sobre el Funcionamiento Diferencial de los Ítems (DIF) han estado centradas en el desarrollo de métodos estadísticos para identificar de forma fiable aquellos ítems que funcionan diferencialmente en personas igualmente capaces, pero miembros de distintas características sociodemográficas, (ANGOFF, 1993; CAMILLI y PENFIELD, 1997; JIANG y STOUT, 1998; KIM y COHEN, 1998; OSHIMA y col., 1997; SCHEUNEMAN y GRIMA, 1997; WILLIAMS, 1997). Por el contrario, la comprensión de las causas del DIF no ha recibido una atención semejante (SCHEUNEMAN, 1982, 1987; SKAGG y LISSITZ, 1992; SCHMITT y col., 1993).

3.2.- FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM

En el estudio del funcionamiento diferencial del ítem (DIF) se comparan las respuestas de sujetos de distintos grupos a un ítem.

El DIF se define como aquella técnica que detecta si un ítem funciona igual o diferente para un grupo que para otro.

Técnicamente, el DIF se define como una diferencia en la ejecución o rendimiento en el ítem entre dos o más grupos que han sido condicionados o igualados con respecto al constructo medido en el test (POTENZA y DORANS, 1995).

Para FIDALGO (1996) un ítem funciona diferencialmente si la probabilidad de acertarlo difiere, a igual nivel de la variable medida, entre distintos subgrupos de una población dada.

HIDALGO y LÓPEZ-PINA (2000) señalaron que un ítem tiene DIF cuando sus propiedades estadísticas varían en función de las características del grupo que lo ha contestado, siempre que los grupos comparados manifiesten un mismo nivel del rasgo medido por el test o hayan sido igualados, en términos estadísticos, en dicho rasgo. Si es

el test en su conjunto el que presenta propiedades estadísticas distintas en cada grupo, se habla de *Funcionamiento Diferencial del Test (Differential Test Functioning, FDT).*

La definición del Funcionamiento Diferencial del Ítem que más se adapta a la terminología de Calidad de Vida fue la propuesta por MORALES (2004). Un ítem presenta Funcionamiento Diferencial cuando individuos con un mismo nivel de habilidad (θ), nivel de Calidad de Vida en nuestro contexto, tienen distinta probabilidad de obtener puntuaciones similares sobre el mismo ítem (o sobre el mismo test, si estudiamos el funcionamiento diferencial de un test), según el grupo al que pertenezcan.

El número de grupos entre los que se establece las comparaciones es variable, aunque la mayoría de las investigaciones se lleva a cabo sobre dos grupos. El grupo objeto de análisis se denomina *grupo focal* (grupo minoritario o el que sea objeto del estudio) y el grupo que sirve como criterio de comparación se conoce como *grupo de referencia* (grupo mayoritario o al que va dirigido el test).

Se distinguen dos tipos de DIF en función de la presencia o ausencia de interacción entre el nivel de habilidad y la variable de agrupamiento de los sujetos (MELLENBERGH, 1982):

DIF Uniforme / o consistente: cuando no existe interacción entre el nivel de habilidad medido y la pertenencia a un determinado grupo. Es decir, al mismo nivel de Calidad de Vida dos grupos diferentes de individuos eligen una respuesta diferente de un determinado ítem.

DIF no uniforme / o inconsistente: cuando se da la interacción entre el nivel de habilidad medido y la pertenencia a un determinado grupo, es decir, para distintos niveles de Calidad de Vida, la respuesta para dos grupos de individuos a un determinado ítem puede o no ser la misma.

Por ejemplo, en los siguiente gráficos (Gráficos 3.1 y 3.2) se representan una respuesta categórica de dos ítems del cuestionario QUALEFFO, el primero presenta un DIF uniforme y el segundo un DIF no uniforme.

Para la segunda respuesta categórica del ítem G28 (*Durante los últimos siete días, ¿se ha sentido solo/a?*) de la dimensión *Estado de Ánimo* del cuestionario QUALEFFO, la curva característica para el grupo de los hombres es más achatada que para el grupo de las mujeres, es decir, la probabilidad de responder a esta categoría (*Raramente*) es más baja para el grupo de los hombres que para el grupo de las mujeres a todos los niveles de Calidad de Vida. Por lo tanto este ítem presenta un DIF Uniforme. (Ver Gráfico 3.1)

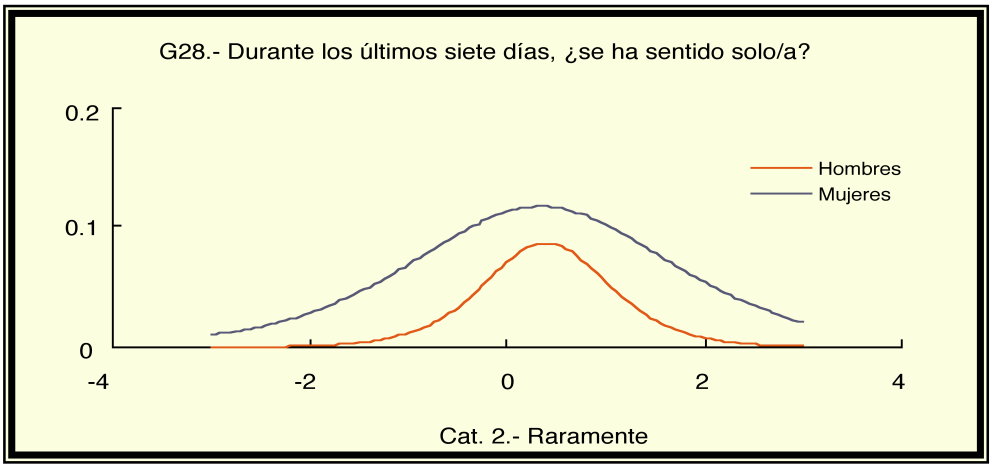

Gráfico 3.1.- Curvas características de la segunda categoría del ítem G28 del cuestionario QUALEFFO para el grupo de los hombres y el grupo de las mujeres

Sin embargo, para la primera respuesta categórica del ítem **G27** (*Durante los últimos siete días, ¿se ha sentido desanimado/a?*) de la dimensión *Estado de Ánimo* del cuestionario QUALEFFO, la curva característica para el grupo de los hombres en niveles de Calidad de Vida inferiores presenta valores más altos que la del grupo de las mujeres, en cambio para niveles superiores de Calidad de Vida presenta valores más bajos en el grupo de los hombres que en el de las mujeres, por lo tanto la probabilidad de responder a esta categoría del ítem es más alta para el grupo de los hombres que para el de las mujeres en ciertos niveles de Calidad de Vida y más baja en otros niveles (Ver Gráfico 3.2). Luego este ítem para la categoría (*No*), presenta Funcionamiento Diferencial No uniforme.

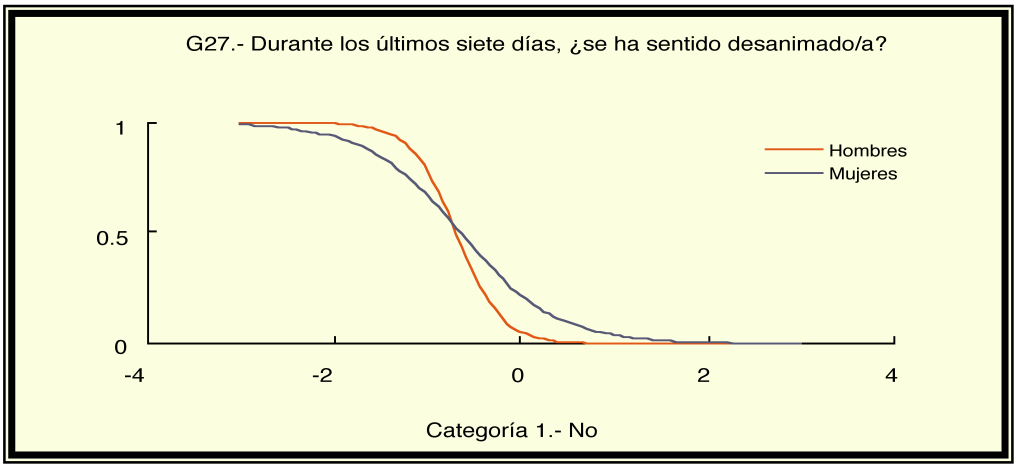

Gráfico 3.2.- Curvas características de la primera categoría del ítem G27 del cuestionario QUALEFFO para el grupo de los hombres y el grupo de las mujeres

3.3.- DIF, IMPACTO Y SESGO

Es importante no confundir DIF con impacto. Este último hace referencia a las diferencias reales existentes en el comportamiento de dos o más grupos, es decir, diferencias en la probabilidad de responder correctamente a un ítem; es decir, en el test de los distintos grupos obedecen a diferencias reales entre los grupos en la característica medida por el test.

Para ACKERMAN (1992), es una diferencia entre grupos en el desempeño en un ítem causada por una diferencia real en la variable medida.

Si un ítem presenta impacto, la probabilidad de responderlo correctamente será mayor para un grupo que para otro, reflejando de esta manera las diferencias entre grupos en la habilidad medida, y la probabilidad de responder correctamente a ese ítem será la misma para sujetos con el mismo nivel de habilidad con independencia del grupo al que pertenezcan.

El impacto de un ítem se puede describir como cualquier disparidad del grupo en el funcionamiento del ítem que refleja diferencias reales del conocimiento y la experiencia del constructo de interés (CLAUSER y MANZOR, 1998).Sin embargo, el DIF está presente en un ítem cuando los individuos de dos grupos diferentes tienen diferente

probabilidad de contestar al ítem correctamente para un nivel de habilidad dado (SHEPARD y col., 1981).

CAMILLI y SHEPARD (1994) definieron el sesgo de un ítem como un error sistemático en la medida del constructo para los miembros de un grupo particular.

Una vez que se ha detectado el DIF, este puede atribuirse o al impacto del ítem o al sesgo del ítem.

Las diferencias entre el impacto de un ítem y el sesgo de un ítem en términos de diferencias de los grupos están basadas sobre las características relevantes e irrelevantes respectivamente. DIF requiere que los miembros de ambos grupos tengan la misma habilidad antes de determinar si se diferencian en la probabilidad para el éxito (ZUMBO, 1999).

El enlace entre DIF, impacto y sesgo del ítem, es en gran parte metodológico: los análisis estadísticos se utilizan para identificar los ítems con DIF y los análisis críticos para determinar si el DIF es atribuible al sesgo o al impacto del ítem para los miembros de un grupo determinado. LINN (1993) propusieron que el análisis de un ítem requiere dos procedimientos: el estadístico y el crítico.

Para analizar la presencia o ausencia de impacto en cada ítem del test se ha llevado a cabo un contraste de hipótesis acerca de la igualdad o desigualdad de las proporciones de éxito obtenidas para cada grupo.

3.4.- TÉCNICAS ESTADÍSTICAS PARA DETECTAR EL DIF

Actualmente, existe una infinidad de métodos estadísticos para detectar el DIF (ANDRIOLA 2000b; 2001c) y muchas formas de clasificarlos, de las cuales comentaremos algunas de ellas. En el contexto de esta investigación se considerarán las técnicas basadas en la TRI para el estudio del DIF.

Según MILLSAP y EVERSON (1993) las técnicas para la detección del DIF podemos clasificarlas en dos grandes grupos:

- Aquellas que utilizan como **variable de igualación entre los grupos** (variables de equiparación) **una puntuación observada en el test**. Para estas técnicas, no es necesario especificar un modelo de medida que relacione la puntuación observada y la variable latente, sino que consideran a la variable observable, usualmente las puntuaciones en el test, como un estimador de la variable que pretende medir el test. Entre éstas técnicas destacamos:

    - Mantel-Haenszel (HOLLAND y THAYER, 1988)
    - Procedimiento estandarizado (DORANS y KULICK, 1986)
    - Los modelos loglineales y los modelos logit (MARASCUILLO y SLAUGHTER, 1981; MELLENBERGH, 1982)
    - La regresión logística (SWAMINATHAN y ROGERS, 1990)
    - el análisis discriminante (MILLER y SPRAY, 1993)

- Las que utilizan como medio de equiparación **la habilidad latente estimada bajo algún modelo de respuesta al ítem.** Con estas técnicas, se comprueba si los parámetros que definen el modelo permanecen invariantes a través de los grupos definidos por la variable que sirve para agrupar a los individuos. Entre estas técnicas destacamos:
    - Los que comparan los parámetros de los ítems estimados a través de los grupos (LORD, 1980) estimando la significación del tamaño de las diferencias entre los mismos
    - Los que comparan las curvas características del ítem estimadas para cada grupo mediante la medición del área comprendida entre las mismas (RAJU, 1988)
    - Los que usan la comparación del ajuste de los modelos.

Según FIDALGO y PAZ (1995), las técnicas podrían agruparse en:

- Aquellas que **no controlan las diferencias en el nivel de habilidad entre grupos**, como:

    - El análisis de la varianza.
    - El método delta, en estas se puede producir una confusión entre DIF e impacto, y concluir que ítems que reflejan la diferencia en la distribución de la variable medida en el grupo presentan DIF.

- Aquellas **que establecen las comparaciones oportunas para iguales niveles de habilidad para controlar las diferencias**, como:

    - Chi-Cuadrado (CAMILLI, 1979; MARASCUILO, 1981; MARASCUILO y SLAUGTHER, 1981)
    - Mantel-Haenszel (HOLLAND y THAYER, 1986; 1988).
    - Modelos loglineales, logit y de clases latentes (KELDERMAN y MACREADY, 1990; KOK y col. 1985; VAN DER FLIER y col., 1991).
    - Regresión logística (SPRAY y CARLSON, 1986; SWAMINATHAN y ROGERS, 1990).
    - Métodos basados en la TRI (KIM y COHEN, 1991; LORD, 1980; RAJU, 1988).

POTENZA y DORANS (1995) dan una doble clasificación de las técnicas para detectar el DIF: distinguen si los ítems del test son dicotómicos o son politómicos y distingue dos tipos de aproximaciones, aproximaciones a las puntuaciones observadas y aproximaciones a las variables latentes. En la siguiente tabla se muestran las distintas técnicas clasificadas según estos autores.

	Ítems Dicotómicos		**Ítems Politómicos**	
	Paramétricos	No Paramétricos	Paramétricos	No Paramétricos
P. Observadas	. Regresión Logística (LRDIF)	.Estandarización (STND) . Mantel-Haenszel (MH)	. Regresión logística a ítems politómicos	. Polytomous STND . HW1 y HW3 . Mantel-Haenszel generalizado (GMH)
P. V. Latentes	. TRI . SIBTEST		. TRI . SIBTEST	

Tabla 3.1.- Técnicas para detectar el DIF según POTENZA y DORANS (1995)

GÓMEZ y NAVAS (1996), dan una clasificación según la cual se dispone de métodos:

- Basados en la **teoría clásica de los tests** (ANGOFF, 1972, 1982; ANGOFF y FORD, 1973)
- Métodos basados en **el análisis factorial** (OORT, 1992, 1993)
- Métodos basados en **la teoría de respuesta a los ítems** (LINN y HARNISCH, 1981; LINN y col., 1981; THISSEN y col., 1986, 1988)
- **Métodos basados en χ^2**, de los que destacamos
    - Métodos basados en la χ^2 *stricto sensu*, como el método de la χ^2 de los aciertos (SCHEUNEMAN, 1979)
    - Método de la χ^2 total (CAMILLI, 1979)
    - Mantel-Haenszel (HOLLAND y THAYER, 1986, 1988)
    - Modelos loglineales (VAN DER FLIER y col., 1984)
    - La regresión logística (ROGERS y SWAMINATHAN, 1993)

HIDALGO MONTESINOS y col., (1997) propusieron la siguiente clasificación:

a) **Métodos de invarianza condicional observada.** Utilizan las puntuaciones observadas en el test desde la perspectiva de la TCT. Dichos autores incluyen los métodos del χ^2 de Scheuneman, el χ^2 de Pearson, el método de Mantel-Haenszel, el método estandarizado y el de la regresión logística.

b) **Métodos de invarianza condicional no observada.** Utilizan las estimaciones de la habilidad (θ) según el modelo TRI más adecuado a los datos. Según ellos, son ejemplos los métodos del χ^2 de Lord, el de las áreas, el de comparación de os parámetros de los ítems y el procedimiento SIBTEST.

Por último, HIDALGO y LÓPEZ PINA (2000) consideran la misma clasificación que MILLSAP y EVERSON (1993) y mencionan las mismas técnicas que estos autores.

3.4.1 MÉTODOS DE DETECCIÓN DEL DIF BASADOS EN LA TEORÍA DE LA RESPUESTA AL ÍTEM

En términos de la TRI el ítem no tiene DIF cuando la CCI es idéntica para los grupos comparados en un mismo nivel o magnitud de la variable latente medida (LORD, 1980; MELLENBERGH, 1989), o lo que es lo mismo, un ítem presenta DIF si la CCI es diferente en cada uno de los grupos donde se calibre el ítem.

En el lenguaje matemático podríamos decir que el ítem no tiene DIF con respecto a la variable G (grupo) dado Z (Capacidad del sujeto o nivel de θ) sí y solamente sí $F(X/g,z) = F(X/z)$ donde: X es la puntuación en el ítem; g es el valor obtenido según la variable G y z es el valor obtenido según la variable Z.

Un ítem presenta DIF si a valores iguales de θ no corresponden valores iguales de probabilidad, $P(\theta)$, en las curvas de los grupos considerados, es decir, cuando

$$(ES_{ij})_R \neq (ES_{ij})_F, \tag{3.1}$$

donde $(ES_{ij})_R$ es la puntuación verdadera para un ítem i y un sujeto j, que pertenece al grupo de referencia; $(ES_{ij})_F$ es la puntuación verdadera para un ítem i y un sujeto j, que pertenece al grupo focal.

Si el ítem es politómico la puntuación verdadera, es la puntuación condicional esperada, (ES_{ij}), donde para un ítem i y un individuo j

$$ES_{ij} = \sum_{r=1}^{m} P_r^i(\theta_j) X_{ir}, \tag{3.2}$$

con X_{ir} la puntuación para una categoría r del ítem i ($r = 1,2,...,m$) y $P_r^i(\theta_j)$ es la probabilidad de responder a esa categoría para un nivel de rasgo latente θ_j.

DOUGLAS y col., en 1996 propusieron los conceptos de DIF "suave" y DIF "adverso". En el caso de que el DIF beneficie al grupo de referencia, es decir, cuando $T_{jR}(\theta) > T_{jF}(\theta)$ o cuando $P_{jR}(\theta) > P_{jF}(\theta)$ se denomina DIF "suave". El DIF adverso

ocurre cuando el DIF es beneficio del grupo focal, es decir, cuando $T_{jR}(\theta) < T_{jF}(\theta)$ o cuando $P_{jR}(\theta) < P_{jF}(\theta)$.

En el contexto de la TRI, la lógica para la detección del DIF consiste en comparar las CCI de los ítems, considerando los grupos de referencia y focal, a través de métodos apropiados (ROUSSOS y col., 1999). Los distintos métodos fueron desarrollados en base a los tipos de DIF (BOCK, 1993). El DIF uniforme o consistente, se observa cuando las CCI del ítem estudiado con respecto a los grupos de referencia y focal son diferentes, pero no se cruzan, o en otras palabras cuando hay una ventaja relativa para uno de los grupos estudiados, cuyo valor es constante a lo largo de todo el rango de la aptitud evaluada. Este caso ocurre cuando el parámetro a tiene el mismo valor en las dos curvas características, es decir cuando ambas son paralelas (ROUSSOS y col., 1999).

Cuando las curvas características del ítem estudiado con respecto a los grupos de referencia y focal son diferentes y, además, se cruzan en algún punto de la escala utilizada para medir la aptitud, el DIF no uniforme o inconsistente Este caso ocurre cuando los parámetros *a, b* ó *c* tiene valores diferentes en las dos CCI, es decir, cuando las CCI no son paralelas (ROUSSOS y col., 1999).

Sin embargo, HANSON (1998), SHEALY y STOUT (1993 ab), habla de tres tipos de DIF:

DIF uniforme: cuando $a_{iR} = a_{iF} \quad \forall k$

DIF ordinal: cuando $a_{iR} \neq a_{iF}$ y $b_{ikR} \neq b_{ikF} \quad \forall k$

DIF no ordinal: cuando $a_{iR} \neq a_{iF}$ y $b_{ikR} = b_{ikF} \quad \forall k$

La TRI ofrece distintos métodos para la detección del DIF que básicamente comparan las respuestas dadas a un ítem por sujetos a los que se les estima un mismo nivel de habilidad en dos grupos distintos (HOLLAND y THAYER, 1988). Entre los procedimientos más utilizados podemos mencionar:

1) Los que comparan los parámetros del ítem estimados en los dos grupos (LORD, 1980); si no existiese DIF los parámetros del ítem en ambos grupos no deberían de diferir entre sí más allá de lo esperable por las fluctuaciones aleatorias del muestreo.

2) Los que comparan las curvas características del ítem, estimadas para cada grupo mediante la medición del área comprendida entre las mismas (RAJU, 1988, 1990). Cuando mayor sea el área entre las curvas mayor DIF habrá.

3) Los que usan la comparación del ajuste de modelos (THISSEN y col., 1988, 1993). En esta estrategia se compara un modelo en el que los parámetros del ítem son iguales en los dos grupos frente a otro en el que alguno o algunos de los parámetros del ítem difiere entre grupos. Si el segundo modelo se ajusta significativamente mejor a los datos que el primero, podemos concluir la presencia de DIF.

Veamos a continuación cada uno de estos procedimientos de evaluación del DIF.

3.4.1.1.- COMPARACIÓN DE PARÁMETROS DE ÍTEMS ESTIMADOS

La lógica de este procedimiento es muy simple: un ítem tiene DIF si los parámetros estimados para dos subpoblaciones no coinciden, esto es, existen diferencias significativas (THISSEN y col., 1993). Para el caso del modelo logístico de un parámetro, la comparación del parámetro b para dos subpoblaciones viene dado por el estadístico:

$$Z = \frac{\hat{b}_R - \hat{b}_F}{\sqrt{S^2(\hat{b}_R) + S^2(\hat{b}_F)}}, \qquad (3.3)$$

donde: b_R *y* b_F son los parámetros de dificultad estimados de los ítems para cada grupo (referencia y focal); $S^2(b_R)$ y $S^2(b_F)$ son las varianzas de b en cada grupo (referencia y focal).

Este valor Z se compara con una distribución normal, correspondiente al nivel de confianza fijado, que permite aceptar o no la hipótesis nula ($H_0 : b_1 = b_2$).

Para los modelos logísticos de dos parámetros y de tres parámetros se comparan los parámetros a y b, considerando para los modelos de tres parámetros que c se mantiene invariante (MUÑIZ 1997). Las fórmulas matemáticas para comparar *a* y *b* son:

$$Z_a = \frac{\hat{a}_R - \hat{a}_F}{\sqrt{S^2(\hat{a}_R) + S^2(\hat{a}_F)}} \;;\; Z_b = \frac{\hat{b}_R - \hat{b}_F}{\sqrt{S^2(\hat{b}_R) + S^2(\hat{b}_F)}}, \tag{3.4}$$

donde b_R *y* b_F son los parámetros de dificultad estimados de los ítems para cada grupo (referencia y focal); $\hat{a}_R$ *y* $\hat{a}_F$ son los parámetros de discriminación estimados de los ítems para cada grupo (referencia y focal); $S^2(b_R)$ y $S^2(b_F)$ son las varianzas de *b* en cada grupo (referencia y focal); $S^2(\hat{a}_R)$ y $S^2(\hat{a}_F)$ son las varianzas de a en cada grupo (referencia y focal).

Una limitación de este método, es que se hace una comparación de los parámetros *a* y *b* por separado.

3.4.1.2.- ESTADÍSTICO DEL LORD

Este autor propone un estadístico para contrastar la hipótesis nula de los vectores que definen los parámetros de los ítems en las poblaciones de referencia y focal, son iguales:

$$H_0 : x'_F = x'_R, \tag{3.5}$$

donde x'_F, x'_R son vectores de dimensión $(1xn)$ que tienen como elementos los parámetros del ítem en el grupo focal y los parámetros del ítem en el grupo de referencia respectivamente.

Para ítems dicotómicos, el estadístico utilizado para someter a prueba la hipótesis nula de ausencia de DIF es:

$$CHI - LORD = \mathbf{V'S^{-1}V}, \tag{3.6}$$

donde **V** es el vector de diferencias entre los parámetros estimados para un ítem en el grupo de referencia y los parámetros estimados para ese mismo ítem en el grupo focal. ($\mathbf{V'} = (\hat{a}_R - \hat{a}_F, \hat{b}_R - \hat{b}_F, \hat{c}_F - \hat{c}_R)$, para el modelo de tres parámetros por ejemplo) y $\mathbf{S^{-1}}$, es la inversa de la matriz de varianza-covarianza de V, de dimensión (nxn)

$$\mathbf{S^{-1}} = (\mathbf{S}_R + \mathbf{S}_F)^{-1}. \tag{3.7}$$

El estadístico sigue una distribución χ^2 con grados de libertad iguales al número de parámetros del modelo (n). Se rechaza la hipótesis nula de ausencia de DIF al nivel de significación elegido sólo si: $\chi^2 \geq \chi^2_{1-\alpha,n}$.

Una crítica al estadístico de Lord es que la hipótesis nula pueda ser rechazada habiendo poca diferencia entre las curvas características del ítem en las regiones donde se encuentran la mayoría de los encuestados (LINN, co1. 981). Para interpretar en este caso los resultados, se recomienda el cálculo de algún índice del área de intervalos cerrados cuando χ^2 de Lord resulte significativo.

COHEN y col. (1993) proponen una extensión de este estadístico para el caso del modelo de respuesta graduada. En este caso $\mathbf{V}'$ es el vector de diferencias entre las estimaciones de los parámetros de un ítem:

$$\mathbf{V}' = (\hat{a}_R - \hat{a}_F,\ \hat{b}_{1R} - \hat{b}_{1F},\ \hat{b}_{2R} - \hat{b}_{2F},\ \hat{b}_{3R} - \hat{b}_{3F},\ldots,\ \hat{b}_{(m-1)R} - \hat{b}_{(m-1)F}), \tag{3.8}$$

y la matriz de varianza-covarianza asintótica para los vectores de diferencias entre parámetros vendría dada por:

$$\hat{\mathbf{S}} = \begin{pmatrix} \hat{\mathbf{S}}^2_{(\hat{a}_R - \hat{a}_F)} & \hat{\mathbf{S}}_{(\hat{a}_R - \hat{a}_F)(\hat{b}_{1R} - \hat{b}_{1F})} & \cdots & \hat{\mathbf{S}}_{(\hat{a}_R - \hat{a}_F)(\hat{b}_{(m-1)R} - \hat{b}_{(m-1)F})} \\ & \hat{\mathbf{S}}^2_{(\hat{b}_{1R} - \hat{b}_{1F})} & \cdots & \hat{\mathbf{S}}_{(\hat{b}_{1R} - \hat{b}_{1F})(\hat{b}_{(m-1)R} - \hat{b}_{(m-1)F})} \\ & & \ddots & \vdots \\ & & & \hat{\mathbf{S}}^2_{(\hat{b}_{(m-1)R} - \hat{b}_{(m-1)F})} \end{pmatrix}. \tag{3.9}$$

Según COLE y MOSS (1989), los métodos basados en la teoría de respuesta a los ítems y en χ^2, abordan el estudio del DIF desde una perspectiva condicional: el funcionamiento diferencial del ítem se estudia a partir de las diferencias en la dificultad del ítem estimada en grupos distintos pero de la misma habilidad.

Estos dos métodos comparten un problema en su forma de proceder ya que el ítem en estudio sirve para definir la variable que se va utilizar para formar los grupos a partir de los cuales se obtienen las estimaciones de la dificultad al ítem. Entonces, si el ítem presenta DIF se está utilizando una medida también contaminada de la habilidad para estudiar el posible DIF. Hay diferentes procedimientos para paliar este problema, centrados en purificar la medida de la habilidad utilizada como criterio o variable para formar los grupos.

3.4.1.3.- MEDIDAS DE ÁREA

Esta técnica consiste en estimar las curvas características del ítem para los grupos de interés y realizar el cálculo del área comprendida entre dichas curvas (WAINER, 1993). El área entre las curvas constituye un índice de discrepancia entre ellas. En consecuencia, indica una posible existencia de DIF, ya que si ambas curvas coinciden el área comprendida entre ellas es cero y no habría DIF. En este método existen diversos procedimientos para determinar el valor comprendido entre las curvas características de los grupos estudiados.

RUDNER y col. (1980) propusieron la siguiente fórmula para su cálculo:

$$A = \sum_{\theta=-4}^{\theta=4} \left| P_{GR}(\theta_j) - P_{GF}(\theta_j) \right| \Delta\theta, \qquad (3.10)$$

donde $P_{GR}(\theta_j)$ es el valor de la probabilidad de acierto de un ítem en el grupo de referencia dado θ_j; $P_{GF}(\theta_j)$ es el valor de la probabilidad de acierto de un ítem en el grupo focal dado θ_j; $\Delta\theta_j$ es el valor de la base de un rectángulo y altura $\left| P_{GR}(\theta_j) - P_{GF}(\theta_j) \right|$.

Cuanto menor es el valor del incremento más preciso es el cálculo del área.

LINN y HARNISCH (1981) propusieron otro procedimiento dado por:

$$A = \sum_{\theta=-3}^{\theta=3} \sqrt{\left|P_{GR}(\theta_j) - P_{GF}(\theta_j)\right|^2 \Delta\theta} \ . \qquad (3.11)$$

Los términos de esta fórmula tienen el mismo significado que la ecuación anterior (3.11).

Otro procedimiento para el cálculo del DIF fue propuesto por RAJU (1988) que define el área entre dos curvas características para modelos de tres parámetros como:

A) Si $c = c_1 = c_2$ y $a_1 \neq a_2$:

Área con signo entre las Curvas Características vale $(1-c)(b_2 - b_1)$

Área sin signo, o absoluta, entre las Curvas Características vale

$$(1-c)\left|\frac{2(a_2 - a_1)}{Da_2a_1}\ln(1 + e^{\frac{Da_1a_2(b_2-b_1)}{a_2-a_1}}) - (b_2 - b_1)\right|, \qquad (3.12)$$

donde el único punto de corte entre las curvas ($c = c_1 = c_2$) es $\theta_0 = \frac{a_2b_2 - a_1b_1}{a_2 - a_1}$

B) Si $c = c_1 = c_2$ y $a_1 = a_2$:

Área con signo las Curvas Características vale $(1-c)(b_2 - b_1)$

Área sin signo o absoluta entre las Curvas Características vale $(1-c)|b_2 - b_1|$

C) Si $c_1 \neq c_2$

El área con signo vale $\pm\infty$ y el área sin signo vale $+\infty$.

El área entre dos curvas características para modelos de dos parámetros la definió como:

A) Si $a_1 \neq a_2$

Área con signo entre las Curvas Características vale $(b_2 - b_1)$

Área sin signo o absoluta entre las Curvas Características vale

$$(\text{UA})=\left|\frac{2(a_2-a_1)}{Da_2a_1}\ln(1+e^{\frac{Da_1a_2(b_2-b_1)}{a_2-a_1}})-(b_2-b_1)\right|. \qquad (3.13)$$

B) Si $a_1 = a_2$

Área con signo entre las Curvas Características vale $(b_2 - b_1)$

Área sin signo o absoluta entre las Curvas Características vale $|b_2 - b_1|$

COHEN y col. (1993) extendieron las medidas de RAJU al caso del modelo de respuesta graduada. El área con signo y sin signo para un ítem *i* vienen dadas por las expresiones

$$SA_i = \int_{-\infty}^{\infty}\sum_{r=1}^{m_j} w_{ir}\left[\hat{P}^i_{rR}(\theta)-\hat{P}^i_{rF}(\theta)\right]d\theta, \qquad (3.14)$$

$$UA_i = \int_{-\infty}^{\infty}\left|\sum_{r=1}^{m_j} w_{ir}\hat{P}^i_{rR}(\theta)-\sum_{r=1}^{m_j} w_{ir}\hat{P}^i_{rF}(\theta)\right|d\theta, \qquad (3.15)$$

donde, w_{ir} son los pesos y $\hat{P}^i_{rR}(\theta)$ y $\hat{P}^i_{rF}(\theta)$ son las estimaciones de las probabilidades para el grupo de referencia y el grupo focal respectivamente.

3.4.1.4.- COMPARACIÓN DE MODELOS

THISSEN, y col. (1988) propusieron el siguiente procedimiento para evaluar el DIF: si el ajuste a los datos de un modelo que incluye parámetros de los ítems diferentes para el grupo focal y el de referencia es significativamente mejor que un modelo en el que los parámetros de los ítems son iguales en ambos grupos, podemos concluir que el ítem presenta DIF. Este procedimiento puede ser utilizado para comprobar la presencia de DIF en varios ítems a la vez.

Tenemos dos modelos a comparar, el modelo compacto (modelo C) y el modelo aumentado (modelo A), que incluye todos los parámetros del modelo C y alguno más. Se trata de decidir si los parámetros adicionales del modelo A son significativamente diferentes a cero. El estadístico utilizado para la comparación de modelos es la razón de

verosimilitud (LR), que sigue una distribución χ^2 con grados de libertad igual a la diferencia en el número de parámetros entre los modelos bajo la hipótesis nula de que los parámetros del modelo aumentado son iguales a cero.

3.4.1.5.- FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM Y DEL TEST (DFIT)

RAJU y col. (1995) propusieron un procedimiento paramétrico basado en la Teoría de Respuesta al Ítem, el Funcionamiento Diferencial del Ítem (*DIF*) y el Funcionamiento Diferencial del Test (*DTF*) conocido como "Funcionamiento Diferencial de Ítems y de Test" (*DFIT*, Differential Functioning of Items and Tests). Este procedimiento será el que consideraremos en esta investigación, por lo que lo desarrollaremos en detalle.

En el marco del *DFIT*, el funcionamiento diferencial se define como una diferencia en el valor esperado del ítem o de la escala para los individuos con la misma situación en el constructo latente θ.

En la *Teoría de Respuesta al Ítem*, la puntuación verdadera para ítems politómicos es la puntuación condicional esperada (ES_{ij}) sobre la posición del rasgo latente. Entonces, para un ítem *i* y un individuo *j* como

$$ES_{ij} = \sum_{r=1}^{m} P_r^i(\theta_j) X_{ir} \,, \tag{3.16}$$

donde X_{ir} es la puntuación para la categoría *r* del ítem *i* ($r = 1,2,..,m$) y $P_r^i(\theta_j)$ es la probabilidad de responder a esa categoría para un nivel de rasgo latente θ_j.

Sumando las puntuaciones esperadas para los *n* ítems de un cuestionario, la puntuación verdadera de un cuestionario para un individuo *j* viene dada por

$$T_j = \sum_{i=1}^{n} ES_{ij} \,. \tag{3.17}$$

Sumando las puntuaciones esperadas, que se han obtenido con las estimaciones de los parámetros de los ítems, obtenemos la puntuación verdadera del cuestionario para el

grupo de referencia (T_R) y análogamente obtenemos la puntuación verdadera para el grupo focal (T_F). La hipótesis nula para el funcionamiento diferencial a nivel de test es $T_R = T_F$.

En el análisis *DFIT*, RAJU y col. (1995) propusieron varios índices para analizar el funcionamiento diferencial, dos índices para evaluar el funcionamiento diferencial del ítem (*CDIF* y *NCDIF*) y uno para evaluar el funcionamiento diferencial del test (*DTF*).

Cuanto mayor es la diferencia entre las dos puntuaciones esperadas, mayor funcionamiento diferencial en el test. De acuerdo con RAJU y col. (1995), una medida del *DTF* a nivel del individuo *j* se puede definir como:

$$D_j^2 = (T_{jF} - T_{jR})^2, \tag{3.18}$$

mientras el índice *DTF* a través de todos los individuos del grupo focal se define como

$$DTF = E_F D_j^2 = E_F (T_{jF} - T_{jR})^2, \tag{3.19}$$

o equivalentemente

$$DTF = D_j^2 = \int_\theta D_j^2 f_F(\theta) d\theta = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_D^2, \tag{3.20}$$

donde $f_F(\theta)$ es la función de densidad de θ en el grupo focal, y μ_{TF} y μ_{TR} representan la media de la puntuación esperada de los individuos en el grupo focal y de referencia, respectivamente.

Para evaluar el funcionamiento diferencial a nivel de ítem, en el marco *DFIT*, se definen dos índices, el índice *CDIF* y el índice *NCDIF*.

Compensatory Differential Ítem Functioning (*CDIF*) es un índice a nivel de ítem que representa la contribución neta de un ítem al *DTF*, es decir, es el cambio en *DFT* asociado a la cancelación del ítem focal del test. Para el cálculo del índice se asume que todos los ítems del test presentan Funcionamiento Diferencial simultáneamente.

Si denotamos por d_{ij} a $ES_{ijF} - ES_{ijR}$, entonces:

$$DTF = E\left[\left(\sum_{i=1}^{n} d_{ij}\right)^2\right] = \sum_{i=1}^{n} [Cov(d_i, D) + \mu_{d_i}\mu_D], \tag{3.21}$$

donde $Cov(d_i, D)$ es la covarianza de la diferencia en puntuaciones esperadas en el ítem *i* y la diferencia de las puntuaciones esperadas en el test (*D*) y μ_{d_i} y μ_D son las medias de d_{ij} y D_j, respectivamente. En este caso el *DIF* puede escribirse como:

$$DIF_i = Cov(d_i, D) + \mu_{d_i}\mu_D, \tag{3.22}$$

y se le denomina *Compensatory D*IF (*CDIF*). De esta definición concluimos que

$$DTF = \sum_{i=1}^{n} CDIF_i \,. \tag{3.23}$$

En otras palabras, el índice *CDIF* ofrece una estimación de la contribución de cada uno de los ítems al funcionamiento diferencial a nivel de test y la suma de los *CDIF* para todos los ítems del test es una medida del funcionamiento diferencial a nivel de test (*DTF*).

Como el índice *CDIF* puede tomar valores positivos y valores negativos, ítems con valores de *CDIF* positivos e ítems con valores de *CDIF* negativos pueden cancelar en funcionamiento diferencial a nivel de test, en este sentido este índice *compensa* el DIF. Por ejemplo, si un ítem de un test tiene un valor de *CDIF* que favorece al grupo focal y otro ítem tiene un valor de *CDIF* que favorece al grupo de referencia y ambos valores son aproximados en magnitud, el efecto es que no haya funcionamiento diferencial a nivel de test. (MORALES y col., 2006).

El índice *Noncompensatory Differential Item Functioning* (*NCDIF*), es un estadístico puramente a nivel de ítem que refleja las diferencias de la puntuación verdadera para los dos grupos de individuos. Para su cálculo se asume que todos los ítems del test, excepto el que se esta estudiando, están libres de funcionamiento diferencial. Matemáticamente, es el cuadrado de la diferencia de la puntuación verdadera para los dos grupos, para un nivel de rasgo latente θ. En el caso dicotómico, este índice esta estrechamente

relacionado otros índices para identificar DIF como el estadístico χ^2 de Lord, o el área sin signo de Raju.

Si todos los ítems están libres de DIF, excepto el ítem en estudio *i*, entonces $d_k = 0$ para todo $k \neq i$.

Por lo tanto $$DIF_i = Cov(d_i, D) + \mu_{d_i}\mu_D = \sigma^2_{d_i} + \mu^2_{d_i}, \tag{3.24}$$

entonces $$NCDIF_i = \sigma^2_{d_i} + \mu^2_{d_i}, \tag{3.25}$$

Destacamos tres aspectos sobre el NCDIF:

1.- *NCDIF*=0 si y solo si los parámetros del ítem *i* son iguales para el grupo focal y para el grupo de referencia, ya que d_i se ha definido como la diferencia de las puntuaciones esperadas para el ítem *i* en el grupo focal y el grupo de referencia.

2.- Si denotamos por $f_F(\theta)$ a la función de densidad de θ en el grupo focal, $NCDIF_i = \int_{-\infty}^{\infty} \left[ES_{ijF} - ES_{ijR} \right]^2 f_F(\theta) d\theta$, que es la definición de DIF dada por WAINER (1993).

3.- $NCDIF_i = \int_{-\infty}^{\infty} \left| ES_{ijF} - ES_{ijR} \right|^2 f_F(\theta) d\theta$, la cual, de acuerdo con la inecuación de CAUCHY-SCHWARTZ puede expresarse como:

$$NCDIF_i \geq \left[\int_{-\infty}^{\infty} \left| ES_{ijF} - ES_{ijR} \right| f_F(\theta) d\theta \right]^2. \tag{3.26}$$

RAJU en 1988 demostró que si $f_F(\theta)$ es rectangular, entonces el lado derecho de la ecuación (3.26) es el cuadrado del valor absoluto o área sin signo entre las dos funciones de respuesta al ítem. Por lo tanto, la definición del área sin signo del DIF puede también ser vista como un caso especial de $NCDIF_i = \sigma^2_{d_i} + \mu^2_{d_i}$.

El índice *NCDIF,* por lo tanto, parece estar relacionado con muchos de los métodos actuales para determinar DIF dentro del contexto de TRI. Este caso especial, sin embargo, asume que todos los ítems en el test, con excepción del ítem que se está estudiando, no funcionan diferencialmente, y esta asunción no se satisface probablemente en la mayoría de las situaciones del desarrollo del test.

Un concepto importante en el marco del *DFIT* es la direccionalidad del funcionamiento diferencial. La condición de funcionamiento bidireccional a nivel de test se da cuando los ítems que favorecen a un grupo se equilibran con los ítems que favorecen al otro grupo. Un funcionamiento unidireccional a nivel de test se da cuando el funcionamiento diferencial del ítem prevalece a favor del grupo de referencia, sobre el grupo focal. Así, dos ítems con funcionamiento diferencial unidireccional podrían ser considerados "sesgados" de acuerdo a las definiciones de los índices *NCDIF* y *CDIF*. Sin embargo dos ítems con funcionamiento diferencial balanceado (un ítem favorece al grupo de referencia en el mismo grado que el otro al focal) serían considerados "sesgados" solamente con la definición de DIF del índice *NCDIF*.

Dos ítems pueden exhibir niveles similares del funcionamiento diferencial, pero en direcciones opuestas (por ejemplo, si hacemos una comparación entre hombres y mujeres, un ítem puede favorecer a los hombres y otro puede favorecer a las mujeres). En este caso, los dos ítems se cancelarían y no se produciría ningún funcionamiento diferencial neto a nivel de test, inversamente, un número de ítems que son no significativos, pero distintos de cero, niveles de *NCDIF* en la misma dirección pueden producir un test con *DTF* significativo debido a la acumulación del nivel del funcionamiento diferencial a nivel de test.

De acuerdo con el marco del *DFIT*, (RAJU y col., 1995) si en un test se detecta Funcionamiento Diferencial se deben ir eliminado ítems del cuestionario hasta que el índice *DTF* sea no significativo o su valor no supere al punto crítico. Para el criterio de exclusión de los ítems no podemos utilizar el índice *CDIF* ya que para su estimación asumimos que el resto de los ítems del cuestionario también presentan Funcionamiento Diferencial. Localizaremos en primer lugar aquel que presente mayor valor $(2CDIF - NCDIF)$ (MORALES y col., 2006) y se eliminará del estudio, repetiremos

este paso hasta conseguir un índice *DTF* no significativo y que no supere el valor critico propuesto por RAJU (1999).

Significación de los índices DTF y NCDIF

En la práctica, los índices *DTF*, *CDIF* y *NCDIF* se calculan usando las estimaciones del rasgo latente θ, y de los parámetros de los ítems *a, b* y *c* (denotados por $\hat{\theta},\hat{a},\hat{b},\hat{c}$)

Las estimaciones de *DTF*, *CDIF* y *NCDIF* (denotadas por $\hat{DTF}$, $\hat{CDIF}$, $\hat{NCDIF}$) se calculan con D_j, d_{ij} para un individuo *j*. Estas estimaciones son

$$\hat{DFT} = \hat{\sigma}_D^2 + \hat{\mu}_D^2, \tag{3.27}$$

$$\hat{CDIF}_i = \hat{Cov}(d_i, D) + \hat{\mu}_{d_i}\hat{\mu}_D, \tag{3.28}$$

$$\hat{NCDIF}_i = \hat{\sigma}_{d_i}^2 + \hat{\mu}_{d_i}^2, \tag{3.29}$$

con $\hat{\sigma},\hat{\mu},\hat{Cov}$ las estimaciones insesgadas de σ,μ,Cov.

Según las definiciones dadas de $\hat{DFT},\hat{CDIF},\hat{NCDIF}$ podemos destacar dos fuentes distintas de error, un error de muestreo y un error en la estimación de los parámetros de los ítems y del nivel de habilidad.

Si el nivel de habilidad y las estimaciones de los parámetros fuesen conocidos, $\hat{DFT}, \hat{CDIF}, \hat{NCDIF}$ incluirían solamente error aleatorio. En este caso, $D_j = 0$ con probabilidad 1 para todo individuo *j* en el grupo focal cuando la hipótesis nula es verdadera (es decir, *DTF*=0).

Hasta hoy no se han propuesto pruebas de significación que expliquen los errores asociados a la estimación de los parámetros de habilidad y del ítem.

Test χ^2 *para* $\hat{DFT}$: Asumiendo que *D* sigue una distribución normal con media μ_D y desviación estándar finita σ_D para el individuo *j*

$$z_j = \frac{D_j - \mu_j}{\sigma_D} \ , \tag{3.30}$$

z_j^2 sigue una distribución χ^2 con un grado de libertad; la suma de z_j^2 para los N_F individuos en el grupo focal, tiene una distribución χ^2 con N_F grados de libertad, donde N_F es el tamaño de muestra del grupo focal y N_R es el tamaño de muestra del grupo de referencia.

Algebraicamente, esto puede expresarse como:

$$\chi^2_{N_F} = \sum_{j=1}^{N_F} z_j^2 = \frac{\sum_{j=1}^{N_F} (D_j - \mu_D)^2}{\sigma_D^2} \ . \tag{3.31}$$

El interés está en minimizar la aproximación de $D\hat{F}T$ o lo que es lo mismo que $E(D\hat{T}F) = \mu_{D^2} = 0$, lo cual implica que μ_D es cero. Notemos que $\mu_D = 0$ es una condición necesaria pero no suficiente. Sustituyendo en la ecuación (28):

$$\chi^2_{N_F} = \sum_{j=1}^{N_F} z_j^2 = \frac{\sum_{j=1}^{N_F} (D_j)^2}{\sigma_D^2}, \tag{3.32}$$

o lo que es lo mismo

$$\chi^2_{N_F} = \frac{N_F(D\hat{T}F)}{\sigma_D^2} \ , \tag{3.33}$$

donde los grados de libertad para esa chi-cuadrado son probablemente menores que N_F.

Test t para $D\hat{F}T$: El estadístico *t*-test utilizado para $D\hat{F}T$ viene expresado por:

$$t = \frac{(N_F)^2(\hat{\mu}_D - \mu_D)}{\hat{\sigma}_D} \ , \tag{3.34}$$

que bajo la hipótesis nula de que $\mu_D = 0$ puede rescribirse como:

$$t = \frac{(N_F)^2(\hat{\mu}_D)}{\hat{\sigma}_D} \quad . \tag{3.35}$$

Este estadístico, según la distribución de D y la normalidad asintótica de $\hat{\mu}_D$ con varianza igual a $\frac{\sigma_D^2}{N_F}$, sigue una distribución t asintótica con $N_F - 1$ grados de libertad. Como el tamaño de muestra (N_F) es generalmente grande en los análisis de la TRI, los tests t y chi-cuadrado conducen a conclusiones similares.

Cuando los tests para $D\hat{F}T$ son estadísticamente significativos, debemos de realizar la búsqueda de ítems que pueden causar la significación de los tests. Después de identificar y de quitar tales ítems del test, $D\hat{F}T$ deben recalcularse con los restantes ítems.

Como el valor de $C\hat{o}v(d_i, D)$ depende, entre otras cosas, del número de ítems que no se mueven en el test, se recomienda que el proceso se realice ítem a ítem hasta que la chi-cuadrado asociada llegue a ser no significativa.

Test χ^2 y t - test *para* NCD̂IF: Como d_i sigue una distribución normal y tiene varianza finita, podemos definir un test chi-cuadrado y un t- test para NCD̂IF para un ítem i de la siguiente forma:

$$\chi^2_{N_F} = \frac{N_F(N\hat{CDIF})}{\hat{\sigma}^2_{d_i}}, \tag{3.36}$$

con N_F grados de libertad, o

$$t = \frac{(N_F)^{\frac{1}{2}}(\hat{\mu}_{d_i})}{\hat{\sigma}_{d_i}} \quad , \tag{3.37}$$

con $N_F - 1$ grados de libertad

No es necesario dar un test de significación para el índice $CDIF$ ya que el índice DTF es la suma de los $CDIF$ para cada ítem i (Ver ecuación 3.23). Cuando se encuentre un $D\hat{T}F$ estadísticamente significativo, los ítems con un índice $C\hat{D}IF$ alto y positivo

deben de eliminarse del estudio (un ítem cada vez) y volver a calcular el índice $\hat{DTF}$ hasta que este no sea significativo. Todos los ítems eliminados del estudio, pueden ser caracterizados como ítems que tienen un índice $\hat{CDIF}$ significativo. Como la eliminación de los ítems se realiza de forma secuencial, este resultado final se puede aprovechar para saber si el tamaño de muestra utilizado es conveniente para este tipo de análisis.

RAJU y col. (1995) recomendaron fijar un punto crítico para los índices *NCDIF* y *DTF*, además de un test χ^2 significativo para un nivel de significación del 0.01 (1% de los ítems son falsamente identificados con funcionamiento diferencial) porque este test era demasiado sensible a grandes tamaños de muestra.

En un examen exploratorio de Monte Carlo del test χ^2 para los índices *DFT* y NCDIF, FLEER (1993) demostró que estos índices eran excesivamente sensibles a valores grandes de N_F (tamaño de muestra del grupo focal). En una condición no DIF (es decir, parámetros de los ítems idénticos en el grupo focal y de referencia), el porcentaje de ítems identificados como sesgados a un nivel de significación 0.01 fue substancialmente superior al 1%. Por lo tanto, después de varias réplicas bajo condición de no DIF, FLEER (1993) encontró que con un punto de corte para ambos índices, en ítem dicotómicos, de 0.006 se identificaron un 1% de los ítems con DIF falsos. Por lo tanto se dio como criterio que para el índice $\hat{\text{NCDIF}}$, ítems con $\text{NCDIF} > 0.006$ y con χ^2 estadísticamente significativa son identificados como ítems que funcionan diferencialmente. Ítems con *DFT* menor ó igual de 0.006 y χ^2no significativa se mantienen en el test para la posterior eliminación de otros ítems.

Basándose en estudios anteriores de simulación, FLEER (1993) y FLOWERS y col. (1995) establecieron puntos de corte para los índices *NCDIF* y *DTF* (similares a los utilizados en un análisis factorial confirmatorio para evaluar el ajuste del modelo) que varían dependiendo del número de opciones de respuesta de los ítems del test.

Así, RAJU en una comunicación (1999), dio como recomendación del valor crítico para la significación de *NCDIF* para un ítem dependiendo del número de respuestas categóricas los que aparecen en la siguiente tabla (Tabla 3.2).

Nº opciones de respuesta	Valor crítico
2	0.006
3	0.024
4	0.054
5	0.096
6	0.150
7	0.216
8	0.294
9	0.384

Tabla 3.2.- Valores críticos para el índice NCDIF

El valor crítico para la significación *DFT* será el valor crítico para *NCDIF* multiplicado por el número de ítems que tenga el test.

Recientemente se ha desarrollado un nuevo método para determinar los valores de corte para el *NCDIF* y *DTF* en el marco del *DFIT* para ítems dicotómicos. Este nuevo método se denomina "réplica del parámetro del ítem" (IPR), esta basado en una técnica de Monte-Carlo en la cual se simulan una gran cantidad de pares de parámetros de los de un conjunto de parámetros de ítems originales, obteniéndose así una distribución para los índices DIF/DTF bajo la condición de no DIF/DTF (OSHIMA y col., 2006).

3.5. EXPRESIÓN DE LOS PARÁMETROS EN UNA MISMA MÉTRICA

En muchas aplicaciones de la TRI no sólo se requiere un ajuste adecuado del modelo, sino también que las estimaciones de los parámetros de los ítems se expresen en la misma métrica. El problema se presenta cuando las estimaciones de los parámetros de los ítems del test se obtienen de forma separada para el grupo focal y el grupo de referencia. Estas estimaciones son diferentes porque la métrica o la escala definida es diferente para cada calibración (estimación de los parámetros) de los ítems.

En los modelos de la TRI, la probabilidad de que un individuo *s*, con un nivel de habilidad θ_s conteste correctamente a un ítem, $P_i(\theta_s, a_i, b_i, c_i)$, es una función de $a_i(\theta - b_i)$ donde recordemos que a_i es el parámetro de discriminación, b_i el de

dificultad y ,c_i es la probabilidad de que un individuo con un nivel de habilidad inferior a θ_s conteste al ítem *i* correctamente.

Si aplicamos una transformación lineal al nivel de habilidad θ_s obtenemos una habilidad θ^*, si la misma transformación lineal es aplicada a b_i obtenemos b_i^* y finalmente, si a_i lo dividimos por la constante multiplicativa de la transformación lineal obtenemos a_i^*; estas transformaciones no cambiarían la probabilidad de contestar correctamente al ítem *i,* es decir: $P_i(\theta_s^*, a_i^*, b_i^*, c_i) = P_i(\theta_s, a_i, b_i, c_i)$. La transformación para c_i no es necesaria porque c_i está sobre la probabilidad métrica.

Supongamos que hay dos tests compuestos por los ítems que miden un único rasgo latente y que hay algunos ítems comunes a ambos test, si un ítem es calibrado como parte de un test y después calibrado como parte del segundo test, los valores reales de las estimaciones de los parámetros se diferenciaran porque las escalas establecidas por las dos calibraciones son diferentes. Sin embargo, la relación entre estas dos escalas será lineal, puesto que se diferencian solamente en el origen y en la unidad de medida.

Si b_{1i} es la estimación del parámetro de dificultad del ítem de la calibración del ítem *i* en el test uno, y b_{2i} la estimación del parámetro de dificultad de la calibración del mismo ítem *i* pero para el segundo test, b_{i2}^* será b_{i2} después de haberle realizado la transformación lineal del primer test, es decir:

$$b_{i2}^* = Ab_{i2} + B, \tag{3.38}$$

si le aplicamos a la habilidad la misma transformación lineal tendríamos:

$$\hat{\theta}_{a2}^* = \hat{\theta}_{a2} + B, \tag{3.39}$$

y por último la estimación del parámetro de discriminación del ítem *i* es transformada por:

$$a_{i2}^* = a_{i2} / A. \tag{3.40}$$

Si en lugar de tener dos tests tenemos un único test el cual se le aplica a dos grupos de individuos diferentes (grupo focal y grupo de referencia) la situación sería semejante, siendo las ecuaciones las siguientes:

$$b^*_{iR} = Ab_{iR} + B, \tag{3.41}$$

$$\hat{\theta}^*_{aR} = \hat{\theta}_{aR} + B, \tag{3.42}$$

$$a^*_{iR} = a_{iR} / A\ , \tag{3.43}$$

donde b_{iR} es la estimación del parámetro de dificultad del ítem *i* para el grupo de referencia; b^*_{iR} es b_{iR} después de haberle realizado la transformación lineal del grupo focal; $\hat{\theta}_{aR}$ es la estimación de la habilidad del individuo *a* del grupo de referencia para el ítem *i*; $\hat{\theta}^*_{aR}$ es $\hat{\theta}_{aR}$ después de haberle realizado la transformación lineal del grupo focal; a_{iR} es la estimación del parámetro de discriminación del ítem *i* para el grupo de referencia; y por último a^*_{iR} es a_{iR} después de haberle realizado la transformación lineal del grupo focal; b^*_{iR} será b_{iR} después de haberle realizado la transformación lineal del grupo focal; b^*_{iR} será b_{iR} después de haberle realizado la transformación lineal del grupo focal.

Por lo tanto, el problema de la igualación de los parámetros de los ítems para el grupo focal y el de referencia queda reducido a encontrar los coeficientes A y B apropiados para la transformación lineal.

Para encontrar una relación lineal entre dos conjuntos de datos se pueden aplicar las técnicas de regresión lineal simple. En nuestro caso, podríamos considerar como variable independiente la estimación de los parámetros de dificultad del ítem (o las habilidades) y estos parámetros de dificultad obtenidos en la segunda calibración como la variable dependiente. Ahora bien, en una regresión se asume que la variable dependiente está medida sin error y esto en nuestro caso no es correcto ya que son valores estimados. Además, pretendemos un procedimiento simétrico y la regresión no lo es, puesto que no hay razón para favorecer o desfavorecer una estimación del parámetro de dificultad del ítem sobre otra estimación del mismo parámetro de dificultad.

Existen diversos métodos para realizar la igualación de las estimaciones de los parámetros de los ítems. Describimos a continuación alguno de ellos:

MARCO (1977) y COOK y col. (1979) aplicaron en sus investigaciones un método simétrico en el que se utilizan los dos primeros momentos de las distribuciones de los parámetros de dificultad del ítem estimados. En este método se buscan los coeficientes de la transformación lineal A y B, de manera que la media y la desviación típica de la distribución transformada de las estimaciones de los parámetros de dificultad de los ítems son iguales a la media y desviación típica de la primera estimación de los parámetros de dificultad de los ítems.

LINN y col. (1981) procuraron reducir la influencia de outliers usando momentos ponderados donde los pesos son inversamente proporcionales a la estimación del error estándar estimado de las estimaciones de las dificultades del ítem.

DIVIGI (1980) eligió los coeficientes A y B de la transformación lineal para reducir al mínimo la diferencia máxima entre la suma de las funciones de respuesta al ítem para primera calibración y la suma de las funciones de la respuesta del ítem para la segunda calibración.

BEJAR y WINGERSKY (1981) desarrollaron un método más elaborado. Los denominados métodos robustos, que dan pequeñas ponderaciones a los puntos periféricos que se utilizan para estimar los momentos; además trataron los outliers de la misma manera, sin importar su error de estándar.

IRONSON (1982) trata los dos conjuntos de parámetros de dificultad estimados del ítem simétricamente utilizando la primera componente principal como la línea que da la transformación de un conjunto de parámetros de dificultad estimados a otro conjunto. En este procedimiento, si todos los parámetros de dificultad del ítem estimados se dividen por una constante, éstos no cambian.

KOLEN y BRENNAN (1995) sintetizaron cuatro métodos de igualación de parámetros: a) Media/Media (LOYD y HOOVER, 1980); b) Media/Sigma (MARCO, 1977); c) STOCKING-LORD (1983); d) HAEBARA (1980). Los métodos media/media y media/sigma usan los momentos de los parámetros de los ítems estimados para producir la transformación lineal; por lo tanto, estos métodos serán referidos como *"métodos de los momentos"*. La diferencia entre los métodos Media/Media y Media/Sigma es que los métodos de Media/Media usan la media de los parámetros de a para computar la pendiente de la escala de transformación lineal, mientras que los métodos Media/Sigma utilizan la desviación típica de los parámetros b para computar la pendiente de la escala de transformación. Los métodos de Stocking-Lord y Haebara hacen referencia a los *"métodos de las curvas características"* porque estos métodos producen una escala de transformación para minimizar la diferencia entre las curvas características del test y las curvas características del ítem, respectivamente.

3.5.1.-MÉTODOS DE LAS CURVAS CARACTERÍSTICAS

Los métodos de transformación de las curvas características consideran la información de todos los parámetros de los ítems simultáneamente para minimizar la diferencia entre las curvas características de los ítems.

Los dos métodos más conocidos son el método de HAEBARA (1980) y el método de STOCKING y LORD (1983). En el método de Haebara, para cada individuo y para cada ítem de test, se considera la diferencia entre las dos curvas características para un rasgo latente θ, estas diferencias se elevan al cuadrado y se suman para todos los ítems comunes del individuo. Finalmente, se repite el proceso para cada uno de los individuos de la muestra y se suman. Se eligen A y B para que esta suma sea mínima.

En el método de STOCKING y LORD (1983) para cada individuo se calcula la diferencia elevada al cuadro de las curvas características de todos los ítems se suman esas diferencias para todos los individuos y se busca el mínimo de esta suma. Vamos a desarrollar dicho método por ser el más utilizado en las investigaciones.

Un individuo *j*, con habilidad θ_j tiene como puntuación verdadera ξ_j definida por

$$\xi_j \equiv \xi(\theta_j) \equiv \sum_{i=1}^{n} P_i(\theta_j; a_i, b_i, c_i), \tag{3.44}$$

donde *n* es el número de ítems en el test. La correcta transformación lineal de escala para dos calibraciones diferentes del mismo test produciría las mismas puntuaciones verdaderas para el individuo *j* si los parámetros a_i, b_i, c_i fuesen conocidos. Si ξ_j^*es la estimación de la puntuación verdadera obtenida de la segunda calibración del test después de que se haya transformado a la escala de la primera, entonces

$$\hat{\xi}_j^* \equiv \hat{\xi}^*(\theta_j) \equiv \sum_{i=1}^{n} P_i^*(\theta), \tag{3.45}$$

donde $P_i^*(\theta_j) \equiv P_i(\theta_j^*; a_i^*, b_i^*, c_i)$. Para un individuo la diferencia $(\hat{\xi}_j - \hat{\xi}_j^*)$ debe ser mínima. En la práctica, se desearía elegir los coeficientes A y B de la transformación lineal tales que para un grupo conveniente de individuos, la media de la diferencia ajustada entre las estimaciones de las puntuaciones verdaderas sea lo más pequeña posible. La función que debe ser minimizada es:

$$F = \frac{1}{N} \sum_{j=1}^{N} (\hat{\xi}_j - \hat{\xi}_j^*)^2, \tag{3.46}$$

donde *N* es el número de individuos en un grupo arbitrario.

Esta función F, considerada como una función de A y B, será mínima cuando

$$\frac{\partial F}{\partial A} = \frac{-2}{N} \sum_{j=1}^{N} (\hat{\xi}_j - \hat{\xi}_j^*) \frac{\partial \hat{\xi}_j^*}{\partial A} = 0,$$

$$\frac{\partial F}{\partial B} = \frac{-2}{N} \sum_{j=1}^{N} (\hat{\xi}_j - \hat{\xi}_j^*) \frac{\partial \hat{\xi}_j^*}{\partial B} = 0, \tag{3.47}$$

Ahora, usando la regla de cadena de la diferenciación en la ecuación (3.44),

$$\frac{\partial \hat{\xi}_s^*}{\partial A} = \sum_{i=1}^{n} \left(\frac{\partial P_i^*(\theta_s)}{\partial b_{i2}^*} \frac{\partial b_{i2}^*}{\partial A} + \frac{\partial P_i^*(\theta_s)}{\partial a_{i2}^*} \frac{\partial a_{i2}^*}{\partial A} \right), \tag{3.48}$$

$$\frac{\partial \hat{\xi}_s^*}{\partial B} = \sum_{i=1}^{n} \left(\frac{\partial P_i^*(\theta_s)}{\partial b_{i2}^*} \frac{\partial b_{i2}^*}{\partial B} \right), \tag{3.49}$$

Diferenciando las ecuaciones (3.48) y (3.49) tenemos: $\frac{\delta b_{i2}^*}{\delta A} = b_{i2}$ y $\frac{\delta a_{i2}^*}{\delta A} = \frac{-a_{i2}}{A^2}$ y sustituyendo en la ecuación (3.44) y (3.45)

$$\frac{\partial \hat{\xi}_s^*}{\partial A} = \sum_{i=1}^{n} \left(b_{i2} \frac{\partial P_i^*(\theta_s)}{\partial b_{i2}^*} - \frac{a_{i2}}{A^2} \frac{\partial P_i^*(\theta_s)}{\partial a_{i2}^*} \right), \tag{3.50}$$

$$\frac{\partial \hat{\xi}_s^*}{\partial B} = \sum_{i=1}^{n} \left(\frac{\partial P_i^*(\theta_s)}{\partial b_{i2}^*} \frac{\partial b_{i2}^*}{\partial B} \right), \tag{3.51}$$

como $\frac{\delta b_{i2}^*}{\delta B} = 1$, sustituyendo en (3.51)

$$\frac{\partial \hat{\xi}_s^*}{\partial B} = \sum_{i=1}^{n} \left(\frac{\partial P_i^*(\theta_s)}{\partial b_{i2}^*} \right). \tag{3.52}$$

La forma funcional de las derivadas parciales de la función de la respuesta del ítem dependen del modelo matemático elegido.

Estas derivadas se sustituyen en las ecuaciones (3.47) y tenemos el mínimo de F en la ecuación (3.46).

3.5.2.- MÉTODOS DE LOS MOMENTOS

Los métodos de los momentos, hacen uso de la media y de la desviación típica de los parámetros comunes a un ítem para estimar las constantes A y B de la transformación lineal.

El método "media y sigma" utiliza un procedimiento robusto para dar pequeñas ponderaciones a los valores donde la distancia perpendicular de la recta inicial es grande, y se estima una nueva recta. Estas ponderaciones robustas se repiten hasta que los cambios en las distancias perpendiculares llegan a ser pequeños. Una desventaja de estos procedimientos de transformación "media y sigma" es que solamente se aplica a

los parámetros de dificultad de los ítems, es decir, para el calculo de las estimaciones de los coeficientes A y B sólo se utiliza el b_i, y la transformación lineal calculada se aplica a $\hat{\theta}_s$ y a_i.

Las dos constantes de comparación se estiman a partir de los dos primeros momentos de las distribuciones de los pesos estimados de los parámetros de dificultad del ítem. El peso para el ítem *i*, es el inverso de la varianza estimada más grande del parámetro dificultad para el grupo de referencia y del parámetro de dificultad para el grupo focal. De esta manera, a los ítems para los cuales el parámetro de dificultad está mal estimado (para cualquiera de los grupos, focal o de referencia) se dan pesos más pequeños en la determinación de las constantes de comparación.

Si $b_{iF}^{w\ *}$ es el parámetro de dificultad con su peso del ítem *i* en el grupo focal después de la igualación y b_{iF}^{w} es el correspondiente valor a priori de la igualación, entonces:

$$b_{iF}^{w\ *} = Ab_{iF}^{w} + B$$

Donde A y B son seleccionados de modo que la media y la desviación típica de las cargas del parámetro de dificultad del ítem en el grupo focal son las mismas que la media y la desviación típica de las cargas del parámetro de dificultad del ítem en el grupo de referencia.

Para esta transformación, $A = \sigma_{b_R^w} / \sigma_{b_F^w}$ y, $B = \mu_{b_R^w} - A\mu_{b_F^w}$ donde $\mu_{b_R^w}$ es la media y $\sigma_{b_R^w}$ es la desviación típica de los pesos de los parámetros de dificultad del grupo de referencia y $\mu_{b_F^w}$ y $\sigma_{b_F^w}$ son los correspondientes valores para el grupo focal.

El método "Media/Media" (LOYD y HOOVER, 1980) incorpora información del parámetro de discriminación. Estima el intercepto B igual que el método Media/Sigma, pero la pendiente A la estima como el cociente de las medias de los parámetros de dificultad.

Varios son los programas que tienen implementados los métodos desarrollados anteriormente. El programa EQUATE v.2.0 desarrollado por BAKER en 1993

implementa el procedimiento de las curvas características del test de puntuaciones dicotómicas o de puntuaciones de respuesta graduadas o nominales. Para ítems dicotómicos o de respuesta graduada, el cálculo de los coeficientes de la transformación lineal (A y B) se basa en igualar las curvas características de los tests.

Para el caso de respuesta nominal se igualan funciones de respuesta categórica. En la versión 2 del EQUATE para el modelo de respuesta graduada, la ponderación de los ítems puede asignarse de mayor a menor o de menor a mayor, y para el modelo nominal el usuario especifica como igualar los ítems en los dos tests (o en los dos grupos de individuos). En ambos casos se pueden utilizar las estimaciones de los parámetros obtenidas del MULTILOG (THISSEN, 1991).

EQUATE 2.0 está escrito en FORTRAN para MS-Dos, y se ejecuta de modo iterativo con una serie de preguntas que se le hacen al usuario. El sistema de preguntas también se pueden almacenar en un archivo y ejecutar con una versión de MS-Dos superior a la 3.3.

KIM y KOLEN en 2003 desarrollaron el programa informático POLYST que tiene implementados los cuatro métodos de transformación de escalas de los parámetros de los ítems (y si es necesario la habilidad) de un grupo de individuos (grupo de referencia) sobre una escala determinada por un grupo base (grupo focal) de individuos. Los cuatro métodos de transformación de escalas fueron primeramente desarrollados para modelos de la TRI dicotómicos y posteriormente se han extendido a modelos de la TRI politómicos.

POLYST tiene la capacidad de realizar las transformaciones para cinco modelos de la TRI: Modelo logístico de tres parámetros (3PL); Modelo de respuesta graduada (GR); Modelo de crédito parcial generalizado (GPC); Modelo de respuesta nominal (NR) y modelo de múltiples opciones (MC). Para los tres primeros modelos, tiene implementados los cuatro métodos de transformación. Para los modelos NR y MC, tiene implementados los tres métodos de transformación de escalas a excepción del método de STOCKING-LORD.

En este trabajo vamos a trabajar con el programa POLYST ya que es más actual y tiene implementados los cuatro métodos de transformación de escalas más conocidos.

3.6.- REGRESIÓN LOGÍSTICA PARA DETECTAR EL DIF

La técnica de la regresión logística se originó en la década de los 60 con el trabajo de CORNFIELD y col. en 1961, WALTER y DUNCAN en 1967 la utilizan ya en la forma que la conocemos actualmente, o sea para estimar la probabilidad de ocurrencia de un proceso en función de ciertas variables. Su uso se incrementa desde principios de los 80 como consecuencia de los adelantos en el campo de la computación.

La regresión logística, al igual que otras técnicas estadísticas, da la posibilidad de evaluar la influencia de cada una de las variables independientes sobre la variable dependiente o de respuesta y controlar el efecto del resto. Tendremos, por tanto, una variable dependiente, Y (dicotómica o politómica) y una o más variables independientes, X (cualitativas o cuantitativas).

La ecuación general para un modelo de regresión logística vendría dada por (HOSMER y LEMESHOW, 1989)

$$P(Y = 1/X) = \frac{e^z}{1 + e^z}, \tag{3.53}$$

donde Y es la variable respuesta, P es la probabilidad de obtener una respuesta correcta (probabilidad de éxito) condicionada a X, X es el vector de variables predictoras y $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$, siendo p el número de variables predictoras.

SWAMINATHAN y ROGERS (1990) y ROGERS y SWAMINATHAN (1993) propusieron el análisis de regresión logística para la detección del DIF uniforme y no-uniforme en ítems dicotómicos. En los últimos años el análisis de regresión logística se ha configurado como una técnica eficaz en la detección del funcionamiento diferencial del ítem. El análisis de regresión logística proporciona un marco común para analizar el funcionamiento diferencial de ítems en diferentes formatos (dicotómicos, politómicos) y para diferentes tipos de DIF (uniforme, no-uniforme simétrico y no-uniforme asimétrico) (ZUMBO, 1999). Estas dos características hacen especialmente

recomendable su uso en situaciones aplicadas.

El análisis de regresión logística es uno de los métodos de comprobada eficacia para su uso en la detección tanto del DIF uniforme como no-uniforme (CLAUSER y MAZOR, 1998).

Se ha demostrado que este procedimiento produce resultados similares a los obtenidos con el estadístico Mantel- Haenszel cuando se trata de detectar DIF uniforme y una mayor potencia estadística para detectar ítems con DIF no-uniforme (CLAUSER y col., 1996; FERRERES y col., 2000; HIDALGO y GÓMEZ, 2000; HIDALGO y LÓPEZ, 2004; NARAYANAN y SWAMINATHAN, 1996; ROGERS y SWAMINATHAN, 1993).

Este modelo tiene ventajas sobre el análisis discriminante al no requerir supuestos como el de normalidad o de homocedasticidad, que en muchos casos son difíciles de cumplir. Por otro lado, la regresión logística tiene semejanzas con la regresión múltiple: cuenta con contrastes estadísticos, puede incorporar efectos no lineales y permite realizar diversos diagnósticos.

El modelo de regresión logística dicotómica se basa en la probabilidad de obtener una respuesta correcta al ítem que se considera función de dos variables: la pertenencia al grupo (referencia o focal) y el nivel de habilidad de los sujetos (puntuación empírica u observada en el test o bien el nivel de habilidad estimado bajo algún modelo de respuesta al ítem). La puntuación total observada en el test se utiliza para igualar a los sujetos respecto de la habilidad medida por el test y, a diferencia de otros procedimientos como los modelos log lineales o el estadístico Mantel-Haenszel, el análisis de regresión logística trata dicha puntuación total de forma continua. El modelo estadístico de regresión logística para la detección del DIF considera la respuesta al ítem como la variable dependiente, mientras que las variables explicativas son la puntuación observada del sujeto en el test, la pertenencia al grupo y la interacción entre puntuación observada y pertenencia a grupo. El efecto de las variables explicativas sobre la variable dependiente puede evaluarse utilizando distintas pruebas de significación y estrategias analíticas (CLAUSER y MAZOR, 1998; GÓMEZ e HIDALGO, 1997a; HOSMER y LEMESHOW, 1989; ROGERS y SWAMINATHAN, 1993).

La expresión del modelo de regresión logística en el análisis del DIF es la siguiente:

$$P(Y = 1/X,G) = \frac{e^z}{1+e^z}, \tag{3.54}$$

donde

$$z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG, \tag{3.55}$$

siendo:

X la puntuación observada de un sujeto en un test, *G* el grupo de pertenencia de los sujetos (grupo focal o grupo de referencia), β_0 el intercepto, el efecto total de la dificultad del ítem, β_1 el coeficiente para la habilidad o puntuación total observada en el test, representa la diferencia en la habilidad, β_2 el coeficiente para la variable grupo de pertenencia (referencia o focal), representa las diferencias entre los grupos en la ejecución del ítem y β_3 la interacción entre la puntuación observada en el test y el grupo.

Según el modelo, un ítem no muestra DIF si ni el efecto del grupo ni la interacción habilidad por grupo resultan estadísticamente significativos ($\beta_2 = \beta_3 = 0$); un ítem muestra DIF uniforme si el efecto del grupo (*G*) resulta estadísticamente significativo, mientras que la interacción habilidad por grupo (*XG*) no ejerce ningún efecto sobre el ítem (si $\beta_2 \neq 0$ y $\beta_3 = 0$). Por el contrario, si la interacción *XG* resulta estadísticamente significativa, el ítem presentaría DIF no-uniforme ($\beta_3 \neq 0$ con independencia del valor que adopte β_2).

Para la detección de DIF se ajustan tres modelos en distintas etapas. En esta estrategia de análisis la evaluación del DIF se efectúa comprobando la significación del efecto de las sucesivas variables que se van introduciendo en el modelo. En la primera etapa, se ajusta el modelo base de ausencia de DIF (Modelo 1), donde se introduce en la ecuación la puntuación total del sujeto en el test (*X*). En la segunda etapa, se añade a la ecuación la variable de agrupamiento (*G*), ajustándose el modelo de DIF uniforme (Modelo 2). Por ultimo, en la tercera etapa se introduce en la ecuación la interacción entre el grupo y la puntuación total en el test, valorándose el ajuste del modelo de DIF no-uniforme o modelo completo (Modelo 3).

En un primer paso, se compara el valor de la razón de verosimilitud del Modelo 1 (que expresa ausencia de DIF ya que la respuesta al ítem solo depende del nivel de la habilidad del sujeto) con el del Modelo 2 (que introduce además el efecto de grupo), se obtiene una prueba para el DIF uniforme; el estadístico de comparación representa el cambio en el ajuste desde una ecuación a la otra, sigue una distribución χ^2 con 1 grado de libertad, y se suele denominar G^2 de diferencia[1] (BISHOP y col., 1975).

Si se compara el valor de verosimilitud[2] del Modelo 2 con el del Modelo 3, se puede probar la existencia de DIF no uniforme; este estadístico de diferencia sigue también una distribución χ^2 con 1 grado de libertad. En este proceso de comparación de modelos lo que se evalúa es la mejora explicativa al introducir un nuevo termino al modelo.

Mediante el análisis de regresión logística también se puede realizar una prueba simultánea de la presencia de DIF uniforme y no-uniforme. Esta hipótesis conjunta se puede someter a comprobación comparando el valor de verosimilitud del modelo sin DIF (Modelo 1) con el del modelo completo (Modelo 3); en este caso el estadístico G^2 de diferencia sigue una distribución χ^2 con 2 grados de libertad.

Si ajustamos el modelo completo (Modelo 3) y comprobamos la significación de los coeficientes del modelo asociados a cada termino, usando el estadístico de Wald[3] que, siempre que la variable de agrupamiento tenga dos niveles, se distribuye según una distribución $\chi 2$ con 1 grado de libertad (solo bajo la hipótesis nula y en muestras grandes) tenemos que:

Si únicamente el coeficiente de la variable grupo (β_2) es significativo, el ítem muestra DIF uniforme;

[1] G^2 = D (modelo con la variable) – D (modelo sin la variable)= -2 log (verosimilitud modelo con la variable / verosimilitud modelo sin la variable)

[2] El valor de la verosimilitud es -2 veces el logaritmo de la verosimilitud, y se suele representar por -2LL o por D (desvianza).

Si solo es significativo el coeficiente de la interacción (β_3), el DIF detectado es no-uniforme simétrico;

Finalmente, si ambos coeficientes son significativos, el ítem esta afectado por DIF no-uniforme asimétrico.

La extensión del modelo de regresión logística dicotómica para variables de respuesta con más de dos categorías (datos politómicos) para la detección del DIF uniforme y no-uniforme fue desarrollada por MILLER y SPRAY (1993), FRENCH y MILLER (1996), y ZUMBO (1999).

Uno de esos métodos consiste en recodificar en *m-1* variables dicotómicas, siendo *m* el número de categorías del ítem, siguiendo uno de los siguientes esquemas AGRESTI, 1984, 1990; FRENCH y MILLER, 1996): categorías adyacentes, categorías continuas y categorías acumulativas. Para cada una de las *m-1* variables codificadas tendremos un modelo de regresión logística dicotómica, debiendo ajustarse *m-1* modelos.

Para categorías adyacentes, la dicotimización se crea usando $Y = k$ vs. $Y = k + 1$, para $k = 0,...,m-1$. Para categorías continuas, $Y = k$ vs. $Y > k$, para $k = 0,...,m-1$ y para categorías acumulativas $Y \leq k$ vs. $Y > k$, para $k = 0,...,m-1$. Un ejemplo de las tres codificaciones para un ítem politómico con cuatro categorías se presenta en la siguiente tabla (PENFIE y CAMILLI, 2007).

	Variables DUMMY	C. codificadas como Missing	C. codificadas como 0	C. codificadas como 1
C. Continuas	1		0	1,2,3
	2	0	1	2,3
	3	0,1	2	3
C.Acumulativas	1		0	1,2,3
	2		0,1	2,3
	3		0,1,2	3
C. Adyacentes	1	2,3	0	1
	2	0,3	1	2
	3	0,1	2	3

Tabla 3.3.- Posible codificaciones del un ítem con cuatro categorías

Si asumimos que la variable de respuesta está medida en una escala ordinal, es posible ajustar un modelo de regresión logística politómica asumiendo pendientes paralelas (bajo el esquema de categorías acumulativas). En este modelo, sólo intercepto es diferente para las *m-1* funciones, y se asume que los efectos de las variables predictoras son constantes a través de las comparaciones.

A continuación vamos a desarrollar la regresión logística para ítems de respuesta ordinal (AGRESTI, 1990, 1996).

Sea Y la puntuación del ítem *i-ésimo*, sin perdida de generalidad, se asume que hay un total de *m* categorías y que las puntuaciones de los ítems son *0, 1, 2,..., m-1*, en lugar de *1, 2,...,m.*

Denotemos por $\pi_0,\pi_1,...,\pi_{m-1}$ las probabilidades de respuesta categórica con $\sum_m \pi_m = 1$. La variable aleatoria Y sigue una distribución multinomial.

Vamos a expresar las probabilidades *k-ésima* acumulada o superior como (AGRESTI, 1996, p.211)

$$P(Y \geq k) = \pi_k + ... + \pi_{m-1} \ , \qquad (3.56)$$

donde $P(Y \geq 0) = 1$ y $P(Y \geq m) = 0$. Las probabilidades para las categorías k = 0,1, , m-1 son: $P(Y = k) = P(Y \geq k) - P(Y \geq k+1)$.

Además:

$$\log it\left[P(Y \geq k)\right] = \log\left[\frac{P(Y \geq k)}{1 - P(Y \geq k)}\right] = \log\left[\frac{\pi_k + ... + \pi_{m-1}}{\pi_0 + ... + \pi_{k-1}}\right], \quad (3.57)$$

Por ejemplo, para el ítem "*¿Puede levantarse de la silla?* " de la escala Movilidad del cuestionario QUALEFFO con cinco categorías (Sin dificultad, Con un poco de dificultad, Con moderada dificultad, Con mucha dificultad, No puedo en absoluto) de respuesta, los modelos serían:

$$\log it\left[P(Y \geq 1)\right] = \log\left[\frac{P(Y \geq 1)}{1 - P(Y \geq 1)}\right] = \log\left[\frac{\pi_1 + \pi_2 + \pi_3 + \pi_4}{\pi_0}\right], \quad (3.58)$$

$$\log it\left[P(Y \geq 2)\right] = \log\left[\frac{P(Y \geq 2)}{1 - P(Y \geq 2)}\right] = \log\left[\frac{\pi_2 + \pi_3 + \pi_4}{\pi_0 + \pi_1}\right], \quad (3.59)$$

$$\log it\left[P(Y \geq 3)\right] = \log\left[\frac{P(Y \geq 3)}{1 - P(Y \geq 3)}\right] = \log\left[\frac{\pi_3 + \pi_4}{\pi_0 + \pi_1 + \pi_2}\right], \quad (3.60)$$

$$\log it\left[P(Y \geq 4)\right] = \log\left[\frac{P(Y \geq 4)}{1 - P(Y \geq 4)}\right] = \log\left[\frac{\pi_4}{\pi_0 + \pi_1 + \pi_2 + \pi_3}\right]. \quad (3.61)$$

Para la detección del DIF se pueden ajustar cuatro modelos usando las puntuaciones observadas de los sujetos (X) y el grupo de pertenencia (G) (SWAMINATHAN y ROGERS, 1990; ZUMBO, 1999)

$$\text{Modelo 0: } \log it\left[P(Y \geq k)\right] = \alpha_k, \quad (3.62)$$

$$\text{Modelo 1: } \log it\left[P(Y \geq k)\right] = \alpha_k + \beta_1(X), \quad (3.63)$$

$$\text{Modelo 2: } \log it\left[P(Y \geq k)\right] = \alpha_k + \beta_1(X) + \beta_2(G), \quad (3.64)$$

$$\text{Modelo 3: } \log it\left[P(Y \geq k)\right] = \alpha_k + \beta_1(X) + \beta_2(G) + \beta_3(XxG), \quad (3.65)$$

siendo X la puntuación observada de un sujeto en un test, G el grupo de pertenencia de los sujetos (grupo focal o grupo de referencia), α_k el intercepto, β_1 el coeficiente para la habilidad o puntuación total observada en el test, representa la diferencia en la habilidad, β_2 el coeficiente para la variable grupo de pertenencia (referencia o focal), representa las diferencias entre los grupos en la ejecución del ítem y β_3 la interacción entre la puntuación observada en el test y el grupo.

La probabilidad puede obtenerse con la estimación $\hat{\pi}_k = \hat{P}(Y = k)$ que se obtiene de sustituir los parámetros del modelo por sus estimaciones de máxima verosimilitud.

La estrategia de análisis para la evaluación del DIF se efectúa igual para la Regresión Logística Dicotómica. Si el efecto del grupo (G) resulta estadísticamente significativo y el efecto de la interacción habilidad x grupo (XG) no, entonces el ítem presentaría DIF uniforme. Por el contrario si la interacción XG resulta estadísticamente significativa el ítem presentaría DIF no-uniforme. Estas hipótesis pueden someterse a comprobación mediante el estadístico de Wald[3] (HOSMER y LEMESHOW, 1989)

DETECCIÓN DEL DIF CON REGRESIÓN LOGÍSTICA PARA DATOS POLITÓMICOS

Etapa 1	Comparación del Modelo 1 y Modelo 2 Comparación del Modelo 2 y Modelo 3	DIF- Uniforme DIF No-Uniforme
Etapa **2**	Comparación del Modelo 1 y Modelo 3	DIF Uniforme y No-Uniforme
Etapa **3**	Coeficiente β_2 significativo en el Modelo 3 Coeficiente β_3 significativo en el Modelo 3 Coeficientes β_2 y β_3 significativos en el Modelo 3	DIF- Uniforme DIF No-Uniforme Simétrico DIF No-Uniforme Asimétrico

Tabla 34.- Resumen de la detección del DIF con Regresión Logística

3.6.1.- TAMAÑO DEL EFECTO DEL DIF

Existen varios índices para valorar la magnitud del DIF. Cuando el tamaño de muestra es grande, las medidas descriptivas del DIF pueden ser alternativas viables junto con los métodos de detección del DIF.

ZUMBO (1999) sugiere el uso del coeficiente R^2 para valorar la magnitud del DIF en regresión logística. Se definen dos medidas de R^2:

[3] El estadístico de Wald es igual al cuadrado del cociente entre el coeficiente y su error estándar.

1) La diferencia entre los dos coeficientes de determinación generalizados (COX y SNELL, 1989), R_d^2.

   El coeficiente R^2 que compara los diferentes modelos, Modelo 1, Modelo 2 y Modelo 3 serían:

$$\begin{aligned} R_d^2 &= R^2(Modelo2) - R^2(Modelo1) \\ R_d^2 &= R^2(Modelo3) - R^2(Modelo2), \\ R_d^2 &= R^2(Modelo3) - R^2(Modelo1) \end{aligned} \tag{3.66}$$

   donde:

$$\begin{aligned} R^2(Modelo1) &= 1 - \left[\frac{L(Modelo0)}{L(Modelo1)}\right]^{2/N} \\ R^2(Modelo2) &= 1 - \left[\frac{L(Modelo0)}{L(Modelo2)}\right]^{2/N}, \\ R^2(Modelo3) &= 1 - \left[\frac{L(Modelo0)}{L(Modelo3)}\right]^{2/N}, \end{aligned} \tag{3.67}$$

con L(.) la función de verosimilitud y *n* el tamaño de muestra.

2) La diferencia entre los dos coeficientes de determinación generalizados que son reescalados por sus valores máximos (NAGELKERTE, 1991), $MaxR_d^2$

$$\begin{aligned} MaxR_d^2 &= MaxR^2(Modelo2) - MaxR^2(Modelo1) \\ MaxR_d^2 &= MaxR^2(Modelo3) - MaxR^2(Modelo2), \\ MaxR_d^2 &= MaxR^2(Modelo3) - MaxR^2(Modelo1) \end{aligned} \tag{3.68}$$

   donde:

$$\begin{aligned} MaxR^2(Modelo1) &= \frac{R^2(Modelo1)}{1 - \left[L(Modelo0)\right]^{2/N}} \\ MaxR^2(Modelo2) &= \frac{R^2(Modelo2)}{1 - \left[L(Modelo0)\right]^{2/N}}, \\ MaxR^2(Modelo3) &= \frac{R^2(Modelo3)}{1 - \left[L(Modelo0)\right]^{2/N}}, \end{aligned} \tag{3.69}$$

Para la clasificación de COHEN (1992) del tamaño del efecto en pequeños, moderados y grandes, ZUMBO y THOMAS (1996) propusieron el siguiente criterio de clasificación del incremento del R^2:

- DIF pequeño (no significativo) si $\Delta R^2 < 0.13$.
- DIF moderado si $0.13 \leq \Delta R^2 \leq 0.26$.
- DIF grande si $\Delta R^2 > 0.26$.

Para las categorías de DIF moderado y DIF grande, es necesario que el ítem sea estadísticamente significativo para el test Chi-cuadrado con dos grados de libertad.

JODOIN y GIERL (2001) propusieron otro criterio de clasificación basado en la medida del tamaño del efecto para el procedimiento SIB (ROUSSOS y STOUT, 1996) como predictor de ΔR^2. Los valores para la clasificación de la magnitud del DIF fueron:

- DIF pequeño (no significativo) si $\Delta R^2 < 0.035$
- DIF moderado si $0.035 \leq \Delta R^2 \leq 0.070$, y la hipótesis nula es rechazada.
- DIF grande si $\Delta R^2 > 0.070$ y la hipótesis nula es rechazada.

En el contexto del DIF, el error Tipo I es una identificación de un ítem con DIF cuando realmente no lo presenta. Los errores Tipo I son problemáticos por dos razones: primero, la identificación incorrecta de los ítems con DIF puede conducir a una eliminación incorrecta del ítem del cuestionario que se esté analizando y en segundo lugar puede interferir en la naturaleza del constructo que se está analizando.

LI y STOUT en 1996, aseguraron que el error Tipo I puede dar lugar a un tercer problema, que la significación de las comparaciones entre los diferentes métodos para la detección del DIF sea errónea.

Una explicación de la inflación del error Tipo I asociado a la Regresión Logística es que el test chi-cuadrado es sensible a los tamaños de muestra grandes. Así, para tamaños de muestra grandes un ítem puede presentar un DIF de pequeña magnitud cuando realmente debe ser nulo.

JODOIN y GIERL en 1999 investigaron sobre la capacidad del incremento del R^2 para reducir el error Tipo I en los procedimientos de Regresión Logística para la detección del DIF como una alternativa de los ajustes alpha sugeridos en NARAYANAN y SWAMINATHAN (1996).