Tema 3: Funcionamiento Diferencial de los Ítems (DIF).

Licenciatura de Psicología:
Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica.
José Antonio Pérez Gil
Dpto. de Psicología Experimental.
Universidad de Sevilla.

Agradecimientos: a Inmaculada Sivianes Monge, Mª Angeles Guerrero Orzáez, Francisco Antonete Oria y Jonas Schrod.

Tema 3

Funcionamiento diferencial de los ítems (DIF),

3.1. INTRODUCCIÓN.

En el proceso de medición de características o habilidades psicológicas están implicados varios elementos. En primer lugar tenemos la variable que queremos medir (por ejemplo, habilidad verbal), en segundo lugar tenemos un instrumento diseñado para medirla (test verbal) compuesto por un número variable de elementos (ítems del test) a los que los examinados, tercer elemento, responden. Estos tests, si se cumplen los requisitos psicométricos de fiabilidad y validez, proporcionarán medidas con poco error de la variable que pretenden medir. Un test válido, situándonos en un nivel más general, dará lugar a medidas idénticas para sujetos o grupos con iguales valores en la variable medida. Por el contrario, si la puntuación obtenida es función no sólo del nivel que los sujetos tienen en la variable medida, sino también de otras características irrelevantes como su pertenencia a diferentes grupos étnicos, culturales, etc., entonces hablamos de funcionamiento diferencial del test (FDT). A nivel molecular, se utiliza la expresión de funcionamiento diferencial de los ítems (DIF) para señalar aquellos ítems cuya probabilidad de acertarlos difiere, a igual nivel en la variable medida, entre distintos subgrupos de una población dada.

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 psicológico (psicométrico) es una clara amenaza a la validez del propio test (Ackerman, 1992); tanto, que la validación del mismo no sólo supone aportar evidencias de validez de constructo (interna y externa), sino también de generalizabilidad de las puntuaciones a través del tiempo y de los grupos (ausencia de DIF). Por todo ello, no es de extrañar que, en la actualidad, el análisis del DIF se considere como un paso más en el proceso de desarrollo y construcción de un test, tal como lo recomiendan los Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), en los que se menciona la importancia de investigar el comportamiento de los sujetos en el test, en función de características tales como la raza, el género y/o el bagaje étnico/cultural, y asegurar que las diferencias encontradas están relacionadas con la habilidad que mide el test y no se deben a otros factores irrelevantes.

Una gran fuente de apoyo a los estudios de DIF es el considerable aumento de los trabajos de investigación transculturales, donde el objetivo central es comparar diferentes culturas con respecto a su comportamiento en distintos constructos psicológicos. El principal problema, en este reciente campo de interés, es asegurar la equivalencia de las medidas (entre otras cosas ausencia de DIF) utilizadas en una y otra cultura, siendo la adaptación y traducción de test de una lengua o cultura a otra una práctica bastante habitual.

3.2. CONCEPTOS BASICOS.

3.2.1. El problema del sesgo

El inicio de la polémica se puede trazar sobre los años cincuenta, a partir de los estudios en la Universidad de Chicago de Eells y colaboradores (1951), siendo en estos primeros trabajos donde se pusieron de manifiesto las disparidades de funcionamiento en algunos tests de inteligencia en función del grupo donde se aplicaban, por lo que se produjo un gran interés en los investigadores por explicar las raíces de estas disparidades. En un primer momento, se adujo que la explicación de este fenómeno se debía a que los grupos podían diferir en función de características tales como nivel cultural, clase social,

raza u otras, que llevaba a un comportamiento diferente ante las tareas reflejadas en el test. Durante estos años se hablaba en principio del sesgo del test y de los ítems (Jensen, 1980) y se pone de manifiesto el posible sesgo cultural de los tests. Estos trabajos tuvieron un importante impacto social y político en USA, de este modo los tests en los que se constataban diferencias en función de características étnicas y culturales o socioeconómicas, se consideraban sesgados e injustos. Los términos sesgo e injusticia, el primero más técnico y el segundo referido a cuestiones éticas y sociales, se equiparaban. En USA el problema fue aún más allá, teniendo en cuenta que estos tests se utilizaban tanto como pruebas de selección para un puesto de trabajo como de admisión en colegios y otras instituciones educativas, discriminando injustamente a sujetos de clases sociales desfavorecidas o pertenecientes a minorías étnicas. Y las protestas sociales no tardaron en llegar, los movimientos por los derechos civiles revindicaron igualdad de derechos para los grupos que eran tratados injustamente por los tests, y reclamaban tests libres de cultura.

Hasta los años setenta la investigación sobre el sesgo era realizada por sociólogos, antropólogos y educadores, y no se disponía de criterios claros para reconocer cuando el comportamiento diferencial en el test era debido a diferencias reales en el rasgo psicológico medido, y cuando estas diferencias eran un artefacto provocado por diferencias en bagaje cultural en los grupos. Hacia esa década, los investigadores en psicometría abordan este tema, estableciendo criterios objetivos para el análisis del mismo, proponiendo las primeras técnicas analíticas y delimitando términos. Así, en 1982 Angoff propone sustituir el término sesgo de los ítems por el más neutro de discrepancia de los ítems. Pero fueron Holland y Thayer (1988) quienes con funcionamiento diferencial de los ítems acuñaron la expresión que finalmente desplazaría al término sesgo de los ítems, que lleva asociado una carga social y política. Hoy, existe un consenso generalizado en reservar el término funcionamiento diferencial para el análisis estadístico de la cuestión y el término sesgo para la inferencia acerca de la naturaleza de las diferencias observadas. Por tanto, solo se puede hablar de sesgo cuando sea posible relacionar el funcionamiento diferencial con el constructo que se pretende medir. El sesgo hay que entenderlo consiguientemente en términos de validez de constructo, mientras que el funcionamiento diferencial solo sería un indicador estadístico que revelaría si un ítem funciona o no de la misma forma en distintos grupos de sujetos.

3.2.2. Algunas distinciones importantes

Impacto

Si un ítem presenta DIF simplemente es señal de las diferentes propiedades estadísticas del ítem entre grupos. Esto implica que cualquier test o ítem que muestre diferencias entre grupos funciona diferencialmente, y esto no es así. Hay que precisar conceptos. Es importante saber distinguir entre el término DIF y el término impacto. Aunque el hecho de que un ítem funcione diferencialmente implica necesariamente una diferencia entre grupos, no de toda diferencia grupal en la respuesta a un ítem se sigue la presencia de DIF. Por ejemplo, supongamos que los hombres tienen una mayor capacidad espacial que las mujeres. Si esto es cierto, los hombres, por término medio, obtendrán puntuaciones superiores en los tests de aptitud espacial, y también tendrán una probabilidad mayor que las mujeres de acertar los ítems que componen estos tests. Sin embargo, los tests sólo muestran diferencias reales en la habilidad medida. Estas diferencias se denominan impacto. Más formalmente, impacto, tomando la

definición de 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. Por el contrario, si la probabilidad de responder correctamente a ese ítem difiere, para sujetos con el mismo nivel de habilidad entre grupos, entonces el ítem presenta DIF. Si el DIF y el impacto son dos cosas diferentes, el primer requisito de las técnicas empleadas para detectar DIF será que distingan las diferencias reales entre grupos (impacto) de las artificiales (DIF).

Para clarificar conceptos, proponemos un ejemplo en el que podría existir impacto. Se ha aplicado un test de compresión lectora a un grupo de niños y niñas de 5º curso de educación primaria. Asimismo, se ha aplicado a los niños un pequeño cuestionario sobre sus hábitos de lectura en el que se pone de manifiesto que las niñas leen más libros que los niños. Por otra parte, los resultados del test indican que el promedio de comprensión lectora es superior en las niñas que en los niños. ¿Existe impacto?. La respuesta es que, probablemente, debido a que las niñas han estado más expuestas a tareas de comprensión con las lecturas, esto ha hecho que desarrollen en mayor grado su comprensión lectora y que esto se vea reflejado en los resultados del test aplicado. Por lo tanto, las diferencias son reales y se puede afirmar que existe impacto. Para asegurarse de que las diferencias se deben básicamente al entrenamiento recibido por las niñas y no a otros posibles factores, se calculan y comparan las proporciones de respuesta correcta a cada ítem de los niños y niñas que han obtenido la misma puntuación en comprensión lectora. Si no existiese funcionamiento diferencial, los niños y las niñas con el mismo nivel de comprensión lectora deberían tener la misma proporción de respuestas correctas en cada uno de los ítems. En tal caso, el ítem presentaría impacto pero funcionaría del mismo modo en niños y niñas.

Un ejemplo en el que puede existir funcionamiento diferencial es el siguiente. Se ha administrado un test de analogías verbales para medir la inteligencia general a una muestra compuesta de estudiantes universitarios y jóvenes que no han finalizado los estudios primarios. Se calculan las medias de ambos grupos y los universitarios superan claramente en promedio a los que no han estudiado. ¿Existe impacto?. En este caso, la respuesta es que no. Los estudiantes universitarios están más acostumbrados a los exámenes de lápiz y papel, conocen mejor los conceptos empleados en los ítems de analogías; en definitiva, han estado más expuestos a las características de la situación y al contenido del test. Eso explicaría que sus puntuaciones medias en el test fuesen diferentes. Sin embargo, para corroborar que las diferencias no son reales, se calcula para cada ítem la proporción de respuestas correctas para los sujetos con el mismo nivel de inteligencia general de cada subgrupo. Probablemente, sujetos igual de inteligentes presenten grandes diferencias en la probabilidad de responder correctamente al ítem, a favor del grupo universitario. En este caso, los ítems sí presentan funcionamiento diferencial: La probabilidad de responder correctamente al ítem no depende solo de la inteligencia del sujeto sino de su formación, siendo más alta en sujetos con mayor formación.

6 *Funcionamiento diferencial de los ítems (DIF).*

3.2.3. El concepto de funcionamiento diferencial de ítems

Utilizamos la expresión *funcionamiento diferencial de los ítems* (differencial ítem functioning, DIF) para señalar aquellos ítems cuya probabilidad de acertarlos difiere, a igual nivel en la variable medida, entre distintos subgrupos de la población dada.

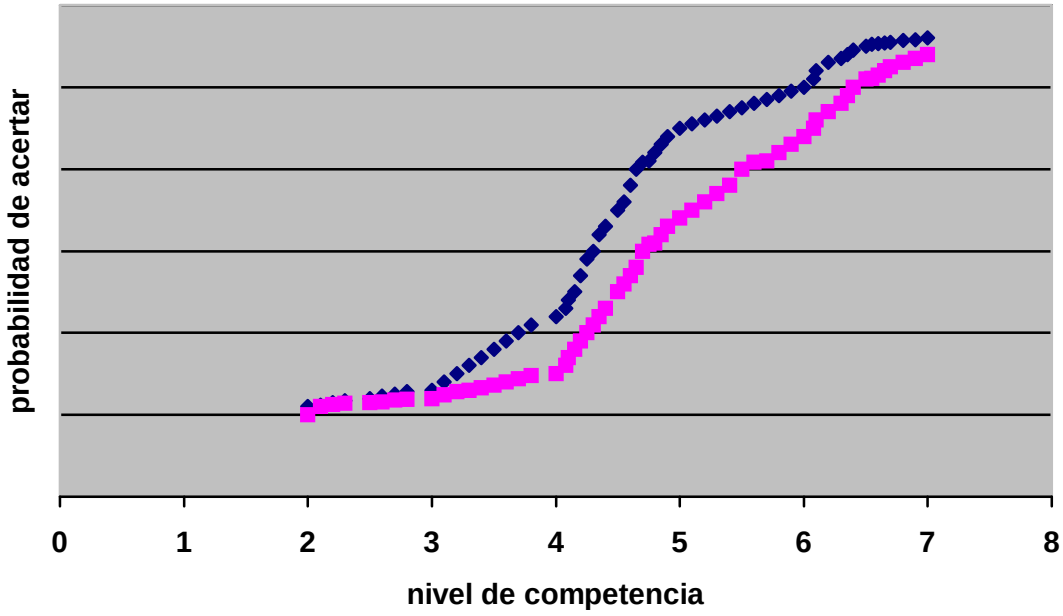
El número de grupos entre los que se establecen las comparaciones es variable, aunque la mayoría de las investigaciones se llevan a cabo sobre dos grupos. Por convención, a estos dos grupos se les denomina *grupo focal* (GF), para señalar aquel grupo que es el foco de interés de los análisis y que suele coincidir con el grupo minoritario, y *grupo de referencia* (GR), que es el grupo que sirve como base de comparación y suele ser el grupo mayoritario.

3.2.4. Tipos de funcionamiento diferencial de ítems

DIF uniforme

El DIF uniforme se produce cuando la probabilidad de contestar correctamente a un ítem es mayor para un grupo que para otro a través de todos los niveles de habilidad. Por ejemplo, en un ítem que funcione diferencialmente en contra de las mujeres éstas tendrán una probabilidad de acertarlo menor, para todos los niveles de la variables medida, que los hombres con igual nivel de habilidad que ellas.

Una representación gráfica de este tipo de DIF sería:

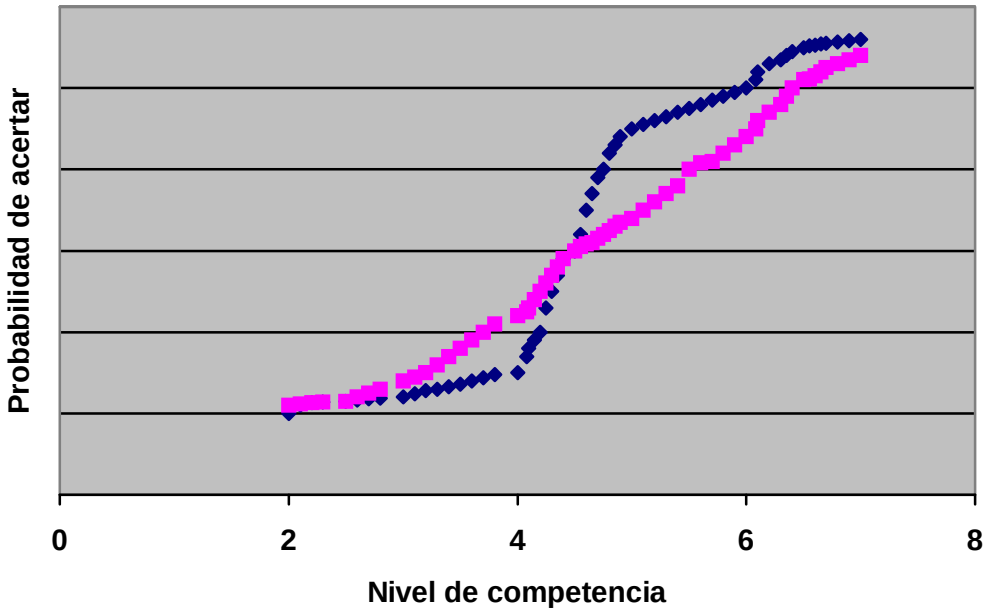


En ella podemos ver u diagrama de dispersión donde cada punto en el plano representa a una persona con determinada probabilidad de acertar el ítem (valor que le corresponde en el eje de ordenadas) dado su nivel de competencia (valor que le corresponde en el eje de abscisas). Como se observa, las mujeres con el mismo nivel en la variable medida que los hombres, sea cual sea éste, tienen siempre una probabilidad menor de acertar el ítem.

DIF no uniforme

El DIF no uniforme se produce cuando la diferencia en la probabilidad de responder correctamente a un ítem entre dos grupos no es la misma en todos los niveles de habilidad. En este caso no cabría hablar de DIF contra un grupo ya que, para determinados niveles de la habilidad, la probabilidad de acertar un ítem, a igual nivel en la variable medida, es mayor para un grupo, en tanto en los otros niveles es mayor para el otro.

Un ejemplo de este tipo de DIF no uniforme:



El ítem es más fácil para las mujeres en los niveles de habilidad bajos y más difícil en los altos. Un ítem no presentará DIF cuando todos los puntos en el plano se disponen en la misma línea con independencia del grupo al que pertenezca el sujeto al que representa cada punto.

Además, en el DIF no uniforme, podemos diferenciar dos subtipos: a) simétrico y b) asimétrico o mixto.

3.3 Teorías explicativas del DIF

Las técnicas estadísticas informan si un items funciona diferencialmente en distintos grupos, pero no ofrecen una explicación de cual es la causa de su presencia en términos de teorías psicológicas (Camilli,1993; Mellenbergh,1989). Es decir, podemos detectar que uno o más items presentan DIF con respecto a la variable de categorización de los grupos (por ejemplo, sexo), pero no sabemos la razón. Las causas del DIF son tanto o más importantes que su detección por lo que un grupo numeroso de investigadores han orientado sus esfuerzos hacia la elaboración de teorías sobre cúales pueden ser sus causas (Ackerman,1992; Camilli,1992; Lord,1980; Shealy y Stout,1993). Estos investigadores afirman que las posibles causas se pueden encontrar en la multidimensionalidad de los items.

Desde esta teoría se distinguen dos tipos de habilidad:

- habilidad principal, aquella que pretende medir el test.

- habilidades espúreas o ruido, aquellos otros rasgos o habilidades que el test no pretende medir pero que interfieren en el rendimiento del mismo, tales como estilos cognitivos, ansiedad, habilidad lectora o cultura.

Aunque la multidimensionalidad de un item no es la causa del DIF *per se*, sino las diferencias en las distribuciones condicionales de las variables espúreas. La multidimensionalidad es condición necesaria pero no suficiente para que el DIF ocurra. Es decir, sólo si dos grupos tienen diferentes distribuciones en las habilidades que no se pretenden medir, y los items del test son capaces de medir esas múltiples dimensiones, y la estimación de esas habilidades se concreta en una única medida (puntuación del test), se dan las condiciones necesarias para que se produzca DIF.

A continuación exponemos un ejemplo:

"Supongamos que tenemos un test de 40 ítems diseñado para medir la habilidad espacial. Lo hemos aplicado a una muestra de 2000 personas, de las que 1000 son hombres y 1000 mujeres. Si el test midiese sin ningún error y además fuese válido, a las personas con el mismo nivel de habilidad espacial les correspondería la misma puntuación en el test, con independencia de cualquier otra característica en la que varíen los individuos. Asimismo, la probabilidad de acertar cada uno de los ítems que componen el test sería sólo función del nivel de habilidad de los examinados, y no de ninguna otra cosa. Tanto el test como los ítems estarín, por tanto, insesgados. Ahora bien, supongamos que las mujeres por término medio tienen mayor habilidad verbal que los hombres, y que alguno de los ítems del test no sólo miden la habilidad espacial sino también la habilidad verbal. En esta ituación la probabilidad de acertar uno de los ítems no depende sólo de la habilidad espacial sino también de la habilidad verbal, y ya que las mujeres tienen por término medio mayor capacidad espacial., será mayor para las mujeres. Si en el test hay varios ítems sesgados las puntuaciones totales en el test también se verán afectadas, resultado de las medias obtenidas en el test también sesgadas".

De esta forma se pueden producir distintas situaciones según Ackerman (1992):

. Se puede producir DIF uniforme si:

a) los grupos tienen diferentes medias en la habilidad principal (impacto), siempre que haya una correlación

significativa entre la habilidad principal y la espúrea

b) si existen diferencias en las medias en la habilidad espúrea entre grupos

. Mientras que el DIF no uniforme puede ocurrir cuando:

a) la varianza de la habilidad espúrea no sea la misma entre los grupos

b)la magnitud de la correlación entre la habilidad principal y la espúrea difiera entre grupos.

No son esta las únicas causas posibles de DIF. Éste no es más que un efecto, las causas del mismo pueden se varias, pero la teoría sobre la que se asientan probablemente la mayoría de las situaciones en las que ocurre el DIF, es la teoría multidimensional.

De Ayala, Kim, Stapleton y Dayton (1999) apuntan como posible explicación del DIF la falta de homogeneidad de la población de referencia, es decir, los datos provienen de múltiples poblaciones latentes o clases. Uno de los requisitos exigibles a cualquier instrumento de medida se refiere a la invarianza de los mismos, es decir, las características de los mismos no deben variar ni por las condiciones de medida ni por los objetos medidos. En el caso de instrumentos de medida de variables

psicológicas y más concretamente de test psicológicos, como ya hemos comentado anteriormente, las características psicométricas de los mismos deben ser invariantes,es decir, los parámetros estimados deben ser los mismos para cada muestra extraída de la población. Cuando los parámetros estimados no son invariantes, esta diferencia se interpreta como una prueba de DIF, sin embargo, en TRI la ausencia de invarianza es evidencia de un mal ajuste del modelo. En efecto, si se observa un mal ajuste a través de submuestras obtenidas aleatoriamente, a partir de la muestra original de sujetos, entonces es muy probable que para uno o más de los items calibrados el modelo de TRI no sea el apropiado.

Teniendo en cuenta lo anterior, para De Ayala (1999) el DIF sería un ejemplo de un mal ajuste del modelo, cuando la muestra original se ha dividido en dos o más submuestras basándose en características manifiestas de la población (por ejemplo, varón/mujer). Aunque, por supuesto, los items que se identifican como DIF pueden ser detenidos en el test si no existe una evidencia lógica de sesgo. Es más, la diferencia entre, por ejemplo, la dificultad de un ítem estimada en un grupo y la del mismo item estimada en otro grupo, pueden ser un reflejo de diferentes "escalas" que subyacen a los datos observados. En concreto, los datos observados no reflejan una población homogénea de individuos sino una mezcla de datos de múltiples poblaciones o clases latentes. Donde hay que tener en cuenta que en una misma clase latente las diferencias individuales existen, y que las clases son cualitativamente diferentes, es decir, si consideramos una clase latente la escala de habilidad de la misma es parcial o totalmente distinta a la escala de habilidad de otra clase latente. En términos de ajuste de modelos, lo que supone es que para unos datos observados hay una o más clases latentes y dentro de cada clase latente ajusta un modelo de TRI distinto. El caso más sencillo sería la existencia de una sola clase latente, donde la muestra calibrada contiene solamente los miembros de esa clase; en esta situación un modelo de TRI simple ajustará adecuadamente a los datos. Sin embargo, cuando en los datos observados encontramos miembros de diferentes clases latentes, y no hay un modelo de TRI que refleje adecuadamente este comportamiento de los datos, es necesario trabajar con modelos más complejos. Donde se consideren distribuciones mixtas (Kelderman y Macready,1990; Rost,1990).

Por último en el contexto de la adaptación transcultural de test y escalas, algunas fuentes que provocan DIF podrían ser:

a) cuando traducimos un ítem, la dificultad de una o más de las palabras del ítem puede ser distinta de una cultura a otra.

b) cambios en el formato del ítem.

c) errores en la traducción.

d) diferencias en relevancia cultural de los términos o situaciones utilizadas en los ítems

Así, los trabajos de Hulin avanzan estableciendo que cuando encontramos DIF uniforme (diferencias en los parámetros de dificultad de los items) puede ser debida a un error en la traducción del ítem. Mientras que la presencia de DIF no uniforme (diferencias en los parámetros de discriminación) puede ser un efecto de las diferencias culturales.

Independientemente de la fuente que provoca el DIF, existen otras dificultades en el proceso de detección que impiden a veces encontrar una explicación clara de la ocurrencia del mismo.

3.4. Métodos de detección

3.4.1.Procedimientos que no especifican ningún modelo de medida

Existen procedimientos para evaluar el DIF que para aplicarlos no es necesario especificar un modelo de medida que relacione las puntuaciones obtenidas en el test y la variable latente. Este tipo de procedimientos no necesitan ningún procedimiento matemático que les permita estimar la variable latente a partir de las puntuaciones obtenidas por los sujetos en los ítems. Y no lo necesitan porque van a tomar la variable observable Z, usualmente las puntuaciones del test, como un estimador de la variable que pretende medir el test. Estos métodos utilizan la variable Z para establecer las comparaciones oportunas entre los examinados con el mismo nivel en la variable latente.

Procedimiento MANTEL-HAENSZEL

El procedimiento Mantel-Haenszel (MH) por su sencillez, bajo costo computacional y buenos resultados, es uno de los métodos más utilizados para detectar DIF. Lo primero que debemos hacer para aplicar el procedimiento MH es seleccionar la variable externa de agrupamiento que se sospeche que pueda estar generando funcionamiento diferencial en ciertos ítems del test. La variables externa debe generar solo dos grupos a comparar: grupo de referencia (GR) y grupo focal (GF). El primero suele coincidir con el grupo mayoritario o socialmente favorecido y es el que teóricamente se beneficia de la presencia de funcionamiento diferencial. El grupo focal, grupo minoritario, es en el que se centra la atención y el que se piensa que resulta perjudicado por la existencia de funcionamiento diferencial. Después se ha de disponer la información que tenemos (las respuestas de los examinados en el ítem, la puntuación de los mismos en el test y su grupo de pertenencia) en k tablas de contingencia 2x2, siendo k el número de intervalos en los que se divide la puntuación en el test. El siguiente paso sería calcular el número de respuestas correctas e incorrectas por cada grupo (GR y GF) y nivel de habilidad k.

GRUPO	RS.CORRECTAS 1	RS.INCORRECTAS 0	TOTAL
GR GF	Ak Ck	Bk Dk	NRk NFk
	N1k	N0k	Nk

Los valores marginales NRk y NFk representan el número de examinados en el grupo de referencia y focal, respectivamente; y N1k y N0k representan el número de examinados que han contestado correcta e incorrectamente el ítem, respectivamente. Finalmente, NK es el número total de examinados en el nivel de puntuación k.

La lógica que subyace al procedimiento MH es la siguiente: si el ítem no presenta DIF, la razón entre el número de personas que aciertan el ítem y las que lo fallan debe ser la misma en los dos grupos comparados a lo largo de todos los niveles de puntuación. Formalmente:

$$H0: (Ak/Bk) = \alpha(Ck/Dk) \text{ siendo } \alpha = 1 \text{ para todo k}$$

$$H1: (Ak/Bk) = \alpha(Ck/Dk) \text{ siendo } \alpha \neq 1 \text{ en algún k}$$

Mantel-Haenszel (1959) proporcionan un estimador de α para poder estimar la cantidad de funcionamiento diferencial mediante la expresión:

$$\alpha MH = \frac{\sum_{k=1}^{m}\left(\frac{Ak.Dk}{Nk}\right)}{\sum_{k=1}^{m}\left(\frac{Bk.Ck}{Nk}\right)}$$

αMH es un estimador de la magnitud del tamaño del DIF en una métrica que varía entre 0 e infinito. Un valor de 1 representa la hipótesis nula de no DIF. Si αMH es mayor que 1, el ítem estudiado favorece al grupo de referencia, por el contrario si αMH es menor que 1 indica que el ítem está favoreciendo al grupo focal.

El principal defecto que se le achaca al procedimiento MH es su incapacidad para detectar el DIF no uniforme (Rogers y Swaminathan, 1993). Este déficit puede soslayarse mediante una variación en el cálculo del procedimiento MH que recientemente han propuesto Mazor, Clauser y Hambleton (1994). Dicha modificación consiste en calcular separadamente los estadísticos MH en el grupo de examinados con puntuaciones más bajas en el test y en el grupo de sujetos con mayores puntuaciones.

Este procedimiento trae además una serie de implicaciones prácticas:

-El tamaño mínimo de muestra para utilizar le procedimiento MH con ciertas garantías es de 200 personas por grupo.

-Los ítems con DIF muy difíciles o muy poco discriminativos tienen una probabilidad muy alta de no ser detectados. Se puede evitar el problema utilizando muestras de examinados que presenten un nivel muy alto en la variable medida.

-Las puntuaciones de los sujetos en el test se deben dividir en k+1 categorías, siendo k el número de ítems en el test. A medida que disminuye el número de categorías aumenta la tasa de error tipo I, es decir, el número de ítems sin DIF que son falsamente identificados.

-Cuando un ítem está siendo evaluado debe ser siempre incluido en el cálculo de la puntuación total en el test, aunque haya presentado DIF en el análisis inicial.

Ejemplo del procedimiento MH:

Con el fin de descartar la posibilidad de funcionamiento diferencial en contra de las niñas en el test A de Álgebra, se llevó a cabo un análisis del funcionamiento diferencial de los ítems mediante el procedimiento Mantel-Haenszel. Por ese motivo, se formaron cuatro grupos de nivel de aptitud a partir de las puntuaciones en el test. En este caso, la variable externa es el género. El grupo de referencia (GR) está formado por los niños y el grupo focal (GF) por las niñas.

Se formaron grupos homogéneos de aptitud, que fueron:

- Grupo I de habilidad (entre 0 y 10)

Grupos	Correctas	Incorrectas
Niños (GR)	1	8
Niñas (GF)	2	8

12 *Funcionamiento diferencial de los ítems (DIF).*

- Grupo II de habilidad (entre 11 y 20)

Grupos	Correctas	Incorrectas	
Niños (GR)	13	58	
Niñas (GF)	10	50	
			131

- Grupo III de habilidad (entre 21 y 30)

Grupos	Correctas	Incorrectas	
Niños (GR)	30	51	
Niñas (GF)	19	84	
			184

- Grupo IV de habilidad (entre 31 y 40)

Grupos	Correctas	Incorrectas	
Niños (GR)	69	15	
Niñas (GF)	47	35	
			166

Realizando los cálculos:

Grupo De Aptitud	Ak DK / Nk	Bk Ck / Nk
I	8 / 19= 0.42	16 / 19= 0.84
II	650 / 131= 4.96	580 / 131= 4.43
III	2520 / 184= 13.70	969 / 184= 5.26
IV	2415 / 166= 14.55	705 / 166= 4.25
Total	33.63	14.78

Sustituyendo en la expresión:

$$\alpha MH = \frac{\sum_{k=1}^{m}\left(\frac{Ak.Dk}{Nk}\right)}{\sum_{k=1}^{m}\left(\frac{Bk.Ck}{Nk}\right)} = \frac{33.63}{14.78} = 2.28$$

Dado que αMH = 2.28 mayor que 1, podemos concluir que existe funcionamiento diferencial que beneficia al grupo de referencia, es decir, favorece a los niños y perjudica a las niñas.

Estandarización

El elemento básico del procedimiento de estandarización es la diferencia en la proporción de sujetos que aciertan el ítem entre el grupo focal y el grupo de referencia en cada nivel de puntuación. Si el ítem no presenta DIF no existirán diferencias entre los grupos comparados a lo largo de todos los niveles. Gráficas de las diferencias en las proporciones condicionales entre el grupo focal y el grupo de referencia proporcionan una indicación de la cantidad de DIF que un ítem exhibe. Además de estas gráficas la estandarización proporciona un índice numérico para cuantificar el DIF, la diferencia en proporciones estandarizadas (DPE):

$$DPE = \frac{\sum_{k=1}^{m} Wk(PFk - PRk)}{\sum_{k=1}^{m} Wk}$$

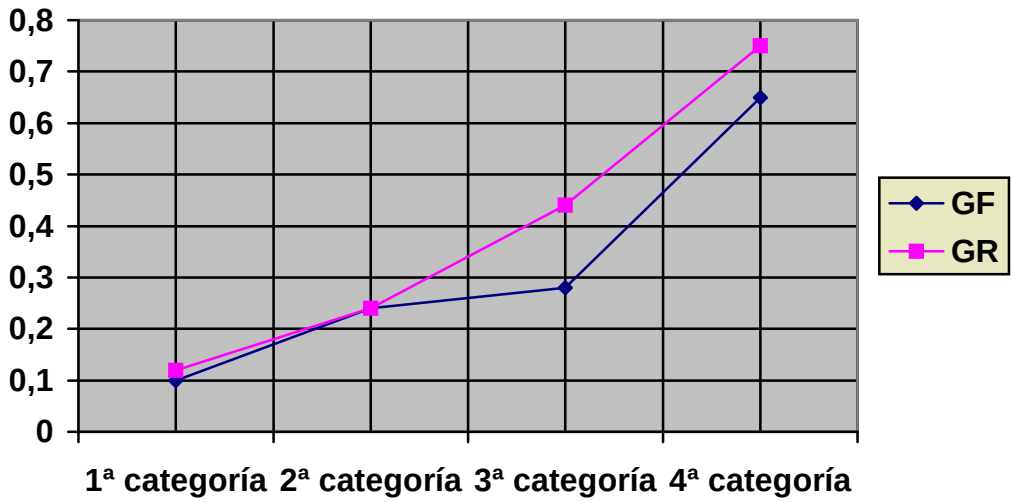
Donde, Wk es el factor de ponderación en el nivel de puntuación k, usado para ponderar las diferencias en la proporción de respuestas correctas entre el grupo focal y el de referencia. PFk y PRk son las proporciones de examinados que responden correctamente al ítem en el nivel de puntuación k en el grupo focal y de referencia, respectivamente. Dichas proporciones son iguales a:

$$PFk = Ck/NFk \qquad y \qquad PRk = Ak/NRk$$

El índice DPE varía entre –1 y 1, obteniéndose valores positivos cuando el ítem favorece al grupo focal y valores negativos cuando favorece al grupo de referencia. Dorans y Holland (1993) proponen una serie de valores que ayudan a interpretar los resultados:

- Valores entre –0,05 y 0,05 indican ausencia de DIF
- Valores entre –0,10 y –0,05 y entre 0,05 y 0,10 aconsejan una inspección de los ítems.
- Valores fuera del rango (-0,10, 0,10) son altamente sospechosos y deben ser examinados detenidamente

Un ejemplo de las gráficas de proporciones condicionales para el grupo focal y para el grupo de referencia sería:



14 *Funcionamiento diferencial de los ítems (DIF).*

Procedimiento SIBTEST

El SIBTEST, desarrollado por Shealy y Stout (1993), es uno de los más recientes y prometedores procedimientos para detectar tanto el funcionamiento diferencial de los ítems como el funcionamiento diferencial del test. Como su nombre indica (SIB significa simultaneous ítem bias), es un procedimiento para estudiar simultáneamente el DIF presente en uno o más ítems del test.

La principal ventaja que conlleva comprobar la presencia o ausencia de DIF en varios ítems a la vez es la posibilidad de determinar si los ítems con DIF actuarán juntamente afectando a las puntuaciones en el test de manera diferente según que los examinados pertenezcan a un grupo u otro, dicho brevemente, si provocarán también funcionamiento diferencial del test (FDT). A este fenómeno se le denomina amplificación. Por cancelación entendemos el fenómeno contrario: los efectos del DIF presentes en varios ítems se anulan unos a otros no llegando a provocar FDT. Por ejemplo, si tenemos un conjunto de ítems sesgados contra las mujeres y otro conjunto sesgado contra los hombres, es posible que los efectos de estos ítems se cancelen unos con otros y no se produzca un funcionamiento diferencial del test.

La teoría multidimensional del sesgo señala que los ítems de un test pueden mostrar DIF cuando, además de la habilidad principal que el test pretende medir, están midiendo otras habilidades adicionales. Los ítems insesgados son, por su parte, aquellos que miden sólo las diferencias de los sujetos en la variable que se pretende medir. Estos ítems insesgados constituyen el que se denomina subtest válido. Los ítems que pueden presenta potencialmente DIF constituirán el subtest estudiado. Para detectar el funcionamiento diferencial de los ítems se debe determinar qué porción del test mide la habilidad principal para poder igualar a los examinados de ambos grupos en esa habilidad. Es decir, se debe determinar qué ítems conforman el subtest válido. De acuerdo con esta lógica, los examinados dentro del grupo focal y de referencia son agrupados de acuerdo a su puntuación total en el subtest válido y se compara después su ejecución en el subtest estudiado.

Igualmente existe un parámetro que refleja la cantidad y dirección del sesgo (□u). Valores positivos indican que los ítems son más fáciles para el grupo de referencia. Valores negativos indican que los ítems son más fáciles para el grupo focal.

El SIBTEST ha sido diseñado para detectar el DIF/FDT uniforme, por lo tanto, una limitación evidente del mismo es su incapacidad para detectar el DIF/FDT no uniforme. Además, este procedimiento sólo puede aplicarse a test de más de 25 ítems y que además no presenten impacto.

Procedimientos basados en Técnicas de Análisis de Tablas de Contingencia Multidimensionales

Existen distintos procedimientos que tienen en común la formulación de modelos que permiten comprobar, mediante la inclusión o no de determinados términos en los mismos, diferentes hipótesis sobre el tipo de DIF o las características de la distribución de las puntuaciones de los examinados en el test. Dichos modelos son: modelos loglineales, modelos logit y la regresión logística.

Modelos loglineales

Los modelos loglineales o lineales logísticos permiten las relaciones existentes entre variables categóricas representadas en tablas de contingencia multidimensionales. Por un lado tenemos los datos obtenidos en una muestra, que son dispuestos en una tabla donde aparece el número de sujetos que hay en cada una de las categorías de las variables consideradas conjuntamente. A estas frecuencias las

denominaremos frecuencias observadas. Por otro lado tenemos una técnica, los modelos loglineales, que nos permite formular distintos modelos sobre las relaciones que mantienen entre sí las variables consideradas. Los modelos especificados son modelos teóricos y de acuerdo con ellos se generan las frecuencias esperadas o teórica, por ser las que deberían ocurrir si el modelo propuesto fuese correcto. Una forma de comprobar la corrección del modelo es ver si las frecuencias esperadas de acuerdo a él coinciden con las observadas en la muestra. Si no existen diferencias estadísticamente significativas entre las frecuencias observadas y las teóricas, debemos concluir que el modelo se ajusta suficientemente a los datos. Centrándonos en el análisis del DIF, la información con las respuestas de los examinados a los ítems son codificadas en una tabla de contingencia de 3 entradas de acuerdo con el grupo, el nivel de habilidad y la respuesta al ítem. Así las tablas de contingencia 2x2 quedarían convertidas en una tabla de contingencia de 4x2x2, compuesta de 16 celdas.

Diferentes tipos de modelos jerárquicos pueden ser ajustados para determinar si los ítems no presentan DIF o, si presentan DIF, si este es uniforme o no uniforme. Se denominan modelos jerárquicos aquellos en los que la inclusión de los términos de más alto orden implica necesariamente la inclusión de los términos de orden menor que forman parte de ellos. Por ejemplo, si en el modelo está incluida la interacción respuesta por habilidad, necesariamente formarán parte del modelo los parámetros referidos a los efectos individuales de las variables respuesta al ítem y nivel de habilidad. Para describir un modelo jerárquico es suficiente con enumerar los términos de orden superior que los componen, a esto se le denomina clase generadora del modelo. En un modelo jerárquico el DIF no uniforme vendría indicado si el modelo saturado es el único que se ajusta a los datos al nivel de significación elegido. El modelo saturado es aquel que incluye parámetros referidos a los efectos principales de cada variables y de todas las posibles interacciones posibles entre ellas:

$$Ln(Fijk) = \lambda + \lambda H(i) + \lambda G(j) + \lambda R(k) + \lambda HG(ij) + \lambda HR(ik) + \lambda GR(jk) + \lambda HGR(ijk)$$

La ausencia de DIF viene indicada por un modelo en que no están presentes ni el parámetro λHGR(ijk) ni el parámetro λGR(jk), esto es:

$$Ln(Fijk) = \lambda + \lambda H(i) + \lambda G(j) + \lambda R(k) + \lambda HG(ij) + \lambda HR(ik)$$

Los parámetros y sus errores típicos para modelos saturados se estiman mediante máxima verosimilitud. Sin embargo, en modelos no saturados o saturados muy complejos se requiere alguna suerte de algoritmo implementado en un ordenador. Los dos procedimientos más usuales son el algoritmo de Newton-Raphson, implementado en el programa MULTIQUAL de Bock y Yates (1973), y el algoritmo de ajuste proporcional iterativo, que se encuentra implementado en el SPSS (Norusis, 1988).

Por lo tanto, la hipótesis nula a probar implica que el modelo se aceptará cuando el grado de significación es superior al nivel de significación previamente establecido. Para comprobar el grado de discrepancia entre las frecuencias esperadas y las observadas podemos utilizar el estadístico ji-cuadrado de Pearson.

El procedimiento Backward (opción dentro del comando Hiloglinear del programa estadístico SPSS) partiendo del modelo saturado, va eliminando parámetros y comprobando si el efecto de esos

parámetros es estadísticamente significativo, es decir, si su presencia contribuye a mejorar el ajuste del modelo. Si un parámetro no es estadísticamente significativo es eliminado del modelo. Así se procede sucesivamente hasta conseguir un modelo en el que todos los términos de orden superior sean estadísticamente significativos al nivel de significación elegido. En nuestro caso, concluiremos que existe DIF si en la clase generadora del modelo que mejor se ajusta está presente la interacción entre las respuestas al ítem y el grupo. Es decir, si para el mismo nivel de habilidad existen diferencias entre los grupos en el número de sujetos que responden correctamente al ítem.

Modelos logit

En los modelos loglineales no se hace distinción entre variables dependientes e independientes. Es esta distinción, sin embargo, lo que diferencia a un modelo logit de un modelo loglineal. Los modelos logit toman una variable dicotómica (por ejemplo, la respuesta al ítem) como dependiente de los efectos inducidos por otras variables. Este tipo de modelos pueden ser considerados como un caso especial de los modelos loglineales, de tal forma que cada modelo logit pude ser reescrito y tiene su equivalente modelo loglineal, e inversamente.

Dada esta equivalencia podemos considerar que aplicar los modelos loglineales o el modelo logit para la detección del DIF conducen a resultados equivalentes. EL SPSS dispone del módulo LOGIT que permite realizar este tipo de análisis. El programa ofrece para cada modelo logit, al igual que con los modelos loglineales, su ajuste y una estimación de los parámetros independientes del modelo, pero no ofrece la posibilidad de comprobar cuál es el modelo que mejor se ajusta a los datos, por lo que será el usuario el encargado de ver si las diferencias entre las LR2 de los modelos son estadísticamente significativas o no.

Regresión logística

En la regresión logística (RL), como en cualquier otro tipo de modelo de regresión, lo que se pretende es predecir los valores de una variable, la variable dependiente (VD), a partir de los valores conocidos de una o varias variables independientes o predictoras (VIs). El modelo de regresión pretende estimar el factor de ponderación correspondiente a cada VI que hace mínimos de errores cometidos en los pronósticos.

Los primeros autores en proponer la regresión logística para el análisis del DIF fueron Spray y Carlson (1986), Bennet, Rock y Kaplan (1987) y Swaminathan y Rogers (1990). Mediante esta técnica se trata de determinar si en la función matemática necesaria para predecir las respuestas dicotómicas a un ítem se debe incluir un término referido a la interacción entre el grupo y la habilidad (DIF no uniforme), o al grupo de pertenencia (DIF uniforme), o simplemente puede predecirse en función del nivel de habilidad de los sujetos con independencia de su grupo de pertenencia (ausencia de DIF).

La estrategia para determinar si un ítem presenta DIF o no se basa, como en los modelos loglineales y logit, en la búsqueda del modelo más parsimonioso que mejor se ajuste a los datos. La elección de un modelo u otro se hace en función de la existencia de diferencias estadísticamente significativas entre las funciones de verosimilitud de los modelos comparados. El estadístico utilizado para comparar los modelos es el logaritmo neperiano de una razón de verosimilitudes (LR), dada por:

$$LR = -2Ln\left[\frac{L(1)}{L(2)}\right]$$

La principal diferencia entre el modelo logit y la LR es que en el modelo de regresión logística las VIs pueden ser continuas, por lo que siempre será preferible al modelo logit, cuando exista alguna variable de este tipo, por usar toda la información disponible en este tipo de variables.

Procedimientos basados en la Teoría de Respuesta al ítem.

Los procedimientos que veremos a continuación establecen un modelo de medida que relaciona las respuestas al ítem con la variable latente que pretende medir el test. Cuando las respuestas al ítem son dicotómicas y la variable latente es unidimensional se han formulado modelos de TRI. La detección del DIF desde la TRI es conceptualmente simple: calcular la curva característica del ítem (CCI) en cada uno de los grupos y determinar si coinciden (ausencia de DIF) o no coinciden (DIF).

MEDIDAS DEL ÁREA

En esta estrategia el DIF es cuantificado en función del área existente entre las CCI de los grupos comparados para determinado intervalo de nivel de competencia del sujeto en la variable medida, es decir, se trata de determinar si el área existente entre las CCI de los grupos es mayor de la que cabría esperar por azar. Aunque esta es la lógica subyacente atodas las medidas de área, se han propuesto índices del área que difieren según que sean:

- Con signo o sin signo.
- El intervalo del nivel de habilidad sea infinito o finito.
- La aproximación sea continua (integrando) o discreta (utilizando el sumatorio).
- Se ponderen o no las diferencias entre las probabilidades.

Se han planteado muchos índices de áreas, pero para aplicar cualquiera de ellos hay que seguir los siguientes pasos que citamos:

1. Elegir el modelo de TRI que proporcione el mejor ajuste a los datos.
2. Estimar los parámetros del ítem y la aptitud para el grupo focal y de referencia por separado. Si además queremos establecer la significación estadística de la diferencia entre las curvas, debemos estimar las varianzas y covarianzas de los parámetros de los items.
3. Establecer una escala común para ambos grupos. La investigación del DIF dentro de la TRI requiere que los parámetros de los items estén en la misma métrica para que se puedan realizar las comparaciones. Este cambio de métrica se realiza sometiendo a los parámetros de los items, sus varianzas y covarianzas, y los valores del nivel de competencia de uno de los grupos a una transformación lineal que los coloque en la misma métrica del otro grupo.
4. Y, por último, el cálculo de las medidas del área pertinentes.

CHI-CUADRADO DE LORD

Si en la anterior estrategia para evaluar el DIF se compara el área existente entre las CCI, en esta se compararán directamente los parámetros del ítem que definen dichas curvas. Si los parámetros del ítem estimados en cada grupo coinciden, concluiremos que el ítem no presenta DIF.

Para aplicar este estadístico hay que seguir los mismos pasos que vimos para aplicar las medidas del área. Aquí también es necesario poner los parámetros de los items de un grupo en la misma métrica que los del otro.

Una de las críticas que se hace a este procedimiento es que la hipótesis nula puede ser rechazada habiendo muy poca diferencia entre las CCI en las regiones donde de encuentran la mayoría de los examinados. A este respecto, el cálculo de algún índice del área de intervalos cerrados cuando JI-cuadrdo de Lord resulte significativo, puede ser una ayuda suplementaria que ayude a interpretar los resultados.

COMPARACIÓN DE MODELOS

Como sabemos ya la hipótesis nula de ausencia de DIF puede definirse dentro de la TRI como una diferencia no significativa entre los parámetros del ítem del grupo focal y de referencia. Una alternativa a JI-cuadrado de Lord para evaluar la igualdad de los parámetros de los ítems entre grupos se basa en la comparación de modelos. La estrategia para evaluar el DIF que se utiliza es la siguiente: si el ajuste a los datos de un modelo que incluye parámetros de los items diferentes para el grupo focal y el de referencia es significativamente mejor que un modelo en el que los parámetros de los items son iguales para ambos grupos, podemos concluir que el ítem presenta DIF.

Se trata de que tenemos dos modelos a comparar, el modelo compacto (ausencia de DIF) y el modelo aumentado (DIF), este último que incluye todos los parámetros del modelo compacto y alguno más. El estadístico utilizado para comparar los modelos es la razón de verosimilitudes (LN). Habrá que determinar si los parámetros adicionales del modelo aumentado son significativamente diferentes de 0, es decir, si su inclusión mejora significativamente el ajuste del modelo a los datos (bajo la hipótesis nula de que los parámetros del modelo aumentado son iguales a 0).

Esta estrategia puede ser utilizada para comprobar la presencia de DIF en varios items simúltaneamente. Así, el rechazo del modelo compacto (ausencia de DIF) implicaría un análisis *post hoc*, ítem por ítem, para ver que ítems presentan DIF.

Ventajas e inconvenientes de las técnicas de detección del DIF

Se han probado varias técnicas estadísticas que han tenido éxito desigual en la detección del DIF y de sus distintos tipos. En términos generales, podemos clasificarlas en aquellas que utilizan como variable de igualación entre los grupos (variable de equiparación) una puntuación observada en el test y las que utilizan como medio de equiparación la habilidad latente estimada bajo algún modelo de respuesta al ítem. Entre las primeras se podrían incluir el estadítico de Mantel-Haenszel, el procedimiento estandarizado; los modelos loglineales y modelos logit, la regresión logística y el análisis discriminante logístico. Entre las segundas encontramos procedimientos como SIBTEST y los propuestos en el marco de la TRI. Entre éstos últimos procedimientos encontramos aquellos que directamente comparan los parámetros de los items a través de los grupos estimando la significación del tamaño de las diferencias entre los mismos; los que estiman estas diferencias en términos del área entre las curvas características de los 2 grupos o los que comparan el ajuste de distintos modelos bajo el supuesto o no de parámetros de items distintos en función del grupo de pertenencia.

En la actualidad, uno de los problemas con que se puede encontrar el investigador a la hora de detectar el DIF en uno o más items, es precisamente seleccionar la técnica más adecuada para su situación

aplicada, tal y como lo han demostrado los numerosos estudios de simulación realizados hasta el momento.

De este modo, si el tamaño muestral de los grupos (focal y de referencia) es pequeño entre 200-250 sujetos por grupo, técnicas tales como Mantel-Haenszel, regresión logística y SIBTEST obtienen una adecuada potencia estadística y controlan bastante bien la tasa de error tipo I. Si el tamaño muestral es elevado (igual o más de 1000), es posible considerar la utilización de alguno de los procedimientos basados en TRI.

Por otro lado, si se sospecha que los items pueden presentar tanto DIF uniforme como no uniforme, las versiones típicas del estadístico del Mantel-Haenszel, del procedimiento de estandarización o SIBTEST serán imprecisas en la detección del DIF no uniforme. En estos casos será necesario utilizar técnicas basadas en TRI, regresión logística o los modelos logit, aunque actualmente también se dispone de modificaciones de SIBTEST.

Por último, las características psicométricas de los items (dificultad y discriminación) también son una limitación en la detección del DIF, en general, encontramos problemas para detectar DIF en items con baja discriminación, y en items con alta y baja dificultad cuando utilizamos algunos de los procedimientos basados en tablas de contingencia como Mantel-Haenszel y los modelos logit.

En una situación aplicada, antes de usar una técnica concreta para evaluar el DIF, es necesario tener en cuenta distintas cuestiones a fin de que definan la implementación de la misma.

En primer lugar, habrá que decidir si el criterio usado para equiparar los grupos será interno o externo; en caso de que el criterio sea interno (los valores de habilidad reportados por el propio test bajo estudio) habrá que considerar la posibilidad de utilizar algún procedimiento de purificación (bietápico o interativo) del mismo, teniendo en cuenta que la presencia de items con DIF contaminan las puntuaciones en el test y puede llevarnos a obtener incrementos del numero de falsas identificaciones, tanto un aumento de la tasa de error tipo I como un descenso de la potencia. No obstante, si utilizamos algún procesamiento de purificación del test lo más oportuno parece incluir en el estudio, incluso si en éste se ha identificado el DIF durante algunos de los pasos anteriores de purificación del test. Además, debemos tener en cuenta otras dos cuestiones adicionales con respecto al criterio de equiparación:

En primer lugar, la medida debe ser fiable, en este sentido, Donognue, Holland y Thayer(1993) señalan que, para obtener resultados adecuados, el test a partir del cual se equiparan los sujetos debe estar compuesto por al menos 10 items. Utilizar un criterio de equiparación poco fiable nos puede llevar a identificar falsos positivos en items que son precisamente los más discriminativos.

En segundo lugar, debemos cerciorarnos de que la puntuación de equiparación con la que estamos trabajando (el test) no es unidimensional o esencialmente unidimensional los riesgos de cometer falsas identificaciones son mayores, independientemente de la técnica de detección del DIF que estemos utilizando. El investigador, en estos casos, deberá seleccionar un subtest válido que asegure la unidimensionalidad del criterio de equiparación de los grupos, siendo SIBTEST uno de los procedimientos más adecuado para esta finalidad. Sin embargo, en el caso de que el test sea intencionadamente multidimensional, entonces será más adecuado trabajar con múltiples criterios de equiparación. Para este caso, el análisis de regresión logística o SIBTEST en su versión multidimensional podrían ser alternativas a tener en cuenta. Indudablemente uno de los aspectos más importantes, previo a la aplicación de cualquier técnica de DIF, será definir qué sujetos formarán los grupos focal y de

referencia, en qué variable o variables interesa asegurar la inexistencia de DIF (decisiones de tipo social y político), y/o estudiar la presencia/ausencia de DIF (decisiones de carácter teórico y técnico relacionadas con el propio proceso de elaboración del test).

Así, teniendo en cuenta lo anterior, y desde el punto de vista técnico-estadístico, ¿qué le debemos exigir a una técnica de detección del DIF?. Un requisito sería que proporcione una medida del DIF en términos claros y sencillos, tanto para ser transmitidos científicamente como para ser comprendidos por los consumidores finales de la misma. Y, además, disponer de un estimador adecuado para dicha medida, que sea potente y eficiente, y para el que se pueda derivar un error típico de estimación que permita construir intervalos de confianza.

Además añadimos una tabla con más información sobre las diferencias entre los distintos procesos:

PROCEDIMIENTO	CONJUNTO ITEMS (a)	ESTIMADOR DIF (b)	TEST SIGNIF. (c)	DIF NO UNIF. (d)	COSTO COMPUTO (e)
Mantel- Haenszel	no	Si	si	si	No
Estandarización	no	Si	si	no	no
SIBTEST	si	Si	si	no	no
Modelos loglineales	si	No	si	si	Si
M. logit	No	No	si	si	Si
Regresión logística	no	No	si	si	si
Medidas de área	no	Si	No	si	si
Ji-cuadrado	no	No	Si	si	si
Comparación de modelos	si	No	Si	si	si

a- Si el procedimiento permite evaluar el DIF no sólo para cada uno de los ítems, sino también para un conjunto de ellos.
b- Si dispone de un estimador de la magnitud y dirección del DIF.
c- Si es posible generalizar los resultados obtenidos mediante algún test de significación estadítica.
d- Si es capaz de detectar DIF no uniforme.
e- Si su aplicación es computacionalmente costosa.

Ítems politómicos.

En la medición psicológica, una gran parte de los test requieren un formato de respuesta en el que existan más de dos categorías, sean nominales u ordinales; para este tipo de items, denominados politómicos, no son adecuados los métodos de detección de DIF anteriormente citados por lo que hay que recurrir a la técnica de dicotomización, con la consecuente pérdida de información.

Ante la cada vez más creciente utilización de formatos de respuesta politómica, y la necesidad de poder detectar el DIF con precisión en este tipo de items, las investigaciones recientes se encaminan a desarrollar procedimientos de evaluación adecuados para items politómicos, generalmente adaptando las técnicas propuestas para items dicotómicos y estableciendo extensiones para el caso de más de dos categorías de respuesta.

Las técnicas de detección de DIF para items politómicos deben tener en cuenta la mayor complejidad que reviste la naturaleza del DIF en este tipo de items. No se tratan solamente de analizar la mayor o menor probabilidad de acertar un items para un grupo u otro de igual nivel de habilidad, sino que además, debe tenerse en cuenta que el DIF puede existir para cada categoría de respuesta, es decir, que ésta puede resultar más atractiva para un grupo que para otro.

Así, se han propuesto extensiones del método de estandarización, del procedimiento Mantel-Haenszel, y del SIBTEST para su uso con items politómicos. Las deficiencias de estos procedimientos coinciden con las señaladas para el caso de items dicotómicos: su incapacidad para detectar el DIF no uniforme y los problemas derivados de utilizar la puntuación total en el test como un estimador de la capacidad de los examinados. Sin embargo, un procedimiento que sí se ha mostrado eficaz en la detección del DIF no uniforme es el análisis discriminante logístico propuesto por Miller y Spray (1993). En un estudio de simulación donde se comparaba la regresión logística, el procedimiento M-H y el análisis discriminante logístico, este último resultó ser tan eficaz como la regresión logística en la detección del DIF no uniforme además de ser menos costoso computacionalmente. los modelos loglineales pueden utilizarse para analizar varios grupos y/o items simultáneamente, además de poder utilizarse con items politómicos.

Dentro de la TRI se han utilizado con items politómicos tanto índices de área, como procedimientos basados en la comparación de los parámetros del ítem, similares a ji-cuadrado de Lord, además de estrategias de comparación de modelos.

3.5. Tendencias de futuro.

La actualidad y relevancia del DIF se puede constatar con una revisión del número creciente de trabajos que hasta la fecha han sido publicados en revistas científicas de corte metodológico y/o aplicado. Sin embargo, todavía quedan temas por resolver. Existen algunos puntos débiles en la detección y evaluación del DIF.

Indudablemente la investigación en el DIF viene guiada por la hipótesis acerca de un funcionamiento diferente del ítem a través de dos o más grupos, y por la sospecha de que el test con el que estamos trabajando favorece a un grupo sobre otro/s. Pero no siempre el investigador es capaz de identificar, a priori, aquellas variables extrañas que están provocando el DIF, ni de identificar los grupos apropiados para los que el test o algunos de sus ítems están sesgados. En otras palabras, si uno lleva a cabo un análisis del DIF teniendo en cuenta solamente los grupos observados y no detecta DIF, La conclusión de que el instrumento no está sesgado es sólo para los grupos establecidos, pero no se pueden extender más allá de los grupos observados y por supuesto no es garantía suficiente de que el test con el que estamos trabajando no está sesgado. Las técnicas tradicionalmente propuestas para detectar DIF presentan esta limitación, sólo resultan útiles cuando el investigador, guiado por la teoría, especifica las variables de agrupamiento en las que se hipotetiza un posible DIF.

Por otro lado, no hay que perder de vista que el objetivo último, cuando se lleva a cabo un estudio de DIF, es identificar aquellos ítems que están sesgados. La eliminación y/o modificación de los mismos es lapráctica habitual para mejorar la validez del test y obtener un instrumento que es justo para todos los grupos examinados. Sin embrago, esta práctica tan común no está exenta de algunos problemas estadísticos (Camilli y Congdon, 1999). En efecto, en el análisis del sesgo, tanto la estimación como la comprobación estadística se abordan individualmente para cada ítem del test, olvidando el test como un todo. En este sentido, se apunta de la necesidad de valorar la cuestión de si el DIF mostrado por un ítem es o no importante cuando se compara con el test en su conjunto.

En resumen, es importante evaluar si los ítems acumulan o cancelan el funcionamiento diferencial cuando consideramos el test en su conjunto, y eneste sentido la investigación sobre este tema también debe aportar mayor evidencia de la eficacia de estos procedimientos frente a los que consideran el estudio individual del ítem, y proporcionando, en general, software más amigable para los investigadores interesados en el uso práctico del mismo.

Por último, la presencia de DIF no es una garantía de la presencia de sesgo en el ítem, el Dif es una condición necesaria pero no suficiente para determinar que existe sesgo; confirmar éste último requiere un análisis más pormenorizado en términos de evaluación empírica y análisis del contenido de los ítems. Además, no podemos olvidar que las técnicas propuestas para detectar el DIF sólo son apropiadas para detectar el sesgo potencial en un ítem y no ofrecen en sí mismas una explicación de laa causas del mismo. Camilli (1993) indica que una forma de abordar la explicación del sesgo del ítem ha sido tratar de conjugar los procedimientos estadísticos (objetivos) y las revisiones o juicios de expertos (Title,1982). Pero todavía no se ha demostrado la fiabilidad de este proceso y más bien parecen servir a funciones distintas, siendo el acuerdo entre ambos procedimientos moderado o escaso (Skaggs y Lissitz, 1992).

En resumen, aunque la literatura científica sobre el funcionamiento diferencial de los ítems y del test es amplia, todavía hay cuestiones abiertas que necesitan demás investigación. Una de las primeras cuestiones a resolver se refiere a la explicación de las causas del DIF, las teorías desarrolladas hasta el momento nos ofrecen una explicación parcial de este problema, y aunque las técnicas del DIF detecten un funcionamiento diferencial, las causas del mismo son a veces oscuras.

Por último, desde un punto de vista aplicado, la detección del DIF necesita una mayor atención , no debemos olvidar que en España la adaptación de los tests de otras lenguas y/o culturas es una práctica cad vez más habitual, hay un gran interés en la realización de estudios transculturales.

3.6. Bibliografía.

Fidalgo, A. M. (1996). Funcionamiento Diferencial de los ítems. En J. Muñiz (Coord), **Psicometría**, Cap. 9, pp. (372-455). Madrid: Ed. Universitas SA.