UNIVERSIDAD DE SEVILLA

DEPARTAMENTO DE PSICOLOGIA EXPERIMENTAL

EVALUACION DEL DIF EN EL TEST DE RECUERDO VERBAL SELECTIVO ENTRE POBLACIÓN ESPAÑOLA, MEXICANA Y DE HABLA HISPANA EN E.U.A.

Tesis presentada por:

Fabiola Peña Cárdenas

Para la obtención del Grado de

Doctor en Aprendizaje y Cognición

Trabajo dirigido por:

Dr. Manuel Morales Ortíz

Profesor Titular de la Universidad de Sevilla

Sevilla, España., abril de 2015

D E D I C A T O R I A

Este trabajo lo dedico a mis seres más queridos:

Mamá y papá, gracias por el apoyo infinito que me han brindado siempre, Por sus enseñanzas y porque sé que han hecho lo imposible por darme mucho más de lo que debían. Espero no defraudarlos y ser por lo menos la mitad de buena como persona, de lo que han sido ustedes, con eso creo que “ya estoy del otro lado”.

Mayra, contigo crecí y compartí gran parte de mi vida y aunque ya no estemos siempre juntas, sabemos que los lazos de sangre son indestructibles. Te debo tanto, tú y yo sabemos qué. No tengo forma de pagarte, solamente agradecerte. Ojalá que la vida te devuelva siempre tus buenas acciones.

Benito, gracias por ser el esposo comprensivo que ha estado conmigo en los altibajos de mi vida, tu apoyo y serenidad son mi refugio “cuando la tempestad azota la ciudad”. Posiblemente sin tu insistencia y visión objetiva de las cosas, este proyecto no lo habría concluido.

Dayanara y Ximena: Sé que algo del tiempo que he invertido en esta Tesis, es tiempo que les he restado a ustedes, sin embargo sepan que, en todo momento mamá piensa y está al pendiente de su bienestar, son mi mayor tesoro, mi motivo de alegría diario, sin ustedes mi vida profesional no tendría sentido. Trabajo y me esfuerzo por ser un ejemplo a seguir para las dos; espero que en unos años más, cuando crezcan, puedan leer y entender este mensaje.

Con amor infinito:

Fabiola

A G R A D E C I M I E N T O S

Quiero agradecer a mi institución de trabajo: la **Universidad Autónoma de Tamaulipas**, por la oportunidad que me brindaron de realizar el Doctorado en Aprendizaje y Cognición, también a mi centro de estudios: la **Universidad de Sevilla** por las facilidades otorgadas.

Un agradecimiento especial a mi asesor: **Dr. Manuel Morales Ortiz,** por no desistir y confiar en mí a pesar de las dificultades que conllevó el trabajar separados por la distancia geográfica; por aclarar perfectamente mis dudas, plasmadas en interminables correos electrónicos. Porque en estos 10 años casi de trabajo juntos, hemos crecido como profesionistas pero también como personas, nos tocó vivir momentos buenos y malos, que hemos compartido a distancia, sin embargo a pesar de que en algunos periodos disminuimos el ritmo de trabajo o incluso lo pospusimos hasta que las cosas mejoraran, no desistimos; y he aquí el producto de nuestra tenacidad. Espero éste sea el principio de una fructífera relación de trabajo en el campo de la investigación psicológica intercultural: España-México, gracias Manuel.

Un agradecimiento especial también, al **Dr. Juan Carlos López**, coordinador del Doctorado en Aprendizaje y Cognición, por su apoyo en los procesos de gestión y administrativos que realizamos al cursar este Doctorado.

Gracias a los Directivos y compañeros de trabajo de la **Unidad Académica Multidisciplinaria Matamoros-UAT**, por la confianza depositada en una servidora y el apoyo brindado en mis funciones laborales que me permitieron realizar mis estudios.

Expreso también mi gratitud a las Instituciones y directivos que me permitieron la recolección de datos con sus usuarios; en México: la **UAMM-UAT**, la Directora del **Gimnasio Multidisciplinario de Matamoros-UAT, Mtra. Ma. Antonia Hernández Saldívar**, y en el **Centro para la Juventud y la Familia (CEPAJUF), la Psicóloga Leticia González Díaz**.

En Estados Unidos de Norteamérica: la **Directora** del **Brownsville Community Health Center**, **Dra. Paula Gomez**, al personal de la Clínica **BCHC Cameron Park** y su Directora: **Lic. Rosalba Fernández.** Mi agradecimiento especial también a la **Dra. Jennifer Salinas profesora de la School of Public Health, University of Texas - Houston Health Science Center**, quien me apoyó enormemente con la gestión en las Instituciones de aquél país.

En España: el Departamento de Psicología Experimental, el Aula de la Experiencia de la Universidad de Sevilla, los centros para el cuidado de adultos.

Gracias también a las pasantes en psicología **Yazmín Sánchez Carreón**, **Araceli Castro Hernández, Dasaeb López Cantú, Nalleli Olivares Delgado,** por su colaboración en el proceso de recolección de datos en México y E.U.A., así como a **Eunice Cortinas** por el contacto con estudiantes de la **Universidad de Texas en Brownsville UTB**, También agradezco al equipo de trabajo del Dr. Manuel Morales Ortíz: **Ignacio Sañudo, Almudena, Silvia** y **César** que lo hicieron en España.

Por supuesto que agradezco a todos los voluntarios que participaron como sujetos de estudio en cada una de las sedes e instituciones, la donación desinteresada de un poco de su tiempo ha sido invaluable para el desarrollo de esta investigación, esperamos que los resultados sirvan para ayudar a otros y el incremento del acervo científico en este campo.

A los **colegas y amigos** más queridos que han estado conmigo en todo momento, ayudándome y motivándome a seguir cuando el cansancio y desmotivación hacían estragos en mí; su aporte, instruyéndome, alentándome o simplemente escuchándome, han sido cruciales, no menciono nombres porque son muchos, pero si te he dado a leer esto, considera estás en la lista.

Sé que podría seguir mencionando personas e instituciones, y que probablemente he omitido a algunos sin desearlo; pero ha sido un trayecto

muy largo desde aquél año de 2004 en que llegó a mí, la oportunidad de estudiar este Doctorado, pensé sin duda que sería un trabajo más fácil y rápido; sin embargo, creo firmemente que la recompensa personal es directamente proporcional al esfuerzo que ha implicado, y que a todos nos llega siempre antes o después.

Por lo anterior: a todos los que con o sin saberlo, hicieron posible esta investigación: **MI GRATITUD INFINITA.**

I N D I C E

PÁGINAS PRELIMINARES

INDICE DE FIGURAS

INDICE DE TABLAS

ANEXOS

PRESENTACION

La práctica neuropsicológica requiere conocimiento, flexibilidad, curiosidad y creatividad en el trabajo cotidiano pues, ésta es útil para identificar trastornos neurológicos en pacientes no psiquiátricos; documentar el estado cognoscitivo del paciente en las distintas condiciones neurológicas, y para identificar y localizar zonas de compromiso funcional, que en ocasiones escapan a los registros para clínicos convencionales. La evaluación neuropsicológica es útil también al estudiar enfermedades que no modifican en sus estadios iniciales la anatomía del cerebro, como las demencias o las encefalopatías tóxicas.

El estudio de la memoria ha sido de gran interés para los investigadores, en parte porque en muchos padecimientos, tanto neurológicos como psiquiátricos, uno de los primeros síntomas que se manifiestan en los pacientes son las alteraciones en esta función. En un principio se avanzó poco en esta área, debido a que se estudiaba la memoria como una función única, pero ahora se sabe que esta habilidad está formada por diferentes subsistemas o procesos relacionados entre sí, que se organizan en distintas regiones del Sistema Nervioso.

Sin embargo, a pesar de la importancia que reviste, en algunos países se cuenta con pocas herramientas para realizar correctamente este trabajo.

La evaluación neuropsicológica de la memoria en los países de habla hispana es un claro ejemplo de ello; es poca la investigación científica encaminada al desarrollo de instrumentos de evaluación confiables para esta población y la deficiencia es aún mayor en los países en vías de desarrollo como México.

Por otra parte, el crecimiento del porcentaje de población de habla hispana en Estados Unidos de Norteamérica, ha incrementado la necesidad de contar con normas y tests apropiados para estudiar a ese grupo poblacional, el USA Census Bureau (1990), ha estimado que para el 2020, habrá más de 52 millones de latinos en E.U.A. lo que se reflejará en un incremento notable en los servicios de salud requeridos por esta población.

El presente documento, sin ser exhaustivo, pretende sumarse a los esfuerzos realizados en el campo de la evaluación neuropsicológica de la memoria episódica verbal en población de habla hispana, mediante la adaptación del Test de Recuerdo Verbal Selectivo desarrollado por Buschke y Fuld (1974), el cual ha sido estandarizado para su uso con población española por Campo y Morales (2004) pretendiendo a partir de esta investigación, ampliar su uso para población mexicana y de habla hispana en E.U.A.

Este documento ha sido dividido en seis apartados para facilitar su lectura, en el capítulo uno, se presenta la revisión bibliográfica en torno al tema de la evaluación neuropsicológica de la memoria, mostrando desde las principales clasificaciones de la memoria, los estratos neurológicos y fisiológicos implicados, pasando por las principales técnicas y pruebas desarrolladas para la evaluación neuropsicológica de la memoria verbal, aquellas que han sido desarrolladas y/o adaptadas en el idioma español, concluyendo con una revisión de la situación actual de la evaluación neuropsicológica en México y los latinos de E.U. A.

En el capítulo dos se presenta una revisión de la evaluación neuropsicológica transcultural, mostrando los principales errores en los cuales se puede caer al realizar esta labor, realizando una revisión desde la situación de los hispanos en E.U.A, una de las muestras focales del presente estudio. También se aborda el tema de la necesidad de adaptar pruebas para su uso en diferentes idiomas y culturas; las recomendaciones y directrices de la International Test Comission (ITC) en ese campo; posteriormente se realiza una revisión de las principales técnicas para comprobar la equivalencia de las mediciones para asegurar una evaluación justa a los clientes/pacientes de distinta procedencia cultural.

El capítulo tres se centra en el tema del Funcionamiento Diferencial de Ítems (DIF), antecedentes históricos, así como las principales ventajas y desventajas de las diferentes técnicas estadísticas para calcularlo,

centrándose principalmente en los procedimientos utilizados en nuestro estudio: las técnicas de Mantel-Haenszel y de Regresión Logística.

En el capítulo cuatro, se presenta el estudio empírico llevado a cabo para la realización de nuestra investigación: objetivos, método, participantes, instrumento utilizado y procedimiento.

En el capítulo cinco se presentan los resultados obtenidos en el estudio del DIF en el vSRT, se presentan las diferentes etapas seguidas en el análisis de resultados, mediante los procedimientos elegidos para tal efecto: Regresión Logística y Mantel-Haenszel.

Por último, en el capítulo seis se presentan las conclusiones generales, discusión, limitantes del estudio y recomendaciones para investigaciones futuras.

Debido a la extensión de los análisis estadísticos realizados, los visores de resultados de los mismos, se han presentado como Anexos, divididos en 18 apartados para su verificación. En documento anexo también se han incluido los instrumentos utilizados para la recolección de datos y evaluación en los tres países.

CAPITULO I

EVALUACION NEUROPSICOLOGICA DE LA MEMORIA

1.1 Introducción

En el presente capítulo se tratarán cuestiones relacionadas con el concepto y clasificaciones de la memoria, sus sustratos fisiológicos y funcionamiento adecuado y patológico.

Posteriormente se hará una revisión de la evaluación neuropsicológica de la memoria verbal, así como de las principales pruebas desarrolladas hasta la actualidad para tal fin.

Se presentan también, datos de la prueba objetivo del presente estudio, el SRT de (Buschke, 1973), sus diferentes versiones y aplicaciones en el campo de la evaluación neuropsicológica.

Se concluirá con una revisión de la situación actual de la evaluación neuropsicológica de la memoria para población de habla hispana, Latinoamérica y México.

1.2 Fundamentos de neuropsicología

La neuropsicología estudia las relaciones existentes entre la función cerebral y la conducta humana, basándose en el análisis sistemático de las alteraciones conductuales asociadas a trastornos de la actividad cerebral, provocados por enfermedad, daño o modificaciones experimentales (Hécaen & Albert, 1986). Esta ciencia interdisciplinaria, recoge las aportaciones de la neurología y de la psicología, para estudiar la base neurológica de los procesos psíquicos (Goldstein, 1992).

Ardila y Ostrosky Solis (1991) presentan los objetivos de la neuropsicología en sus inicios mediante los planteamientos de uno de los precursores de la neuropsicología:

> "Luria (1970) señala que la neuropsicología persigue dos objetivos fundamentales: 1. Al delimitar las lesiones cerebrales causantes de las alteraciones conductuales específicas, se pueden

> desarrollar métodos de diagnóstico tempranos y efectuar la localización exacta del daño, a fin de que éste pueda tratarse lo antes posible. 2. La investigación en neuropsicología aporta un análisis factorial que conduce a un mejor entendimiento de los componentes de las funciones psicológicas complejas, las cuales son producto de la actividad integrada de diferentes partes del cerebro" (Alfredo Ardila & Ostrosky Solis, 1991).

De acuerdo a Hannay (1998), la neuropsicología clínica es "la aplicación de principios de evaluación e intervención basados en el estudio científico de la conducta humana a lo largo del periodo de una vida, en la medida en que se relacionan con el funcionamiento normal y anormal del sistema nervioso central".

La neuropsicología clínica es extremadamente útil para completar el diagnóstico neurológico, para la evaluación de los efectos de un tratamiento médico o quirúrgico, para establecer la línea base previa a la readaptación funcional del enfermo con un síndrome orgánico cerebral, así como para la investigación aplicada (Lezak, 2004).

Para los fines anteriores se utilizan pruebas psicológicas estandarizadas, las cuales son diseñadas para evaluar diferentes funciones cognitivas,

capacidades y habilidades humanas con el fin de proporcionar datos acerca de una diversidad de cuestiones clínicas acerca del sistema nervioso central y la conducta.

Como plantean Hebben & Milberg (2011), se podría afirmar que inicialmente las pruebas utilizadas por los neuropsicólogos no habían sido construidas con el propósito de evaluar la disfunción cerebral, y en muchos de los casos eran producto de tradiciones en la evaluación clínica, más que investigación básica en cognición o ciencias neuronales; ejemplo de ello son las Escalas Wechsler de inteligencia para adultos (Wechsler Adult Intelligence Scale; WAIS; (Wechsler, 1955); otras más fueron creadas como pruebas de inteligencia, en el campo militar, vocacional o académico para la identificación del retardo mental (Kaufman & Litchenberg 1999; Matarazzo, 1972; en Hebben & Milberg, 2011).

En la actualidad, los requerimientos mínimos para pruebas neuropsicológicas son la sensibilidad a la presencia de disfunción cerebral, y a la capacidad de distinguir correctamente entre la presencia de función cerebral anormal y el funcionamiento cerebral normal. A lo largo de los años, estos criterios han ido aumentando, para incluir la capacidad de predecir la ubicación y severidad de la disfunción cerebral y, en algunas ocasiones incluso la de predecir la causa específica o etiología de esa disfunción. Durante la introducción de las primeras pruebas neuropsicológicas formalmente construidas y validadas, la sensibilidad de

éstas se consideraba por su correspondencia con los juicios clínicos de los neurólogos (Reitan & Davison, 1974, citados por Hebben & Milberg, 2011). Actualmente, con el avance por un lado en el campo de la psicometría, la Teoría Clásica de los Test, la Teoría de Respuesta al Ítem (Franzen, 2000) y por otra parte con el avance de la tecnología y surgimiento de las técnicas de imágenes neuronales, ha aumentado también la expectativa de sensibilidad de las pruebas neuropsicológicas para detectar estos cambios en la fisiología y estructura cerebral, contrastándolos con los resultados obtenidos con las técnicas de neuroimágen.

En la práctica clínica, algunos recomiendan el uso de una batería de pruebas fijas para poder comparar las observaciones a través de diferentes grupos poblacionales de pacientes; mientras que otros recomiendan que la batería de pruebas sea flexible y adecuada a la problemática y necesidades del paciente.

1.2.1 Antecedentes históricos de la neuropsicología

Los problemas que son el foco de atención de la neuropsicología clínica moderna han sido descritos desde hace siglos y han tratado de ser resueltos por médicos y filósofos; aunque no es el objetivo del presente estudio, a continuación se presentará un panorama general histórico de los hechos que han influido en el estado actual de la neuropsicología.

Como menciona López de Ibáñez (1998), aunque la neuropsicología como ciencia es relativamente reciente, los intentos por tratar de explicar la relación cerebro-conducta se encuentran desde épocas muy antiguas; sin embargo, no es hasta comienzos del siglo XVII, que la idea de que los pensamientos, recuerdos y pensamientos se originan en el cerebro. Tenemos por ejemplo los trabajos del anatomista alemán Charles Meyer quien hacia 1700, postulaba en su tratado sobre anatomía y fisiología del cerebro que en la corteza gris estaba localizada la memoria, en la sustancia blanca la imaginación y la razón y en las porciones basales la percepción y la voluntad.

Sin embargo, de acuerdo a López de Ibáñez (1998) fue Franz Joseph Gall (1758-1828), un anatomista del cerebro que continuó con este esfuerzo por localizar las distintas funciones mentales en estructuras del cerebro. De acuerdo con los argumentos de Gall (1835), los órganos separados dentro del cerebro controlaban funciones como la sabiduría, la religiosidad, lenguaje y memoria. Realizó una descripción detallada de las comisuras, señaló que los nervios craneales no se originan en el cerebro sino en la médula y presentó las funciones de la sustancia gris y su relación con la sustancia blanca. Podríamos, menciona López de Ibáñez, considerar su “mapa frenológico” publicado hace 200 años, como la primera formulación del *localizacionismo* estricto (López de Ibáñez, 1998).

Esas y otras investigaciones posteriores, permitieron que el mapa de la corteza cerebral se llenara de esquemas que proyectaban sobre ella las ideas de la psicología asociacionista predominante en la época y que se verían cristalizadas en el que aún hoy es un punto de referencia para las neurociencias: el famoso mapa de las localizaciones de Broadmann, creado en 1909.

Sin embargo, posteriormente surgió otro enfoque conocido como *holismo*, término empleado por Goldstein en 1939 (K. Goldstein, 1939; citado por Hebben & Milberg, 2011), cuya postura se puede entender con el siguiente ejemplo: “A pesar de que un tornillo flojo pudiera ser responsable de un mal funcionamiento que evitaría que el motor de un automóvil arrancara, sería erróneo localizar la función de la locomoción en el propio tornillo. Un síntoma puede surgir debido a la interrupción de un componente importante de una red mayor de funciones, o a que únicamente la más complicada y susceptible o débil función de muchas otras funciones favorecidas por la misma área está interrumpida (Hebben & Milberg, 2011).

1.2.2 La neuropsicología y el psicodiagnóstico

En la misma época de estos hallazgos psicofisiológicos surge el *psicodiagnóstico*; en 1890, Catell acuña el término “test mental” (Cattell, 1890) y en 1905 Alfred Binet publica el primer “test” en el sentido estricto, denominado

“Escala Métrica para la Evaluación de la Inteligencia” (Binet & Simon, 1905). En este periodo comenzaron a desarrollarse una gran cantidad de instrumentos destinados a evaluar constructos cognitivos y la personalidad, otros se empezaron a encaminar hacia la evaluación de los cambios que se presentaban en las personas cuando sufrían una lesión o enfermedad cerebral (Binet & Simon, 1948). Surgieron entonces dos escuelas que fueron las más importantes en su época: la rusa con los trabajos de Alexander Luria y su enfoque cualitativo, clínico e informal y la americana con un enfoque mucho más psicométrico, basado en la estandarización de pruebas y procedimientos, teniendo como promotores a Ward Halstead y sus discípulos, quienes crearon la primera batería sólida de pruebas neuropsicológicas psicométricas y construyeron las bases para muchos de los instrumentos y estándares para la construcción de pruebas utilizadas en la actualidad. Ralph Reitan empezó a trabajar con Halstead en los años cuarenta y extendió el trabajo de Halstead ensamblando una batería de tests para la evaluación completa de los individuos con daño cerebral: el Halstead Reitan Neuropsychological Test Battery, la cual se basaba ampliamente en supuestos no localizacionistas (Reitan & Wolfson, 1985).

1.2.3 La neuropsicología en la actualidad

La lista de nombres influyentes en el desarrollo de la neuropsicología hasta la actualidad sería infinita. Hoy podemos afirmar que la unión de los conocimientos

de neuropsicología y la aparición de procedimientos de alta tecnología en neuroimagen como la resonancia magnética MRI (Magnetic Resonance Imaging) o la tomografía por emisión de positrones PET (Positron Emission Tomography), en ocasiones llamada neurociencia cognitiva, han permitido la localización cada vez más detallada de cambios en la actividad neuronal asociada con mediciones de cognición experimentales cada vez más específicas. Por otra parte, también hay una gran proliferación de pruebas que pretenden evaluar el amplio rango de habilidades y funciones cognitivas; sin embargo, con los avances en materia de psicometría, las exigencias han aumentado en relación al procedimiento utilizado para su construcción, fundamento teórico y utilidad como instrumento diagnóstico o pronóstico de la situación neurológica del paciente (Franzen, 2000; López de Ibáñez, 1998) así como los aspectos relacionados con la evaluación de las poblaciones minoritarias, para las que no han sido inicialmente desarrolladas las pruebas, y que pueden resultar en evaluaciones sesgadas, inválidas o injustas (Elbulok-Charcape, Rabin, Spadaccini, & Barr, 2014), aspectos que veremos a mayor detalle en el siguiente capítulo.

1.3 La memoria, conceptos fundamentales

La memoria es la capacidad de recuperar información y utilizarla para propósitos adaptativos (Fuster, 2005); comúnmente se concibe a la memoria como aquellas impresiones mentales de la experiencia, de acontecimientos pasados y

de los aprendizajes; tomando en cuenta básicamente solo aquella información que es adquirida y recuperada conscientemente (memoria explicita). Sin embargo, la neuropsicología y psicología cognitiva tratan de ampliar esta información al incluir también toda aquella información que es adquirida, recuperada y utilizada sin conocimiento consciente (memoria implícita); por ejemplo, las habilidades motoras, el conocimiento perceptual. Por lo tanto, una definición más completa de la memoria debería incluir un enorme cúmulo de experiencias que el organismo ha almacenado a lo largo de la vida en su sistema nervioso para adaptarse a su entorno, sea o no consciente de ello en algún momento (Andrewes, 2013).

Como plantea Fuster (2005), la distinción entre memoria y conocimiento es muy sutil; fenomenológicamente, el conocimiento es la memoria de hechos y de relaciones entre hechos, como que toda la memoria es adquirida mediante la experiencia. Las memorias nuevas conllevan un proceso temporal de consolidación antes de ser almacenadas permanentemente o convertirse en conocimiento. Y esta falta de diferencias significativas es obvia al observar la interacción funcional que existe entre memoria y conocimiento tanto a nivel cognitivo como neurológico. Una primera cuestión es que no existe una memoria totalmente nueva como tal, sino que es una expansión de un conocimiento viejo. Tomando como base que toda percepción requiere del recuerdo, que a su vez es una interpretación del medio de acuerdo a conocimientos previos, y en ese proceso de interpretación, las percepciones sensoriales son instantáneamente clasificadas de acuerdo a experiencias pasadas; entonces, una nueva percepción

se convierte en memoria nueva al construirse sobre la memoria pasada. Podemos afirmar por tanto, que sin un conocimiento previo, una nueva percepción es ininterpretable y por consiguiente, es imposible de codificar como una memoria nueva.

Al ver la interacción entre estas funciones podemos darnos cuenta, que los sustratos neurológicos de la percepción, el conocimiento y la memoria, también son muy similares; un acoplamiento de conexiones corticales o cogniciones, que contienen en su conjunto de estructuras la información contenida de las tres funciones.

Un aspecto muy general que vale la pena mencionar es que la memoria es fundamentalmente una función asociativa. El proceso biofísico básico que permite la formación de la memoria es la modulación de transmisión de información a través de las sinapsis, lo cual permite "asociar" una célula con otra (Hebb, 1949; Kandel, 1991; citado en Fuster, 2003). La memoria individual es formada y almacenada en las redes neuronales de la corteza de asociación, llamado así debido a que en él se conectan los sistemas sensoriales y motores. Las redes de la memoria cortical, están formados por acumulaciones de neuronas por un proceso asociativo autónomo (Kohonen, 1984; Edelman, 1987; citado en Fuster, 2003) y por último, la recuperación de la memoria es esencialmente un proceso también asociativo (Fuster, 2003).

1.3.1 Tipos de memoria

Llegar a un acuerdo sobre los diferentes tipos de memoria y su ubicación dentro de las distintas regiones cerebrales, parece ser una tarea muy complicada, debido sobre todo a que como ya se vio, es un proceso complejo que implica la activación de una gran cantidad de funciones. Sin embargo, parece haber algunas clasificaciones que son más o menos constantes y defendibles a través de distintos estudios.

La investigación del funcionamiento del cerebro y las regiones implicadas en la memoria es posible en parte gracias al estudio de sujetos que por accidentes, enfermedades y diversas circunstancias tienen lesiones en esas estructuras y que posteriormente presentan dificultad para realizar algunos procesos implicados en la memoria, véase el caso del paciente H.M. que se menciona más adelante.

Un hecho ampliamente estudiado y que cuenta con gran aceptación, es la función del hipocampo en la memoria. Se ha encontrado que la formación de asociaciones entre las agrupaciones de células corticales tiene lugar gracias al control de las estructuras que conforman el sistema límbico, y especialmente del hipocampo (Squire, 1992).

Las investigaciones precursoras de este campo se iniciaron gracias a las observaciones que se realizaron del famoso paciente H. M., sujeto que como tratamiento para la epilepsia que padecía fue sometido a una cirugía de ablación bilateral de las estructuras del lóbulo medio-temporal (Scoville & Milner, 1957) después de la cual perdió la capacidad para formar nuevas memorias, a pesar de mantener intacta su memoria autobiográfica de largo plazo. Hoy se sabe que las estructuras medio-temporales y especialmente el hipocampo tienen una función primordial para el depósito de la memoria nueva posiblemente en la neo-corteza.

Otras operaciones como la atención, la práctica y la repetición parece que influyen al reforzar las sinapsis que forman las redes de recuerdos en la corteza permitiendo la consolidación de la memoria; proceso que a nivel cognitivo se conoce como *codificación* de la memoria, aunque algunos teóricos prefieren utilizar este término para el primer momento de adquisición de la memoria. Como cualquier otra función neuronal, el proceso de consolidación de la memoria implica la activación de algunos grupos neuronales, así como la inhibición recíproca de otras (Squire, 1992).

Los estímulos que construyen nuevas memorias pueden provenir de muchas fuentes algunas externas y otras internas. Las externas son estímulos sensoriales o salidas de áreas de procesamiento sensorial. Las internas pueden

ser de aquellas redes corticales asociadas preexistentes, que han sido activadas por un estímulo sensorial en un acto de percepción.

Un hecho que vale la pena mencionar es que la información contenida de una cognición o memoria está compuesta por relaciones y no solo por la suma de sus componentes, por lo tanto, cualquier neurona o conjunto de neuronas en cualquier jerarquía cognitiva, puede formar parte de muchas memorias. Estas múltiples relaciones y superposiciones pueden producir redundancia en la representación o degeneración; es decir, permitir que diferentes estructuras produzcan un mismo resultado o memoria, lo cual por una parte es positivo, debido a que permite recuperar memoria después del daño en algunas áreas específicas, puede ser la clave además para comprender la poca especificidad de ciertas amnesias al producirse una lesión a nivel neuronal, pero también puede ser la causa de que se produzca el fenómeno conocido como *"falsa memoria"* (M. K. Johnson, Raye, Mitchell, & Ankudowich, 2012), tema de gran controversia en países como EUA dentro del sistema legal, ya que pone en entredicho la confiabilidad de los llamados "testigos visuales" al observarse incongruencias entre los datos reportados por el recuerdo o descripción de los hechos de los sujetos implicados y lo concluido mediante otro tipo de evidencias, como por ejemplo restos de ADN (Belli, 2011).

1.3.1.1 Memoria a corto y a largo plazo

Una de las principales clasificaciones que suele hacerse sobre la memoria es aquella que distingue entre memoria a corto plazo (MCP) y memoria a largo plazo (MLP).

Las primeras observaciones que se hicieron en este campo fueron realizadas por Ebbinghaus (1885), quien experimentando consigo mismo, puso a prueba su habilidad para recordar una lista de sílabas sin sentido en distintos intervalos temporales, para posteriormente analizar sus resultados, encontrando que el descenso en sus recuerdos durante las primeras horas fue muy alto, y que en los días subsecuentes la disminución del recuerdo fue más lenta, hasta llegar casi a estabilizarse la *tasa de olvido* al paso de un mes. A partir de sus observaciones, expuso la existencia de *dos formas o estados de la memoria*; sugirió que aquellas sílabas que había olvidado se hubieran quedado en la *"memoria a corto plazo"* mientras que las sílabas que podía recordar después de un mes, hubieran pasado a un almacén mucho más estable que previene de la pérdida de los recuerdos: *"la memoria a largo plazo"* (Baddeley, 1999; Ebbinghaus, 2014).

En 1968 Atkinson y Shiffrin proponen la existencia de dos almacenes separados para los dos tipos de memoria: *"el modelo lineal de los dos almacenes"* apoyados en estudios en donde los sujetos tenían diferentes grados de eficiencia al recordar una lista de palabras que se les presentaba, dependiendo el tiempo que pasaba entre la presentación y el recuerdo. También al presentárseles una lista de palabras y pedírseles que las recordaran independientemente del orden, la eficiencia en el recuerdo era mayor para aquellas palabras que se encontraban al principio de la lista (*efecto de primicia*) y aquellas que estaban al final de la lista (*efecto de recencia*), mientras que las palabras del medio eran poco recordadas. Otro estudio realizado es el de insertar un estímulo distractor o actividad entre la presentación de las palabras y el recuerdo, lo cual interfiere con el efecto de recencia, es decir el recuerdo de las últimas palabras de la lista, pero no con el efecto de primicia, sugiriendo que las primeras palabras de la lista han pasado a la memoria a largo plazo y por ello son menos susceptibles de ser olvidadas por efecto de la *interferencia*, mientras que las últimas palabras de la lista son más susceptibles debido a que aún se encuentran en la memoria a corto plazo (Atkinson & Shiffrin, 1968).

Existen numerosos estudios con sujetos con daños a nivel hipocampal que han apoyado la "teoría de los dos almacenes"; pacientes que sufren de amnesia anterógrada, es decir que pueden recordar eventos recientes, pero que son incapaces de mantenerlos a largo plazo, lo que puede ser producto de que la memoria a corto plazo se guarda en un almacén provisional y que el hipocampo

tiene la función de transportarlo a un almacén permanente: la memoria a largo plazo; recuérdese el caso del paciente H.M mencionado con anterioridad (Scoville & Milner, 1957). Algunos otros sujetos amnésicos pierden el efecto de primicia en pruebas de recuerdo libre, mientras que mantienen el efecto de recencia.

A pesar de la evidencia acumulada a favor del modelo modal de Atkinson & Shiffrin (1968), que concebía a los almacenes de memoria como estructuras esencialmente unitarias, cada una con funciones bien diferenciadas, a finales de los años sesenta, surgieron otros modelos que cuestionaban el carácter unitario del Almacén a Corto Plazo (ACP), así como también, investigaciones que, explorando el aprendizaje y la memoria de sujetos amnésicos, dieron sustento a la idea de una memoria de largo plazo compuesta de múltiples componentes. Es así como surgieron modelos teóricos alternativos al *modelo modal*, tales como el Enfoque de los Niveles de Procesamiento desarrollado por Craik & Lockhart (1972) y el Modelo de Memoria Operativa desarrollado por Baddeley & Hitch (1974). Estos modelos en su conjunto se caracterizaban por remarcar los aspectos funcionales antes que los estructurales de la memoria, revisando cuestiones tales como la forma en que se procesa la información sobre el recuerdo posterior y la función del ACP. Por otra parte, como plantea Baddeley (1999), los estudios de Warrington & Weiskrantz (1968) y Warrington & Weiskrantz (1970) mostraron que bajo algunas condiciones de recuperación, los amnésicos podían recordar información estimular pasada, la cual, sin embargo, parecía encontrarse por completo disociada de la conciencia. Estos descubrimientos y algunos otros

similares, tanto con sujetos amnésicos como normales, parecían demostrar la existencia de una forma de memoria esencialmente no consciente o “implícita” (Schacter, 1987) lo cual, posteriormente, redundó en una fragmentación del entonces unitario Almacén de Largo Plazo (ALP), (Schacter, Chiu, & Ochsner, 1993).

1.3.1.2 Memoria perceptual

La *memoria perceptual* es aquella adquirida a través de los sentidos en cualquier momento de la vida del individuo, es decir, adquirida por medio de la experiencia sensorial y solamente por fracciones de segundo (o uno o dos segundos) en lo que el mecanismo de atención selecciona y da paso a un procesamiento posterior (Varela Ruíz, Fortoul, Ávila Acosta, & Fortoul van derGoes, 2005)

Si tratamos de buscar la ubicación de este tipo de memoria, podemos dividir el cerebro en dos partes, a través de la fisura de Rolando, la corteza cerebral ubicada en la parte posterior de la misma es principalmente *sensorial,* encargada de recibir y almacenar la información proveniente de los sentidos, mientras que la parte anterior a la fisura es mayoritariamente *motora*; es decir,

controladora de la realización de acciones del organismo y demás funciones cognitivas relacionadas (Fuster, 2001).

1.3.1.3 Memoria ejecutiva

El estudio de sujetos con daño en el lóbulo frontal, ha permitido apreciar la función de la corteza frontal como un almacén de la *memoria ejecutiva* o *motora*. En sujetos con daños en la corteza frontal lateral es común observar que presentan dificultad para realizar actividades de manera ordenada, así como para formular o realizar nuevos planes o "*esquemas de acción"*. Es decir, que los déficits se presentan en ambas direcciones temporales: en el pasado se presenta la incapacidad para la realización de secuencias conductuales y hacia el futuro, se presentan *"déficits en la planeación"* o en la *"memoria prospectiva",* es por esa razón que la corteza prefrontal suele denominarse el ejecutivo o administrador del cerebro, órgano de la creatividad etc., además de que juega un papel crucial en la inteligencia (Fuster, 2003).

La repetición, el ensayo y la práctica tienen efectos en la memoria ejecutiva, similares a aquellos que se presentan en la memoria perceptual, mediante esas operaciones se consolidan las redes de la memoria ejecutiva en la corteza frontal. La representación de las secuencias de acción más concretas y automatizadas

son relegadas a las estructuras inferiores del sistema motor: las cortezas motora y premotora, los ganglios basales, el sistema piramidal y el cerebelo. Por el contrario, la representación de los componentes de las acciones más generales o abstractas es consolidada en la corteza prefrontal (Fuster, 2001).

1.3.1.4 Memoria explicita e implícita

Otra clasificación que se suele realizar sobre la memoria es aquella que la divide por el grado de conciencia de la misma; se denomina *memoria explicita* a aquella que es voluntaria y que generalmente permite almacenar recuerdos de palabras, hechos y lugares y a la vez se subdivide en *memoria semántica* y *memoria episódica*. *La semántica* es el almacén de conocimientos acerca de los significados de las palabras y las relaciones entre estos significados, constituyendo una especie de diccionario mental. La *memoria episódica* representa eventos o sucesos que reflejan detalles de la situación vivida y no solamente el significado, está sujeta a parámetros espacio-temporales; esto es, los eventos que se recuerdan representan los momentos y lugares en que se presentaron.

La *memoria implícita* por otro lado es involuntaria, en otras palabras; es una respuesta compulsiva hacia un estímulo o situación: y se subdivide en procedural, automática y emocional.

En cuanto a sus sustratos neurobiológicos, como plantea Squire (1992), parece haber una gran implicación del hipocampo junto con las estructuras relacionadas anatómicamente (estructuras medio-temporales), en la formación de la memoria declarativa o explícita, mientras que la implícita o no declarativa, parece no requerir del hipocampo para su formación, codificación y mantenimiento.

1.3.2 Memoria y funciones cognitivas

El funcionamiento eficiente de la memoria requiere de la función intacta de un gran número de regiones cerebrales, algunas de las cuales son especialmente susceptibles de daño. Además, muchas condiciones neurológicas y psiquiátricas producen déficit en estas funciones. De hecho, entre los pacientes neurológicos los problemas de memoria se consideran la queja inicial más común. Uno de cada tres individuos mayores de 75 años sin demencia, presentan déficit en la memoria (Riedel-Heller, Matschinger, Schork, & Angermeyer, 1999).

El uso del término “memoria”, generalmente crea confusión debido a la gran cantidad de actividades mentales que engloba. Es muy común el que los pacientes y en ocasiones los mismos clínicos, cataloguen ciertos tipos de disfunciones cognitivas bajo el concepto de “déficit en la memoria”. Por el contrario, en algunos sujetos que presentan un déficit en la capacidad de aprendizaje, indican que tienen una buena memoria debido a que poseen vivos recuerdos de sucesos de los primeros años de su vida y que pueden evocarlos fácilmente. Otros, como sucede con muchos de los adultos mayores, pueden manifestar problemas al tratar de recordar palabras comunes o los nombres de las cosas consistentemente (*disnomia*), lo cual se puede presentar con o sin problemas en el área de la memoria episódica (Fuster, 2003; Lezak, 2004).

Además, los déficits en otros procesos cognitivos como la atención, concentración, velocidad de procesamiento de la información, organización, por mencionar algunos, también pueden producir alteraciones en el desempeño de la memoria (Ganor-Stern, Seamon, & Carrasco, 1998)

El tener una distinción adecuada sobre la terminología implicada y sus alcances, dentro del campo de la memoria, brinda a los clínicos una mayor oportunidad de realizar una buena evaluación y conceptualización de los diferentes trastornos de la memoria (Lezak, 2004).

1.4 Evaluación neuropsicológica de la memoria

Debido a la diversidad de formas en que se pueden presentar los déficits de la memoria, se requiere contar con pruebas confiables, válidas y con datos normativos que permitan tener una impresión clínica sobre la severidad y naturaleza de los trastornos de la memoria y el aprendizaje (Bauer, Tobias, Valenstein, & Heilman, 1993; Mayes, 1986; Squire & Shimamura, 1996).

Las técnicas para evaluar los trastornos de memoria y aprendizaje son importantes por dos motivos; en primer lugar, son fundamentales si se quiere lograr un cuadro clínico preciso, favoreciendo que el paciente pueda ser tratado correctamente, optimizándose el uso de los tratamientos cognitivos, conductuales y farmacológicos pertinentes. Por otra, juegan un papel esencial en la investigación de las causas funcionales de los trastornos de aprendizaje y memoria, por ejemplo, identificando los déficits de procesamiento y almacenaje que subyacen a tales trastornos (Mayes, 1986). De forma general, los tests de memoria deberían indicar no sólo la gravedad del trastorno, sino también especificar los tipos de memoria afectados así como los procesos que se ven comprometidos. Cuando los trastornos de memoria se evalúan con métodos cuantitativos, se hace posible identificar las distintas alteraciones que pueden ocurrir, así como comprender las semejanzas y diferencias entre ellas.

Como plantean Hebben & Milberg (2011): “Evaluar la memoria significa evaluar codificación y adquisición de información, retención y recuperación, índice de deterioro y susceptibilidad a la interferencia, así como memoria de reconocimiento contrapuesta a memoria espontánea” (Hebben & Milberg, 2011). Existen algunos instrumentos que incorporan elementos para evaluar una gran cantidad de esos procesos, mientras que otros miden un solo aspecto.

La Wechsler Memory Scale-Fourth Edition (WMS-IV), es una de las pruebas más completas utilizadas ampliamente en Estados Unidos para la evaluación exhaustiva de la memoria, en coordinación con el WAIS-IV, ya que permite obtener una evaluación de la memoria conjunta con el nivel intelectual en adultos de 16-90 años (Wechsler, 2009). Evalúa cinco subescalas: Auditory memory index, visual memory index, immediate memory index, delayed memory index y visual working memory index.

Test of memory and Learning, second edition (TOMAL-2), evalúa aprendizaje y memoria en niños y adultos de 5 a 59 años, provee índices de memoria verbal, no verbal y compuesta. Incluye además índices complementarios de memoria para historias, aprendizaje verbal, memoria de oraciones, de diseños, memoria secuencial, recuerdo libre, atención y concentración (Reynolds & Voress, 2007).

1.4.1 Pruebas neuropsicológicas de memoria verbal

Con respecto a las pruebas específicas de memoria verbal, que son ampliamente utilizadas en todo el mundo; tenemos el caso del California Verbal Learning Test (CVLT; Delis, Kramer, Kaplan, & Ober, 1987). Se basa en el aprendizaje cuatro listas de 16 palabras cada una, siendo estas divididas en categorías. Se tienen cinco intentos para repetición y recuerdo, posteriormente se introduce una lista de palabras de interferencia. Se evalúa memorización después de una demora breve y pasados 20 minutos; al final de la prueba se da opción para reconocimiento. Actualmente en Estados Unidos se cuenta con la segunda edición de la prueba, para adultos entre 16 y 89 (CVLT-II; Delis & Kramer, 2000), mientras que en español tenemos la adaptación de Jacobs y colaboradores para la evaluación de adultos (Jacobs, Winston, & Polanco, 1997).

El Rey Auditory Verbal Learning Test (RAVLT; Rey, 1941; 1958), esta prueba requiere que el sujeto memorice 15 palabras a lo largo de cinco ensayos; posteriormente se introduce una segunda lista, después de la cual se solicita una repetición de memoria luego de una demora breve, una repetición de memoria después de una demora larga y por último reconocimiento de la primera lista.

El Hopkins Verbal Learning Test (Brandt, 1991), consiste en la memorización de 12 palabras de tres categorías semánticas distintas, teniendo tres ensayos libres de claves para su memorización y al final se da la opción de reconocimiento por respuesta “si” y “no” (estaba o no estaba en la lista). La versión revisada (HVLT-R), cuenta con datos para adultos de 16 a 80 años o mayores, se puede administrar una repetición de memoria 20 o 25 minutos después y al tener seis conjuntos equivalentes de listas de palabras, se puede repetir la administración de la prueba para medir cambios a lo largo del tiempo (Brandt & Benedict, 2001).

1.4.1.1 El Test de Recuerdo Verbal Selectivo (vSRT)

El SRT desde la creación de su versión original en 1974 por Buschke y Fuld y en las adaptaciones que se han realizado posteriormente, ha sido una de las pruebas más utilizadas para valorar los desórdenes de la memoria asociados con patologías como epilepsia del lóbulo temporal (Plenger et al., 1996), esclerosis múltiple (Boringa et al., 2001; Rao, Leo, & Aubin-Faubert, 1989) demencia tipo Alzheimer (Campo, Morales, & Martínez-Castillo, 2003; Masur, Fuld, Blau, Crystal, & Aronson, 1990), e incluso aquellos sujetos con deterioro cognitivo leve quienes aún no cumplen con los criterios para el diagnóstico de demencia (Tabert et al., 2006) debido a que permite evaluar retención, almacenamiento y recuperación de información verbal nueva, diferenciando entre memoria a corto plazo y memoria a

largo plazo; además de dos estancias diferentes en el proceso de recuperación: el recuerdo aleatorio y el recuerdo consistente; permitiendo a los clínicos formarse un panorama sobre la situación del paciente en cuanto a esta capacidad cognitiva.

La versión original de esta prueba consiste en una serie de diez ítems, todos de la misma categoría (animales, ropa), los cuales son leídos al sujeto a tasa de dos por segundo; y se le pide su recuerdo inmediato. En los intentos subsiguientes se le repiten solamente aquellas palabras que haya omitido en el intento anterior. El procedimiento se continúa hasta que el sujeto recuerde todas las palabras o hasta el décimo segundo intento. Se obtienen medidas de aquellas palabras recordadas consistentemente sin necesidad de dar recordatorio, lo cual se califica como *recuerdo consistente de largo plazo*, se puede obtener también *memoria a largo plazo* o *almacenamiento a largo plazo* a través del número de palabras recordadas en dos o más intentos consecutivos sin necesidad de recordatorio; *recuerdo de corto plazo* son las palabras nombradas solo después de que se ha brindado el recordatorio: el *recuerdo aleatorio a largo* plazo se refiere a palabras en el almacenamiento a largo plazo, que no reaparecen consistentemente sino que requieren recordatorio (Buschke, 1973; Buschke & Fuld, 1974).

Hannay & Levin (1985), describieron y evaluaron las propiedades psicométricas del vSRT, basados en el mismo paradigma de Buschke y Fuld

(1974), para cuatro versiones., pero utilizando una serie de 12 palabras, las cuales a diferencia del original no estaban relacionadas; contando con un máximo de 12 ensayos para intentar memorizarlas o de acuerdo al criterio de Larrabe y colaboradores (Larrabee, Trahan, Curtiss, & Levin, 1988), suspender el procedimiento cuando después de 3 ensayos consecutivos se haya repetido la lista completa sin errores, en cuyo caso, se considerará como aprendida (Mitrushina, 2005).

Buschke (1984) agregó el componente de recuerdo con clave, denominando a esta prueba Free and Cued Selective Reminding Test (FCSRT), utilizado posteriormente para sujetos con sanos, con demencia de diversos tipos, amnesia tanto en la vejez como en adultez (Buschke, 1984; Degenszajn, Caramelli, Caixeta, & Nitrini, 2001; Grober et al., 1999; Grober et al., 2008; Ivnik et al., 1997; Peña-Casanova, Gramunt-Fombuena, et al., 2009; Traykov et al., 2005).

En cuanto al vSRT, se han presentado datos normativos, propiedades psicométricas y descrito su uso con muestras de sujetos sanos sin demencia en una gran cantidad de estudios (Larrabee et al., 1988; Ruff, Light, & Quayhagen, 1989; Trahan & Larrabee, 1993). También se han evaluado la validez de una versión del test con 6 ensayos tanto en sujetos sanos como con patologías (Bell, Fine, Dow, Seidenberg, & Hermann, 2005; Drane, Loring, Lee, & Meador, 1998; Smith, Goode, La Marche, & Boll, 1995) y se han presentado datos normativos

para dos versiones de la prueba con 6 ensayos en español (Morales et al., 2010). En todos estos estudios se han reportado propiedades psicométricas similares entre las versiones de 12 y 6 ensayos, sin embargo, al disminuir los ensayos, se ahorra tiempo en la aplicación, lo cual parece ser productivo ya que produce menos fatiga y oposición a la evaluación por parte de los sujetos.

Como apunta Mitrushina (2005) también se ha presentado evidencia en una gran cantidad de estudios respecto a la eficacia del SRT original y en sus versiones modificadas para la evaluación de pacientes clínicos con patologías diversas (O'Connell & Tuokko, 2002; Noe et al., 2004), demencias de tipo Alzheimer (Bartok et al., 1997; Boeve et al., 2003; Devanand et al., 2007; Masur et al., 1990; Stern, Albert, Tang, & Tsai, 1999) o Parkinson (Jacobs et al., 1995; Levy et al., 2002) epilepsia del lóbulo temporal (Drane et al., 1998; Umfleet et al., 2014; Westerveld, Sass, Sass, & Henry, 1994), epilepsia de lóbulo frontal (Johnson-Markve, Lee, Loring, & Viner, 2011), esclerosis múltiple (Beatty, Krull, et al., 1996; Beatty, Wilbanks, et al., 1996; Chiaravalloti, Demaree, Gaudino, & DeLuca, 2003; Chiaravalloti, DeLuca, Moore, & Ricker, 2005), tumores cerebrales (Torres et al., 2003), traumatismos craneoencefálicos (Zec et al., 2001).

Una de las principales ventajas o diferencias de ésta prueba, frente a las similares de memorización de lista de palabras es que al ser una lista de palabras independientes (no divididas por categorías semánticas como el HVLT o el CVLT,

que podría utilizarse como mnemotécnica o ayuda para la codificación de la información), requiere un esfuerzo extra por parte del sujeto para tratar de memorizar y recuperar la lista de palabras. Además, el hecho de que solamente se repitan o recuerden al sujeto las palabras que ha olvidado en el ensayo anterior, es otra característica distintiva de esta prueba.

1.4.1.2 Adaptación del SRT para su uso en diferentes idiomas y grupos culturales

Rao y Andrade realizaron una adaptación del SRT para usarse de manera verbal y visual en población Indú, además de validarlo con una muestra de sujetos normales (S. L. Rao & Andrade, 1998). Gigi y colaboradores realizaron una adaptación para el idioma hebreo, con tres formas paralelas (Gigi, Michal Schnaider-Beeri, Davidson, & Prohovnik, 1999).

En cuanto a Latinoamérica, en el 2002, Xavier y colaboradores, usaron una versión del SRT, con una muestra de sujetos brasileños, adultos mayores de 80-95 años de edad que no padecían demencia (Xavier, Ferraz, Trentini, Freitas, & Moriguchi, 2002). En Argentina, se realizó un estudio para la detección temprana de la demencia mediante una batería que incluía la versión del SRT de Buschke (Butman, 2001).

Peña-Casanova, Gramunt-Fombuena, et al., (2009) presentaron datos normativos del *Free and Cued Selective Reminding Test (FCSRT)* dentro del proyecto NEURONORMA, el cual pretende normativizar una serie de pruebas neuropsicológicas para su uso con adultos mayores de 49 años, en este estudio, se muestran datos de 356 sujetos adultos mayores de 49 años, cognitivamente sanos y agrupados por edad, sexo y nivel educativo (Peña-Casanova, Blesa, et al., 2009).

Muy recientemente, Palomo y colaboradores, como parte del mismo proyecto NEURONORMA, también han presentado datos normativos con población española joven, contando con una muestra de 179 sujetos sanos de entre 18 y 49 años de edad agrupados por escolaridad y edad, e indican será uno de los objetivos, el ampliar la muestra con población de estas edades para futuros estudios (Palomo et al., 2013).

La selección de los estímulos siguió los mismos criterios que la versión inglesa, aunque considerando la frecuencia y prototipicidad de las palabras en lengua española, pero solo realizan 3 ensayos de recuerdo inmediato libre, cada uno seguido de un recuerdo facilitado, y además un recuerdo diferido a los 30 min, también de forma libre y posteriormente facilitado.

Una versión pictórica que fue denominada pFCSRT+IR, traducida al español de la versión en inglés del *Pictoric Free and Cued Selective Reminding Test* sido recientemente desarrollada por el grupo de sus creadores originales para el diagnóstico de demencia en hispanos (Grober, Ehrlich, Troche, Hahn, & Lipton, 2014).

Un grupo de investigadores del departamento de psicología experimental de la Universidad de Sevilla han estado realizando trabajos en torno a esta prueba, el vSRT ha sido adaptada al idioma español se han estudiado sus propiedades psicométricas de confiabilidad (Campo, Morales, & Juan-Malpartida, 2000a), validez (Campo et al., 2003), se han obtenido datos normativos (Campo & Morales, 2004) en población española y como se mencionó anteriormente, se han desarrollado y presentado datos normativos de dos versiones de la prueba con 6 ensayos (Morales et al., 2010).

1.4.2 Otras pruebas neuropsicológicas desarrolladas en español

Ejemplo de ello son los datos normativos para la escala de memoria de la prueba (Wechsler, 1945) obtenidos por Alfredo Ardila & Rosselli (1994), datos normativos en una muestra de adultos mayores para dos versiones en castellano del California Verbal Learning Test (CVLT) estudiados por Jacobs y colaboradores

(Jacobs et al., 1997) y posteriormente datos normativos también de una versión para españoles de esta misma prueba, obtenidos por Benedet y Alejandre, que fue denominado TAVEC Test Auditivo Verbal España-Complutense (Benedet & Alejandre, 1998). En el campo de los estudios enfocados a población bilingüe y su impacto en el aprendizaje verbal y la memoria, encontramos el trabajo realizado por Harris y colaboradores, quienes construyeron dos pruebas de listas de palabras equivalentes en inglés y español, similares al CVLT (Harris, Cullum, & Puente, 1995). Por otra parte, en el 2005, González y colaboradores, desarrollaron y presentaron datos normativos de una prueba de aprendizaje verbal con versiones en inglés y español (Gonzalez, Mungas, & Haan, 2005).

Peraita, Diaz, & Gonzalez-Labra (2000), construyeron una batería para la evaluación de la memoria semántica en adultos con Demencia tipo Alzheimer, y muy recientemente, ha sido adaptado para su uso en Argentina por Grasso y Peraita (Grasso & Peraita, 2011a; Grasso & Peraita, 2011b).

Como hemos mencionado con anterioridad, el grupo de Peña-Casanova, tienen en España, un trabajo extenso y reconocido en la estandarización de una gran cantidad de pruebas (Tamayo et al., 2012), que son actualmente ampliamente utilizadas en la comunidad clínica española, incluidas entre ellas la *FCSRT* recuerdo libre y con claves (Palomo et al., 2013).

Con respecto a Latinoamérica, recientemente un grupo de investigadores obtuvo datos normativos de una batería de evaluación neuropsicológica integral para hispanos denominada: NEUROPSI (Ostrosky-Solis, Ardila, & Rosselli, 1999) y otra específica de la atención y la memoria en sujetos desde 6 hasta los 65 años (Ostrosky-Solis, Gomez-Perez, et al., 2007). Estas pruebas están actualmente comercializándose en México, ya que permite obtener índices tanto independientes como una puntuación global de atención y memoria; ha sido utilizada en niños, adultos y población geriátrica. Evalúa memoria de trabajo, memoria a corto y largo plazo para material verbal y visuoespacial, también se han medido los efectos de la educación en la ejecución del test en los diferentes rangos de edades, encontrándose una gran influencia del nivel educativo en la mayoría de las pruebas (Ostrosky-Solis et al., 1999), se evaluaron sus propiedades psicométricas para población hispana en Estados Unidos de Norteamérica (Abwender & Sfikouris, 2005) y una versión abreviada del mismo, como instrumento diagnóstico en la demencia tipo Alzheimer (Abrisqueta-Gomez, Ostrosky-Solis, Bertolucci, & Bueno, 2008).

Un estudio muy reciente con sujetos mayores de 70 años, presenta datos normativos para el Mini Mental State Exam (MMSE), una versión del Free and Cued Selective Reminding Test (FCSRT) y del Isaacs Set Test, con población adulta mayor de la ciudad de México (Mokri, Alberto Avila-Funes, Meillon, Gutierrez Robledo, & Amieva, 2013).

1.4.3 Conclusiones

En la práctica clínica el neuropsicólogo debe seleccionar, de entre los instrumentos de evaluación disponibles, aquellos que resulten los más apropiados para valorar a cada paciente. Como hemos visto, el estudio de la memoria, sus componentes, alteraciones así como formas de evaluación, son de gran importancia para formarse una impresión clínica global de la naturaleza y severidad de los problemas de memoria.

Como vimos con anterioridad actualmente existen numerosas pruebas que evalúan la memoria verbal y que son ampliamente utilizadas: California Verbal Learning Test (CVLT; Delis et al., 1987), el Rey Auditory Verbal Learning Test (RAVLT; Rey, 1941) y el Selective Reminding Test (SRT; Buschke, 1973) por mencionar solo algunas, las cuales han demostrado poseer confiabilidad y validez; sin embargo, la mayor parte de ellas están enfocadas para la población de habla inglesa y solo recientemente se han realizado trabajos con la finalidad de desarrollar pruebas y establecer normas adecuadas para la población de habla hispana.

Siendo mucho más precisos en México, ante la falta de instrumentos de evaluación, al igual que en otros países, se ha convertido en una práctica común

el hecho de utilizar pruebas que han sido desarrolladas en otros idiomas/culturas, sin embargo esta práctica no es recomendable, debido a que las diferencias culturales, idiomáticas, contextuales, etc. pueden producir diferencias significativas en los resultados obtenidos en las pruebas, disminuyendo la confiabilidad y validez de los mismos; cuestiones que serán revisadas a detalle en el siguiente capítulo.

De acuerdo a la revisión de la literatura realizada por la autora, son mínimas las investigaciones que se han realizado para la construcción o adaptación de pruebas neuropsicológicas que midan los diferentes aspectos de la memoria para poblaciones no solo de México, sino de Latinoamérica en general.

De antemano se sabe que el trabajo y camino por recorrer es extenso, pero también es innegable la necesidad de que la comunidad científica de estos países haga algo para tratar de eliminar este vacío en la práctica neuropsicológica, la cual requiere acciones inmediatas enfocadas a la construcción y/o adaptación de pruebas neuropsicológicas no sólo para la atención de la población de habla hispana, sino para cada grupo cultural.

Como se mencionó con anterioridad, un grupo de investigadores de la Universidad de Sevilla han estado realizando trabajos en torno al vSRT (Campo et al., 2000a; Campo, Morales, & Juan-Malpartida, 2000b; Campo et al., 2003;

Campo & Morales, 2004; Morales et al., 2010); de ese grupo de investigación se desprendió un primer proyecto de investigación neuropsicológica transcultural, encaminado a evaluar el Funcionamiento Diferencial del Ítem en el vSRT entre población mexicana y española, el cual fue presentado como tesina para la obtención del DEA de la autora, en ese primer trabajo, no se encontró DIF en los Ítems entre ambas poblaciones, sin embargo la muestra era muy pequeña; al momento de redactar este documento, no existen datos normativos del vSRT, para población Mexicana e incluso podría afirmarse que para ninguna población de habla hispana en América.

CAPITULO II

EVALUACIÓN NEUROPSICOLÓGICA TRANSCULTURAL

2.1 Introducción

En el presente capítulo se revisarán los aspectos más relevantes de la neuropsicología transcultural, el impacto del trabajo clínico con minorías, la situación actual de los hispanos en E.U.A, así como las principales fuentes de error en el trabajo con minorías y culturas diferentes (Byrne et al., 2009). Posteriormente se revisará el tema de la adaptación de tests en idiomas y culturas, los procesos necesarios para establecer equivalencias entre pruebas. Por último, se presentan las principales fuentes de error o invalidez clasificados en tres grandes apartados: diferencias culturales e idiomáticas que afectan a las puntuaciones, aspectos técnicos y métodos, e interpretación de los resultados.

2.2 Neuropsicología transcultural

La neuropsicología transcultural, retoma los conceptos de la psicología transcultural tradicional, para evaluar como un grupo cultural denominado

generalmente minoritario, es evaluado desventajosamente en comparación con uno más grande, para el cual fueron creadas originalmente la mayoría de las pruebas o tests neuropsicológicos disponibles. Como plantea Alfredo Ardila (1995), en vías de eliminar estos fallos, se requiere trabajar no solo en la traducción o estandarización de los tests existentes; sino que es necesario evaluar además el impacto de diversas variables culturales en el desempeño en los test neuropsicológicos, por ejemplo, la aculturación, transculturación, educación, etnia, raza, bilingüismo, por mencionar algunas; así como su relación con la función cognitiva y el rendimiento en las pruebas psicológicas y neuropsicológicas, antes de recomendar el uso de una prueba de forma indistinta para la nueva población.

2.2.1 Minorías e hispanos en E.U.A.

Los cambios poblacionales en todo el mundo son producto de múltiples factores, como son la globalización, economía, política, migración, educación, entre otros. En Estados Unidos de Norteamérica, esto no es la excepción; de acuerdo a datos proporcionados por el US Census Bureau, (D. I. S. U. S. Census Bureau, 2005; P. U. S. Census Bureau, 2010) para el año 2050, casi un 50% de la población de los E.U.A. estará conformada por minorías étnicas, de los cuales casi un 25% serán hispanos. Según Llorente, los hispanos se han convertido en una de las minorías más fuertes en los E.U.A., actualmente más del 11% de la población en ese país es hispana, lo que representa más de 32 millones de individuos, si se

considera solamente a los que están viviendo legalmente (Llorente, 2008) las cifras aumentarían al incluir a la población con situaciones migratorias complicadas; ello obliga a pensar en la necesidad de establecer medidas para garantizar su bienestar en todos los aspectos, establecer cambios en la prestación de los servicios de salud, educación, justicia, empleo; y esto es una cuestión no sólo de política, sino de la ciencia también (Poreh & Sultan, 2009; Puente & Ardila, 2000).

En la era de la diversidad cultural, la cultura por si misma se ha convertido en una variable difícil de explicar, la incursión de la psicología transcultural, con la revisión de los aspectos de la aculturación, cultura, sesgo, e incluso con el tema de la psicología de las minorías (Wong & Fujii, 2004), ha propiciado que la evaluación neuropsicológica considere aspectos que antes no se pensaban relevantes e involucran una gran cantidad de ambigüedades que van desde variaciones inexplicables en los resultados obtenidos en la investigación, hasta el rendimiento variable en perfiles neuropsicológicos (Pontón & Carrión, 2001). Es de reconocer el impacto de trabajos como el de la American Psychological Association, al establecer las Guías para proveedores de servicios psicológicos a poblaciones étnica, lingüística y culturalmente diversa *"APA Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations"*, en sus diferentes años de presentación: 1991, 2001, 2003, impulsando entre los clínicos el capacitarse, entender y aceptar la diversidad cultural a fin de brindar un servicio de calidad a estas poblaciones, mientras que a

los investigadores deja la responsabilidad de explorar en nuevas líneas de investigación en el campo, a fin de observar el impacto de diversos factores culturales en la salud mental de estos grupos (Puente, 1993).

2.2.2 La cultura y fuentes de error en la evaluación neuropsicológica

La cultura es un concepto que hace referencia a un cuerpo de creencias habituales y normas sociales que son compartidas por un grupo particular de personas, la cual incluye conductas, ideas, valores y otros elementos compartidos. También podría definirse simplemente como la forma de vida específica de un grupo humano. El término cultura es un concepto complejo que muy frecuentemente es usado de manera indistinta por los términos de raza, grupo étnico y en menor proporción con el de lenguaje (Ardila, 2013; Gasquoine, 1999). Los principales problemas a los que se enfrentan los profesionales que evalúan a sujetos de diferentes culturas incluyen un pobre *rapport* entre profesional y cliente, sobreestimación o subestimación del significado de los síntomas, atribuir erróneamente la presencia de síntomas, la actitud del cliente hacia la actividad de la evaluación, entre otros; ya que como el ejemplo que presenta Alfredo Ardila (1995), mientras que para la mayoría de los norteamericanos y europeos, la evaluación es una práctica bien aceptada, no sucede lo mismo en otros países y culturas.

Hambleton (1996) indica que los errores en la aplicación de tests, se presentan cuando existen problemas de comunicación entre el aplicador y los examinados o cuando las instrucciones escritas no se han dado con claridad suficiente, dando margen a las explicaciones verbales. Para evitar este problema es conveniente que la selección de los aplicadores sea lo más adecuada posible. Deben proceder de la población a la que se va a aplicar el test, estar familiarizados con la cultura, idioma y dialectos, tener experiencia y habilidades adecuadas para la aplicación de tests y conocer la importancia de seguir los procedimientos estandarizados con el test; estas deficiencias se pueden minimizar con un entrenamiento básico a todos los aplicadores como parte del proceso de desarrollo del test.

A continuación veremos algunas variables, cuestiones teóricas y prácticas así como su influencia en la evaluación neuropsicológica a la que se enfrentan los clínicos en el proceso de evaluación neuropsicológica con grupos minoritarios o culturas distintas; Wong y colaboradores plantean que las principales diferencias interculturales son producidos por la cultura, etnicidad, educación y lenguaje (Wong, Strickland, Fletcher-Janzen, Ardila, & Reynolds, 2000), no obstante que Ardila (2013) agrega a éstas, los patrones de habilidad; otros más, incluyen el nivel de aculturación (Boone, Victor, Wen, Razani, & Ponton, 2007; Coffey, Marmol, Schock, & Adams, 2005).

2.2.2.1 Etnicidad

Mientras que el término cultura hace referencia a las creencias y costumbres asociados a un grupo, etnicidad se refiere a las agrupaciones que se hacen de acuerdo a la descendencia, características físicas y cuestiones hereditarias que comparten un grupo de personas, frecuentemente este término es utilizado como “raza”, tenemos por ejemplo en Estados Unidos de Norteamérica, un país con una gran diversidad cultural, que se suele dividir al origen étnico en Caucásicos (que son considerados la mayoría), los hispanos/latinos, indios americanos o nativos, asiáticos y afroamericanos (Boone et al., 2007). Estas clasificaciones claramente presentan dificultades conceptuales y no representan la totalidad de la diversidad cultural, a manera de ejemplo, tenemos el caso de los hispanos o latinos, en donde son incluidos tanto los mexicanos como el resto de los sujetos de Centro y Sudamérica; no obstante, podemos entender que entre uno y otro país existe una gran diversidad cultural, e incluso idiomática como sucede con los brasileños cuyo primer idioma es el portugués o los puertorriqueños que al ser un país/estado perteneciente a E.U.A. tienen tanto inglés como español por primer idioma. Otro problema referente a la clasificación en “razas” es el de las personas con herencia racial o étnica mixta y que por tanto tienen una mezcla de hábitos, costumbres e idiomas. En un estudio realizado por Boone y colaboradores, se evaluó la relación entre la raza y el desempeño en pruebas cognitivas de pacientes de una clínica neuropsicológica, encontrando diferencias significativas entre todos los grupos para las diversas pruebas: copia

de la Figura Compleja de Rey-Osterrieth, Test de Wisconsin, Trail Making Test, Boston Naming Test. Se encontraron diferencias también entre aquellos que tienen el idioma inglés como primer lenguaje y aquellos que lo tienen como un segundo idioma; años de educación en los E.U.A.; tiempo de residir en aquél país y edad en la que se aprendió el inglés conversacional (Boone et al., 2007; Gasquoine, 1999; 2001).

Sin embargo, a pesar de la necesidad de contar con datos normativos para las diferentes poblaciones en la investigación neuropsicológica; la composición étnica de las muestras raramente se identifica. Como apuntan Stanczak, Stanczak, & Awadalla (2001), quienes realizaron un metaanálisis de los artículos publicados de 1988 a 1994 en Archives of clinical neuropsychology y Journal of clinical and experimental neuropsychology; de los 567 estudios publicados, sólo en 83 de ellos (14.6%) se mostraba la composición étnica de las muestras simplemente como blancos vs. no blancos y solamente en 6 (7%) de los 83 estudios había algún tipo de análisis usando la etnicidad como una variable de estudio.

2.2.2.2 Educación

Este es un factor que tiene un gran impacto no solo en el conocimiento general del individuo, sino que también permite desarrollar ciertas habilidades actitudes que son comúnmente evaluadas con la mayoría de las pruebas o tests neuropsicológicos (Manly, 2008). Este es un factor que habría que considerar en los estudios transculturales, y principalmente cuando se trabaja con minorías y migrantes, ya que como menciona Wong et al., (2000), es mucho más común encontrar niveles bajos de educación formal (p. ej. menos de ocho años) en inmigrantes de ciertas partes del mundo. Se suele utilizar como indicador de esta variable el número de años de asistencia reglada a la escuela, aunque algunos autores afirman que más importante que el número de años de escolaridad, sería la calidad de la enseñanza recibida (Manly, 2008).

Al evaluar la influencia del nivel educativo en las pruebas neuropsicológicas, encontramos el trabajo de Rosselli et al. (1999) por ejemplo, quienes plantean que las puntuaciones globales en las pruebas de fluidez verbal fonética se ven ampliamente influidos por el nivel educativo, encontrándose diferencias aproximadas de un 20% entre los resultados de los sujetos con educación a nivel universitario y aquellos sujetos analfabetos (Ardila, Ostrosky-Solis, Rosselli, & Gomez, 2000; Gonzalez da Silva, Petersson, Faísca, Ingvar, & Reis, 2004; Ostrosky-Solis, Gutierrez, Flores, & Ardila, 2007), también presentan

diferencias en la capacidad para nombrar palabras, que tienen una frecuencia baja, siendo más difícil para los sujetos analfabetos, lo mismo que para nombrar imágenes (Reis, Guerreiro, & Castro-Caldas, 1994) y en la memoria de trabajo (Silva, Faísca, Ingvar, Petersson, & Reis, 2012).

Por otra parte, como plantean Reis y colaboradores, la educación no siempre influye a favor, las diferencias entre niveles educativos no se presentan en las tareas de fluidez verbal cuando se utilizan objetos reales, en lugar de dibujos bidimensionales (Reis, Petersson, Castro-Caldas, & Ingvar, 2001), ni para identificación visual de aquellas palabras comunes (Reis et al., 1994) o cuando se evalúa la habilidad para resolver problemas de la vida cotidiana. Incluso, los sujetos con niveles educativos bajos parecen tener ventajas sobre aquellos con nivel educativo alto en actividades motrices. Al parecer el mayor impacto del aspecto socioeducativo, se da en aquellas pruebas que requieren aspectos del lenguaje complejos y habilidades motoras (Puente & McCaffrey, 1992; Rosselli & Ardila, 2003).

2.2.2.3 Lenguaje

Artiola I Fortuny, Heaton, & Hermosillo (1998) plantean que el lenguaje es una herramienta para la evaluación, por lo que un evaluador con poca habilidad en

el lenguaje del examinado sería menos capaz de detectar una prosodia anormal, sintaxis poco común o algún otro síntoma que pudiera indicar un desorden en el lenguaje con base neurológica (Wong et al., 2000). En el caso de los Estados Unidos de Norteamérica, el incremento en la inmigración de población que habla idiomas diferentes al inglés, ha hecho que existan dificultades en la evaluación no solo inherentes a las diferencias culturales y étnicas, sino también, errores producidos por la diferencia lingüística y semántica entre evaluador y paciente, traduciéndose en barreras que imposibilitan o merman la evaluación neuropsicológica (Gasquoine, Cavazos, Cantu, & Weimer, 2010; Harris et al., 1995). Incluso, se ha encontrado que además de las diferencias producidas por el instrumento en sí, se pueden observar diferencias por factores inherentes al propio sujeto debido a algún factor lingüístico: la familiaridad con los términos o incluso con la mayor o menor utilización del idioma a pesar de que sean considerados bilingües. Véase el estudio de Gasquoine, Croyle, Cavazos-Gonzalez y Sandoval (2007), quienes evaluaron el funcionamiento de sujetos bilingües (español-inglés) neurológicamente sanos mediante baterías neuropsicológicas en ambos idiomas, encontrando diferencias significativas entre los resultados de los sujetos en las subpruebas con mayor carga verbal, dependiendo del mayor o menor dominio de cada idioma.

2.2.2.4 Patrones de habilidades

Ardila y Ramos (2007) apuntan que, las habilidades que son evaluadas frecuentemente en una prueba psicológica, tienen una gran carga cultural, siendo un factor desarrollado de forma cualitativamente distinta entre una cultura y otra, teniendo una gran cantidad de habilidades aprendidas. Esta variable parece muy similar al factor educativo mencionado anteriormente, sin embargo; aquí, habría que considerar no sólo la educación escolar, sino las diferentes oportunidades de aprendizaje y experiencias contextuales o ambientales. Además, habría que puntualizar que no se trata de establecer que esas diferencias son innatas para cada grupo o cultura, tema de gran controversia en décadas pasadas (innatismo vs. ambientalismo), podríamos afirmar que mientras que las habilidades cognitivas básicas son universales, el contexto establece en qué situaciones particulares es más apropiado utilizar un determinado proceso cognitivo. La cultura dicta que cosas deben ser aprendidas, a qué edad, por qué género, en qué forma; es decir, el ambiente cultural permite el desarrollo o aplicación de determinados patrones de habilidades (Ardila & Ramos, 2007; Ardila, 2013; Berry, 1979), formas de pensar, actuar y sentir (Ardila, Rosselli, & Ostrosky-Solis, 1992; Ardila, 1995).

2.2.2.5 Aculturación

Otra de las variables que ha sido sometida a investigación es el nivel de aculturación, que cuantifica el nivel en que los miembros de una minoría participan de las tradiciones culturales, valores, creencias y prácticas de la cultura dominante; como plantean Poreh y Sultan (2009) al referirse a la clasificación realizada por Dana (1996), la aculturación se podría dividir en cuatro formas: (a) tradicional, (b) marginal, (c) bicultural, (d) asimilación; sin embargo, no es la única clasificación, la cual puede ser evaluada de manera subjetiva o por medio de algún instrumento (Berry, 1979; 2005) como son los autoinformes, cuestionarios específicos o por el cumplimiento de ciertos indicadores (Moran et al., 2007). Ésta ha sido estudiada en Estados Unidos de Norteamérica, con los diferentes grupos étnicos, culturas y subculturas; se ha estudiado también su relación con el desempeño en pruebas neuropsicológicas (Coffey et al., 2005; Manly, Byrd, Touradji, & Stern, 2004; Touradji, Manly, Jacobs, & Stern, 2001).

Cabassa (2003), realiza una revisión de las formas de evaluación de la aculturación en hispanos y mexicanos en E. U. A. mediante escalas subjetivas como la *Bidimensional Acculturation Scale for Hispanics (BAS)* y la *Acculturation Rating Scale for Mexican Americans-Revised (ARSMA-II).* Otra situación importante es la aculturación que ocurre en las fronteras, en donde la cercanía geográfica, intercambio económico, flujo de personas de un país a otro genera una cultura similar y muy influyente entre dos países, tal es el caso de las personas

hispanas o mexico-americanas que viven en las fronteras de Estados Unidos con México (Guinn, Vincent, Wang, & Villas, 2011) y que son una de las poblaciones focales del presente estudio. En el estudio realizado por Arnold, Montgomery, Castañeda, y Longoria (1994) se evaluó el funcionamiento de personas de esta población en la batería neuropsicológica Halstead Reitan, encontrando un funcionamiento bajo en casi todas las sub-pruebas.

2.2.3 Uso de tests adaptados para diferentes idiomas y/o culturas

Como se mencionó anteriormente, la evaluación psicológica es una herramienta fundamental en el trabajo de los psicólogos, que ayuda a comprender a los pacientes, proveyendo información importante acerca de sus habilidades, inteligencia, personalidad e incluso funcionamiento cognitivo y neurológico (Judd et al., 2009). Sin embargo, la evaluación psicológica para ser efectiva requiere desarrollar instrumentos y procedimientos que sean válidos y confiables para asegurar que las conclusiones a las que lleguemos una vez que los utilicemos, sean aceptadas o por lo menos medianamente aceptadas en el ámbito científico y clínico.

Sin embargo, a pesar del trabajo que su construcción representa; el uso de tests o pruebas es universal, se utilizan prácticamente en todos los países, con individuos desde recién nacidos hasta la edad adulta. Su aplicación en las ciencias

de la conducta permite describir una gran cantidad de cualidades, conductas presentes e incluso prever conductas futuras. Es de esperar que en un mundo globalizado esto permite realizar estudios comparativos entre diferentes poblaciones, para conocer sus diferencias y similitudes no solamente en el campo de la psicología, sino educativo (Hambleton, 1996), salud (Arnold & Matus, 2000; Dunckley, Hughes, Addington-Hall, & Higginson, 2003; Hilton & Skrutkowski, 2002; Wang, Lee, & Fetzer, 2006; Wild et al., 2005), de mercados (Dolnicar & Grün, 2013; Squires et al., 2014), que son considerados como fuente de información para la toma de decisiones políticas (Quaranta, 2013), sociales, económicas, no sólo a nivel local, sino internacional (Harkness, 2006; International Standards Organization, 2006), por lo que se requiere contar con instrumentos que se puedan aplicar en las diferentes poblaciones, partiendo de los mismos presupuestos o conceptos teóricos, para así poder realizar las contrastaciones más fácilmente que si se tuviera una prueba distinta en cada idioma o cultura. Para una revisión a fondo de otras metodologías para estudios transculturales en diversos ámbitos se puede referir al libro: *"Survey methods in multicultural, multinational, and multiregional contexts"* (Harkness et al., 2010).

2.3 Directrices para seleccionar, construir o aplicar tests psicológicos y educativos

En vías de eliminar los sesgos en la evaluación psicológica y educativa transcultural, la Asociación Americana de Investigación Educativa (*American Educational Research Association*, AERA), la Asociación Americana de Psicología (*American Psychological Association*, APA) y el Consejo Nacional de Evaluación en la Educación (*Nacional Council on Measurement in Education* NCME), de los Estados Unidos, en 1953 elaboraron directrices para aquellos psicólogos encargados de seleccionar, construir o aplicar tests psicológicos y educativos *Standards for educational and psychological tests* (AERA, APA, NCME,& JCSEPT, 1999) en su versión más reciente, son de gran importancia para el contexto de la adaptación y uso de los mismos.

Ese documento se encuentra dividido en tres apartados, el primero de ellos referente a los requisitos para la construcción, evaluación y documentación de pruebas incluyendo en uno de los capítulos lo referente la equiparabilidad de las puntuaciones; en la parte dos, lo concerniente a la evaluación justa; incluyendo un capítulo sobre la evaluación de individuos de diversos ambientes lingüísticos, en el tercer apartado las aplicaciones de los test para diferentes campos: educativo, psicológico, selección, (AERA, APA, NCME,& JCSEPT, 1999).

Por otra parte, en los esfuerzos internacionales, en 1994, la Comisión Internacional de Tests (*International Test Commission*), mejor conocida como ITC, en conjunto con otras siete organizaciones internacionales publicaron las Normas para la adaptación de instrumentos de evaluación psicológica de un idioma y cultura a otro idioma y cultura. Estos lineamientos actualmente en su segunda edición (International Test Commission, 2010), se han convertido en un referente indispensable cuando se utilizan pruebas para tratar de evaluar individuos de culturas diferentes, y las consideraciones que se deben tener para garantizar que las pruebas adaptadas cumplen los requisitos para una evaluación equitativa.

Estos organismos establecen que cuando se hacen cambios sustanciales en el formato del test, modo de aplicación, instrucciones, idioma o contenido, el usuario debería de revalidar el test para las nuevas condiciones, o en su defecto tener argumentos que apoyen que no es necesaria o posible una validación adicional. También determinan que cuando se pretenda que dos versiones de un test en idiomas distintos sean comparables, es preciso aportar pruebas acerca de la comparabilidad de los tests, equivalencia de las mediciones y puntuaciones, entre otras.

Los estándares de la ITC (2010), proporcionan un marco de referencia al analizar las fuentes de error o invalidez que pueden surgir en la adaptación de los test y presenta un total de 22 directrices divididas en 4 apartados: a) Contexto, b)

Construcción y/o adaptación de pruebas, c) Administración e d) Interpretación de los resultados.

Para efectos del presente documento, presentaremos las recomendaciones para adaptación de pruebas para diferentes culturas, en cuatro aspectos: 1) Evitar errores relacionados con las diferencias culturales e idiomáticas, 2) Interpretación justa de los resultados y 3) Aspectos técnicos y métodos, 4) Comprobar la equivalencia de las mediciones. A continuación, haremos una descripción de los mismos, centrándonos principalmente en el de los aspectos técnicos y métodos, ya que son objetivo del presente estudio.

2.3.1 Evitar errores relacionados con las diferencias culturales e idiomáticas

La evaluación e interpretación de los resultados interculturales no se debe limitar a la traducción o adaptación de los test, sino que se debe ampliar a todo el proceso de evaluación, incluyendo en esto la aplicación de la prueba, la selección de los evaluadores, diferencias lingüísticas, etc.

En un estudio reciente, realizado por Ostrosky-Solis, Gutierrez, Flores, y Ardila (2007), en el que se encontraron diferencias en el rendimiento de sujetos en

las pruebas de fluidez verbal, proponen que estas diferencias pueden deberse a los procedimientos de aplicación y evaluación, y no solamente a cuestiones lingüísticas en sujetos que utilizan el mismo idioma pero que pertenecen a países distintos, por lo que recomiendan la adaptación de cada prueba a utilizar para cada población en la que se vaya a utilizar.

2.3.2 Interpretación de los resultados justa

En los estudios interculturales la finalidad de la evaluación es proporcionar herramientas para establecer comparaciones entre grupos culturales e idiomáticos; empero, la información disponible no debería ser utilizada para apoyar ideologías de superioridad e inferioridad entre países o con fines de competencia (Westbury, 1992). En caso de encontrar diferencias entre culturas o idiomas, deberían analizarse los diversos factores que pueden estar influyendo ajenos al test o medidas de evaluación, entre ellos podríamos enumerar las diferencias entre países en cuanto a currícula, políticas educativas, nivel de vida, valores, nivel socioeconómico, por mencionar algunas (Stedman, 1994)

Van de Vijver & Poortinga (1991), afirman que en ocasiones el rendimiento obtenido tiene poca relación con las aptitudes en una prueba de conocimientos, afirmando que no se puede asumir de primera instancia que todos los examinados

tratan de obtener siempre las mejores puntuaciones; ellos observaron, por ejemplo, que los estudiantes negros en Sudáfrica intentan solo obtener la puntuación mínima necesaria para aprobar un test, esto debido probablemente a que el sistema educativo estatal impuesto es percibido por ellos como discriminativo con los negros, por lo cual los estudiantes solo aspiran a lograr lo mínimo que se les exige.

Si tratáramos de comparar los resultados en pruebas de conocimientos en países desarrollados y en vías de desarrollo, probablemente encontraríamos que los resultados obtenidos no estén relacionados con sus aptitudes, sino con el acceso que tengan a los recursos, calidad de los servicios educativos, la currícula (van de Vijver & Tanzer, 2004)

Para una revisión a profundidad, sobre los usos varios de las puntuaciones de las pruebas, así como la relación con las cuestiones de sesgo y evaluación justa, ver Hambleton (1996; 2001) y Hambleton, Merenda y Spielberger (2004)

2.3.3 Aspectos técnicos y métodos que garanticen la validez

Es de suma importancia tomar en cuenta todos los factores técnicos que pueden influir en la validez de una prueba adaptada, Hambleton (2001) tener

especial cuidado en cinco aspectos: el propio test, el proceso de traducción**,** la elección y capacitación de los traductores, la metodología para la adaptación de pruebas y los procedimientos para establecer su equivalencia.

2.3.3.1 Características de la prueba

La elección de los ítems, su formato, el vocabulario utilizado, el tipo de materiales utilizado, etc. pueden ser piezas decisivas para la validez de una prueba que pretende ser utilizada en diferentes culturas, por lo cual, si se pretende de antemano extender su uso en diferentes grupos idiomáticos y/o culturales es conveniente minimizar estos efectos desde el momento de su construcción para evitar problemas posteriores.

2.3.3.2 El proceso de la traducción

Realizar una buena traducción es un factor fundamental para conseguir un instrumento que se corresponda con el instrumento original. La calidad de la traducción es evaluada por diferentes tipos de equivalencia entre la versión original y adaptada del cuestionario (Behling & Law, 2000). Mientras más riguroso es el proceso de traducción, más probable será que resulten equivalentes las dos pruebas (Maneesriwongul & Dixon, 2004; Stansfield, 2003). Conviene documentar

cómo se ha realizado el proceso de traducción. Hay que incluir todas las cuestiones referidas al formato utilizado, la forma de administración y la selección de los traductores. En la traducción inversa (*back translation*), se considera que existe equivalencia entre las dos pruebas cuando no existen diferencias sustanciales entre ambas versiones, esta es la técnica más habitual para garantizar una traducción exitosa, aunque por sí sola no constituye una garantía, existiendo otros diseños alternativos. En el proceso de traducción hay que cuidar los aspectos relacionados con la selección y competencia de los traductores y las estrategias de traducción:

2.3.3.3 Selección y competencia de los traductores

A pesar de la obvia necesidad de utilizar traductores competentes, en muchas investigaciones por diversos factores se utiliza a un solo traductor o traductores a los que se tiene fácil acceso: un amigo, conocido, etc.; sin embargo, si lo que se quiere es tener un buen resultado Hilton y Skrutkowski (2002) recomiendan al menos dos traductores, uno de los cuáles traduce de la versión original al lenguaje objetivo; el otro traductor devolverá la versión traducida al idioma original.

2.3.3.4 Estrategias de traducción

En la literatura se suele hacer referencia a dos tipos de traducciones: directa e inversa; la traducción directa es la forma más sencilla de realizar el proceso de traducción. Consiste en que un traductor traslade un test desde el lenguaje original a otro lenguaje en el que dicho traductor es competente. Este procedimiento tiene la ventaja de que es menos costoso y mucho más rápido. Sin embargo, no aporta ninguna prueba que garantice la equivalencia entre las pruebas. Por ello es que es más aceptado el procedimiento de traducción inversa (Brislin, 1986), debido a que el investigador obtiene dos versiones de la prueba al final del proceso y puede establecer comparaciones con objeto de identificar inconsistencias, significados distintos, entre otros. Cualquier diferencia que se encuentre debe consultarse con los traductores con objeto de discernir cuáles son las razones del cambio y cómo puede corregirse el instrumento (Brislin, 1970; Sireci, Yang, Harter, & Ehrlich, 2006).

2.3.4 Comprobar la equivalencia de las mediciones

Una traducción correcta solo es el punto de partida para realizar una buena adaptación de la prueba, lo que es un proceso caro en tiempo y esfuerzo; en este capítulo se ha preferido el término de adaptación al de traducción, ya que como

indica Hambleton (1996) es más amplio y representativo del proceso que se realiza cuando se utiliza una prueba en otro país y/o idioma. La adaptación de un test incluye varias tareas, tales como la realización de estudios piloto para identificar el concepto en la población a la que se quiere adaptar, evaluar si los ítems de la prueba presentan el mismo significado en la cultura original y en la objetivo, hasta la selección de traductores, etc. Por lo tanto, el proceso de traducción es sólo una parte del procedimiento general de adaptación (Hambleton et al., 2004; Sireci et al., 2006).

Se podría afirmar que un alto porcentaje de investigaciones transculturales son defectuosas, e incluso no válidas, debido a una adaptación deficiente de las pruebas utilizadas. Las pruebas transculturales ideales deberían separar la varianza debido a las diferencias verdaderas en el fenómeno de interés de la varianza debida a las diferencias culturales y lingüísticas (Pedraza & Mungas, 2008; Siedlecki et al., 2010).

Muñiz, Elosua y Hambleton (2013), afirman que en los últimos años se han dado avances importantes en el campo de la adaptación de los tests, tanto metodológica como psicométricamente (Hambleton, Merenda, & Spielberger, 2005; Matsumoto & Van de Vijver, 2010; Van De Vijver, 2013), siendo un campo importante el relacionado a los aspectos metodológicos y técnicos en el

establecimiento de la equivalencia intercultural (Byrne & Campbell, 1999; Byrne et al., 2009; Elosua & López-Jáuregui, 2008).

Las directrices de la ITC 2010, reúnen las pautas a seguir para asegurar el máximo nivel de equivalencia entre las versiones original y adaptada de un test, incluyendo entre muchos otros: la valoración del constructo en la población diana, los estudios de equivalencia y la delimitación del grado de comparabilidad entre puntuaciones (Muñiz et al., 2013). Aunque existe un repertorio amplio de técnicas estadísticas, las más habituales son el análisis factorial (confirmatorio y exploratorio) para determinar la equivalencia conceptual y distintas técnicas para evaluar el funcionamiento diferencial de los ítems en los distintos grupos de comparación; sin embargo, como plantea Xi (2010), la mayoría de los estudios para la adaptación de pruebas interculturales, se enfocan en sólo un aspecto de los que conforman la equidad de un test.

2.3.4.1 Equivalencia, equivalencia de medida e invarianza

El concepto de equivalencia no tiene un significado unánime dentro de la literatura; siguiendo a T. P. Johnson (1998), éste enumera un total de 52 tipos de equivalencias que han sido descritos en el campo de la evaluación intercultural. Como mencionan Peters y Passchier (2006) por otra parte, en el trabajo de

Herdman y colaboradores, se revisan hasta 19 tipos diferentes de equivalencia, lo que nos llevaría a pensar en que una equivalencia total, no se puede lograr; sin embargo, al adaptar o traducir instrumentos se debe tratar de llegar a la mayor equiparabilid posible (Herdman, Fox-Rushby, & Badia, 1997). Dentro del campo de la adaptación de pruebas, se han propuesto distintas clasificaciones (Guillemin, Bombardier, & Beaton, 1993; Hilton & Skrutkowski, 2002); no obstante, nos centraremos en la equivalencia conceptual y métrica.

La equivalencia conceptual hace referencia a la existencia de los conceptos que mide el test en las distintas culturas a estudiar. Por tanto, es necesario presentar evidencia empírica de que una escala mide el mismo constructo, de la misma forma (métrica) cuando se administra a dos o más grupos distintos (Cheung & Rensvold, 1999), esto es conocido en el lenguaje técnico como “equivalencia de medida” o “invarianza de las mediciones”.

Para comparar grupos de individuos en cuanto a sus niveles en algún constructo, o respecto a las relaciones entre esos constructos, se debe asumir que los instrumentos utilizados en la evaluación tienen “equivalencia de medida” o “invarianza” entre los grupos (Drasgow, Levine, & McLaughlin, 1987). De no ser así, las diferencias entre los grupos en medias o en los patrones de las correlaciones son potencialmente artificiales y pueden ser sustantivamente erróneas.

2.3.4.2 Justicia en las mediciones, invarianza, sesgo y DIF

En la literatura se utilizan algunos términos, que si bien no corresponden exactamente al mismo concepto de justicia en las evaluaciones, sí tienen una clara relación con la misma, por ejemplo equivalencia de medida, el sesgo y Funcionamiento Diferencial del Ítem, conceptos íntimamente ligados y que serán ampliamente descritos en el siguiente capítulo (Byrne et al., 2009; Kunnan, 2007; Zumbo, 2007).

Como mencionábamos al principio del capítulo, la evaluación psicológica como cualquier otra, debe ser lo más objetiva y justa posible, ya que los resultados obtenidos pueden utilizarse en una diversidad de escenarios. Sin embargo, el garantizar que una prueba esté libre de sesgos y sobre todo que sea equitativa cuando de tratar de evaluar a personas de distinta procedencia cultural se trate, la tarea se vuelve aún más compleja. Una perspectiva que puede resultar fructífera a la hora de analizar los motivos por los que sucede el funcionamiento diferencial es la perspectiva multidimensional, que considera que el DIF se produce cuando hay ítems multidimensionales en un test que pretende ser unidimensional y existen diferentes distribuciones entre grupos en alguno de los constructos que no se pretenden medir (Ackerman, 1992; Shealy & Stout, 1993).

El análisis de la equivalencia de medida de un test, por tanto, es parte sustancial del análisis de la validación de las puntuaciones al aplicar el instrumento de medida en cuestión. Para asegurar la equidad de las puntuaciones de sujetos que pertenecen a distintos grupos, éstas tienen que depender únicamente del nivel del sujeto en el constructo medido, si se encuentra que existen variaciones una vez que se han controlado las variables mediadoras, es porque está midiendo algo más que el constructo objetivo; es decir, el test o el ítem no es unidimensional.

Como plantean Prieto y colaboradores: “Sería imposible comparar la estatura de los varones y las mujeres si el metro no tuviese las mismas propiedades en ambas poblaciones... La ausencia de funcionamiento diferencial de los ítems (DIF) es una condición de la invarianza métrica entre poblaciones. ” (Prieto, Delgado, Perea, & Ladera, 2011).

2.4 Ventajas de la adaptación de tests

Si pensamos en la rapidez con la que circula la información, a través de los medios de comunicación internacionales y de internet, nos daremos cuenta que ya no hay fronteras para el conocimiento y la información; esto ha influido para que en que los últimos años exista un interés creciente por los estudios internacionales

comparativos en los ámbitos educativos (Hambleton, 1996), en el campo de la salud (Hilton & Skrutkowski, 2002; Wang et al., 2006), laborales y de mercados (Dolnicar & Grün, 2013; T. P. Johnson, 1998; Squires et al., 2014), etc.

Hay varias razones para adaptar los tests, pero la más importante es el interés que han mostrado distintos países por establecer estándares educativos, de salud, políticos y económicos internacionales; el objetivo es comparar sus progresos con los de otros, que les permita una toma de decisiones rápida y eficaz. Para llevar a cabo de forma rigurosa estos trabajos inter-idioma o cultura es imprescindible contar con instrumentos de medición que permitan establecer las equivalencias correspondientes (Barbero García, Vila Abad, & Holgado Tello, 2008; Harkness, 2006; Quaranta, 2013). La forma más común de hacerlo es adaptar las pruebas ya existentes a los idiomas o culturas objetivo que se desea investigar (Grigorenko, 2009).

También es cierto que existen ocasiones en las que resulta más barato y rápido adaptar un test ya existente que construir uno nuevo. Tal es el caso de la escala de ansiedad-rasgo de Spielberg, la cual ha sido adaptada para su uso en más de cincuenta países. El uso de un instrumento previamente desarrollado y validado tiene la ventaja de ahorrar tiempo y energía, pero además facilita la construcción de conocimiento trans-cultural. El uso del mismo instrumento puede unificar la conceptualización del fenómeno estudiado entre investigaciones

distintas, pudiéndose comparar los resultados (Brislin, 1986). Si se utiliza el instrumento en una población lingüísticamente diferente, la adaptación de dicho test se torna en un paso crítico. Los errores en la adaptación pueden distorsionar la idea original del test y comprometer la validez y fiabilidad del nuevo instrumento (Arffman, 2013; Chang, Chau, & Holroyd, 1999; Sireci, 1997).

Otra razón para adaptar los tests se relaciona con la competencia lingüística de los sujetos evaluados; Oller y otros (Gunnarsson, 1978; Oller & Perkins, 1978; Robinson, 2010) han defendido que los tests de inteligencia, personalidad y actitudes plantean cuestiones muy similares y que realmente evalúan las habilidades lingüísticas de una persona que no es completamente competente con el idioma. Los estudios para seleccionar alumnos en secundaria en USA (Alderman, 1982), han mostrado que la capacidad lingüística es una variable moderadora de las puntuaciones obtenidas en el test. Esto es, la puntuación obtenida por los estudiantes en la prueba es principalmente dependiente de su competencia en el manejo del idioma. Por tanto, mientras no se controle esta variable, lo que miden muchas pruebas es la competencia lingüística de los sujetos y no el constructo que se pretenda medir con el test (Sireci & Allalouf, 2003).

Una razón más mencionada por Hambleton (1996) y Hambleton y Patsula (1999) es que en el idioma focal no siempre existe la experiencia técnica para

construir un nuevo test validado convenientemente, lo cual suele suceder en los países del tercer mundo, por lo cual los profesionales e investigadores de estos países se sientan más seguros adaptando una prueba que ya goza de prestigio.

Como plantean Muñiz, Elosúa y Hambleton (2013):

> "… una revisión de los veinticinco tests más utilizados en la práctica profesional española (clínica, educativa u organizacional) deja patente que de ellos diecisiete son adaptaciones de versiones construidas en otro idioma, en su mayoría del inglés" (Muñiz et al., 2013).

2.5 Conclusiones

El proceso de adaptación de una prueba psicológica tiene distintas interpretaciones. Hay ocasiones en las que se entiende que este proceso supone simplemente una mera traducción de la prueba desde el idioma original a otro. Sin embargo, la adaptación de un test implica que el test creado en el lenguaje objetivo es equivalente al que previamente se había creado en el lenguaje fuente. Cualquier diferencia entre la cultura de los sujetos o el manejo del idioma puede provocar sesgos que invaliden los resultados.

Al utilizar un test, se debe estar consciente de los aspectos culturales específicos que pueden influir en las puntuaciones de las pruebas y tenerlos en cuenta antes de utilizarlo en una nueva cultura. Si se quiere lograr una internacionalización de la psicología o como plantea Van de Vijver (2013) hacer de la psicología una disciplina global e inclusiva, es necesario realizar una revisión de las teorías, enfoques, metodologías de evaluación, a fin de garantizar que estas sean aplicables y libres de sesgos en diferentes países (Van de Vijver & Poortinga, 2002).

Como vimos, es un hecho que existe la necesidad de adaptar o traducir pruebas de un idioma o cultura a otro; hemos analizado las principales ventajas que ello implica: vimos que generalmente es más barato adaptar que elaborar una prueba nueva, que permite realizar comparaciones más justas entre culturas, que si se utilizaran pruebas distintas, etc; sin embargo, a pesar de las ventajas que ello representa esto es algo que no siempre se cumple en la práctica.

En el siguiente capítulo se hará una revisión de los distintos procedimientos para la evaluación del Funcionamiento Diferencial del Ítem (DIF, por sus siglas en inglés: *Differential Item Functioning*), procedimiento estadístico, ampliamente utilizado para evaluar si una prueba presenta funcionamiento distinto entre diferentes poblaciones.

CAPITULO III

FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM (DIF)

3.1 Introducción

En el presente capítulo se hará una revisión del tema Funcionamiento Diferencial del Ítem o DIF por sus siglas en inglés: *Differential Item Functioning.* Se iniciará con una introducción conceptual e histórica del surgimiento de esta metodología, desde sus inicios como parte de los estudios del "sesgo" y la discriminación, en una época de gran auge de la psicometría y la Teoría Clásica de los Tests, las técnicas pioneras en este campo, para ir avanzando en el entendimiento de las principales técnicas estadísticas que se utilizan para su estudio, así como sus principales ventajas y desventajas.

Posteriormente se hará una revisión un poco más detallada de los procedimientos a utilizar en el presente estudio: el Mantel-Haenszel y el de Regresión Logística.

Se pasará a una revisión breve de la Teoría de Respuesta al ítem y los principales procedimientos que de ella se desprenden en el estudio del DIF.

Para cerrar el capítulo se hará una revisión de algunos problemas y cuestiones en la práctica de la evaluación del DIF, el tamaño de la muestra, procesos de purificación y procedimiento general, que pueden ayudar al investigador a decidir el procedimiento de evaluación del DIF que sea más conveniente utilizar en su estudio.

3.2 Historia y semántica

El Funcionamiento Diferencial del Ítem (DIF) es uno de los campos de investigación que más interés ha suscitado en los últimos años en el campo de la psicometría. Se suele señalar el inicio de este interés en los años cincuenta, con las investigaciones realizadas en la Universidad de Chicago por Eells y colaboradores en 1951, ya que fueron de los primeros estudios en los que se mostraban las diferencias en los resultados obtenidos en diversas pruebas de inteligencia por sujetos de diferentes grupos (Eells, Davis, Havighurst, Herrick, & Tyler, 1951).

Inicialmente se intentó dar explicación a éstas diferencias basándose en que los diferentes grupos evaluados tenían características diferentes en cuanto a nivel cultural, clase social, raza u otras y que esas diferencias podrían ser las responsables de las variaciones en las pruebas. Sin embargo, es hasta los años sesenta, coincidiendo con los movimientos de los derechos civiles en los E.U.A. cuando se dio mayor importancia a la posibilidad de existencia de *"sesgo cultural de los tests"*, esto debido a que los tests eran utilizados como principal fuente de información para la toma de decisiones en una gran cantidad de escenarios, por ejemplo: para la admisión en la educación superior, oportunidades de empleo, promoción en el trabajo y demás cuestiones en las cuales se buscaba lograr la igualdad de derechos y oportunidades para los grupos sociales menos favorecidos y minorías. Es en ese momento cuando los tests que estaban tan de moda, dejaron de ser vistos como instrumentos neutrales y se cuestionó si no serían instrumentos para discriminar a ciertos grupos socioeconómicos y raciales, ya que eran fabricados por la clase y raza dominante económica y políticamente.

En este contexto, el término sesgo *(bias)* se equiparaba frecuentemente con el concepto de injusticia *(unfair),* partiendo del supuesto trascendental de que todos los hombres son iguales, al comprobarse empíricamente que no todos los grupos humanos tienen el mismo desempeño promedio en las pruebas, se concluía por tanto que, los test son injustos al hacer a los hombres que son iguales, desiguales. De esta forma, cualquier test en el que se encontrasen

diferencias entre grupos étnicos, culturales o socioeconómicos, era considerado sesgado e injusto.

Hasta los años setenta, la investigación sobre el sesgo era realizada por sociólogos, antropólogos y educadores, se carecía de reglas claras para distinguir cuando el funcionamiento diferencial en las pruebas era producto de las diferencias reales en el rasgo psicológico evaluado y cuando esas variaciones eran provocadas por las diferencias culturales de los grupos. Es en esa década, cuando la psicometría aborda esta problemática, estableciendo términos y proponiendo técnicas analíticas para intentar solucionar estas interrogantes. Es así como el término *"sesgo"*, ligado a una carga social y política es sustituido por el nombre técnico preferido actualmente, el cual fue acuñado por Holland y Thayer en 1988: "*Differential Item Functioning (DIF)",* que en español se ha traducido a *"Funcionamiento Diferencial del Ítem"* (Holland & Thayer, 1988).

En la actualidad al decir que un test presenta DIF, se hace referencia a que presenta propiedades estadísticas distintas en cada grupo. Cuando es el test en su conjunto el que muestra propiedades estadísticas distintas entre grupos, se habla de "*Differential Test Functioning"* DFT o "*Funcionamiento Diferencial del Test"* (FDT) por su traducción en español. Obviamente este proceso implicó para la comunidad psicométrica un gran trabajo en torno al desarrollo de procedimientos y técnicas estadísticas que permitieran comprobar la existencia de

sesgo en los ítems y en los tests; algunos de los cuales se presentarán más adelante (Holland & Wainer, 2012).

3.3 ¿Por qué se presenta el DIF?

El *sesgo* ocurre cuando los ítems que son planteados a los sujetos evaluados de un grupo específico presentan dificultades, fundamentalmente de carácter lingüístico o culturales, que son irrelevantes de cara a la medición del constructo que se pretende estudiar, pero que influyen en la menor posibilidad de contestar acertadamente el ítem y por tanto influyen en la puntuación total de la prueba (Camilli & Shepard, 1994).

El sesgo también es una cuestión a tener en cuenta cuando un test es traducido o adaptado desde el lenguaje de una cultura al lenguaje de otra como se vio en el capítulo anterior, una traducción literal del ítem o directa no es una garantía de que en el idioma objetivo tendrá el mismo significado para los evaluados.

A partir de lo dicho con anterioridad, se podría pensar que todos los tests o ítems que muestren diferencias entre grupos funcionan diferencialmente y esto no es así. Una vez que se ha identificado, el DIF puede atribuirse al *sesgo del ítem* o

al *impacto del ítem*. Se define el *sesgo del ítem* como la invalidez o el error sistemático producido por un test a la hora de medir a un grupo particular (Camilli & Shepard, 1994), este error es sistemático porque distorsiona constantemente la actuación de los miembros del grupo. Se considera que un ítem está sesgado cuando favorece a un determinado grupo gracias a factores distintos a los presentes en el constructo que se pretende medir. Por el contrario, las diferencias obtenidas entre grupos debidas al nivel del constructo que se está midiendo es el *impacto del ítem*. El impacto es constante para los miembros de un determinado grupo, pero estos efectos reflejan diferencias reales en la variable que se intenta medir (Dorans, 2013). Suponiendo por ejemplo, que los hombres tienen una mayor capacidad espacial que las mujeres, de ser cierto, los hombres tendrían por término medio, puntuaciones superiores en las pruebas de aptitud espacial y tendrían también una probabilidad mayor que las mujeres de acertar los ítems que conforman estas pruebas.

La relación entre DIF, sesgo e impacto es metodológica: se utilizan diferentes análisis estadísticos para identificar ítems con DIF (Shealy & Stout, 1993), que es lo que veremos en el apartado siguiente; sin embargo, los evaluadores han de decidir si ese DIF es atribuible al sesgo o al impacto del ítem en un grupo específico, evaluando en ese sentido la validez de constructo de cada prueba, procedimiento que se apoya principalmente en el sustento teórico en la construcción del instrumento.

3.4 Tipos de DIF

Dentro del estudio del funcionamiento diferencial de los ítems cabe distinguir entre *DIF uniforme y no uniforme* (Mellenbergh, 1982). El *DIF uniforme* se presenta cuando la probabilidad de contestar correctamente un ítem es mayor para un grupo que para otro a través de todos los niveles de habilidad. Por otra parte, el *DIF no uniforme* se produce cuando la diferencia en la probabilidad de responder acertadamente a un ítem entre dos grupos no es la misma en todos los niveles de habilidad. En estos casos no se puede hablar de DIF contra un grupo, ya que, para determinados niveles de habilidad, la probabilidad de acertar el ítem, a igual nivel en la variable medida, es mayor para un grupo, en tanto en los otros niveles es mayor para el otro.

Siguiendo la explicación de Pedraza y Mungas (2008) desde la Teoría de Respuesta al ítem, se propone el concepto de la curva característica del ítem (CCI), que permite entender de una manera gráfica muy sencilla los distintos tipos de DIF, cuestiones que se verán a detalle más adelante. En la **figura 1** tenemos unas CCIs características de dos ítems sin DIF, en la **figura 2** del DIF uniforme, mientras que en la **figura 3** tenemos un caso de DIF no uniforme. En ellas se observa la habilidad, dificultad y la discriminación del ítem.

Figura 1. *CCIs ítems sin DIF*

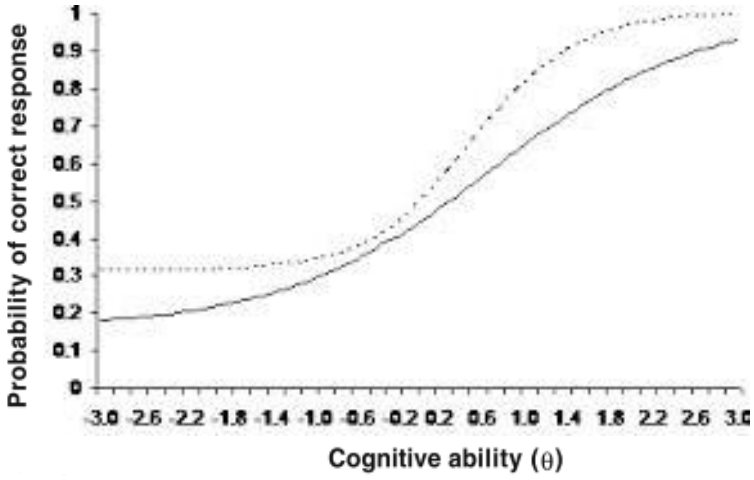


Figura 2. *CCIs con DIF uniforme*

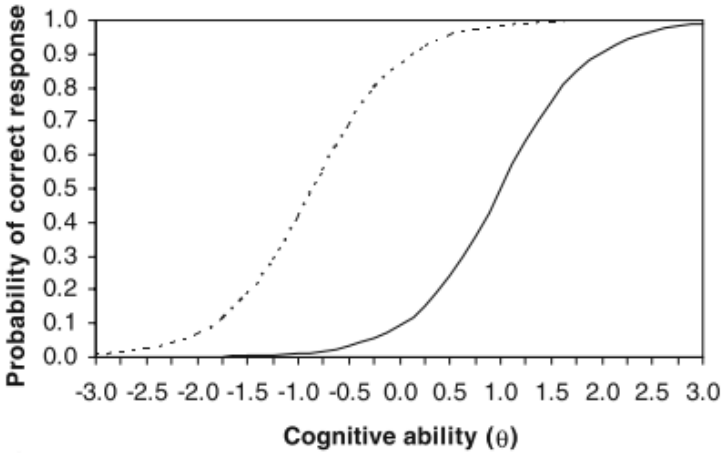
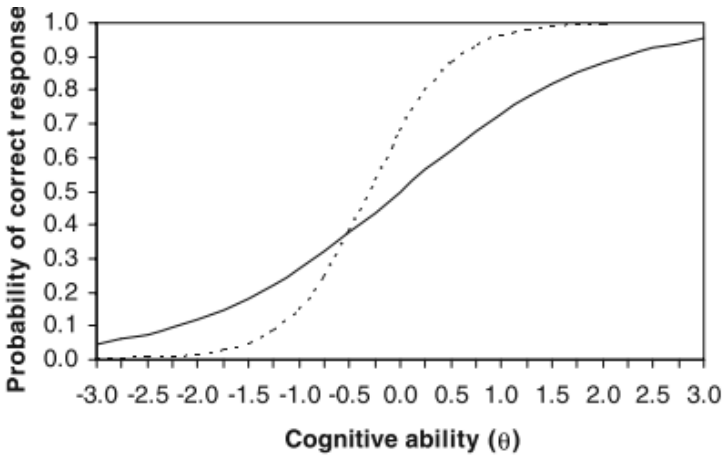


Figura 3. *CCIs con DIF no uniforme*



Figuras 2 y 3. CCI´s con DIF uniforme y CCI´s con DIF no uniforme, reproducidas con autorización de la editorial de: Pedraza & Mungas, 2008. Measurement in Cross-Cultural Neuropsychology. *Neuropsychology review.* Copyright 2014.

En los casos en que no existe funcionamiento diferencial del ítem estas curvas deberán tener una forma muy similar como lo muestra la Figura 1, mientras que para aquellos ítems con DIF uniforme, sujetos con un mismo nivel de habilidad tendrán menos probabilidad de acertar al ítem de manera constante que los sujetos con ese mismo nivel de habilidad del otro grupo lo indicado por la Figura 2, por lo que se observarán curvas separadas, por último en la figura 3 vemos que las curvas se cruzan, indicando que mientras en algunos niveles de habilidad los sujetos de un grupo tienen mayor probabilidad de acertar al ítem,

mientras que en otros niveles tienen menor probabilidad, como se puede observar en la Figura 3 (Pedraza & Mungas, 2008).

El DIF no uniforme a su vez se ha subclasificado en dos tipos: el *DIF no uniforme simétrico* el cuál, siguiendo con el ejemplo de las CCI´s, se vería representado por un cruce central de la curva característica del ítem (CCI) del grupo de referencia y del grupo focal dentro del rango establecido, mientras que el *DIF no uniforme mixto*, por una intersección no simétrica de la CCI del grupo de referencia y del grupo focal. Posteriormente, Li y Stout acuñaron los términos "*no direccional*" (DIF no uniforme simétrico) y "*unidireccional*" (DIF no uniforme mixto) (Li & Stout, 1993), para las dos formas de DIF no uniforme (Finch & French, 2007; Narayanan & Swaminathan, 1996).

3.5 Clasificación de las técnicas para detectar el DIF

3.5.1 Técnicas para ítems dicotómicos o politómicos

Una forma de clasificación de las técnicas DIF, es separarlas entre aquellas aplicables para ítems dicotómicos y aquellas para escalas likert o politómicos. Como plantean Elosua y López-Jáuregui (2007), la literatura está repleta de estudios sobre métodos de detección del DIF en ítems de respuesta dicotómica

Sin embargo, es menor el número de trabajos destinados a profundizar en los métodos de detección aplicables a formatos de respuesta ordinal o escala Likert.

Los procedimientos de detección de DIF en ítems dicotómicos comparan las respuestas dadas a un ítem por sujetos que provienen de dos grupos (referencia/focal) y tienen el mismo nivel en la variable medida (puntuación total o nivel de habilidad estimado); sin embargo, como solamente existen dos opciones posibles (acierta o no acierta, obtiene puntaje o no lo obtiene), los resultados y puntuaciones globales son más sencillos de evaluar.

Los procedimientos aplicables a ítems politómicos son más complejos, debido en primer lugar a que el formato de respuesta ordinal tiene más categorías que el dicotómico, lo cual dificulta la comparación de las respuestas dadas al ítem; esta comparación podría llevarse a cabo teniendo en cuenta la media aritmética del ítem, o teniendo en cuenta las diferencias asociadas con cada una de las opciones de respuesta o sobre todas ellas conjuntamente (French & Miller, 1996; Zwick & Thayer, 1994). En segundo lugar, la utilización de ítems de respuesta ordenada amplía el rango de puntuaciones utilizado para crear los niveles de habilidad necesarios para emparejar sujetos antes de poder ser comparados (Elosua & Wells, 2013).

Como apuntan Barbero García, Prieto y San Luis (2000), el tipo de formato de respuesta, ha motivado que en los últimos años aparezcan nuevas propuestas para la evaluación del DIF, o extensiones de los procedimientos ya existentes ajustados al nuevo tipo de formato. Ejemplo de ello son Zwick y colaboradores (1992), quienes presentaron dos extensiones del estadístico Mantel-Haenszel: el procedimiento Mantel-Haenszel generalizado y el procedimiento Mantel-Haenszel politómico, los trabajos de Penfield (2007) con los procedimientos P1 y P2, como él mismo los denominó, considerándolos equivalentes a los modelos de Mantel-Haenszel propuestos para ítems dicotómicos por la *Educational Testing Service.* Dentro del marco de la TRI, se proponen generalizaciones del estadístico χ^2 de Lord y de las medidas exactas de área de Raju propuestas por Cohen, Kim y Baker (1993), para el caso en que los ítems se ajusten al modelo de respuesta graduada (Kim & Cohen, 1995; Kim, Cohen, & Park, 1995).

En la presente revisión teórica, se presentarán principalmente las técnicas para ítems dicotómicos, ya que son las que se utilizarán para efectos de esta investigación.

3.5.2 Métodos condicionales o incondicionales

Esta categorización fue aportada por Mellenbergh (1989), al revisar la evolución histórica del DIF, las técnicas pioneras, se les clasifica dentro de los

métodos incondicionales, estaban basados en las diferencias en dificultad del ítem en cuestión; en ellas no se igualaban los grupos con respecto al nivel del rasgo medido, por lo que solían confundir el DIF con la discriminación del ítem, razón por la cual se ha restringido su uso. Ejemplo de éstas técnicas son los procedimientos basados en análisis de la varianza (ANOVA) o el método Delta-plot de Angoff, reemplazándose por los métodos condicionales en los que, al emparejar los niveles del rasgo evaluado en los grupos, éstos pueden considerarse realmente comparables y por tanto, permiten distinguir entre DIF e impacto (Van Der Flier, Mellenbergh, Adèr, & Wijn, 1984).

3.5.2.1 Métodos de invarianza condicional observada o no observada

Los métodos condicionales pueden dividirse en función de si las comparaciones entre grupos se llevan a cabo con respecto a una variable latente, denominados *métodos de invarianza condicional no observada*, o a una variable observable conocidos como *métodos de invarianza condicional observada*, (Gómez Benito & Hidalgo Montesinos, 1997; Millsap & Everson, 1993).

Los métodos de invarianza condicional no observada, especifican un modelo de medida y contrastan si los parámetros de dicho modelo permanecen invariantes para los distintos grupos evaluados. Por su parte, los métodos de

invarianza condicional observada, utilizan la puntuación observada del test como un estimador del rasgo medido, sin especificación formal de modelo de medida, y verifican si las distribuciones de puntuaciones del ítem entre sujetos con valores iguales en la puntuación total del test son independientes del grupo de pertenencia (Gómez Benito & Hidalgo Montesinos, 1997).

3.6 Descripción de los principales métodos para el estudio del DIF

Como vimos con anterioridad, debido a la gran cantidad de técnicas para calcular el DIF que existen en la actualidad, se han propuesto diferentes clasificaciones para tratar de agruparlos. Para efectos de la presente investigación, seguiremos la clasificación propuesta por Camilli & Shepard (1994) en tres categorías: a) los métodos basados en el análisis de varianza y en la Teoría Clásica de los Test (TCT), b) los que se basan en la Teoría de Respuesta al Ítem (TRI) y c) los que se basan en el análisis de tablas de contingencia (TC), describiendo principalmente aquellos que serán utilizados en el presente estudio.

3.6.1 Procedimientos basados en análisis de la varianza y en la teoría clásica de los tests (TCT)

3.6.1.1 Procedimiento de análisis de la varianza (ANOVA)

Uno de los primeros procedimientos formales tomados para evaluar lo que en ese tiempo se conocía como ítems sesgados en un test es el procedimiento utilizado por Cardall y Coffman en 1964; con grupos de “negros” y “blancos” que habían tomado el SAT (“Scholastic Aptitude Test” por sus siglas en inglés) en 1963 para evaluar la interacción de los ítems entre ambos grupos, mediante procedimientos de análisis de varianza (ANOVA). Bajo esta perspectiva un ítem funciona diferencialmente en los miembros de un grupo si las diferencia en términos absolutos entre la media de dicho grupo y las medias de los otros grupos sometidos a comparación son mayores que lo esperado en función del comportamiento de los otros ítems del test (Cleary & Hilton, 1968). Esta metodología, es poco recomendable (Camilli & Shepard, 1987) ya que no detecta una buena parte de los ítems que realmente tienen funcionamiento diferencial y ocasiona una elevada tasa de falsos positivos al no tener en cuenta la capacidad discriminativa del ítem, dejándose afectar por la dificultad media de los ítems y por el impacto entre grupos, por esta razón actualmente son preferibles los métodos condicionales, en los cuales al equipararse los niveles de habilidad entre los grupos, se puede distinguir realmente entre DIF e impacto (Gómez Benito & Hidalgo Montesinos, 1997). Sin embargo, puede ser utilizada como una técnica

inicial, que se puede utilizar complementada por otras técnicas más potentes como el MH o RL, para observar concordancias entre las diferentes técnicas.

3.6.1.2 Método delta-plot

Posteriormente, Angoff (1972) presentó un método que se popularizó rápidamente por su facilidad y simplicidad en el estudio de diferencias culturales, conocido como método delta-plot (Holland & Wainer, 2012); la idea de éste y otros métodos denominados *Transformed Item Difficult* (TID), es que cuando los ítems son jerarquizados de acuerdo a su dificultad, los ítems sin sesgo deberían presentar el mismo orden en los dos grupos de comparación, por lo que cualquier alteración en el orden de dificultad de los ítems sería considerado como posible DIF. Es un procedimiento sencillo mediante gráficos de dispersión (delta-plot), en ausencia de DIF, el delta-plot deberá formar una elipse a lo largo de un eje mayor o principal, y la correlación entre los delta-valores pareados será alta (0.98 o más), por otra parte, si existen ítems con DIF, el delta-plot, se alejará de esta elipse, por lo que los ítems con DIF, serán visibles fácilmente; sin embargo, tampoco es muy utilizada actualmente, sustituyéndose por los procedimientos basados en tablas de contingencia que veremos a continuación o por los sofisticados procedimientos basados en TRI.

3.6.2 Procedimientos basados en tablas de contingencia

Este procedimiento engloba a varias técnicas que tienen en común la forma de tabular los datos mediante tablas de contingencia. Tienen la ventaja con respecto a los modelos IRT de que son más sencillos de calcular y no exigen los enormes tamaños muestrales requeridos para la estimación de parámetros en la IRT.

Para construir las tablas de contingencia que habrán de analizarse es necesario disponer para cada sujeto de un código de pertenencia al grupo, su respuesta al ítem (codificada como acierto, 1, o error, 0), y la puntuación total del test. La puntuación total se calcula como el número de respuestas correctas, siendo el rango de 0 a *n*. Esta se subdivide normalmente en *k* intervalos. Una adaptación de la notación general de la tabla, dada por Holland y Thayer (1988) se presenta en la tabla Figura 4:

Figura 4. *Notación de las tablas de contingencia en el análisis DIF.*

Grupo	Acierto (1)	(Fallo 0)	Total
R	A_j	B_j	$\mathbf{n_{Rj}}$
F	C_j	D_j	$\mathbf{n_{Fj}}$
Total	$\mathbf{M_{1j}}$	$\mathbf{M_{0j}}$	$\mathbf{n_j}$

Puesto que todos los examinados de cada tabla se encuentran dentro del mismo intervalo, pueden considerarse comparables los dos grupos y la pregunta a la que se intenta responder es ¿tienen los sujetos del grupo focal y del grupo de referencia la misma probabilidad de acertar el ítem? Para responder a esta pregunta pueden utilizarse distintos métodos (medidas de diferencia de proporciones, chi-cuadrado, Mantel-Haenszel o la regresión logística).

3.6.2.1 Estadísticos χ^2 tradicionales

Como hemos visto, en cualquier estudio de DIF el objetivo es evaluar si la probabilidad de responder correctamente a un mismo ítem en dos grupos es diferente; en otras palabras, si la proporción de aciertos es distinta a través de los grupos que están siendo comparados. Para conocer la significación de dichas diferencias es posible aplicar una prueba χ^2 tradicional de igualdad de

proporciones. Siendo el nivel de habilidad una variable a controlar mediante la igualación de los grupos en cuanto al rasgo a evaluar, generalmente la puntuación global del test, se elaborarán tantas tablas de contingencia como niveles de *k* se tengan, para cada uno de los ítems. Son dos por tanto las variables que conforman la tabla de contingencia a estudiar: la respuesta al ítem en el caso de ítems dicotómicos, codificada en 1 para aciertos y 0 para el fallo y el grupo de pertenencia (*G*) siendo el de referencia *R* o focal *F*, quedando conformada como el ejemplo de la Figura 4.

En esta tabla bidimensional n_{Rj} y n_{Fj} son las frecuencias de sujetos que han contestado el ítem *j* en el intervalo *k* en los grupos de referencia (R) y focal (F), respectivamente; en dónde, m_{1j} y m_{0j}, corresponden al número de sujetos que han acertado y que han fallado el ítem *j* en ambos grupos; A_j y C_j son las frecuencias relativas correspondientes al número de sujetos que han acertado el ítem en cada grupo (R y F), respectivamente, y B_j y D_j son las frecuencias relativas correspondientes al número de sujetos que han fallado el ítem también en cada uno de los grupos. Por último, n_j es el número total de sujetos que han contestado al ítem *j* en el intervalo *k*.

χ^2 de Schneuman. Si se utiliza la proporción de individuos que aciertan a un ítem (*p*) en un intervalo *k,* como una estimación de la probabilidad de acertar el ítem en dicho intervalo, en el caso de un ítem unidimensional, no deberían existir

diferencias en esa probabilidad de acierto debido a la variable grupo de pertenencia: Focal o Referencia. Por tanto, afirmar que un ítem *j* no presenta DIF es probar la hipótesis de:

$$p_{jFk} = p_{jRk}$$

Es decir, la probabilidad de acierto del ítem *j* es igual entre el grupo focal o de referencia (F o R) en cada uno de los intervalos de habilidad *k* (Scheuneman, 1979).

Una de las fallas que presenta este procedimiento es que al considerar solamente la proporción de aciertos, distorsiona los resultados, especialmente si existe impacto entre grupos, por lo que se suele ver afectado por los tamaños muestrales de los grupos comparados y es poco probable tener una distribución χ^2 (Baker, 1981).

χ^2 modificada o sumada de Camilli: El método ideado por Camilli en 1979 denominado χ^2 total o modificada, pretende eliminar ese error al sumar la χ^2 tanto de los aciertos como de los errores con *k* (R-1) grados de libertad, también se parte de la construcción de tablas de contingencia para cada nivel de

capacidad *k*, en la que se presentan las frecuencias de acierto y error para los grupos focal y de referencia, (Ver Figura 4, p. 117), siguiendo con el caso de ítems dicotómicos, a partir de tabla es posible calcular el estadístico χ^2.(Camilli, 1979; Camilli & Hopkins, 1979; Shepard, Camilli, & Averill, 1981).

3.6.3 Procedimientos basados en la Teoría de la Respuesta a los ítems (IRT)

Dentro del marco de la IRT, hay esencialmente dos enfoques para analizar el DIF que se diferencian en el método de relacionar los dos grupos estudiados y en la identificación del DIF significativo (Orlando & Marshall, 2002). En el primer enfoque, los parámetros de los ítems obtenidos desde una calibración separada de los dos grupos son igualados utilizando la puntuación total como base para relacionar los grupos y el DIF es identificado para cada ítem usando uno o más índices DIF.

El segundo enfoque, calibra los ítems para ambos grupos conjuntamente usando mecanismos de relación de la IRT (Embretson, 1996) sobre la base de un subconjunto de ítems, considerados como ítems de anclaje, que a priori se consideran insesgados. Hay varias formas de hacerlo. En un estudio a gran escala, los ítems de anclaje son a menudo seleccionados de un grupo de ítems insesgados. Cuando no se dispone de información previa sobre los ítems, pueden

seleccionarse en base a resultados basados en la teoría clásica de los tests tales como el test de Mantel-Haenszel o el análisis de regresión logística. Una tercera forma de establecer un conjunto de ítems es a través del diseño de estudio. Dependiendo de los grupos que se comparan, es a menudo posible administrar algunos ítems a ambos grupos en el mismo formato. Una vez que el conjunto de ítems de anclaje se ha establecido, el DIF es identificado entre los ítems sobre la base del modelo basado en las pruebas de razón de verosimilitud.

Como se vio anteriormente, estas técnicas se encuentran dentro de los métodos condicionales, ya que permiten trabajar con grupos comparables debido a que igualan los grupos respecto al rasgo medido; sin embargo, se consideran de invarianza condicional no observada, debido a que en estos la variable de igualación es un *variable latente,* es decir, que *no se puede observar* pero *se estima.* Algunos de estos métodos se basan en la estimación de diferencias en las CCI´s de los ítems, conocidos como técnicas de *medidas de áreas* como las propuestas por Raju (1988) o Kim y Cohen (1991), otros son basados en la *comparación de los parámetros* como la X^2 de Lord, o en la *comparación de los parámetros* como los propuestos por Thissen, Steinberg y Gerrard (1986) o la propuesta de Kelderman (1989) utilizando un modelo loglineal basado en el modelo de Rasch. La principal dificultad de todas estas técnicas es que son mucho más costosos para realizar desde el punto de vista computacional, aunado a que requieren tamaños muestrales mucho más amplios que los métodos que hemos revisado hasta el momento (Gómez Benito & Hidalgo Montesinos, 1997).

3.6.3.1 Teoría de Respuesta al Ítem TRI (Ítem Response Theory IRT)

Inicialmente desarrollada en los años 50, pero modificada y ampliada hasta la actualidad, la Teoría de Respuesta al Ítem engloba una teoría matemática para las características del ítem y las escalas de medición y los correspondientes métodos numéricos para estimar los parámetros del ítem en correspondencia con la habilidad de los examinados. La característica distintiva de estos procedimientos es que relacionan las respuestas al ítem con la variable latente que pretende medir; considerando que la variable latente sea unidimensional y las respuestas al ítem dicotómicas, el procedimiento básico 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 (presencia de DIF); sin embargo, aunque conceptualmente y visualmente parece algo relativamente sencillo, procedimentalmente y computacionalmente es muy complejo (Ver Figuras 1, 2 y 3 p.p. 107-108).

3.6.3.2 Conceptos a considerar en la TRI

Como sugieren Pedraza y Mungas (2008), la *habilidad* es un concepto central en la TRI, que se refiere a la capacidad de responder correctamente a las preguntas de la prueba y representa la influencia de todos los aspectos

ambientales, experienciales y genéticos que pueden influir en el rendimiento del sujeto en la misma, sin suposiciones sobre las contribuciones relativas de estos factores, es decir; un individuo con una mayor habilidad simplemente tiene una mayor probabilidad de responder correctamente a un ítem en una prueba determinada.

Otros dos parámetros fundamentales de los modelos de TRI son la *dificultad del ítem* y *discriminación del ítem*. Para un ítem dicotómico, la *dificultad* se refiere al nivel de habilidad asociado con una probabilidad del 50% de acertar el ítem, lo que en la Teoría Clásica de los Test (TCT) equivaldría a la *proporción de respuestas correctas*, mientras que la *discriminación del ítem* se refiere al grado en que las pequeñas diferencias de capacidad se asocian con diferentes probabilidades de acertar el ítem, que en la TCT se conoce como *correlación total e ítem.*

Un modelo basado en la TRI, expresa la asociación probabilística entre las respuestas observables al ítem de una persona y su nivel de *habilidad* θ (theta) que **no** se puede observar, pero se estima; como vimos anteriormente, por ello se consideran técnicas de *invarianza condicional no observada.*

Un requisito para considerar la evaluación del DIF desde el TRI, es obtener las *estimaciones de los parámetros de los ítems*, existen varios, aunque algunos de los más utilizados son los de *máxima verosimilitud conjunta* (MVC), sin embargo como apuntan Gómez Benito & Hidalgo Montesinos (1997), cuando se tienen tamaños muestrales pequeños (menores a 500 sujetos), es más apropiado un método de *máxima verosimilitud marginal* (MVM).

Otro aspecto fundamental es el de *invarianza de los parámetros*, ya sean de los ítems o de la habilidad (θ), que implica que los parámetros estimados deben ser los mismos en cada muestra obtenida en una misma población y que cuando se realicen comparaciones entre parámetros estimados entre diferentes grupos, es necesario igualarlos. Existen varias formas de obtención de las constantes de igualación, sin embargo lo más común es mediante una curva característica del ítem o CCIs, en donde la dificultad está representada por la ubicación a lo largo del eje de las abscisas (X) en cuyo punto la probabilidad de un respuesta correcta para un ítem dicotómico es 50 por ciento, y la *discriminación* está representada por la pendiente de la línea en esa ubicación del parámetro; una pendiente más pronunciada refleja un mayor grado de discriminación (Crane, Narasimhalu, Gibbons, Mungas, et al., 2008; Pedraza & Mungas, 2008).

La curva característica del ítem puede estar definida por el modelo logístico de un parámetro (1-*p*), de dos parámetros (2-*p*) o tres parámetros (3-*p*). El modelo

en TRI, representa la asociación probabilística entre las respuestas al ítem observables y su nivel de habilidad que no es observable, pero se estima. Los modelos de un parámetro estiman libremente la dificultad del ítem y requiere que se asuma que la discriminación del ítem es igual para todos los ítems. Los modelos de dos parámetros obtienen estimaciones tanto para la dificultad del ítem como para la discriminación. Por último, un modelo de tres parámetros se puede usar para ítems con opción múltiple, en los cuales se puede presentar una respuesta acertada por efecto de la adivinación; el tercer parámetro entonces estima esa probabilidad (likelihood) de acertar al ítem por azar (Crane, Narasimhalu, Gibbons, Pedraza, et al., 2008; Pedraza & Mungas, 2008).

Una ventaja adicional de la TRI es que se puede obtener la curva del *funcionamiento global de la prueba*, que cuantifica el nivel de fiabilidad del test en cada punto del continuo de la habilidad. El concepto de *información* corresponde a la fiabilidad de la medición, de modo que el error estándar de medición en un valor de capacidad específica es la raíz cuadrada inversa de la información en esa capacidad (Hambleton & Swaminathan, 1985; citados por Pedraza & Mungas, 2008). Por lo tanto, la confiabilidad no se reduce a un solo coeficiente de confiabilidad para toda la escala, sino que varía a lo largo del continuo de la habilidad completa. Esto permite la identificación de las regiones del espectro donde la capacidad de una escala es particularmente fiable, o por el contrario, tiene escasa confiabilidad y sensibilidad limitada para detectar diferencias en la habilidad (Pedraza & Mungas, 2008).

Si los supuestos básicos se conocen y las muestras incluyen un amplio rango de variabilidad, los parámetros del ítem son invariantes entre muestras y los estimadores de habilidad son invariantes entre ítems, esto significa en un lenguaje pragmático que las diferencias en las habilidades cognitivas de la población estudiada, no deberían afectar los resultados en los resultados de las pruebas, (Williams, 1997), mientras que esa es una deficiencia que presenta la Teoría Clásica de los Tests (TCT); las características de la muestra estudiada, influyen determinantemente en las puntaciones psicométricas de las pruebas

Una de las aplicaciones de la Teoría de Respuesta al ítem, es como podremos imaginar las pruebas adaptadas automatizadas. Al tener una gran cantidad de ítems disponibles (bancos de ítems), y sobre los que se han evaluado sus propiedades de dificultad y discriminación entre diferentes sujetos, considerando características muy heterogéneas: cultura, educación, sexo, edad, habilidad, por mencionar algunas. En este caso se podrían tener pruebas específicas para cada sujeto evaluado con base en sus características y “libres de sesgo” (Walker, Beretvas, & Ackerman, 2001; Williams, 1997). Empero, aquí podríamos encontrar la contraparte de los estudios o pruebas “a medida”, y la controversia de si es “justo” aplicar pruebas libres de “sesgo”, en lugar del mismo “instrumento de medida” para todos los aspirantes en un determinado momento.

Por tanto, no todo es positivo en la TRI, y es que como se ha mencionado repetidamente, el procedimiento computacionalmente costoso y el alto tamaño muestral requerido tanto para la construcción de las pruebas, como para el estudio del DIF desde ella, la hacen mucho más compleja y dificultan su utilización, fuera del campo de los estudios de simulación.

3.7 Procedimiento de Mantel-Haenszel (MH)

La prueba de Mantel-Haenszel (MH) es una adaptación de los enfoques tradicionales de X^2 descritos anteriormente, como plantean Guilera, Gómez Benito, Hidalgo Montesinos y Sánchez Meca (2007); a finales de los años 50, Mantel y Haenszel (1959) propusieron un método para el análisis de tablas de contingencia tridimensionales que permitía estudiar diferencias entre grupos comparables; posteriormente Holland (1985) y Holland y Thayer (1988) utilizaron este procedimiento adaptado como técnica de detección del DIF. De todos los estadísticos descritos basados en X^2**,** el procedimiento MH, es el mejor aceptado por el *Educational Testing Service,* contando con una gran cantidad de estudios de adaptación para utilizarla en diversas situaciones y es por ello que es uno de los más utilizados actualmente (Fidalgo Aliste, 2011; Magis & de Boeck, 2014; Zwick, 2012).

El procedimiento consiste en comparar la ejecución de un ítem en el grupo de referencia y el grupo focal a través de distintos niveles de una determinada variable de equiparación (criterio o habilidad), en este sentido, se asume que en cada uno de estos niveles los individuos de uno y otro grupo son comparables, y si un ítem no presenta DIF lo ejecutarían por igual, presentando la misma probabilidad de acierto o error. Con el propósito de comparar las probabilidades de acierto a un ítem, los datos del grupo de referencia y del grupo focal se distribuyen en tantas tablas de contingencia bidimensionales (2x2) como niveles en la habilidad de los sujetos; dichos niveles normalmente se diseñan a partir de la puntuación total de los sujetos en el test, utilizando una estructura similar a la mostrada en la Figura 4 (p. 117).

La hipótesis nula de ausencia de DIF postula que la probabilidad de acertar el ítem bajo estudio en el nivel de habilidad *j* es la misma para el grupo de referencia que para el grupo focal, mientras que la hipótesis alternativa de presencia de DIF formula que esa probabilidad de acierto en el grupo de referencia equivale a la probabilidad de acierto en el grupo focal multiplicado por un cociente de razones común, denominado alfa (α), siguiendo a Guilera et al., (2007) y utilizando nuestra tabla de contingencia ejemplo, la expresión de (α) se obtiene mediante la siguiente fórmula:

$$\alpha_{MH} = \frac{\sum A_j D_j / n_j}{\sum B_j C_j / n_j}$$

donde:

n_j= número de sujetos de los grupos de referencia y focal que han contestado al ítem, en el nivel de capacidad k.

Aj= Frecuencia observada de sujetos del GR que han acertado al ítem j.

Bj= Frecuencia observada de sujetos del GR que han fallado al ítem j.

Cj= Frecuencia observada de sujetos del GF, que han acertado al ítem.

Dj= Frecuencia observada de sujetos del GF que han fallado al ítem.

(α) es un estimador de la magnitud del DIF; es decir, no sólo nos indica su presencia o ausencia, sino que nos indica su magnitud y hacia donde se orienta; es decir, si favorece al grupo focal o de referencia α puede adoptar valores entre 0 e infinito. Como plantea Bandeira Andriola & Gaviria Soto (2002), la hipótesis nula de ausencia de DIF está representada por un alfa con valor 1, mientras que un valor mayor que 1 indica que el grupo de referencia presenta una probabilidad más elevada de acertar el ítem que el grupo focal; por el contrario, un valor menor que 1 indica que el grupo aventajado es el focal frente al de referencia. Esta es una de las ventajas por las que el MH ha venido a sustituir el uso de los otros procedimientos basados en χ^2.

El estadístico de contraste para el α_{MH} es dado por:

$$\chi^2_{MH} = \frac{\left\{\left|\sum_{j=1}^{K} A_j - \sum_{j=1}^{K} E(A_j)\right| - 0{,}05\right\}^2}{\sum_{j=1}^{K} Var\,(A_j)}$$

donde:

$$E(A_j) = \frac{n_{Rj} n_{Aj}}{n_j}$$

y

$$Var\,(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{(n_j)^2 (n_j - 1)}$$

En un estudio metaanalítico realizado por Guilera et al., (2007), sobre el MH y el Error Tipo I en la detección del DIF (proporción de veces que un ítem es identificado con DIF cuando en realidad no lo presenta), y su potencia estadística (proporción de veces que un ítem es identificado correctamente con presencia de DIF), encontró que el procedimiento MH como técnica de detección del DIF cumple ambos criterios (Error Tipo I < 0.055 y Potencia ≥ 0.80) en un gran abanico de situaciones; sin embargo, el balance entre criterios se alcanza en un mayor porcentaje trabajando con modelos logísticos de 3-p, valores medios del parámetro de discriminación, valores medios también del parámetro de dificultad,

en presencia de DIF uniforme y con una cantidad de DIF baja. En cuanto a las variables continuas se encontró que el MH funciona mejor cuando el grupo de referencia contiene más sujetos que el grupo focal. Además, el valor medio de impacto se sitúa en torno a una diferencia entre medias de ambos grupos de 0.50 unidades de desviación típica y, por último, el porcentaje medio de ítems con DIF en el test entre 2% y 15%, aproximadamente.

El procedimiento de MH es no paramétrico y según algunos autores, muy similar en funcionamiento al SIBTEST (*Simultaneous Item Bias Test*), ampliamente utilizado también para la detección del DIF uniforme; sin embargo, al igual que el SIBTEST, una de las deficiencias del Mantel-Haenszel es su poca capacidad para identificar el DIF no uniforme (Narayanan & Swaminathan, 1994). Es por ello que Mazor, Clauser y Hambleton (1994), propusieron una variante del procedimiento que consiste en calcular las puntuaciones medias de los sujetos y dividirlos en dos grupos: el de sujetos con las puntuaciones más bajas y aquellos con las puntuaciones más altas, calculando el χ^2_{MH} de forma separada con cada grupo. Algunos estudios de simulación han aplicado esta variación del MH (MH-modificada) para detectar el DIF no uniforme, y parece ser una alternativa viable y eficaz en la detección de este tipo de DIF no uniforme simétrico, como asimétrico, aunque a costa de incrementar la tasa de error de Tipo I (Fidalgo Aliste, Mellenberg, & Muñiz, 2000; Fidalgo Aliste, Ferreres Traver, & Muñiz, 2004).

Sin embargo, otros autores recomiendan utilizar esta técnica como un primer paso, y en caso de detectar ítems con DIF, complementarlo con un procedimiento como el de Regresión Logística, que ha probado ser bueno en la detección del DIF no uniforme, pero aumenta la cantidad de falsos positivos o error tipo II (Ferreres Traver, Fidalgo Aliste, & Muñiz, 2000) y para observar el grado de concordancia entre los resultados obtenidos mediante las diversas técnicas (Abedalaziz, 2010; Fidalgo Aliste, Alavi, & Amirian, 2014).

3.8 Regresión Logística (RL):

Swaminathan y Rogers (1990) propusieron la regresión logística como método de análisis del DIF uniforme y no uniforme como una opción intermedia entre el MH tradicional y las costosas técnicas basadas en la Teoría de Respuesta al Ítem. En esta aproximación, el modelo permite predecir la probabilidad de una respuesta correcta a un determinado ítem en función del nivel de habilidad del examinado (θ: la puntuación total en el test); su grupo de pertenencia (G), y el término θ x G, que representa la interacción entre el nivel de habilidad del examinado y su grupo de pertenencia, siendo este último término el que en caso de resultar significativo indicaría la presencia de DIF no uniforme (Ferreres Traver, González Romá, & Gómez Benito, 2000).

Al igual que otras técnicas estadísticas multivariadas, la regresión logística, brinda la posibilidad de evaluar la influencia de cada una de las variables independientes sobre la variable dependiente y controlar el efecto del resto; es decir, teniendo una variable dependiente (*y),* que puede ser dicotómica o politómica y una o más variables independientes, que pueden ser de cualquier naturaleza, cualitativas o cuantitativas.

Su fórmula es:

$$P\left(y = \frac{1}{X}\right) = \frac{e^z}{1 + e^z}$$

donde:

P(y=1/X)= es la probabilidad de obtener una respuesta correcta condicionado a X

X= puntuación observada del sujeto en el test

z= representa la combinación lineal de las variables predictoras.

En el análisis del DIF, el modelo de regresión logística se parametriza en los siguientes términos:

$$z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$$

donde:

X = es la puntuación observada de un sujeto en un test.

G = la variable de grupo de pertenencia de los sujetos.

β_0 = es la intercepción

β_1 = es el coeficiente para la habilidad o puntuación total observada en el test

β_2 = es el coeficiente para la variable grupo de pertenencia (referencia o focal) y

β_3 = es la interacción entre la puntuación observada en el test y el grupo (Hidalgo Montesinos, Gómez Benito, & Padilla García, 2005).

Bajo esta formulación, 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. Por el contrario, si la interacción *XG* resulta estadísticamente significativa, el ítem presentaría DIF no-uniforme.

Como explican Hidalgo Montesinos et al. (2005), la aplicación de la regresión logística para la detección del DIF puede realizarse con distintas

estrategias. La primera estrategia se basa en la comparación de modelos anidados. Se ajustan tres modelos en distintas etapas. 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 último, se incluye en la fórmula la interacción entre el grupo y la puntuación total en el test, evaluándose el ajuste del modelo de DIF no-uniforme o completo (modelo 3).

Siguiendo a Hidalgo Montesinos et al. (2005) la segunda estrategia consiste en 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 presenta una distribución χ^2 con 2 grados de libertad; aunque, el uso de esta estrategia de análisis no permite evaluar el tipo de DIF que presenta el ítem; pero comparado con el procedimiento anterior, es más económico en tiempo computacional. Este último es el procedimiento inicial propuesto por Swaminathan y Rogers (1990); sin embargo, es menos recomendable, ya que al incluir todos los parámetros en una sola comparación, aumenta la tasa de error tipo II, al tener 2 grados de libertad.

Las dos estrategias anteriores obligan a ajustar más de un modelo de regresión logística; la tercera estrategia consiste en ajustar solamente el modelo completo, incluyendo todos los términos (*X, G* y *XG)*, y comprobar la significación de los coeficientes del modelo asociados a cada término utilizando el estadístico de Wald; no obstante, como plantean Hidalgo Montesinos y colaboradores, para seguir una distribución X^2 con un grado de libertad, se requiere tener muestras grandes (Hidalgo Montesinos et al., 2005).

3.9 Comparación de las técnica de MH y RL

Como mencionan Ferreres Traver, González Romá, et al. (2000) varios estudios de simulación (Ferreres Traver, Fidalgo Aliste, et al., 2000; Fidalgo Aliste, 1996), parecen apuntar a que cuando el ítem presenta DIF uniforme la potencia de prueba del estadístico MH disminuye a medida que aumenta la dificultad del ítem; por el contrario, cuando el DIF es no uniforme, los ítems peor identificados por el MH son los ítems de dificultad medio-baja y baja discriminación. Por su parte, la Regresión lineal ofrece un comportamiento aceptable en la detección del DIF de los ítems con baja o moderada dificultad y elevada discriminación. No obstante, los ítems peor identificados por ésta son los ítems con niveles medios de dificultad y baja discriminación.

Sin embargo, como encontraron Hidalgo Montesinos y Lopez Pina (2004), en un estudio de datos simulados, el procedimiento de MH-modificado (Mazor et al., 1994) presentó un funcionamiento muy similar al de la Regresión logística (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990) en la identificación del DIF uniforme (76% y 78% respectivamente), comparado con el procedimiento MH original o estándar (Holland & Thayer, 1988) con solo un 68% de identificación de ítems con DIF.

La ventaja de ambas técnicas es que ambas funcionan muy bien cuando se trabaja con tamaños muestrales pequeños 200 o 300 sujetos por grupo, lo que las convierte en técnicas ampliamente utilizadas y que pueden ser usadas en combinación, ya que ambas presentan resultados similares con una correlación hasta del 80% (Abedalaziz, 2010), con tasas de error tipo I del 5% si son escalas amplias como plantea Scott et al. (2009), para escalas de extensión considerable e ítems dicotómicos; y que se pueden utilizar en conjunto para ver el acuerdo en la identificación de ítems con DIF (Fidalgo Aliste et al., 2014).

3.10 Problemas prácticos en la detección del DIF

3.10.1 El problema del tamaño de la muestra

Es importante recalcar que un aspecto de gran importancia, antes de establecer si existe DIF o no en algunos ítems, es verificar que esto no sea debido al tamaño de la muestra. Esto debido a que, el *poder* de algunas pruebas aumenta en muestras mayores, por lo que en ocasiones, teniendo muestras muy grandes el poder del estadístico puede detectar DIF con variaciones muy pequeñas e incluso donde no lo hay, es decir, producir *error tipo I*. Por el contrario, una muestra muy pequeña, suele disminuir el poder de las pruebas estadísticas, no identificando ítems que si tienen DIF, lo que se denomina *error tipo II*.

Como vimos con anterioridad un factor decisivo en los procedimientos basados en la TRI es el contar muestras muy amplias. Sin embargo, como plantean algunos autores, aunque en teoría esto es importante, y relativamente sencillo de lograr en estudios de simulación. No obstante, como mencionan Lei y Li (2013), tener muestras amplias en estudios de traducción o adaptación de pruebas interculturales es difícil, por lo que generalmente las pruebas recomendadas en estos casos son el MH (Holland & Thayer, 1988) y sus diversas modificaciones, el SIBTEST, que funciona similar al MH en cuanto a error tipo I, cuando no hay impacto (es decir las puntuaciones medias de los sujetos de ambos

grupos son similares) y mejor que el MH cuando hay impacto (Roussos & Stout, 1996) y al decir muestras pequeñas nos estamos refiriendo a un mínimo de 200 o 250 sujetos por grupo focal y 300 para el de referencia; es por ello que frecuentemente los procedimientos basados en la TRI, quedan fuera de las posibilidades en este tipo de estudios (Lei & Li, 2013).

Sin embargo, las recomendaciones de la Educational Testing Service, han ido cambiando a través del tiempo, con respecto al tamaño de la muestra, al ir desarrollándose procedimientos más específicos, que han permitido, como plantea Zwick “relajar un poco el tema del tamaño de la muestra” para una revisión detallada ver Zwick (2012).

3.10.2 Procesos de purificación

Independientemente del procedimiento estadístico utilizado, cuando no se utiliza un criterio externo al test evaluado (p. ej. puntuaciones en otras pruebas) sino que se utiliza la puntuación total en el test como un estimador de la habilidad para la equiparación de los grupos, la identificación de ítems con DIF, independientemente de la técnica utilizada, se ve afectada por la presencia en el test de otros ítems con DIF, siendo el principal problema un incremento del número de falsos positivos, en esos casos es recomendable la utilización de algún

procedimiento de *purificación* para eliminar ese problema de *circularidad* (Fidalgo Aliste, Mellenbergh, & Muñiz, 1999).

Se suele recomendar realizar mediciones en dos etapas (*método bietápico*), en la primera se realizan las mediciones correspondientes para cada uno de los ítems, y una vez que se han identificado aquellos con posible DIF, se les *marca* (*flag*) y se les elimina temporalmente de la prueba y se vuelven a correr los análisis para ver si el DIF persiste.

Otra forma de realizarlo es el *método iterativo*, que consiste básicamente en ir eliminando ítems del test marcados como "flag" o con DIF, poco a poco (uno o dos a la vez), dependiendo de la necesidad detectada y volver a correr los análisis. Este procedimiento es más conservador, pero también más costoso computacionalmente; no obstante, tiene la ventaja de que permite ir observando la cantidad de influencia que pudieran tener cada grupo de ítems marcados "flag" o sesgados en la puntuación global, en los ítems que por consecuencia estuvieran siendo detectados erróneamente con DIF (Error tipo I).

Para una descripción más detallada de los distintos métodos de purificación, su fundamentación y usos, con referencia a los diferentes procedimientos estadísticos en la detección del DIF el lector podría consultar en:

Fidalgo Aliste et al., (1999); Fidalgo Aliste & Paz (1995); Gómez Benito & Hidalgo Montesinos (1997); Holland & Thayer (1988); Khalid & Glas (2014); Magis & Facon, 2013; Núñez Núñez, Hidalgo Montesinos, & López Pina (2000).

En el caso del procedimiento Mantel Haenszel, el método bietápico e iterativo parece producir resultados muy similares según los datos obtenidos en estudios de simulación (Clauser, Mazor, & Hambleton, 1993) cuando la cantidad de ítems con DIF "flag" es pequeña en la primera etapa; en este caso el proceso de purificación cualquiera que haya sido elegido parece ayudar a obtener resultados más precisos, sin comprometer la tasa de error tipo I, que cuando se realiza la evaluación del DIF en una sola etapa; sin embargo, cuando la cantidad de ítems identificados con DIF, en la primera etapa es muy alta, el modelo iterativo parece ser más adecuado que el bietápico, ya que al eliminar una gran cantidad de ítems, en ese proceso de purificación, los ítems identificados con DIF, pudieran no ser los mismos, entre la primera evaluación y la segunda, comprometiendo entonces las tasas de error tipo I y/o II (Zenisky, Hambleton, & Robin, 2003).

3.10.3 Elección del método estadístico más apropiado

A partir de la revisión de las principales técnicas de detección de DIF, nos podemos percatar de que no existe un método ideal. Como plantean Gómez-Benito, Hidalgo, & Guilera (2010), la decisión de aplicar un procedimiento

determinado en una situación concreta todavía está llena de interrogantes, la decisión de que técnica aplicar a nuestros datos suele basarse en diversos aspectos, tales como el tamaño muestral de los grupos de referencia y focal, las diferencias en las distribuciones de habilidad de los grupos, el tipo de DIF, la simplicidad computacional y disponibilidad de programas informáticos, y el criterio de igualación de los grupos, principalmente.

Por lo anterior es que varios autores aconsejan aplicar diversas técnicas de detección del DIF y tomar la decisión última de mantener, reformular o eliminar el ítem en función del acuerdo o desacuerdo entre diferentes métodos de detección sobre la presencia o ausencia de DIF.

3.10.4 Decisión final con respecto al ítem y al test

Una vez que se han identificado los ítems con DIF, vienen las preguntas ¿A qué se debe el DIF? y ¿Qué se hará con el ítem?, ¿Los ítems con DIF o marcados "flag" afectarán el funcionamiento del test? ¿El test completo presenta funcionamiento diferencial? ¿Qué se hará con el test?, y eso es un procedimiento aparte, que dependerá ya no de análisis y pruebas estadísticas, sino de la opinión del constructor o adaptador del test, decidir qué hacer con esos ítems, si dejarlos, eliminarlos o modificarlos.

Decidir eliminar un ítem puede ser una decisión difícil de tomar, sobre todo cuando no se tiene una explicación plausible al por qué funciona diferencialmente un ítem. En caso de decidir dejar el o los ítems, tendría que advertirse sobre el funcionamiento diferencial no del ítem, sino de la prueba incluso, o recomendar incluso realizar una nueva estandarización para cada población.

3.10.5 Procedimiento general para el estudio del DIF

A manera de resumen, podemos presentar la siguiente metodología general para los casos de estudio del DIF, con datos reales:

a) Identificar los grupos focal y de referencia.
b) Diseñar el estudio para tener las muestras lo más amplias posibles.
c) Seleccionar las técnicas estadísticas DIF más apropiadas para los datos.
d) Realizar los análisis estadísticos.
e) Interpretar los estadísticos DIF, si este no existe, el proceso termina, en caso contrario continuar con los incisos f al h.
f) En caso de detectar ítems con DIF, realizar algún proceso de *purificación*.
g) Repetir el proceso desde el inciso d, tantas veces como sea necesario.
h) Decidir que ítems se eliminan y cuáles se mantienen.

3.11 Conclusiones

Hasta el momento hemos realizado una revisión en el Capítulo I concerniente al campo de la neuropsicología y evaluación psicológica, hemos visto las dificultades que implica trabajar con personas de diferente procedencia cultural y pruebas que han sido adaptadas de otros idiomas.

En el capítulo II hemos revisado las principales cuestiones a considerar para la adaptación de pruebas para su uso con diferentes idiomas y culturas, así como las cuestiones referentes al establecimiento de la equivalencia en las mediciones entre diferentes grupos.

En el presente capítulo III, hemos desarrollado el tema del DIF, procedimiento ampliamente utilizado para establecer la equivalencia en las mediciones, no solamente en el campo de la evaluación neuropsicológica, sino en el educativo, lingüístico, de salud, entre otros.

Como hemos visto, la institución que sigue siendo pionera en el campo de estudio del DIF, es la Educational Testing Service, sin embargo, habría que considerar que el DIF y los procedimientos que han sido creados y utilizados parten del sistema educativo en Estados Unidos de Norteamérica, los cuales

tuvieron su auge en una época cargada de problemas políticos relacionados con la discriminación racial y de clases económicas; y que, lo que se ha intentado a lo largo de estos años es adaptarlos a otras áreas como la salud, de mercados, político, para tratar de garantizar una evaluación libre de “sesgo”.

Como hemos examinado, algunos de los procedimientos más utilizados para medir el DIF actualmente son los de Mantel-Haenszel y sus adaptaciones, ya que, estadísticamente son muy eficientes y baratos computacionalmente (Dorans & Holland, 1992); sin embargo, tienen la desventaja de no tener un funcionamiento óptimo para la detección del DIF no uniforme. Los procedimientos basados en la Teoría de Respuesta al ítem, suelen ser preferidos en los estudios a gran escala y con tamaños muestrales grandes. No obstante, en el campo de la adaptación de pruebas y trabajo con minorías de diferente procedencia cultural, el tener muestras amplias, principalmente en las muestras focales, es algo difícil de lograr. La Regresión Logística, es una opción intermedia entre el MH y los procedimientos basados en la TRI, con un buen funcionamiento en la detección del DIF uniforme como no uniforme, ofreciendo la ventaja de poderse utilizar con tamaños muestrales pequeños.

De esta forma hemos concluido con la revisión teórica de este documento; en el siguiente capítulo se presentan los datos del estudio empírico seguido en la presente investigación.

CAPITULO IV

ESTUDIO EMPÍRICO

4.1 Introducción

En este capítulo se hará una descripción de los aspectos metodológicos que se siguieron para la elaboración de la investigación. En la primera parte se presenta los alcances del estudio, se presentan los objetivos, posteriormente la metodología del estudio, descripción de los participantes y procesos de selección de los mismos en las muestras de referencia y focales. Se describen los materiales utilizados (Ver Anexos 1, 2 y 3) y el procedimiento de recolección de datos seguido en los tres países en que se realizó el estudio comparativo.

En un estudio exploratorio previo, realizado como parte de los trabajos para la obtención del DEA de la autora, se evaluó el funcionamiento diferencial del ítem entre población española y mexicana, utilizando para tal efecto en la muestra focal (mexicana) una muestra de 90 participantes. En aquel estudio se encontró un Funcionamiento Diferencial del Ítem “bajo” entre estas dos poblaciones. En el presente estudio se espera corroborar estos resultados contando con unas

muestras más grandes; y además ampliarlos a la población de habla hispana en E.U.A. La pregunta de investigación a la que se tratará de dar respuesta es: ¿Existe Funcionamiento Diferencial del Ítem en el TRVS entre poblaciones española, mexicana y de habla hispana en E.U.A.?

Para tratar de responder la pregunta de investigación, se establecieron los objetivos general y específicos, así como la metodología del estudio que a continuación se detallan.

4.2 Objetivo general

El objetivo general del presente estudio es evaluar el Funcionamiento Diferencial del Test de Recuerdo Verbal Selectivo entre población española, mexicana y de habla hispana en E.U.A.

4.3 Objetivos específicos

Como objetivos específicos de esta investigación nos hemos planteado los siguientes:

- Estudiar si las puntuaciones globales del Test de Recuerdo Verbal Selectivo presentan DIF entre población española y mexicana.

- Estudiar si las puntuaciones globales del Test de Recuerdo Verbal Selectivo presentan DIF entre población española e hispana de E.U.A.

- Estudiar si las puntuaciones globales del Test de Recuerdo Verbal Selectivo presentan DIF entre población mexicana y de habla hispana en E.U.A.

- Explorar si existe Funcionamiento Diferencial en cada uno de los ítems del Test de Recuerdo Verbal Selectivo entre población española y mexicana.

- Estudiar si existe DIF en cada uno de los ítems que conforman el Test de Recuerdo Verbal Selectivo, entre población española y de habla hispana en E.U.A.

- Estudiar la existencia de DIF en cada uno de los ítems del Test de Recuerdo Verbal Selectivo entre población mexicana y de habla hispana en E.U.A.

4.4 Método

Este estudio sigue un diseño no experimental, de tipo transversal y comparativo (intercultural) con un enfoque cuantitativo.

4.4.1 Participantes

4.4.1.1 Muestra de referencia española

Los datos de los participantes de la muestra de referencia fueron 211 españoles seleccionados de entre los participantes del estudio para obtener datos normativos para dos versiones en español de la prueba vSRT con 6 ensayos. (Morales et al., 2010). El primer requisito fue ser de nacionalidad española. Los criterios de inclusión en el proceso de estandarización fueron:

a) No tener antecedentes de enfermedad neurológica.

b) No presentar antecedentes de hospitalización por cuestiones psicopatológicas (esquizofrenia, trastorno depresivo mayor).

c) No antecedentes de desarrollo psicomotriz anormal.

d) No tener uso de psicofármacos que pudieran interferir con cuestiones como concentración, atención o producir somnolencia.

e) No tener historial de consumo de drogas o alcoholismo.

f) Tener el idioma español como primera lengua.

Los individuos con condiciones médicas crónicas como hipertensión, diabetes, pérdida auditiva leve, no fueron excluidos del estudio. Todos los participantes fueron considerados como capacitados cognitivamente para el funcionamiento independiente. La mayoría de los participantes fueron reclutados en el sur y sureste de España, intentando obtener datos tanto de áreas rurales como urbanas. De los 804 sujetos evaluados en aquel país se procedió a seleccionar mediante un muestreo no probabilístico por conveniencia a los 211 sujetos, que fueron equiparados con las muestras de las poblaciones focales: mexicanas y estadounidenses en cuanto a las variables, edad, escolaridad y sexo.

4.4.1.2 Muestra focal mexicana

La muestra del primer grupo focal (mexicano) está conformada por un total de 201 sujetos adultos mexicanos voluntarios, seleccionados en la ciudad de H. Matamoros, Tamaulipas. El tipo de muestreo utilizado fue intencional, seleccionando a aquellos sujetos que aceptaron participar en el estudio y que cumplieron con los mismos criterios de inclusión que la muestra española pero con nacionalidad mexicana.

4.4.1.3 Muestra focal estadounidense

La muestra del segundo grupo focal: población de habla hispana en E.U.A., estuvo conformada por un total de 205 sujetos adultos voluntarios, seleccionados en la ciudad de Brownsville, Texas. El tipo de muestreo fue intencional, seleccionando a aquellos sujetos que aceptaron participar en el estudio y que cumplieron con los mismos *criterios de inclusión* que las muestras anteriores más los siguientes:

a) Ser de origen hispano.

b) Estar viviendo de manera permanente en E.U.A.

c) Tener el idioma español como lengua.

Es importante mencionar que fueron incluidos tanto sujetos nacidos en Estados Unidos, como personas con diversa situación migratoria: residentes, ciudadanos nacionalizados, e incluso inmigrantes cuya situación migratoria estuviera en trámites. Esto debido a que por la cercanía, muchas personas de origen Latinoamericano emigran a estas ciudades fronterizas, ya sea de manera temporal o permanente, esto se preguntó a los participantes, pero debido a lo delicado del tema no se indagó más al respecto, pero la mayoría de los participantes fueron nacidos en México (con diferentes situaciones migratorias) o hijos de padres mexicanos (hispanos de segunda generación) con nacionalidad norteamericana.

La muestra de referencia española, tiene una edad media $\bar{X}$= 42.09 años, con una desviación estándar (DE)=14.41, en un rango de edad de 15 a 77 años. En la muestra mexicana la media de edad es de $\bar{X}$= 43.77 años, DE= 18.33, en un rango de los 15-79 años. Mientras que para la muestra estadounidense, la media de edad es $\bar{X}$= 41.96 años, DE= 14.33, en un rango de los 16-77 años. Se comprobó que no existieran diferencias significativas en cuanto a esta variable mediante el análisis de varianza unidireccional (ANOVA unifactorial) realizado; encontrando para la variable edad: F(2,614)= 0.835, p= .434, se realizó análisis de

la varianza univariante (UNIANOVA) para estudiar el tamaño de los efectos r^2corregida= .00, por tanto el efecto es bajo.

En cuanto al nivel educativo; éste fue evaluado mediante el número de años de asistencia reglada a la escuela, teniendo la muestra de España una media $\bar{X}$ = 11.25 años, con una desviación estándar DE= 3.63; mientras que para la de México $\bar{\bar{X}}$ = 11.46 años, con una desviación estándar DE= 5.60, y para la de EUA $\bar{X}$ = 11.00 y DE= 4.10. Se realizó prueba de ANOVA unifactorial y se encontró que no hubo diferencias significativas en cuanto a medias (F(2,611)= 0.531 y p= .588).

(Ver Tabla 1).

Se requirió utilizar la prueba robusta de igualdad medias de Brown-Forsythe encontrando que F(2,572.02)= 0.830, con una significación asociada de p=.437 para la variable edad r^2corregida= .00; mientras que para escolaridad el estadístico fue: F(2,523.57)= 0.526, p=.591; r^2 corregida = .01 calculada con valor de alfa .05; en ambos casos el tamaño del efecto es bajo. Se puede afirmar que no existen diferencias significativas entre los grupos poblacionales en cuanto a las variables dependientes de edad y nivel educativo. (Ver Tabla 1).

Tabla 1

Distribución de las muestras Edad y Escolarización

Variable	N	Media	DE	Error típico	Intervalo de confianza para la media al 95% Límite inferior	Límite superior	Mínimo	Máximo
Edad								
España	211	42.09	14.411	.992	40.13	44.04	15	77
México	201	43.77	18.327	1.293	41.22	46.32	15	79
EUA	205	41.96	14.337	1.001	39.99	43.94	16	77
Total	617	42.59	15.767	.635	41.35	43.84	15	79
Escolarización								
España	211	11.25	3.635	.250	10.75	11.74	2	21
México	201	11.46	5.602	.395	10.68	12.24	0	23
EUA	202	11.00	4.095	.288	10.43	11.57	0	21
Total	614	11.24	4.505	.182	10.88	11.59	0	23

En la Tabla 2, se presenta la tabla de contingencia de población x sexo. En las muestras predominaron mujeres con un 71%, 73% y 72% de la española, mexicana y estadounidense, respectivamente. No hubo diferencias significativas en cuanto a género, verificado mediante el estadístico Chi cuadrado de Pearson, encontrando que $X^2\ (2,\ N = 617) =$ 0,33, p= 0,84; al estudiar el tamaño de los efectos de esta variable, se encontró que r^2 corregida= .00; por lo que se deduce un tamaño del efecto bajo.

Tabla 2

Distribución de muestras x Sexo

Población	Sexo				Total	
	Masculino		Femenino			
	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje
España	62	27%	149	71%	211	100%
México	54	27%	147	73%	201	100%
EUA	57	28%	148	72%	205	100%
Total	173	27%	444	72%	617	100%

Analizando la interacción de las tres variables en el modelo completo, se encontró r^2=.03 y r^2corregida=.01; es decir el tamaño del efecto es bajo. Teniendo para la interacción sexo*edad F(2,614)=.753 y p=.471; para sexo*escolaridad F(2,614)=1.692 y p=.185; para edad*escolaridad F(4,614)=.708 con p=.587 y en la triple interacción: sexo*edad*escolaridad F(4,614)=1.055 y p=.378. Expuesto lo anterior, se puede concluir que las muestras focal y de referencia, son equivalentes en cuanto a las variables edad, escolarización y sexo (Ver Anexo 4).

4.4.2 Materiales

Para la realización de la presente investigación se utilizó el Test de Recuerdo Verbal Selectivo (TRVS) en su versión en español, desarrollado por

Campo et al. (2000). Como ya se mencionó en el capítulo uno, existen dos formas del test, pero para efectos de la presente investigación sólo se utilizó la forma uno (Anexo 2), también se aplicó un cuestionario de datos personales para la identificación de los participantes en cuanto a las variables sociodemográficas (Anexo 1).

4.4.2.1 Descripción de la prueba

La prueba consta de dos fases: la fase del *recuerdo inmediato* y la del *recuerdo demorado*, en la primera de ellas, se realizan a su vez dos evaluaciones, la primera evaluación que es la de *aprendizaje*, consiste en que el evaluador lee al sujeto la lista de palabras completa (*dado, cinta, norte, jarro, pollo, frente, llave, cruz, fuego, pena, modelo, oído);* se realiza a una velocidad de una palabra cada 2 segundos; para que posteriormente intente recordar la mayor cantidad posible de ellas, teniendo para tal efecto seis ensayos como máximo y siendo recordado sistemáticamente con aquellas palabras que no ha dicho en el ensayo precedente.

Posteriormente se realiza la evaluación de *reconocimiento en elección múltiple,* en la cual se le presentan al sujeto doce tarjetas blancas separadas (ver Anexo 2) con cuatro palabras escritas en color negro, una en cada esquina, cada tarjeta contiene una palabra de la lista y tres distractores: un distractor fonético,

otro semántico y uno no relacionado, de los cuales el sujeto tiene que seleccionar la palabra que considera pertenece a la lista, los estímulos de cada tarjeta que están escritos en el orden de arriba a la izquierda, arriba a la derecha, abajo a la izquierda y abajo a la derecha, son: a) dado, ficha, lado, fácil; b) pipa, lazo, pinta, cinta; c) norte, bar, oeste, corte; d) tinaja, carro, jarro, tiesto; e) duque, pollo, bollo, gallina; f) fuente, costa, cara, frente; g) cerradura, sudor, llave, clave; h) cruz, luz, perro, medalla; i) incendio, juego, fuego, ley; j) vena pena, feliz, llanto; k) tía, patrón, pomelo, modelo; l) caído, oído, olfato, cierto.

La segunda fase que es *recuerdo demorado* se evalúa 30 minutos después de haber concluido las dos primeras actividades del recuerdo inmediato. Consiste en solicitarle al sujeto que mencione la mayor cantidad de palabras que recuerde de la lista, pero teniendo en esta ocasión un solo ensayo, posteriormente se pasa al reconocimiento de elección múltiple siguiendo el mismo proceso que en la fase de recuerdo inmediato.

4.4.2.2 Material del tRVS

El tRVS consta de un protocolo de aplicación y doce tarjetas para reconocimiento por opción múltiple (Ver Anexo 2). El protocolo está conformado por cuatro apartados y los datos generales para identificación.

El primer apartado corresponde a la prueba de aprendizaje tanto inmediato (a la izquierda) como demorado (a la derecha). Para la prueba de aprendizaje inmediato, se cuenta con el listado de palabras que hay que leer al sujeto en orden descendente y en las columnas de la derecha se cuenta con los espacios para anotar el orden en que el sujeto los mencione (si es que lo hace) en cada ensayo (6). A la derecha de este apartado se cuenta nuevamente con el listado de palabras y una columna a la derecha para anotar el orden en el que el sujeto las recuerda en el aprendizaje demorado.

El segundo apartado está destinado para la anotación de las *adiciones* o *intrusiones* que cometa el sujeto (palabras que el sujeto mencione y que no pertenecen a la lista) tanto para los ensayos de recuerdo inmediato como de recuerdo demorado.

El tercer apartado se utiliza para la prueba de reconocimiento con opción múltiple tanto inmediato como demorado. En este apartado se encuentra la lista de palabras y enfrente de ellas los cuatro estímulos que se presentan en las tarjetas, para poder marcar la respuesta dada por el sujeto.

El cuarto apartado está dedicado a la calificación de la prueba en cuanto a los principales factores que mide, por facilidad en cuanto al manejo de términos,

se han mantenido las abreviaturas de la prueba en inglés: En la prueba de aprendizaje inicial se evalúan seis medidas distintas, Recuerdo Total (*Recall*), Recuerdo a corto plazo (*Short Term Retrieval,* STR), Almacenamiento a largo plazo (*Long Term Storage,* LTS), Recuerdo a largo Plazo (*Long Term Retrieval,* LTR), Recuerdo consistente a largo plazo (*Consistent Long Term Retrieval,* CLTR), Recuerdo aleatorio a largo plazo (Random Long Term Retrieval, RLTR). Además de reconocimiento con opción múltiple tanto inmediato (*Multiple Choice Recognition Trial,* MCRT) como demorado (*Delayed Multiple Choice Recognition Trial,* DMCRT) adiciones o intrusiones (*Intrusions,* INT) y Recuerdo Demorado (*Delayed Recall,* DR).

En cuanto a los datos generales de identificación, en la parte superior del protocolo se anotan: nombre del sujeto evaluado y fecha de aplicación y en la parte inferior el nombre y firma del evaluador.

4.4.3 Procedimiento

Con aquellos sujetos que aceptaron participar y que contaban con los criterios de inclusión, se concertó la cita para proceder a la aplicación del instrumento.

La aplicación del instrumento se llevó a cabo en las áreas que fueron facilitadas por las instituciones. En España: despachos del Departamento de Psicología Experimental de la Facultad de Psicología y el Aula de la Experiencia de la Universidad de Sevilla, así como en diversos centros de mayores. En México, en el Centro de Atención Psicológica y la Cámara de Gesell de la Unidad Académica Multidisciplinaria Matamoros-UAT, en el departamento de psicología del Centro para la Juventud y la Familia y en el consultorio de enfermería del Gimnasio Multidisciplinario de la UAT. En el caso de la muestra Estadounidense, la mayoría de las evaluaciones se realizaron en la *Cameron Park Clinic of the Brownsville Community Health Center* (Centro Comunitario de Salud de Brownsville, Clínica Cameron Park), se utilizó el consultorio destinado a *Social Work* (Trabajo social). Para el caso de los estudiantes universitarios de la *University of Texas at Brownsville* (Universidad de Texas en Brownsville), se utilizaron espacios de la biblioteca y aulas de la institución. En todos los casos se buscó que la aplicación se llevara a cabo en cubículos privados que contaran con sillas y escritorio, que hubiera iluminación y ventilación adecuada, con la menor cantidad posible de ruido proveniente del exterior, que pudiera interferir con la atención de los sujetos.

La aplicación se realizó de manera individual con cada sujeto, con una duración promedio de 50 a 60 minutos; la ubicación del evaluador y del sujeto fue sentados cara a cara, uno a cada lado del escritorio. La primera actividad dentro del proceso de evaluación consistió en obtener el consentimiento informado por

parte del sujeto, para ello, se elaboró un formato por escrito (Ver Anexo 3) en el cual se describió a grandes rasgos la investigación, los objetivos del estudio, los procedimientos, implicaciones para los participantes, etc. Se le entregó a cada sujeto para que lo leyera detenidamente y una vez que concluyó, se procedió a aclarar cualquier duda que le pudiera haber surgido durante la lectura, aquellos que decidieron participar en el estudio se les pidió que firmaran el consentimiento informado para poder iniciar la aplicación del protocolo, el evaluador también firmó en el espacio correspondiente.

Una vez leído y firmado el consentimiento informado por parte del evaluador y el evaluado, se procedió a aplicar el TRVS, acorde al siguiente procedimiento:

Se inició con la prueba de recuerdo libre dándole al sujeto la siguiente instrucción:

> *Mire (Nombre del sujeto evaluado), le voy a leer una lista de palabras; quiero que escuche atentamente porque después me gustaría que me dijera todas las palabras que recuerde ¿de acuerdo?*

A continuación el evaluador leyó la lista de 12 palabras, a intervalos de una cada dos segundos.

Las respuestas del sujeto en cada ensayo se registraron en la columna correspondiente al ensayo que se estuviera realizando en el protocolo de respuestas, anotando el orden en que el sujeto mencionó cada palabra.

Si se presentaba el caso de que el sujeto diera una palabra en cuatro ensayos consecutivos y después fallara en darla en las siguientes, se le daba la siguiente indicación:

"Hay una palabra que me ha dicho en varias ocasiones y que no me ha dicho en este intento".

El examinador también tenía la opción de pedirle al sujeto que repitiera la lista completa nuevamente para asegurarse de que no hubiera dejado fuera ninguna palabra por error. Cabe la pena mencionar que el examinador nunca debía deletrear o reafirmar las palabras así como tampoco indicar al sujeto el número total de palabras.

También se registraron las *intrusiones o adiciones* emitidas en cada ensayo; es decir, aquellas palabras que el sujeto daba y que no pertenecían a la lista. Cuando el sujeto cometía una intrusión el examinador le decía lo siguiente: *"esa palabra no está en la lista"* y procedía a registrarlo en el espacio destinado a la anotación de adiciones del protocolo, en el ensayo en el cual se emitió. El mismo procedimiento se siguió con cada intrusión que el sujeto pronunció.

La prueba se dio por terminada cuando: a) las 12 palabras fueron recordadas en tres ensayos consecutivos sin ningún recordatorio, o b) cuando se concluyeron los 6 ensayos.

Después de la prueba de recuerdo libre se pasó a la prueba de *reconocimiento con opción múltiple*. La instrucción que se dio al sujeto es:

> *"Le voy a enseñar unas tarjetas en las que aparecen cuatro palabras, pero sólo una de ellas pertenece a la lista que le leí. Dígame cuál es".*

Posteriormente se le presentaron una a una las tarjetas y se registraron las respuestas en el protocolo de aplicación de la prueba. A partir de que se concluyó esta prueba, se esperaron los 30 minutos necesarios antes de poder evaluar el

recuerdo demorado. Para efectos de la investigación, durante este intervalo de tiempo, se aplicó una batería de pruebas que miden diversos constructos, esto con el fin de que el sujeto pasara los 30 minutos ocupado en otras actividades no relacionadas con la lista de palabras. Cabe hacer mención que para interferir lo menos posible con los resultados del TRVS, ninguna de las pruebas utilizadas midió memoria semántica verbal.

Es importante mencionar, que las pruebas aplicadas fueron distintas para los sujetos, aplicando sólo las que fueron necesarias para cubrir los 30 minutos, dependiendo de la velocidad de respuesta de cada sujeto, para influir lo menos posible en los resultados, se procuró suspender las pruebas complementarias verificando que el tiempo pasado entre el término del vSRT no se alejara de los 30 minutos necesarios para la re-evaluación.

Una vez transcurridos los 30 minutos, se inició la aplicación de la prueba de aprendizaje demorado, dándole al sujeto la siguiente indicación:

"¿Recuerda la lista de palabras que le leí antes? Me gustaría que me dijera todas las palabras que recuerde de esa lista".

Posteriormente se procedió a registrar las palabras emitidas por el sujeto en el espacio indicado dentro del protocolo, se realizó un solo ensayo, en caso de que existieran intrusiones o adiciones, se procedió de la misma forma que se explicó anteriormente en la evaluación del recuerdo inmediato.

La prueba de *reconocimiento múltiple demorado* se realizó ulteriormente, siguiendo el mismo procedimiento que el descrito en la etapa de recuerdo inmediato y se registró de igual forma en el espacio correspondiente dentro del protocolo. Una vez hecho esto se dio por terminada la evaluación.

4.4.4 Evaluación de la prueba

En la prueba de aprendizaje inicial se evalúan seis medidas distintas, Recuerdo Total (Recall), Recuerdo a corto plazo (STR), Almacenamiento a largo plazo (LTS), Recuerdo a largo Plazo (LTR), Recuerdo consistente a largo plazo (CLTR), Recuerdo aleatorio a largo plazo (RLTR). Además de reconocimiento con opción múltiple tanto inmediato como demorado y adiciones o intrusiones.

Para la revisión de la prueba se considera *recuerdo total* a la sumatoria de palabras recordadas en los seis ensayos, aquellas palabras recordadas consistentemente sin necesidad de dar recordatorio, lo cual se califica como

recuerdo consistente de largo plazo, se puede obtener también *memoria a largo plazo* o *almacenamiento a largo plazo* a través del número de palabras recordadas en dos o más intentos consecutivos sin necesidad de recordatorio; *Recuerdo a largo plazo* son palabras que el sujeto recuerda después de que han entrado en el almacenamiento a largo plazo. *Recuerdo de corto plazo* son las palabras nombradas solo después de que se ha brindado el recordatorio: el *recuerdo aleatorio a largo* plazo se refiere a palabras en el almacenamiento a largo plazo, que no reaparecen consistentemente sino que requieren recordatorio.

Para evaluar reconocimiento con opción múltiple inmediato y demorado se hace una sumatoria de las palabras que fueron nombradas correctamente de entre las cuatro posibilidades presentadas, en la primera evaluación y después de transcurridos treinta minutos, respectivamente. Por último se puede obtener una sumatoria también de las palabras que hayan sido mencionadas en los distintos ensayos y que no pertenecen a la lista de palabras de la prueba, lo que se califica como *adiciones o intrusiones.*

4.5 Análisis de los resultados

Para el análisis de los resultados se prepararon los protocolos identificando aquellos que cumplieran con los criterios de inclusión establecidos para cada una de las muestras.

Una vez verificado que los protocolos estuvieran completos se procedió a su revisión, para posteriormente capturar los mismos en la base de datos. Los análisis estadísticos generales se realizaron con SPSS versión 20.

Para el análisis del DIF mediante Mantel-Haenszel se utilizó el paquete estadístico DIFAS (Penfield, 2005), mientras que el análisis del DIF mediante Regresión Logística también se realizó con SPSS versión 20.

CAPÍTULO V

RESULTADOS

5.1 Introducción

En este capítulo se presentan los principales resultados del estudio del DIF realizado, para una mejor comprensión de los mismos se han dividido los resultados con base en las distintas comparaciones dos a dos entre las muestras de los tres países, presentando los resultados para los procedimientos de detección de DIF utilizados: en una primera etapa se presentan los resultados obtenidos de las comparaciones mediante ANOVA univariante realizado a las puntuaciones globales de la prueba.

En la segunda etapa se realiza análisis del DIF para ítems dicotómicos a cada unos de los ítems del vSRT en los distintos ensayos, se presentan resultados del análisis Mantel-Haenszel, utilizando el programa estadístico DIFAS, para cada una de las comparaciones dos a dos de las muestras.

En una tercera etapa se presentan resultados de la aplicación de la técnica de Regresión Logística para ítems dicotómicos a cada uno de los ítems en los 6 ensayos del vSRT, también para cada una de las comparaciones dos a dos entre las muestras del estudio.

En la cuarta etapa se realizaron las purificaciones bietápicas para cada una de las comparaciones y se volvieron a realizar los análisis DIF, mediante Mantel-Haenszel.

En la quinta etapa se realiza una fusión de las muestras focales: mexicana y estadounidense y se realiza análisis del DIF mediante Mantel-Haenszel a cada uno de los ítems del vSRT en los seis ensayos.

En la sexta y última etapa se evalúa el Funcionamiento Diferencial del Test mediante DIFAS para las diferentes comparaciones.

PRIMERA ETAPA

5.2 Análisis de las medias cuantitativas mediante ANOVA entre las tres muestras

Cómo primer paso del estudio del DIF se compararon entre sí las tres muestras mediante procedimiento de ANOVA unifactorial de efectos fijos, en cada una de las medidas cuantitativas globales que componen el vSRT (variables dependientes) RECALL, LTR, STR, LTS, CLTR, RLTR, INT, DR, MCRT y DMCRT, utilizando la variable población como variable independiente. Se utilizó el estadístico de Levene para estudiar la homocedasticidad. Para las variables RECALL, LTR, STR, RLTR, DR e INT, además fue necesario utilizar la prueba robusta de igualdad de medias de Brown-Forsythe, ya que no se cumplió el principio de homocedasticidad; encontrando diferencias significativas en todos los casos (Ver tabla 3).

Tabla 3

Análisis de la varianza puntuaciones globales del vSRT.

Variables	ANOVA		Estadístico de Levene		Brown-Forsythe	
	Valor de F	p	Valor de F	p	Valor de F*	p
RECALL	-	-	7.717	**.000***	23.783	**.000***
LTR	-	-	5.894	**.003***	24.260	**.000***
STR	-	-	123.198	.**000***	58.934	**.000***
LTS	64.576	**.000***	1.602	.202	-	-
CLTR	13.079	**.000***	.620	.538	-	-
RLTR	-	-	3.070	**.047***	15.498	**.000***
DR	63.477	**.000***	2.798	.062	-	-
INT	-	-	23.047	.**000***	25.413	**.000***
MCRT	1.893	.152	1.008	.366	-	-
DMCRT	-	-	3.771	.**024***	1.785	**.169***

Como se presenta en la Tabla 4, en todos los casos en que se encontraron diferencias significativas en las puntuaciones medias, éstas fueron a favor de la población española, con excepción de la variable STR, en donde la mayor puntuación fue para los estadounidenses, seguidos por los mexicanos; sin embargo, estas puntuaciones son menos deseables o puntuaciones que nos indican una mayor memoria a corto plazo en lugar de a largo plazo. En la variable adiciones (INT) los mexicanos tuvieron la puntuación más alta seguidos por los estadounidenses; esta variable también es negativa, ya que las adiciones, son

indicador de confusión con otras palabras que no están en la lista, por lo que tampoco es deseable tener puntuaciones altas. Por lo anterior, se puede afirmar, que la muestra española tuvo un funcionamiento más alto en todas las sub-escalas del vSRT seguido por la mexicana y estadounidense (Ver Anexo 5).

Tabla 4

Puntuaciones globales del vSRT por muestras.

Variables	Media	DE	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
				Límite inferior	Límite superior		
RECALL							
España	47.55	8.752	.602	46.37	48.74	23	69
México	44.47	9.810	.692	43.11	45.84	20	66
EUA	40.82	11.153	.779	39.29	42.36	18	71
Total	44.31	10.306	.415	43.50	45.13	18	71
LTR							
ESPAÑA	36.35	13.052	.899	34.58	38.12	2	69
MÉXICO	32.81	14.309	1.009	30.82	34.80	0	66
E.U.A.	26.48	16.281	1.137	24.24	28.72	0	66
Total	31.92	15.138	.609	30.72	33.12	0	69
STR							
España	11.30	5.659	.390	10.53	12.07	0	27
México	11.74	5.910	.417	10.91	12.56	0	28
EUA	19.94	13.447	.939	18.09	21.79	0	65
Total	14.31	9.898	.398	13.53	15.10	0	65
LTS							
España	39.28	13.112	.903	37.50	41.06	3	69
México	35.56	14.537	1.025	33.54	37.58	0	67
EUA	23.74	15.808	1.104	21.56	25.92	0	68
Total	32.91	15.942	.642	31.65	34.17	0	69
CLTR							
España	25.81	14.808	1.019	23.80	27.82	0	69
México	26.20	14.056	.991	24.24	28.15	0	64
EUA	19.46	15.979	1.116	17.26	21.66	0	64
Total	23.83	15.263	.614	22.62	25.03	0	69

RLTR							
España	10.53	7.147	.492	9.56	11.50	0	42
México	8.46	6.042	.426	7.62	9.30	0	36
EUA	7.05	5.994	.419	6.23	7.88	0	31
Total	8.70	6.575	.265	8.18	9.22	0	42
DR							
España	9.47	2.684	.185	9.10	9.83	0	12
México	7.53	2.563	.181	7.18	7.89	0	12
EUA	6.74	2.369	.165	6.42	7.07	0	12
Total	7.93	2.789	.112	7.71	8.15	0	12
INT							
España	.95	1.241	.085	.78	1.12	0	5
México	2.20	2.055	.145	1.92	2.49	0	10
EUA	1.47	1.947	.136	1.20	1.74	0	10
Total	1.53	1.848	.074	1.39	1.68	0	10
MCRT							
España	12.35	6.910	.476	11.41	13.29	8	12
México	11.61	1.095	.077	11.45	11.76	4	12
EUA	11.77	.563	.039	11.69	11.84	9	12
Total	11.91	4.108	.165	11.59	12.24	4	12
DMCRT							
España	11.75	.767	.053	11.64	11.85	6	12
México	11.60	.981	.069	11.46	11.73	4	12
EUA	11.63	.822	.057	11.52	11.74	6	12
Total	11.66	.861	.035	11.59	11.73	4	12

N=617, *n* España=211, *n* México=201, *n* E.U.A.=205

SEGUNDA ETAPA

5.3 Análisis del DIF para ítems dicotómicos mediante la técnica de Mantel-Haenszel

Para analizar los datos mediante esta técnica se construyó una base de datos en la que se registró si cada sujeto recordó o no cada ítem en cada ensayo. Por tanto, se analizaron 72 variables (12 ítems por seis ensayos). Los grupos fueron equiparados en el nivel de habilidad, utilizando el puntaje total de recuerdo (RECALL) de la prueba como indicador de habilidad (θ). Las comparaciones en este caso siempre se hacen sólo entre dos grupos (G) a la vez (uno de referencia y otro focal), por lo que se hicieron tres evaluaciones mediante este procedimiento. La cantidad de niveles de habilidad se suelen designar de manera arbitraria, aunque lo común es entre dos y cinco. En el caso del paquete DIFAS, en caso de no utilizar la estratificación por una variable externa, el paquete realiza la evaluación del DIF, utilizando la puntuación de todos los ítems que componen el test.

En nuestro caso, se eligieron tres niveles de habilidad, debido a que el tamaño de las muestras es el mínimo recomendado para utilizar este procedimiento. Se utilizó una base de datos con cada uno de los ítems por cada

ensayo, puntuando 0 para no acierto y 1 el acierto. Se corrieron los análisis mediante el comando de *modelos no paramétricos en ítems dicotómicos (DIF analysis: Nonparametric tests for dichotomous ítems)*. Para determinar el grado de DIF se utilizaron los siguientes estadísticos:

***Chi cuadrado de Mantel-Haenszel (MH CHI)**:* Estadístico distribuido según una ley de chi-cuadrado con un grado de libertad. El valor crítico es 3,84 para una tasa de error tipo I del 5%.

Mantel-Haenszel Commom log-odds ratio (MH LOR): Este estadístico se encuentra asintóticamente distribuido según una ley normal. Valores positivos indican DIF a favor del grupo de referencia y negativos DIF a favor del grupo focal.

Logaritmo de la razón de ventajas estandarizado del estadístico Mantel-Haenszel (LOR Z): Este estadístico es el logaritmo del estadístico MH LOR dividido por su error estándar estimado (LOR SE). Valores mayores que 2 o menores que -2 indican existencia de DIF.

Chi cuadrado de Breslow-day (BD): Mide la heterogeneidad del estadístico razón de ventajas. Se distribuye como una ley de chi-cuadrado con un

grado de libertad. Este estadístico ha resultado útil para identificar el DIF no uniforme.

Regla de decisión combinada (CDR): Con este criterio se marcan ("flag") aquellos ítems que presentan algún estadístico significativo (MH LOR Y/O BD).

Criterio propuesto por Ziecky (1993): distingue tres categorías:

Categoría A.- El valor del estadístico Mantel-Haenszel en puntuaciones delta (MH D-DIF), no es significativamente distinto de cero ó su valor absoluto es menor de 1. Su DIF es pequeño.

Categoría B.- El estadístico MH D-DIF es significativamente distinto de cero y su valor absoluto es al menos de 1 y menor de 1,5. Su DIF es moderado.

Categoría C.- El estadístico MH D-DIF es significativamente mayor que 1 y su valor absoluto es mayor o igual que 1,5. Su DIF es alto.

En documento anexo se presentan los resultados descriptivos de cada uno de los ítems, para cada uno de los pares de comparaciones; sin embargo para mayor rapidez se presentan los compendios mediante tablas en este apartado.

5.3.1 Análisis del DIF entre población española y estadounidense mediante MH

Se formaron tres niveles de habilidad de acuerdo a las puntuaciones de recuerdo total (RECALL), como se ve en la Tabla 5; el grupo de Referencia conformado por la muestra Española y Focal EUA. Se presenta tabla de contingencia con frecuencias para tres Niveles habilidad (θ) por (G) Grupo.

Tabla 5

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia España y focal EUA.

Habilidad	Grupo		Total (n)
	Referencia (n)	Focal (n)	
1 (<36)	24	73	97
2 (36-48)	88	83	171
3 (>48)	99	49	148
Total	211	205	416

En las Tablas 6-10 se presentan los valores de los estadísticos del procedimiento Mantel-Haenszel en cada uno de los ensayos del vSRT; se

consideró como Grupo de Referencia a la población española y como Grupo Focal a la población de E.U.A. (Ver Anexo 6).

Utilizando el criterio de Ziecky (1993), en el ensayo 1 tenemos los ítems DADO y PENA con un DIF moderado “B” a favor del grupo de referencia y con DIF alto “C” el ítem OIDO, también para el grupo de referencia. Mientras que los ítems NORTE y POLLO, tuvieron DIF moderado “B” a favor del grupo Focal según los resultados del estadístico MH-LOR. Con el estadístico Breslow-Day, encontramos posible DIF no uniforme en el ítem FUEGO (Ver Tabla 6).

Tabla 6

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	5.2784	1.3066	0.5401	2.4192	0.658	Flag	B
CINTA	2.1765	0.3379	0.2132	1.5849	3.024	OK	A
NORTE	5.0638	-0.5917	0.2525	-2.3434	1.208	Flag	B
JARRO	0.7908	-0.2399	0.2376	-1.0097	0.046	OK	A
POLLO	8.3472	-0.6603	0.2231	-2.9597	1.261	Flag	B
FRENTE	0.3997	0.1645	0.2211	0.744	0.306	OK	A
LLAVE	1.7781	-0.3305	0.2276	-1.4521	0.052	OK	A
CRUZ	2.2917	0.3549	0.2212	1.6044	2.763	OK	A
FUEGO	0.4519	-0.1807	0.233	-0.7755	4.879	OK	A
PENA	6.7947	0.6176	0.2259	2.734	0.008	Flag	B
MODELO	0.1056	-0.1044	0.2344	-0.4454	0.119	OK	A
OIDO	22.5288	1.0267	0.2133	4.8134	1.542	Flag	C

En el segundo ensayo (Tabla 7), se encontraron los ítems DADO y POLLO con DIF alto y moderado, respectivamente a favor de la población de referencia, mientras que los ítems LLAVE y OÍDO, ambos con DIF moderado fueron a favor del grupo focal.

Tabla 7

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	51.8985	1.7196	0.2448	7.0245	0.145	Flag	C
CINTA	1.01	0.2469	0.2208	1.1182	0.103	OK	A
NORTE	0.044	-0.0729	0.2259	-0.3227	0.433	OK	A
JARRO	0.0373	-0.0689	0.2256	-0.3054	0.021	OK	A
POLLO	6.4144	0.5655	0.2143	2.6388	0	Flag	B
FRENTE	0.0558	0.077	0.2212	0.3481	0.074	OK	A
LLAVE	5.2431	-0.5304	0.222	-2.3892	0.842	Flag	B
CRUZ	0.9981	0.2467	0.2215	1.1138	0.791	OK	A
FUEGO	0.0437	0.0667	0.2121	0.3145	2.817	OK	A
PENA	2.1309	0.3323	0.2119	1.5682	0.017	OK	A
MODELO	1.2347	0.2633	0.2156	1.2212	0.003	OK	A
OIDO	4.2564	-0.4656	0.2153	-2.1626	0.031	OK	B

En el tercer ensayo los ítems CINTA, CRUZ y OIDO, todos con DIF moderado “B” a favor del grupo de referencia de acuerdo al MH-LOR (Tabla 8).

Tabla 8

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	2.2497	-0.4947	0.2994	-1.6523	0.721	OK	A
CINTA	4.4174	0.5012	0.227	2.2079	1.046	OK	B
NORTE	0.2847	-0.1472	0.2282	-0.645	0.43	OK	A
JARRO	0.0074	-0.0419	0.2154	-0.1945	0.418	OK	A
POLLO	0.0131	0.0495	0.2205	0.2245	2.252	OK	A
FRENTE	0.003	-0.0135	0.2279	-0.0592	0.002	OK	A
LLAVE	0.8245	0.2278	0.2222	1.0252	1.515	OK	A
CRUZ	7.2722	0.6148	0.22	2.7945	0.062	Flag	B
FUEGO	0.0175	-0.0522	0.2169	-0.2407	1.44	OK	A
PENA	0.7401	0.2081	0.2141	0.972	0.114	OK	A
MODELO	1.7765	0.3087	0.2151	1.4351	0.446	OK	A
OIDO	5.2253	0.5161	0.2156	2.3938	0.039	Flag	B

En el cuarto ensayo, de acuerdo al BD solamente se encuentra un posible DIF no uniforme nuevamente el ítem FUEGO. Mientras que con DIF uniforme moderado a favor del grupo focal el ítem OIDO (Ver Tabla 9).

Tabla 9

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	2.4147	0.4407	0.2597	1.697	2.661	OK	A
CINTA	0.1689	-0.1351	0.2506	-0.5391	0.368	OK	A
NORTE	0.6858	-0.2174	0.2325	-0.9351	0.875	OK	A
JARRO	0.1008	-0.1067	0.2419	-0.4411	2.802	OK	A
POLLO	0.0267	0.0656	0.234	0.2803	0.587	OK	A
FRENTE	0.4367	0.1803	0.232	0.7772	0.893	OK	A
LLAVE	0.2101	0.1249	0.2191	0.5701	2.734	OK	A
CRUZ	0	-0.0294	0.2425	-0.1212	0.027	OK	A
FUEGO	0.6277	0.2011	0.2217	0.9071	4.387	OK	A
PENA	0.0012	-0.0334	0.2259	-0.1479	1.265	OK	A
MODELO	0.0014	-0.0372	0.2376	-0.1566	0.195	OK	A
OIDO	6.3221	-0.5758	0.2201	-2.6161	0.03	Flag	B

En el quinto ensayo como se aprecia en la Tabla 10, no aparecen ítems con DIF.

Tabla 10

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.026	-0.0924	0.2975	-0.3106	0.053	OK	A
CINTA	2.2628	0.4207	0.258	1.6306	0.088	OK	A
NORTE	0.0531	-0.0899	0.2522	-0.3565	0.155	OK	A
JARRO	1.1873	0.2817	0.2341	1.2033	1.204	OK	A
POLLO	0.0075	0.0084	0.2438	0.0345	1.329	OK	A
FRENTE	0.0016	-0.0413	0.2508	-0.1647	1.067	OK	A
LLAVE	0.2551	-0.1484	0.2381	-0.6233	0.029	OK	A
CRUZ	2.1139	0.375	0.2387	1.571	1.513	OK	A
FUEGO	3.07	-0.4196	0.2253	-1.8624	0.649	OK	A
PENA	0.3697	0.1624	0.2246	0.7231	0.877	OK	A
MODELO	0.5182	0.2	0.2376	0.8418	0.04	OK	A
OIDO	3.2534	0.4426	0.231	1.916	0.145	OK	A

En el sexto ensayo, como se presenta en la Tabla 11, de acuerdo al BD un posible DIF no uniforme en el ítem POLLO. Y con DIF moderado a favor del grupo focal el ítem OÍDO.

Tabla 11

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.6948	0.2882	0.292	0.987	1.686	OK	A
CINTA	0.0106	-0.0647	0.2717	-0.2381	0.442	OK	A
NORTE	0.7266	-0.252	0.2566	-0.9821	1.939	OK	A
JARRO	3.0573	-0.4852	0.2588	-1.8748	0.015	OK	A
POLLO	0.0163	-0.003	0.2785	-0.0108	4.484	OK	A
FRENTE	2.4987	0.4195	0.2454	1.7095	0.343	OK	A
LLAVE	3.6134	-0.4953	0.2451	-2.0208	0.029	OK	A
CRUZ	0.0167	-0.001	0.2695	-0.0037	1.923	OK	A
FUEGO	0.5606	0.208	0.2376	0.8754	1.431	OK	A
PENA	0.1763	-0.1312	0.2421	-0.5419	0.06	OK	A
MODELO	0.485	0.2146	0.2602	0.8248	0.16	OK	A
OIDO	5.2661	-0.577	0.2385	-2.4193	1.394	Flag	B

Siguiendo la regla de decisión combinada CDR, los ítems marcados con “flag” son los que se han identificado con mayor problema de DIF. Siendo un total de 6 en el primer ensayo, 3 en el segundo, 2 en el tercero, 1 en el cuarto, ninguno en el quinto y 1 en el sexto.

5.3.2 Análisis del DIF entre población española y mexicana mediante MH

En la Tabla 12 se presentan la tabla de contingencias con las frecuencias para los tres niveles de habilidad de los grupos de referencia (España) y focal (México) para el análisis del DIF, mediante procedimiento Mantel-Haenszel con el paquete estadístico DIFAS (Ver Anexo 7).

Tabla 12

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia España y Focal México.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<36)	24	42	66
2 (36-48)	88	84	172
3 (>48)	99	75	174
Total	211	201	412

En las Tablas 13 a la 18, se presentan los resultados del procedimiento MH aplicado a los ítems, con la población española utilizada como muestra de referencia y la mexicana como focal.

Cómo podemos observar, en el ensayo 1 se encontraron los ítems DADO, PENA, MODELO con DIF Moderado “B” a favor del grupo de referencia, de los cuales el ítem PENA, por el valor del estadístico Breslow-Day es un posible DIF no uniforme. El ítem OÍDO con DIF alto “C” a favor del grupo de referencia, mientras que los ítems POLLO y CRUZ, se detectaron con DIF moderado “B” a favor del grupo focal (Ver Tabla 13).

Tabla 13

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	6.2077	1.3504	0.5218	2.588	1.217	Flag	B
CINTA	0.3354	0.1404	0.2062	0.6809	0.957	OK	A
NORTE	2.6441	-0.4108	0.2358	-1.7422	0.124	OK	A
JARRO	0.0024	-0.0153	0.2304	-0.0664	0.126	OK	A
POLLO	3.8916	-0.4438	0.2154	-2.0604	0.894	OK	B
FRENTE	0	-0.0203	0.2084	-0.0974	3.024	OK	A
LLAVE	1.1008	-0.2507	0.216	-1.1606	1.907	OK	A
CRUZ	4.6053	0.479	0.2128	2.2509	0.21	OK	B
FUEGO	0.249	-0.1357	0.2222	-0.6107	0.189	OK	A
PENA	12.6512	0.7749	0.2124	3.6483	3.939	Flag	B
MODELO	5.2868	0.5715	0.237	2.4114	0.043	Flag	B
OIDO	28.2271	1.0951	0.2062	5.3109	1.056	Flag	C

En el ensayo 2, como se puede apreciar en la Tabla 14; sólo el ítem DADO y CRUZ se identifican como DIF uniforme a favor del grupo de referencia con un nivel moderado “B”.

Tabla 14

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	8.9728	0.7799	0.2526	3.0875	0.652	Flag	B
CINTA	2.4485	0.3585	0.2146	1.6705	0.017	OK	A
NORTE	2.2069	0.341	0.214	1.5935	0.148	OK	A
JARRO	0.032	0.0632	0.219	0.2886	0.078	OK	A
POLLO	0.7754	0.2097	0.2118	0.9901	0.754	OK	A
FRENTE	0.1422	-0.1011	0.21	-0.4814	2.535	OK	A
LLAVE	1.4724	-0.2805	0.2121	-1.3225	0.193	OK	A
CRUZ	6.3831	0.5494	0.2092	2.6262	2.839	Flag	B
FUEGO	2.2738	0.3305	0.2055	1.6083	2.368	OK	A
PENA	1.5427	0.2774	0.2061	1.3459	0.933	OK	A
MODELO	3.2509	0.3975	0.2083	1.9083	0.244	OK	A
OIDO	0.8351	-0.2098	0.2069	-1.014	0.129	OK	A

En el ensayo 3 el ítem OIDO, con DIF uniforme y moderado “B” a favor del grupo de referencia (Ver Tabla 15).

Tabla 15

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.1311	-0.1403	0.2797	-0.5016	0.214	OK	A
CINTA	0.2128	0.1327	0.2301	0.5767	2.416	OK	A
NORTE	0.7085	-0.2084	0.2193	-0.9503	0.462	OK	A
JARRO	0.0055	0.0367	0.2067	0.1776	0.214	OK	A
POLLO	0.8424	-0.224	0.2186	-1.0247	0.522	OK	A
FRENTE	1.6653	0.3086	0.2198	1.404	0.506	OK	A
LLAVE	2.7127	0.3753	0.2139	1.7546	0.939	OK	A
CRUZ	0.5339	0.1861	0.2207	0.8432	0.481	OK	A
FUEGO	2.7381	-0.3757	0.2143	-1.7531	2.23	OK	A
PENA	0.2309	0.1245	0.2116	0.5884	0.575	OK	A
MODELO	4.9626	0.4889	0.2098	2.3303	0.106	OK	B
OIDO	10.3082	0.6745	0.2059	3.2759	1.6	Flag	B

En el ensayo 4 sólo los ítems JARRO y CRUZ con DIF no uniforme de acuerdo al estadístico Breslow-Day y de nivel bajo “A” (Ver Tabla 16).

Tabla 16

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.2419	0.1632	0.2616	0.6239	0.042	OK	A
CINTA	0.229	-0.151	0.2495	-0.6052	2.067	OK	A
NORTE	0.2051	-0.1281	0.227	-0.5643	0.549	OK	A
JARRO	0.2359	-0.1444	0.2387	-0.6049	4.026	OK	A
POLLO	0.2081	-0.135	0.2354	-0.5735	0.541	OK	A
FRENTE	0.1762	-0.1252	0.2338	-0.5355	1.972	OK	A
LLAVE	1.6184	-0.3054	0.2213	-1.38	3.062	OK	A
CRUZ	0.0037	-0.0123	0.2292	-0.0537	4.072	OK	A
FUEGO	0.1284	0.099	0.2126	0.4657	0.083	OK	A
PENA	0.0917	0.0882	0.2144	0.4114	0.002	OK	A
MODELO	1.249	0.284	0.2297	1.2364	0.969	OK	A
OIDO	0.0027	-0.0106	0.2063	-0.0514	0.04	OK	A

En el ensayo 5 los ítems POLLO y FUEGO con DIF moderado “B” a favor del grupo focal, mientras que el ítem OÍDO, nuevamente con DIF moderado “B” a favor del grupo de referencia. El ítem CRUZ, se mantiene como en el ensayo anterior con posible DIF no uniforme y bajo “A” (Tabla 17).

Tabla 17

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.8261	-0.3362	0.3147	-1.0683	1.706	OK	A
CINTA	0.9855	0.2908	0.2588	1.1236	0.134	OK	A
NORTE	0.0779	0.0936	0.2361	0.3964	2.53	OK	A
JARRO	0.0232	-0.0665	0.2428	-0.2739	0.048	OK	A
POLLO	6.2281	-0.7047	0.2706	-2.6042	2.111	Flag	B
FRENTE	0.2098	-0.1466	0.2508	-0.5845	0.164	OK	A
LLAVE	2.379	-0.3879	0.2346	-1.6535	0.889	OK	A
CRUZ	0.1221	0.1117	0.2393	0.4668	4.57	OK	A
FUEGO	13.2462	-0.865	0.2344	-3.6903	3.761	Flag	B
PENA	0.0877	0.0927	0.2259	0.4104	0.152	OK	A
MODELO	0.214	0.1335	0.231	0.5779	1.664	OK	A
OIDO	4.4602	0.5022	0.2261	2.2211	0.192	OK	B

En el último ensayo, como se observa en la Tabla 18, los ítems JARRO y LLAVE son identificados con DIF alto “C” y moderado “B” respectivamente a favor del grupo focal.

Tabla 18

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.0003	0.049	0.2955	0.1658	0.016	OK	A
CINTA	0.6026	-0.2541	0.2787	-0.9117	1.071	OK	A
NORTE	2.0708	-0.3924	0.251	-1.5633	0.013	OK	A
JARRO	11.7432	-1.0104	0.2889	-3.4974	1.219	Flag	C
POLLO	0.077	0.112	0.2703	0.4144	1.101	OK	A
FRENTE	0.0316	-0.0793	0.2583	-0.307	0.738	OK	A
LLAVE	6.8218	-0.6647	0.2448	-2.7153	0.112	Flag	B
CRUZ	0.0948	0.1135	0.2589	0.4384	0.022	OK	A
FUEGO	0.5532	0.1979	0.2301	0.8601	0.005	OK	A
PENA	0.004	0.0414	0.231	0.1792	1.117	OK	A
MODELO	0.9603	0.2855	0.2578	1.1074	0.719	OK	A
OIDO	0.669	-0.2066	0.2222	-0.9298	1.833	OK	A

5.3.3 Análisis del DIF entre población mexicana y estadounidense mediante MH

En la tabla 19 se presentan la tabla de contingencias con las frecuencias (n) para los tres niveles de habilidad (θ) de los grupos (G) de referencia (México) y focal (EUA) para el análisis del DIF, mediante procedimiento Mantel-Haenszel con el paquete estadístico DIFAS (Ver Anexo 8).

Tabla 19

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia México y focal EUA.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<36)	42	73	115
2 (36-48)	84	83	167
3 (>48)	75	49	124
Total	201	205	406

Los resultados de los estadísticos para el análisis MH en los ítems se presentan en las Tablas 20 a 25 presentando los resultados por cada uno de los 6 ensayos.

Como se puede observar en la Tabla 20 en el Ensayo 1 se detectó DIF moderado “B” y uniforme a favor del grupo focal el ítem MODELO.

Tabla 20

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.1594	-0.1912	0.3377	-0.5662	0.569	OK	A
CINTA	0.3951	0.1542	0.2095	0.736	0.086	OK	A
NORTE	0.1143	-0.1073	0.236	-0.4547	1.897	OK	A
JARRO	0.7977	-0.2394	0.2364	-1.0127	0	OK	A
POLLO	1.3666	-0.2754	0.2151	-1.2803	0.153	OK	A
FRENTE	0.1619	0.1152	0.2235	0.5154	0.557	OK	A
LLAVE	0.0044	0.037	0.2139	0.173	2.397	OK	A
CRUZ	0.0675	-0.0854	0.2294	-0.3723	2.359	OK	A
FUEGO	0.0082	0.006	0.231	0.026	3.431	OK	A
PENA	0.0706	-0.086	0.2277	-0.3777	3.359	OK	A
MODELO	5.965	-0.6344	0.2508	-2.5295	0.445	Flag	B
OIDO	0.0246	-0.0565	0.214	-0.264	0.002	OK	A

En el Ensayo 2, con DIF alto “C”, el ítem DADO, a favor del grupo de referencia de acuerdo al estadístico MH-LOR (Ver Tabla 21).

Tabla 21

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	17.8199	0.958	0.2232	4.2921	0.168	Flag	C
CINTA	0.1109	-0.0919	0.2103	-0.437	0.007	OK	A
NORTE	3.4243	-0.4352	0.2219	-1.9612	0.02	OK	A
JARRO	0.1765	-0.115	0.2172	-0.5295	0.001	OK	A
POLLO	1.8989	0.3103	0.2093	1.4826	0.748	OK	A
FRENTE	1.1199	0.2442	0.2094	1.1662	1.115	OK	A
LLAVE	1.0393	-0.2403	0.2135	-1.1255	0.123	OK	A
CRUZ	0.7466	-0.2039	0.2111	-0.9659	1.362	OK	A
FUEGO	1.7673	-0.3116	0.2171	-1.4353	0.095	OK	A
PENA	0.0098	0.0015	0.2131	0.007	1.136	OK	A
MODELO	0.2606	-0.1294	0.2099	-0.6165	0.325	OK	A
OIDO	1.8567	-0.3071	0.2089	-1.4701	0.955	OK	A

En los Ensayos 3 y 6 no se detectaron ítems con DIF (Tablas 22 y 25 respectivamente).

Tabla 22

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.4573	-0.3877	0.2869	-1.3513	0.009	OK	A
CINTA	1.653	0.3134	0.2239	1.3997	0.415	OK	A
NORTE	0.3027	0.1447	0.219	0.6607	1.698	OK	A
JARRO	0.1027	-0.0903	0.2114	-0.4272	0.104	OK	A
POLLO	1.393	0.2857	0.2207	1.2945	0.363	OK	A
FRENTE	2.756	-0.3891	0.22	-1.7686	0.654	OK	A
LLAVE	0.5928	-0.1906	0.2163	-0.8812	0.203	OK	A
CRUZ	3.0703	0.3961	0.213	1.8596	0.289	OK	A
FUEGO	1.0642	0.2519	0.2196	1.1471	0.326	OK	A
PENA	0.0029	0.0109	0.2114	0.0516	1.302	OK	A
MODELO	0.4548	-0.1644	0.2108	-0.7799	0.107	OK	A
OIDO	0.1151	-0.0899	0.2038	-0.4411	2.002	OK	A

En el Ensayo 4 se identificó al ítem OIDO, con DIF moderado “B” a favor del grupo focal. Mientras que con un posible DIF no uniforme el ítem CRUZ (Tabla 23).

Tabla 23

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.232	0.3005	0.2427	1.2382	1.328	OK	A
CINTA	0.0168	-0.0613	0.2438	-0.2514	0.339	OK	A
NORTE	0.0705	-0.0864	0.2275	-0.3798	0.288	OK	A
JARRO	0.0281	-0.0693	0.2405	-0.2881	0.231	OK	A
POLLO	0.452	0.1832	0.2319	0.79	0.015	OK	A
FRENTE	2.2102	0.3585	0.2236	1.6033	0.257	OK	A
LLAVE	2.4343	0.3731	0.2227	1.6753	0.08	OK	A
CRUZ	0.1629	0.1165	0.2249	0.518	3.982	OK	A
FUEGO	0.2565	0.1329	0.2157	0.6161	2.752	OK	A
PENA	0.0916	-0.0902	0.2185	-0.4128	1.323	OK	A
MODELO	2.894	-0.4154	0.2292	-1.8124	0.308	OK	A
OIDO	7.3678	-0.6033	0.2144	-2.8139	0.349	Flag	B

En el Ensayo 5 (Ver Tabla 24) el ítem POLLO, detectado con DIF uniforme y moderado “B” a favor del grupo de referencia.

Tabla 24

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.066	0.1184	0.2946	0.4019	1.581	OK	A
CINTA	0.1357	0.1147	0.236	0.486	0.145	OK	A
NORTE	0.0068	-0.0453	0.2299	-0.197	2.16	OK	A
JARRO	1.6549	0.3162	0.2262	1.3979	0.806	OK	A
POLLO	5.079	0.6091	0.2581	2.3599	0.796	Flag	B
FRENTE	0.1857	0.1347	0.2433	0.5536	0.439	OK	A
LLAVE	2.0062	0.3517	0.2303	1.5271	1.37	OK	A
CRUZ	1.7797	0.3221	0.2222	1.4496	0.586	OK	A
FUEGO	1.4574	0.31	0.2341	1.3242	1.786	OK	A
PENA	0.0029	0.0121	0.2181	0.0555	1.464	OK	A
MODELO	0.2217	0.1302	0.2226	0.5849	0.986	OK	A
OIDO	0.0352	-0.0633	0.2145	-0.2951	0.002	OK	A

Tabla 25

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.1155	0.323	0.2701	1.1959	2.282	OK	A
CINTA	0.6281	0.2461	0.2652	0.928	0.132	OK	A
NORTE	0.806	0.2538	0.2478	1.0242	2.823	OK	A
JARRO	1.4788	0.3724	0.2777	1.341	1.883	OK	A
POLLO	0.035	-0.0843	0.2641	-0.3192	1.456	OK	A
FRENTE	2.889	0.4351	0.2394	1.8175	0.407	OK	A
LLAVE	0.4093	0.1834	0.2404	0.7629	0.046	OK	A
CRUZ	0.0062	-0.051	0.2501	-0.2039	2.44	OK	A
FUEGO	0	0.0261	0.2242	0.1164	1.05	OK	A
PENA	0.1617	-0.1182	0.2283	-0.5177	0.364	OK	A
MODELO	0.2122	-0.1394	0.2404	-0.5799	0.823	OK	A
OIDO	3.4918	-0.4632	0.2335	-1.9837	0.038	OK	A

TERCERA ETAPA

5.4 Análisis del DIF mediante Regresión Logística

Para este análisis se realizaron tres regresiones logísticas para cada uno de los 12 ítems en los 6 ensayos, utilizando el procedimiento de comparación de modelos anidados descrito por (Hidalgo Montesinos et al., 2005). Se probaron tres modelos distintos en cada uno de ellos, utilizando el acierto al ítem o no acierto al ítem como variable dependiente (puntuando 0 en no aciertos y 1 el acierto). En el primer modelo se incluye la variable RECALL como variable independiente y considerado como nivel de habilidad (θ); en el segundo se agrega otra variable independiente que es la población o grupo (focal o referencia) a la que se le asigna valor de 0 al focal y 1 al de referencia. En el tercer modelo se agrega la variable interacción entre la variable RECALL y el grupo. Una vez concluidas las tres regresiones, se comparan los aportes de cada variable a la razón de verosimilitud en cada modelo. En el primer Modelo solo se evalúa ausencia de DIF, ya que solo contempla la relación entre el acierto al ítem y el nivel de habilidad (θ). En un segundo momento se compara la diferencia entre el Modelo 1 (ausencia de DIF), con el Modelo 2 (DIF uniforme) ya que se agrega al modelo la variable Grupo (G), se observa el cambio en el ajuste, que siguiendo a (Hidalgo Montesinos et al., 2005) se conoce como G^2 de diferencia basándose en el valor norma de X^2 con 1*gl* . En un tercer momento se compara la razón de verosimilitud

del Modelo 2 con el del Modelo 3 (DIF No uniforme) para detectar el aporte de la interacción Grupo por Habilidad (Gθ) al Modelo completo.

Cuando no existen valores significativos, nos indica ausencia de DIF, valores de cambios significativos en el modelo 2 nos indican un posible DIF uniforme, valores significativos en el modelo 3 nos indicarían posible DIF no uniforme simétrico. Por último, si ambos coeficientes (modelo 2 y 3) son significativos el ítem podría estar afectado por DIF no-uniforme asimétrico.

En la Tabla 26 se presenta un compendio de los valores para cada uno de los ítems, probando los 3 modelos en cada uno a lo largo de los 6 ensayos, (debido a la extensión del procedimiento, los Anexos 9-11 que contienen los visores de los análisis realizados se presentan por separado).

5.4.1 Análisis del DIF entre población española y estadounidense mediante RL

Como se puede observar en la Tabla 26 en el ensayo 1, se encontró DIF uniforme en los ítems: DADO, NORTE, POLLO, PENA y OIDO, y DIF no uniforme para el ítem FUEGO. En el ensayo 2 se encontró DIF uniforme para DADO, POLLO LLAVE y OIDO. En el ensayo 3, sólo DIF uniforme para DADO, CRUZ y

OIDO. En los ensayos 4 y 5 sólo DIF uniforme en OÍDO y FUEGO respectivamente. En el ensayo 6 DIF uniforme en los ítems NORTE, JARRO, LLAVE y OIDO (Ver Anexo 9).

Tabla 26

Análisis del DIF mediante regresión logística en ítems del vSRT, entre España y EUA.

Ítem Modelo	En. 1	En. 2	En. 3	En. 4	En. 5	En. 6
DADO						
M 1	9.846	71.405	31.687	49.75	28.167	67.656
M 2	**17.088***	**118.492***	**35.630***	51.447	28.754	67.920
M 3	17.344	118.548	35.999	51.454	28.929	68.916
CINTA						
M 1	30.215	32.969	57.346	38.951	44.871	39.654
M 2	31.758	33.658	60.477	39.924	46.314	40.134
M 3	31.806	33.87	61.162	39.943	46.506	43.8
NORTE						
M 1	45.847	60.255	79.42	42.78	75.916	86.236
M 2	**53.397***	60.821	81.052	44.594	76.728	**90.162***
M 3	53.57	60.912	81.062	44.743	76.76	91.45
JARRO						
M 1	14.079	34.767	36.416	40.958	49.163	27.333
M 2	15.409	35.119	36.747	41.613	49.731	**32.993***
M 3	15.619	35.923	37.065	43.274	52.187	33.002
POLLO						
M 1	2.551	26.347	40.429	42.291	31.785	40.643
M 2	**11.304***	**31.311***	40.436	42.423	31.913	40.731
M 3	12.021	31.845	42.847	42.622	32.595	42.36
FRENTE						
M 1	35.293	51.16	32.021	46.071	58.504	54.69
M 2	35.389	51.171	32.093	46.192	58.824	56.33

M 3	36.306	53.559	32.803	48.073	59.365	56.505
LLAVE						
M 1	10.942	39.261	36.884	43.799	56.493	54.652
M 2	14.348	**45.784***	37.118	43.825	57.581	**61.389***
M 3	14.777	48.17	37.344	45.654	58.2	61.419
CRUZ						
M 1	52.83	70.068	47.517	80.682	70.86	69.253
M 2	54.589	70.793	**52.814***	81.371	71.722	69.725
M 3	57.368	72.996	52.936	81.509	74.081	69.97
FUEGO						
M 1	23.797	48.505	46.676	65.018	42.014	57.75
M 2	24.562	48.607	46.939	65.103	**47.709***	57.772
M 3	**28.952***	49.172	47.363	66.504	48.386	57.787
PENA						
M 1	57.287	37.496	36.688	70.382	58.598	64.581
M 2	**64.414***	38.837	37.063	70.971	58.607	65.915
M 3	64.463	40.206	37.73	71.005	61.643	65.952
MODELO						
M 1	41.26	35.358	47.252	42.367	86	41.38
M 2	41.627	36.099	48.694	42.741	86.074	41.467
M 3	41.906	37.142	50.596	43.523	87.764	41.852
OIDO						
M 1	20.871	14.689	22.226	24.465	30.038	26.365
M 2	**44.71***	**21.144***	**27.418***	**34.438***	32.52	**33.98***
M 3	45.819	24.047	27.877	34.855	35.039	34.004

*Valores de cambio significativos para comparaciones con 1*gl* y tasa de error <.05

Nota: En= Ensayos; M 1= Modelo 1, M 2= Modelo 2, M 3= Modelo 3

5.4.2 Análisis del DIF entre población española y mexicana mediante RL

Al aplicar la técnica de RL en los ítems del vSRT entre población española y mexicana, se encontró en el Ensayo 1 DIF uniforme en los ítems: DADO, NORTE, POLLO, CRUZ, PENA, MODELO y OIDO. En el ensayo 2 DIF uniforme para DADO y FRENTE y posible DIF no uniforme para CRUZ. En el Ensayo 3 DIF

uniforme en FUEGO, MODELO y OIDO. En el ensayo 4 con DIF no uniforme en JARRO y CRUZ. En el Ensayo 5, posible DIF uniforme en los ítems: POLLO, LLAVE, FUEGO y OIDO; y DIF no uniforme en el ítem CRUZ. Por último en el Ensayo 6, DIF uniforme en los ítems NORTE, JARRO y LLAVE (Ver Tabla 27, Anexo 10).

Tabla 27

Análisis del DIF mediante regresión logística en ítems del vSRT, entre España y México.

Ítem Modelo	En. 1	En. 2	En. 3	En. 4	En. 5	En. 6
DADO						
M 1	12.966	41.507	38.459	25.359	26.377	28.987
M 2	**20.582***	**49.657***	39.105	25.509	27.934	28.991
M 3	21.756	50.045	39.846	25.789	28.765	29.125
CINTA						
M 1	20.728	35.453	40.513	41.921	44.271	32.124
M 2	20.952	37.166	40.638	42.742	44.816	33.732
M 3	20.989	37.258	41.712	43.009	44.994	37.09
NORTE						
M 1	31.568	64.111	45.384	38.877	39.639	41.59
M 2	**36.102***	65.642	47.012	39.47	39.674	**45.563***
M 3	37.303	66.435	48.709	39.627	42.198	46.551
JARRO						
M 1	17.768	39.702	27.825	43.086	55.93	30.76
M 2	17.85	39.708	27.826	43.869	54.402	**46.167***
M 3	17.887	39.742	27.919	**48.106***	54.441	48.122
POLLO						
M 1	4.048	27.123	24.62	33.549	27.397	26.304
M 2	**8.359***	27.586	26.269	34.522	**36.119***	26.317
M 3	9.377	27.683	27.742	34.647	40.459	26.703

FRENTE						
M 1	26.303	33.053	42.559	27.357	36.218	53.772
M 2	26.379	**33.642***	44.055	27.914	36.891	54.294
M 3	27.19	38.053	44.072	31.301	36.891	55.334
LLAVE						
M 1	2.035	44.054	27.155	33.061	30.573	39.525
M 2	3.799	46.383	29.495	35.789	**34.515***	**48.358***
M 3	8.771	47.737	29.53	38.749	37.712	48.54
CRUZ						
M 1	25.682	48.486	39.057	42.303	41.701	46.062
M 2	**29.557***	54.104	39.297	42.444	41.75	46.065
M 3	29.666	**58.433***	40.459	**46.513***	**46.602***	46.264
FUEGO						
M 1	8.346	39.25	49.129	29.621	44.259	43.144
M 2	8.958	40.765	**53.414***	29.663	**61.04***	43.334
M 3	9.45	41.163	56.779	29.844	64.901	43.351
PENA						
M 1	29.401	49.328	45.048	48.428	69.717	44.74
M 2	**41.898***	50.268	45.173	48.434	69.718	44.743
M 3	43.759	50.488	45.418	48.843	69.883	46.009
MODELO						
M 1	30.743	27.393	45.993	56.123	59.179	44.065
M 2	**35.703***	30.411	**50.252***	57.07	59.237	44.803
M 3	36.403	31.701	50.867	57.094	62.673	44.803
OIDO						
M 1	17.341	32.353	12.653	33.76	38.922	42.849
M 2	**45.782***	34.245	**22.43***	33.842	**42.424***	44.376
M 3	45.961	34.28	22.597	33.983	42.552	45.098

*Valores de cambio significativos para comparaciones con 1*gl* y tasa de error <.05

Nota: En= Ensayos; M 1= Modelo 1, M 2= Modelo 2, M 3= Modelo 3

5.4.3 Análisis del DIF entre población mexicana y estadounidense mediante RL

En la Tabla 28 se presentan los resultados del análisis del DIF mediante Regresión Logística a los ítems del vSRT entre México (población de referencia) y

EUA (población focal). Se puede observar una disminución significativa de ítems identificados con DIF, en comparación a los encontrados en las comparaciones hechas con los españoles. En el ensayo 1 sólo se encontró un ítem con DIF uniforme: MODELO. En el ensayo 2: DADO y NORTE. En el ensayo 3 ninguno. En el Ensayo 4 los ítems MODELO y OIDO con DIF uniforme. En el ensayo 5 el ítem POLLO y en el Ensayo 6 DIF no uniforme para NORTE y DIF uniforme para el ítem OIDO (Ver Anexo 11).

Tabla 28

Análisis del DIF mediante regresión logística en ítems del vSRT, entre México y EUA.

Ítem Modelo	En. 1	En. 2	En. 3	En. 4	En. 5	En. 6
DADO						
M 1	13.943	64.912	47.601	40.965	39.935	60.923
M 2	14.169	**81.974***	50.064	42.465	39.992	62.033
M 3	15.084	82.214	50.121	42.952	40.28	64.233
CINTA						
M 1	29.27	33.538	67.002	49.208	49.264	25.309
M 2	29.889	33.96	68.678	49.337	49.317	26.075
M 3	29.89	34.661	68.753	49.828	50.25	26.075
NORTE						
M 1	45.369	70.893	66.31	43.117	55.479	78.926
M 2	45.617	**76.062***	66.549	43.432	55.648	79.444
M 3	46.285	76.479	68.855	43.433	59.351	**84.334***
JARRO						
M 1	16.81	35.3	40.589	71.097	48.425	48.065
M 2	17.838	35.746	40.909	71.347	49.866	49.314
M 3	18.315	36.309	40.977	72.083	51.96	51.113

POLLO						
M 1	10.947	25.535	57.139	41.061	56.822	48.647
M 2	12.28	27.407	58.346	41.509	**61.549***	48.803
M 3	12.345	28.685	58.439	41.515	63.591	49.293
FRENTE						
M 1	44.582	30.618	33.264	28.075	60.848	66.758
M 2	44.801	31.784	36.236	30.383	61.039	69.43
M 3	44.801	32.263	37.346	30.749	61.666	71.93
LLAVE						
M 1	3.36	36.506	35.381	68.057	40.288	60.187
M 2	3.362	37.564	36.391	70.332	42.474	60.484
M 3	6.852	37.688	36.485	70.556	43.752	60.875
CRUZ						
M 1	47.608	37.749	55.728	55.067	42.768	65.635
M 2	47.8	38.72	58.284	55.139	44.332	65.863
M 3	49.514	39.262	58.982	58.388	45.009	66.887
FUEGO						
M 1	32.286	52.291	76.576	56.843	80.795	55.966
M 2	32.287	54.856	77.598	57.084	82.012	56.01
M 3	34.3	54.863	79.336	60.046	83.59	56.085
PENA						
M 1	27.559	38.18	41.009	63.065	56.13	53.194
M 2	27.714	38.185	41.011	63.41	56.15	53.733
M 3	29.25	41.16	43.001	64.205	58.147	54.758
MODELO						
M 1	25.785	21.325	33.538	46.044	60.718	42.95
M 2	**32.222***	21.721	34.121	**49.558***	60.943	43.504
M 3	32.392	21.753	34.471	50.819	61.371	44
OIDO						
M 1	6.793	19.504	8.008	26.862	24.889	42.99
M 2	6.881	21.869	9.08	**35.679***	25.103	**47.231***
M 3	7.263	24.419	9.329	35.754	26.943	48.344

*Valores de cambios significativos para comparaciones con 1gl y tasa de error <.05

Nota: En= Ensayos; M 1= Modelo 1, M 2= Modelo 2, M 3= Modelo 3

5.4.4 Comparación entre técnicas Mantel-Haenszel y Regresión logística

Como se puede observar existe una gran similitud en la detección de DIF entre ambas técnicas, sin embargo hay una mayor cantidad de detecciones por parte del procedimiento de Regresión Logística, para el caso de la comparación entre la muestra española y mexicana, se observa que el MH, mediante el programa DIFAS, ha detectado como "flag" un total de 12 ítems, mientras que la Regresión Logística un total de 23 (incluyendo todas las detectadas por el MH), esto es algo esperado, ya que sabemos que la Regresión Logística tiene una mayor sensibilidad, a costa de elevar la proporción de errores tipo II (falsos positivos), mientras que el MH es bueno detectando DIF, con menor tasa de error tipo II pero a costa de una mayor tasa de error tipo I (falsos negativos).

Sabemos que una de las ventajas de la RL sobre el MH, es su mayor sensibilidad para detectar DIF no uniforme, sin embargo, al hacer esta primera revisión nos damos cuenta que solamente se ha detectado un ítem con DIF de este tipo: en el primer ensayo el ítem FUEGO. Por lo que podríamos pensar que este tipo de DIF, es muy bajo, casi nulo en este estudio.

Sin embargo el procedimiento de Regresión Logística no nos detecta la direccionalidad del DIF, que en este caso parece ser una situación más frecuente

e inconstante en distintos ítems a lo largo de los ensayos. Lo anterior nos llevaría a sospechar que el DIF encontrado puede estarse confundiendo con el proceso de aprendizaje en los ensayos 2 al 6 y el recuerdo de ítems en los ensayos previos, ya que una de las características de esta prueba es que se brinda recordatorio para los ítems no recordados en el ensayo anterior. Esto explicaría el por qué vemos no solamente que la tasa de identificación de ítems con DIF, disminuye en todas las comparaciones realizadas entre las diferentes muestras conforme avanzan los ensayos, sino también; contrario a lo que generalmente sucede, la direccionalidad del DIF cambia para un mismo ítem de un ensayo a otro.

Vemos por ejemplo en el ensayo 1 los ítems DADO y OÍDO que son identificados con DIF, por ambos procedimientos y con MH vemos que el DIF es a favor del grupo de referencia, sin embargo en el ensayo 2 también son identificados por ambas técnicas, pero mediante el estadístico MH-LOR, detectamos que en el ensayo 2 el DIF es a favor del grupo Focal.

CUARTA ETAPA

5.5 Purificación bietápica en el análisis del DIF mediante Mantel-Haenszel

Tomando en consideración los hallazgos de las primeras cuatro etapas y siguiendo las recomendaciones del estudio del DIF, se ha optado por llevar a cabo un proceso de purificación bietápico.

En esta fase se han eliminado temporalmente los ítems detectados con DIF de las puntuaciones totales, por considerar que estas pudieran estar sesgadas por la presencia de ítems con funcionamiento diferencial.

Se realizaron dos purificaciones distintas, la primera utilizando los ítems detectados por el procedimiento de Regresión Logística y la segunda utilizando las detectadas por el procedimiento de Mantel-Haenszel. En ambos casos se volvieron a correr los análisis solamente con el procedimiento de Mantel-Haenszel que de acuerdo a los análisis previos parece ser el procedimiento más adecuado.

Para tal efecto, el procedimiento seguido fue eliminar de la puntuación RECALL los ítems marcados “flag” de acuerdo a la Combined Decision

Rule (CDR Regla de Decisión Combinada) por el procedimiento de Mantel-Haenszel, y volver a correr los análisis de MH con el paquete estadístico DIFAS, que parece en este caso el procedimiento estadístico más conveniente.

5.5.1 Purificaciones utilizando resultados de Regresión Logística

5.5.1.1 Proceso de purificación con resultados de RL muestras española y mexicana

Se eliminaron un total de 23 ítems detectados por la Regresión Logística con posible DIF, y se obtuvo la puntuación directa para habilidad purificada, esta puntuación se recategorizó en tres niveles de habilidad, para su análisis mediante MH, utilizándose los percentiles 33.33 y 66.66 como puntos de corte para esta nueva categorización, en la Tabla 29 se muestran las frecuencias en cada grupo (Ver Anexo 12).

Tabla 29

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia España y focal México.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<29)	59	82	141
2 (29-35)	77	60	137
3 (>35)	75	59	134
Total	211	201	412

5.5.1.2 Detección del DIF mediante MH en muestras española y mexicana posterior a purificación

Como se puede observar el porcentaje de ítems detectados con DIF, se ha reducido, sin embargo se han mantenido constantes el primer ítem DADO y los últimos tres ítems de la lista en el primer ensayo. El ítem OIDO se presenta con DIF a favor del grupo de referencia en el primero y tercer ensayo. Mientras que los ítems detectados con DIF en los ensayos cinco y seis son a favor del grupo Focal (Ver Tablas 30 a la 35).

Tabla 30

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	7.1847	1.4102	0.5176	2.7245	0.716	Flag	B
CINTA	0.4604	0.162	0.2072	0.7819	0.035	OK	A
NORTE	2.6038	-0.4037	0.2342	-1.7237	0	OK	A
JARRO	0.0001	-0.0287	0.2308	-0.1244	0.312	OK	A
POLLO	3.5217	-0.4261	0.2158	-1.9745	0.846	OK	A
FRENTE	0.0008	0.0162	0.2108	0.0769	1.586	OK	A
LLAVE	1.0831	-0.2481	0.2153	-1.1523	2.09	OK	A
CRUZ	5.0754	0.4984	0.2121	2.3498	0.562	Flag	B
FUEGO	0.2254	-0.1319	0.2244	-0.5878	0.441	OK	A
PENA	14.2092	0.8082	0.2097	3.8541	0.367	Flag	B
MODELO	6.0471	0.5993	0.2335	2.5666	0.006	Flag	B
OIDO	30.5475	1.1259	0.2048	5.4976	0.615	Flag	C

Tabla 31

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	9.5102	0.8043	0.2526	3.1841	2.109	Flag	B
CINTA	2.2898	0.3514	0.2163	1.6246	0.6	OK	A
NORTE	2.1953	0.3461	0.2174	1.592	0.525	OK	A
JARRO	0.158	0.1111	0.2186	0.5082	1.052	OK	A
POLLO	0.6195	0.189	0.2111	0.8953	1.028	OK	A
FRENTE	0.0982	-0.0872	0.2085	-0.4182	0.413	OK	A
LLAVE	1.1753	-0.2508	0.2105	-1.1914	0.655	OK	A
CRUZ	7.8782	0.5986	0.2062	2.903	0.797	Flag	B
FUEGO	2.1496	0.3222	0.2056	1.5671	1.147	OK	A
PENA	1.403	0.274	0.2119	1.2931	0.119	OK	A
MODELO	3.7212	0.4222	0.2075	2.0347	0.097	OK	A
OIDO	1.2975	-0.2665	0.2141	-1.2447	0.493	OK	A

Tabla 32

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.0176	-0.074	0.2744	-0.2697	0.119	OK	A
CINTA	0.3951	0.172	0.2304	0.7465	1.191	OK	A
NORTE	0.5201	-0.1813	0.2181	-0.8313	0.136	OK	A
JARRO	0.0066	0.0386	0.2078	0.1858	0.223	OK	A
POLLO	0.8221	-0.2232	0.2202	-1.0136	2.265	OK	A
FRENTE	1.9311	0.3307	0.2202	1.5018	0.066	OK	A
LLAVE	2.6361	0.3717	0.2145	1.7329	1.266	OK	A
CRUZ	0.6769	0.2076	0.2219	0.9356	1.389	OK	A
FUEGO	2.0308	-0.3225	0.2114	-1.5255	4.658	OK	A
PENA	0.37	0.1502	0.21	0.7152	0.507	OK	A
MODELO	5.7158	0.5185	0.2084	2.488	1.725	Flag	B
OIDO	11.1368	0.7037	0.2063	3.4111	0.418	Flag	B

Tabla 33

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.3519	0.1883	0.2605	0.7228	0.756	OK	A
CINTA	0.1577	-0.1298	0.2483	-0.5228	1.302	OK	A
NORTE	0.0717	-0.0855	0.2248	-0.3803	0.436	OK	A
JARRO	0.0583	-0.0834	0.233	-0.3579	7.577	Flag	A
POLLO	0.3722	-0.1745	0.2395	-0.7286	0.179	OK	A
FRENTE	0.0726	-0.0899	0.2328	-0.3862	0.245	OK	A
LLAVE	1.4154	-0.2875	0.221	-1.3009	0.828	OK	A
CRUZ	0.0102	0.0478	0.2244	0.213	2.238	OK	A
FUEGO	0.1973	0.1164	0.2115	0.5504	0.627	OK	A
PENA	0.0994	0.0914	0.2156	0.4239	0.089	OK	A
MODELO	1.4141	0.2988	0.2287	1.3065	0.968	OK	A
OIDO	0	-0.0233	0.2104	-0.1107	0.344	OK	A

Tabla 34

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.5006	-0.2669	0.3088	-0.8643	0.053	OK	A
CINTA	1.1119	0.3087	0.2599	1.1878	0.2	OK	A
NORTE	0.3704	0.1697	0.234	0.7252	0.874	OK	A
JARRO	0.0975	-0.11	0.2509	-0.4384	0.01	OK	A
POLLO	4.4666	-0.5918	0.2638	-2.2434	3.159	OK	B
FRENTE	0.1878	-0.1406	0.2516	-0.5588	0.331	OK	A
LLAVE	2.3626	-0.3855	0.2344	-1.6446	1.812	OK	A
CRUZ	0.382	0.1747	0.2375	0.7356	1.016	OK	A
FUEGO	10.7453	-0.7567	0.2254	-3.3571	3.746	Flag	B
PENA	0.2502	0.1366	0.2229	0.6128	0.029	OK	A
MODELO	0.4958	0.1845	0.2266	0.8142	3.296	OK	A
OIDO	5.2615	0.5328	0.2225	2.3946	0.046	Flag	B

Tabla 35

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.019	0.0821	0.2908	0.2823	0.104	OK	A
CINTA	0.6546	-0.2585	0.2764	-0.9352	2.73	OK	A
NORTE	1.5685	-0.3335	0.2436	-1.369	0.52	OK	A
JARRO	9.432	-0.8831	0.2783	-3.1732	2.481	Flag	B
POLLO	0.0627	0.1055	0.2725	0.3872	0.025	OK	A
FRENTE	0.0006	-0.0389	0.2554	-0.1523	2.039	OK	A
LLAVE	4.2473	-0.5033	0.2318	-2.1713	0.096	OK	B
CRUZ	0.2449	0.1602	0.2564	0.6248	0.006	OK	A
FUEGO	0.5453	0.1981	0.2316	0.8554	0	OK	A
PENA	0.1079	0.1012	0.2284	0.4431	0.391	OK	A
MODELO	1.2245	0.3155	0.2558	1.2334	0.887	OK	A
OIDO	0.5911	-0.1975	0.224	-0.8817	3.317	OK	A

5.5.1.3 Proceso de purificación con resultados de RL en muestras Española y E.U.A.

Se eliminaron un total de 19 ítems detectados por la Regresión Logística con posible DIF, y se obtuvo la puntuación para habilidad purificada, quedando ubicados con puntuaciones brutas en un rango de 13-50, se recategorizaron en tres niveles de habilidad, utilizándose los percentiles 33.33 y 66.66 como puntos

de corte para esta nueva categorización, quedando distribuidos como se muestra en la Tabla 36.

Tabla 36

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia España y focal E.U.A.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<29)	41	108	149
2 (29-36)	86	58	144
3 (>36)	84	39	123
Total	211	205	416

5.5.1.4 Detección del DIF mediante MH en muestras española y E.U.A. posterior a purificación

Una vez eliminados los ítems detectados por Regresión Logística se volvieron a correr los análisis de MH, mediante el paquete estadístico DIFAS (Ver Anexo 13). Como se puede observar en las tablas 37 a la 42 se identificaron con DIF en el Ensayo 1 los ítems: DADO, PENA y OIDO con DIF a favor del grupo de referencia, mientras que a favor del grupo focal los ítems: NORTE, POLLO.

Observando el estadístico BD el ítem FUEGO parece presentar DIF no uniforme (Ver Tabla 37).

Tabla 37

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense posterior a purificación Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	7.4667	1.5096	0.5437	2.7765	1.366	Flag	C
CINTA	1.1305	0.2609	0.221	1.1805	1.075	OK	A
NORTE	5.3655	-0.6088	0.2534	-2.4025	0.497	Flag	B
JARRO	0.859	-0.2557	0.2422	-1.0557	0	OK	A
POLLO	6.6398	-0.594	0.2242	-2.6494	2.938	Flag	B
FRENTE	0.1943	0.1213	0.2218	0.5469	3.229	OK	A
LLAVE	2.5588	-0.3982	0.2316	-1.7193	0.058	OK	A
CRUZ	1.257	0.2772	0.2265	1.2238	0.814	OK	A
FUEGO	0.5384	-0.1969	0.2359	-0.8347	5.578	Flag	A
PENA	7.9998	0.6531	0.2229	2.93	0.361	Flag	B
MODELO	0.2861	-0.1562	0.2379	-0.6566	0.01	OK	A
OIDO	23.4773	1.0579	0.2164	4.8886	3.002	Flag	C

En el Ensayo 2 como se puede apreciar en la Tabla 38, se marcaron “flag” los ítems: DADO y POLLO con DIF a favor del grupo de referencia y el ítem OÍDO a favor del focal.

Tabla 38

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense posterior a purificación Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	47.5459	1.6905	0.251	6.7351	1.941	Flag	C
CINTA	1.7648	0.3183	0.2208	1.4416	0	OK	A
NORTE	0	-0.0241	0.2248	-0.1072	1.737	OK	A
JARRO	0.3402	-0.1612	0.2319	-0.6951	0.085	OK	A
POLLO	7.4355	0.607	0.2144	2.8312	0.2	Flag	B
FRENTE	0.0055	0.0414	0.2223	0.1862	0.766	OK	A
LLAVE	3.7933	-0.4535	0.2206	-2.0558	1.498	OK	A
CRUZ	0.3756	0.1655	0.2267	0.73	1.393	OK	A
FUEGO	0.0109	-0.0011	0.2203	-0.005	0.804	OK	A
PENA	0.6168	0.199	0.2204	0.9029	1.737	OK	A
MODELO	1.2329	0.2682	0.2191	1.2241	0.374	OK	A
OIDO	5.0305	-0.5156	0.2196	-2.3479	0.125	Flag	B

En la tabla 39 se presentan los datos estadísticos del Ensayo 3, pudiéndose observar los ítems CRUZ y OÍDO con DIF moderado a favor del grupo de referencia.

Tabla 39

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense posterior a purificación Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.5187	-0.4184	0.2989	-1.3998	1.232	OK	A
CINTA	2.8496	0.4228	0.2333	1.8123	0.5	OK	A
NORTE	0.6185	-0.2101	0.2334	-0.9002	0.16	OK	A
JARRO	0.027	-0.0598	0.2184	-0.2738	1.574	OK	A
POLLO	0.0083	0.0049	0.2265	0.0216	3.744	OK	A
FRENTE	0.0019	-0.0368	0.231	-0.1593	0.078	OK	A
LLAVE	0.0266	0.0647	0.231	0.2801	1.879	OK	A
CRUZ	6.6847	0.5948	0.2213	2.6878	0.669	Flag	B
FUEGO	0.104	-0.0949	0.2198	-0.4318	1.61	OK	A
PENA	0.88	0.2272	0.2163	1.0504	0.243	OK	A
MODELO	1.2103	0.2656	0.2196	1.2095	1.293	OK	A
OIDO	5.4819	0.5259	0.2163	2.4313	1.009	Flag	B

En el Ensayo 4 solamente el ítem FUEGO, fue identificado con posible DIF no uniforme al observar el estadístico BD, similar a lo ocurrido con este mismo ítem en el Ensayo 1, aunque no fue marcado “flag” de acuerdo a la CDR (Ver Tabla 40).

Tabla 40

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense posterior a purificación Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.3695	0.3514	0.2665	1.3186	4.747	OK	A
CINTA	0.5015	-0.2173	0.2574	-0.8442	1.611	OK	A
NORTE	0.4323	-0.1809	0.2341	-0.7727	0.286	OK	A
JARRO	0.1545	-0.1266	0.2444	-0.518	2.717	OK	A
POLLO	0.0024	0.0403	0.2386	0.1689	0.003	OK	A
FRENTE	0.1895	0.1308	0.2363	0.5535	0.461	OK	A
LLAVE	0.203	0.125	0.2216	0.5641	3.164	OK	A
CRUZ	0.3867	-0.1895	0.253	-0.749	0.004	OK	A
FUEGO	0.0524	0.0776	0.2268	0.3422	6.275	Flag	A
PENA	0.2748	-0.1504	0.2339	-0.643	0.028	OK	A
MODELO	0.3325	-0.1741	0.2475	-0.7034	0.464	OK	A
OIDO	4.2436	-0.4732	0.2183	-2.1677	0.028	OK	B

En los Ensayos 5 y 6 ningún ítem identificado con DIF (Ver Tablas 41 y 42). De tal forma que se puede observar una reducción de los ítems identificados con DIF a un total de 12.

Tabla 41

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense posterior a purificación Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.0038	-0.0653	0.3039	-0.2149	1.067	OK	A
CINTA	1.3696	0.3429	0.2628	1.3048	0.057	OK	A
NORTE	0.209	-0.1483	0.2547	-0.5823	1.062	OK	A
JARRO	1.5766	0.3241	0.2365	1.3704	0.417	OK	A
POLLO	0.0146	0.0003	0.2475	0.0012	4.32	OK	A
FRENTE	0.001	0.0238	0.2519	0.0945	2.031	OK	A
LLAVE	0.5287	-0.2043	0.2414	-0.8463	0.094	OK	A
CRUZ	2.3397	0.3918	0.2387	1.6414	0.789	OK	A
FUEGO	2.9164	-0.4121	0.2264	-1.8202	0.514	OK	A
PENA	0.297	0.1507	0.2283	0.6601	0.477	OK	A
MODELO	0.3404	0.169	0.2395	0.7056	0.015	OK	A
OIDO	2.7348	0.3999	0.2299	1.7395	2.54	OK	A

Tabla 42

Análisis del DIF mediante Mantel-Haenszel entre población española y estadounidense posterior a purificación Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.6292	0.2786	0.2926	0.9522	4.317	OK	A
CINTA	0.201	-0.1654	0.2803	-0.5901	0.016	OK	A
NORTE	0.2667	-0.1632	0.2523	-0.6468	1.594	OK	A
JARRO	3.2436	-0.5038	0.2616	-1.9258	0.017	OK	A
POLLO	0.0161	-0.0033	0.2814	-0.0117	1.739	OK	A
FRENTE	1.6201	0.3512	0.2501	1.4042	0.346	OK	A
LLAVE	2.1339	-0.3803	0.2398	-1.5859	0.859	OK	A
CRUZ	0.0546	-0.1027	0.2751	-0.3733	0.771	OK	A
FUEGO	0.0319	0.0745	0.2454	0.3036	1.771	OK	A
PENA	0.3057	-0.167	0.246	-0.6789	1.476	OK	A
MODELO	0.1958	0.1497	0.2624	0.5705	0.612	OK	A
OIDO	3.8169	-0.4876	0.2352	-2.0731	0.439	OK	A

5.5.1.5 Proceso de purificación con resultados de RL en muestras mexicana y E.U.A.

Se eliminaron un total de 6 ítems detectados por la Regresión Logística con posible DIF, obteniendo la puntuación directa para habilidad purificada, esta puntuación se recategorizó en tres niveles de habilidad, para su análisis mediante MH, utilizándose los percentiles 33.33 y 66.66 como puntos de corte para esta nueva categorización, en la Tabla 43 se muestran las frecuencias en cada nivel de habilidad por grupo (focal o de referencia).

Tabla 43

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia México y focal E.U.A.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<34)	50	88	138
2 (34-43)	72	71	143
3 (>43)	79	46	125
Total	201	205	406

5.5.1.6 Detección del DIF mediante MH en muestras mexicana y E.U.A. posterior a purificación

Una vez eliminados los ítems detectados por Regresión Logística se volvieron a correr los análisis de MH, mediante el paquete estadístico DIFAS (Ver Anexo 14). Como se puede observar solamente se identificaron con DIF cuatro ítems, en el ensayo 1 el ítem PENA, en el ensayo 2 el ítem DADO, en el ensayo 4 el ítem OIDO y en el 5 el ítem NORTE. Los resultados de los estadísticos obtenidos para este análisis se presentan divididos por ensayos en las Tablas 44 a la 49.

Tabla 44

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.2166	-0.2149	0.3386	-0.6347	0.092	OK	A
CINTA	0.2875	0.1352	0.2101	0.6435	0.595	OK	A
NORTE	0.4935	-0.1971	0.2405	-0.8195	1.736	OK	A
JARRO	0.9773	-0.266	0.2394	-1.1111	0.058	OK	A
POLLO	1.448	-0.2826	0.2153	-1.3126	0.041	OK	A
FRENTE	0.0083	0.0466	0.2268	0.2055	0.64	OK	A
LLAVE	0.0056	0.0392	0.2153	0.1821	2.082	OK	A
CRUZ	0.3225	-0.1574	0.2322	-0.6779	3.378	OK	A
FUEGO	0.0081	-0.0476	0.2324	-0.2048	4.653	OK	A
PENA	0.2584	-0.1425	0.2304	-0.6185	4.249	OK	A
MODELO	5.8425	-0.6371	0.2529	-2.5192	0.073	Flag	B
OIDO	0.1115	-0.0957	0.2162	-0.4426	0.024	OK	A

Tabla 45

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	14.7601	0.8737	0.2228	3.9215	0.135	Flag	C
CINTA	0.5034	-0.1756	0.215	-0.8167	0.1	OK	A
NORTE	3.4376	-0.4307	0.2193	-1.964	0.015	OK	A
JARRO	0.3295	-0.1494	0.2184	-0.6841	0.071	OK	A
POLLO	1.1745	0.2515	0.2112	1.1908	0.877	OK	A
FRENTE	0.6858	0.1974	0.2105	0.9378	0.194	OK	A
LLAVE	1.2369	-0.2592	0.2131	-1.2163	0	OK	A
CRUZ	0.8385	-0.2137	0.2106	-1.0147	1.978	OK	A
FUEGO	2.8083	-0.3997	0.2237	-1.7868	0.131	OK	A
PENA	0.0027	-0.0343	0.2149	-0.1596	1.591	OK	A
MODELO	0.7224	-0.2048	0.2136	-0.9588	0.997	OK	A
OIDO	2.2705	-0.34	0.2107	-1.6137	1.08	OK	A

Tabla 46

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.8882	-0.4398	0.2886	-1.5239	0.461	OK	A
CINTA	1.4004	0.288	0.2219	1.2979	0.026	OK	A
NORTE	0.024	0.0595	0.2231	0.2667	2.459	OK	A
JARRO	0.3099	-0.1411	0.2127	-0.6634	0.427	OK	A
POLLO	0.9706	0.2434	0.2215	1.0989	0.972	OK	A
FRENTE	3.3631	-0.4309	0.2218	-1.9427	0.197	OK	A
LLAVE	1.4153	-0.2871	0.2202	-1.3038	0.8	OK	A
CRUZ	2.0112	0.3301	0.2156	1.5311	0.007	OK	A
FUEGO	0.4826	0.1815	0.2239	0.8106	0.391	OK	A
PENA	0.0276	-0.0587	0.2145	-0.2737	1.187	OK	A
MODELO	0.49	-0.1699	0.2107	-0.8064	1.161	OK	A
OIDO	0.3585	-0.1452	0.2066	-0.7028	1.32	OK	A

Tabla 47

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.9058	0.263	0.2432	1.0814	1.702	OK	A
CINTA	0.2119	-0.144	0.2469	-0.5832	0	OK	A
NORTE	0.0868	-0.0921	0.2258	-0.4079	0.254	OK	A
JARRO	0.3019	-0.1649	0.2453	-0.6722	0.411	OK	A
POLLO	0.2571	0.1453	0.2326	0.6247	0.104	OK	A
FRENTE	1.9661	0.3408	0.2241	1.5207	0.53	OK	A
LLAVE	1.8714	0.3299	0.2225	1.4827	0.112	OK	A
CRUZ	0.1011	0.0971	0.2245	0.4325	4.384	OK	A
FUEGO	0.0487	0.0718	0.2175	0.3301	3.749	OK	A
PENA	0.2234	-0.1279	0.2189	-0.5843	0.794	OK	A
MODELO	4.2062	-0.508	0.2341	-2.17	0.193	OK	B
OIDO	6.9429	-0.5806	0.2126	-2.731	0.542	Flag	B

Tabla 48

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.02	0.0856	0.2962	0.289	0.226	OK	A
CINTA	0.0558	0.084	0.2368	0.3547	0.167	OK	A
NORTE	0.1094	-0.1019	0.2305	-0.4421	5.456	Flag	A
JARRO	1.3139	0.2843	0.2262	1.2569	0.589	OK	A
POLLO	4.4632	0.5778	0.2591	2.23	0.071	OK	B
FRENTE	0.0554	0.0864	0.2426	0.3561	0.385	OK	A
LLAVE	1.7882	0.3333	0.2297	1.451	1.841	OK	A
CRUZ	1.2233	0.2744	0.2242	1.2239	0.77	OK	A
FUEGO	0.9766	0.2577	0.2341	1.1008	1.659	OK	A
PENA	0.021	-0.0565	0.2211	-0.2555	0.542	OK	A
MODELO	0.0549	0.0777	0.2236	0.3475	0.387	OK	A
OIDO	0.2616	-0.1346	0.2173	-0.6194	0.234	OK	A

Tabla 49

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.6064	0.2506	0.2729	0.9183	2.968	OK	A
CINTA	0.4307	0.2107	0.2661	0.7918	0.232	OK	A
NORTE	0.8176	0.2518	0.244	1.032	3.471	OK	A
JARRO	0.9594	0.3105	0.2799	1.1093	1.65	OK	A
POLLO	0.0924	-0.1143	0.2625	-0.4354	1.724	OK	A
FRENTE	2.3657	0.3944	0.2387	1.6523	0.957	OK	A
LLAVE	0.2033	0.1377	0.2406	0.5723	0.043	OK	A
CRUZ	0.0946	-0.1096	0.2517	-0.4354	3.858	OK	A
FUEGO	0.0025	-0.0371	0.2263	-0.1639	0.595	OK	A
PENA	0.224	-0.1344	0.2279	-0.5897	0.609	OK	A
MODELO	0.4409	-0.1896	0.2418	-0.7841	0.464	OK	A
OIDO	3.1281	-0.4311	0.2291	-1.8817	0.046	OK	A

5.5.2 Purificaciones utilizando resultados de Mantel-Haenszel

Para tal efecto, el procedimiento seguido fue eliminar de la puntuación RECALL los ítems marcados “flag” de acuerdo a la Combined Decision Rule (CDR Regla de Decisión Combinada) por el procedimiento de Mantel-Haenszel, y volver a correr los análisis de MH con el paquete estadístico DIFAS.

5.5.2.1 Proceso de purificación con resultados de MH muestras española y mexicana

Se eliminaron un total de 11 ítems detectados en el primer análisis de Mantel-Haenszel con posible DIF, y se obtuvo la puntuación directa para habilidad purificada, quedando comprendida en el rango de 14-59, esta puntuación se recategorizó en tres niveles de habilidad, para su análisis mediante MH, utilizándose los percentiles 33.33 y 66.66 como puntos de corte para esta nueva categorización, en la Tabla 50 se muestran las frecuencias en cada grupo.

Tabla 50

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia España y focal México.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<36)	62	84	146
2 (36-43)	84	60	144
3 (>43)	65	57	122
Total	211	201	412

5.5.2.2 Detección del DIF mediante MH en muestras española y mexicana posterior a purificación

Como se puede observar el porcentaje de ítems detectados con DIF, se ha mantenido muy similar con respecto a la evaluación del DIF sin purificar las puntuaciones. Esto se puede constatar en las Tablas 51 a 56 en donde se presentan los resultados estadísticos del análisis realizado, divididos por ensayos (Ver Anexo 15).

Vemos que se han mantenido constantes en el Ensayo 1 y 2 los ítems DADO y CRUZ con DIF moderado a favor del grupo de referencia (Ver Tablas 51 y 52). En el ensayo 3 (Tabla 53), los ítems MODELO y OIDO también con DIF a favor del grupo de referencia. En el ensayo 4 el ítem JARRO con DIF no homogéneo de acuerdo al estadístico BD (Ver Tabla 54). En el ensayo 5 el ítem FUEGO con DIF moderado a favor del grupo focal, mientras que OIDO se mantiene a favor del grupo de referencia (Tabla 55). En el Ensayo 6 como se observa en la Tabla 56, solamente el ítem JARRO, con DIF moderado a favor del grupo de referencia.

Tabla 51

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	6.994	1.3943	0.5172	2.6959	1.111	Flag	B
CINTA	0.5094	0.1701	0.2082	0.817	0.032	OK	A
NORTE	2.4101	-0.3975	0.2389	-1.6639	0.107	OK	A
JARRO	0.0012	-0.0185	0.2298	-0.0805	0.396	OK	A
POLLO	3.7149	-0.4351	0.2155	-2.019	1.75	OK	A
FRENTE	0.0261	0.0555	0.2088	0.2658	1.797	OK	A
LLAVE	0.761	-0.212	0.2156	-0.9833	1.974	OK	A
CRUZ	5.1321	0.5115	0.2162	2.3659	0.116	Flag	B
FUEGO	0.1127	-0.1003	0.2237	-0.4484	0.016	OK	A
PENA	14.3295	0.8078	0.2091	3.8632	1.71	Flag	B
MODELO	6.185	0.6053	0.2337	2.5901	0.054	Flag	B
OIDO	30.0846	1.1267	0.2058	5.4747	0.126	Flag	C

Tabla 52

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	10.605	0.8399	0.251	3.3462	2.2	Flag	B
CINTA	2.7027	0.3773	0.2152	1.7533	0.515	OK	A
NORTE	3.0156	0.4019	0.2177	1.8461	0.458	OK	A
JARRO	0.1432	0.1084	0.2211	0.4903	0.08	OK	A
POLLO	0.7062	0.1999	0.2108	0.9483	1.031	OK	A
FRENTE	0.1329	-0.0995	0.212	-0.4693	1.577	OK	A
LLAVE	0.9422	-0.2278	0.2112	-1.0786	0.73	OK	A
CRUZ	8.3771	0.6168	0.2068	2.9826	1.916	Flag	B
FUEGO	2.4457	0.3442	0.2062	1.6693	0.235	OK	A
PENA	2.0546	0.3292	0.2133	1.5434	0.096	OK	A
MODELO	4.0687	0.4384	0.207	2.1179	0.008	OK	B
OIDO	0.9502	-0.2325	0.2149	-1.0819	0.274	OK	A

Tabla 53

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.0003	-0.0429	0.2771	-0.1548	0.347	OK	A
CINTA	0.6207	0.21	0.2316	0.9067	0.398	OK	A
NORTE	0.2539	-0.1338	0.2181	-0.6135	0.19	OK	A
JARRO	0.0495	0.0671	0.2057	0.3262	0.343	OK	A
POLLO	0.652	-0.2039	0.2219	-0.9189	1.227	OK	A
FRENTE	2.5391	0.3714	0.218	1.7037	0.065	OK	A
LLAVE	2.8852	0.389	0.2151	1.8085	0.335	OK	A
CRUZ	0.859	0.2312	0.2222	1.0405	1.74	OK	A
FUEGO	1.7551	-0.3061	0.2143	-1.4284	3.237	OK	A
PENA	0.5381	0.1759	0.2092	0.8408	0.288	OK	A
MODELO	6.4487	0.5479	0.2088	2.624	1.842	Flag	B
OIDO	11.0332	0.7009	0.2062	3.3991	0.003	Flag	B

Tabla 54

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.5984	0.2343	0.2597	0.9022	1.222	OK	A
CINTA	0.0478	-0.084	0.246	-0.3415	1.061	OK	A
NORTE	0.0576	-0.0798	0.2257	-0.3536	0.019	OK	A
JARRO	0.065	-0.0879	0.2356	-0.3731	4.066	OK	A
POLLO	0.258	-0.1519	0.2412	-0.6298	0.268	OK	A
FRENTE	0.0491	-0.0794	0.2345	-0.3386	2.403	OK	A
LLAVE	0.9052	-0.2325	0.219	-1.0616	2.651	OK	A
CRUZ	0.0175	0.0549	0.225	0.244	2.98	OK	A
FUEGO	0.3352	0.1436	0.2104	0.6825	1.123	OK	A
PENA	0.1529	0.1082	0.2164	0.5	0.038	OK	A
MODELO	1.7864	0.3307	0.2275	1.4536	1.159	OK	A
OIDO	0	0.0228	0.2099	0.1086	0.2	OK	A

Tabla 55

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.306	-0.2208	0.31	-0.7123	1.103	OK	A
CINTA	1.1286	0.3113	0.2604	1.1955	0.029	OK	A
NORTE	0.6939	0.2208	0.2326	0.9493	0.851	OK	A
JARRO	0.0011	-0.0387	0.2471	-0.1566	0.048	OK	A
POLLO	3.607	-0.533	0.2628	-2.0282	5.14	Flag	A
FRENTE	0.0345	-0.0774	0.2492	-0.3106	0.472	OK	A
LLAVE	2.1191	-0.3594	0.232	-1.5491	4.302	OK	A
CRUZ	0.4566	0.1901	0.2394	0.7941	2.19	OK	A
FUEGO	10.1606	-0.7366	0.2255	-3.2665	4.151	Flag	B
PENA	0.3669	0.16	0.223	0.7175	0.008	OK	A
MODELO	0.7193	0.2183	0.2281	0.957	3.951	OK	A
OIDO	4.7662	0.5129	0.2247	2.2826	0.092	OK	B

Tabla 56

Análisis del DIF mediante Mantel-Haenszel entre población española y mexicana posterior a purificación Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.069	0.1194	0.2921	0.4088	0.051	OK	A
CINTA	0.4002	-0.2095	0.2751	-0.7615	1.377	OK	A
NORTE	1.4961	-0.3341	0.2479	-1.3477	0.079	OK	A
JARRO	9.6287	-0.901	0.2812	-3.2041	1.365	Flag	B
POLLO	0.1597	0.1451	0.2708	0.5358	0.077	OK	A
FRENTE	0.0096	-0.0075	0.2558	-0.0293	0.97	OK	A
LLAVE	3.5005	-0.4583	0.2307	-1.9866	0.06	OK	A
CRUZ	0.3833	0.1917	0.256	0.7488	0.034	OK	A
FUEGO	0.7409	0.2246	0.2307	0.9736	0.041	OK	A
PENA	0.1976	0.1275	0.228	0.5592	0.856	OK	A
MODELO	1.4569	0.3386	0.2544	1.331	0.586	OK	A
OIDO	0.4344	-0.1725	0.2237	-0.7711	1.846	OK	A

5.5.2.3 Proceso de purificación con resultados de MH en muestras española y E.U.A.

Se eliminaron un total de 12 ítems detectados “flag” con el primer procedimiento MH, indicando posible DIF de la puntuación de recuerdo total RECALL, utilizada como estimador de habilidad. Se obtuvo la puntuación para habilidad purificada, quedando ubicados con puntuaciones brutas en un rango de 15-59, se recategorizaron en tres niveles de habilidad como se muestra en la

Tabla 57, utilizándose los percentiles 33.33 y 66.66 como puntos de corte para esta recodificación.

Tabla 57

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia España y focal E.U.A.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<34)	40	105	145
2 (34-42)	89	58	147
3 (>42)	82	42	124
Total	211	205	416

5.5.2.4 Detección del DIF mediante MH en muestras española y E.U.A. posterior a purificación

En las Tablas 58 a 63 se presentan los datos del análisis del DIF, en esta etapa; como se puede observar se ha disminuido la cantidad de ítems detectados con DIF de 12 a 10 siguiendo la "CDR" (Ver Anexo 16).

En la Tabla 58 se presentan los estadísticos del Ensayo 1; se han identificado con DIF en un grado alto (“C”) a favor del grupo de referencia los ítems DADO y OIDO y con DIF moderado (“B”) el ítem PENA. Mientras que los ítems NORTE y POLLO con DIF moderado, ambos a favor del grupo focal.

Tabla 58

Análisis del DIF mediante Mantel-Haenszel entre población española y E.U.A. posterior a purificación Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	7.5209	1.5023	0.5401	2.7815	1.396	Flag	C
CINTA	1.8995	0.3268	0.2187	1.4943	1.855	OK	A
NORTE	5.0361	-0.6012	0.2556	-2.3521	0.323	Flag	B
JARRO	0.6112	-0.2172	0.2391	-0.9084	0.001	OK	A
POLLO	6.9597	-0.6087	0.224	-2.7174	1.931	Flag	B
FRENTE	0.2369	0.1336	0.2234	0.598	0.461	OK	A
LLAVE	2.1603	-0.3628	0.2291	-1.5836	0.071	OK	A
CRUZ	1.6643	0.3128	0.225	1.3902	2.516	OK	A
FUEGO	0.537	-0.1974	0.2366	-0.8343	4.733	OK	A
PENA	8.7903	0.6821	0.2223	3.0684	0.396	Flag	B
MODELO	0.0553	-0.0827	0.2341	-0.3533	0.44	OK	A
OIDO	23.6889	1.0515	0.2149	4.893	4.16	Flag	C

En el ensayo 2 solamente los ítems DADO y POLLO con DIF alto y moderado respectivamente, a favor de la muestra española (Tabla 59).

Tabla 59

Análisis del DIF mediante Mantel-Haenszel entre población española y E.U.A. posterior a purificación Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	49.7819	1.7035	0.2481	6.8662	2.177	Flag	C
CINTA	1.6197	0.3056	0.2208	1.3841	0.134	OK	A
NORTE	0.0819	-0.0916	0.2288	-0.4003	2.158	OK	A
JARRO	0.0451	-0.0747	0.2289	-0.3263	0.42	OK	A
POLLO	7.7713	0.6183	0.2139	2.8906	0.249	Flag	B
FRENTE	0.0218	0.058	0.2228	0.2603	0.768	OK	A
LLAVE	3.7083	-0.4487	0.2206	-2.034	1.989	OK	A
CRUZ	0.6521	0.2071	0.2244	0.9229	1.897	OK	A
FUEGO	0.0554	0.074	0.2154	0.3435	1.365	OK	A
PENA	0.7122	0.2099	0.2188	0.9593	1.142	OK	A
MODELO	1.5486	0.296	0.2186	1.3541	0.128	OK	A
OIDO	5.0024	-0.5177	0.2209	-2.3436	0.367	OK	B

En el ensayo 3 con DIF moderado a favor del grupo de referencia (Ver Tabla 60) los ítems CRUZ y OIDO.

Tabla 60

Análisis del DIF mediante Mantel-Haenszel entre población española y E.U.A. posterior a purificación Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.8769	-0.459	0.2998	-1.531	0.549	OK	A
CINTA	4.0175	0.4882	0.2302	2.1208	0.006	OK	B
NORTE	0.4535	-0.185	0.2338	-0.7913	0.029	OK	A
JARRO	0.0089	-0.0029	0.2159	-0.0134	2.271	OK	A
POLLO	0.0135	0.0512	0.2239	0.2287	3.297	OK	A
FRENTE	0.0001	-0.0247	0.2304	-0.1072	0.005	OK	A
LLAVE	0.1205	0.106	0.2285	0.4639	2.47	OK	A
CRUZ	7.6228	0.6356	0.2219	2.8644	0.329	Flag	B
FUEGO	0.1751	-0.1174	0.2219	-0.5291	1.557	OK	A
PENA	1.4262	0.2804	0.2147	1.306	0.003	OK	A
MODELO	1.4708	0.2914	0.2199	1.3251	0.09	OK	A
OIDO	6.3456	0.5569	0.2143	2.5987	0.326	Flag	B

Observando el estadístico BD del ítem POLLO con DIF pequeño y no uniforme en el ensayo 5 (Tabla 62). En la Tabla 61 y 63 correspondientes a los ensayos 4 y 6 no se detectó ningún ítem.

Tabla 61

Análisis del DIF mediante Mantel-Haenszel entre población española y E.U.A. posterior a purificación Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	2.1764	0.4281	0.2641	1.621	3.628	OK	A
CINTA	0.4632	-0.2085	0.2565	-0.8129	0.546	OK	A
NORTE	0.5384	-0.1997	0.235	-0.8498	0.121	OK	A
JARRO	0.1473	-0.1246	0.2447	-0.5092	2.303	OK	A
POLLO	0.0028	0.0412	0.239	0.1724	0.007	OK	A
FRENTE	0.4879	0.191	0.2344	0.8148	0.589	OK	A
LLAVE	0.2513	0.1357	0.221	0.614	2.564	OK	A
CRUZ	0.0982	-0.1089	0.2485	-0.4382	0.031	OK	A
FUEGO	0.2275	0.133	0.2253	0.5903	2.737	OK	A
PENA	0.0847	-0.0944	0.2315	-0.4078	0.917	OK	A
MODELO	0.184	-0.1352	0.2449	-0.5521	0.014	OK	A
OIDO	4.6579	-0.4982	0.22	-2.2645	0	OK	B

Tabla 62

Análisis del DIF mediante Mantel-Haenszel entre población española y E.U.A. posterior a purificación Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.0182	-0.0878	0.3042	-0.2886	0.443	OK	A
CINTA	2.1419	0.4109	0.259	1.5865	0.264	OK	A
NORTE	0.3049	-0.1757	0.2585	-0.6797	0.839	OK	A
JARRO	2.3222	0.3823	0.2345	1.6303	1.14	OK	A
POLLO	0.0006	-0.0249	0.2485	-0.1002	6.194	Flag	A
FRENTE	0.0101	0.0063	0.2526	0.0249	0.691	OK	A
LLAVE	0.1712	-0.1262	0.2373	-0.5318	0.098	OK	A
CRUZ	2.6181	0.4135	0.239	1.7301	0.458	OK	A
FUEGO	3.8229	-0.4742	0.2296	-2.0653	0.404	OK	A
PENA	0.4945	0.1834	0.225	0.8151	0.976	OK	A
MODELO	0.5986	0.2144	0.2392	0.8963	0.158	OK	A
OIDO	2.9093	0.4079	0.2285	1.7851	1.907	OK	A

Tabla 63

Análisis del DIF mediante Mantel-Haenszel entre población española y E.U.A. posterior a purificación Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.6597	0.2854	0.2934	0.9727	4.607	OK	A
CINTA	0.0764	-0.1159	0.2784	-0.4163	0.035	OK	A
NORTE	0.576	-0.2295	0.2573	-0.892	2.901	OK	A
JARRO	3.2469	-0.5027	0.2613	-1.9238	0.016	OK	A
POLLO	0.0187	0.0007	0.2823	0.0025	1.921	OK	A
FRENTE	2.144	0.396	0.249	1.5904	0.004	OK	A
LLAVE	3.3785	-0.483	0.2463	-1.961	0.864	OK	A
CRUZ	0.0408	-0.0943	0.2762	-0.3414	0.846	OK	A
FUEGO	0.1515	0.1249	0.2429	0.5142	0.828	OK	A
PENA	0.1408	-0.1226	0.245	-0.5004	1.934	OK	A
MODELO	0.1521	0.1356	0.2623	0.517	0.96	OK	A
OIDO	3.1561	-0.4454	0.2344	-1.9002	1.523	OK	A

5.5.2.5 Proceso de purificación con resultados de MH en muestras mexicana y E.U.A.

Se eliminaron de la puntuación RECALL, considerada como estimador de habilidad, un total de 4 ítems detectados en la primera etapa de análisis del DIF con Mantel-Haenszel con posible DIF, obteniendo la puntuación directa para habilidad purificada en un rango de 17 a 63. Esta puntuación se recategorizó en tres niveles de habilidad, para su segundo análisis mediante MH, utilizándose los

percentiles 33.33 y 66.66 como puntos de corte para esta nueva categorización. En la Tabla 64 se presentan las frecuencias en cada nivel de habilidad por grupo (focal o de referencia).

Tabla 64

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia México y focal E.U.A.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<36)	49	89	138
2 (36-45)	74	66	140
3 (>45)	78	50	128
Total	201	205	406

5.5.2.6 Detección del DIF mediante MH en muestras mexicana y E.U.A. posterior a purificación

Una vez eliminados los ítems detectados por la primera etapa de MH, se volvieron a correr los análisis del DIF mediante el paquete estadístico DIFAS. Como se puede observar en las Tablas 65 a la 70, en donde se presentan los resultados de los estadísticos para cada ítem, divididos por Ensayos 1-6. Igual que

en la primera etapa del análisis con Mantel-Haenszel, únicamente fueron detectados “flag” cuatro ítems (Ver Anexo 17).

En la Tabla 65 se observa al ítem MODELO con DIF moderado a favor del grupo focal.

Tabla 65

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.2311	-0.2167	0.3365	-0.644	0.821	OK	A
CINTA	0.2873	0.1351	0.2102	0.6427	0.108	OK	A
NORTE	0.3976	-0.18	0.2406	-0.7481	1.224	OK	A
JARRO	0.7077	-0.2283	0.2372	-0.9625	0.018	OK	A
POLLO	1.3915	-0.2777	0.2151	-1.291	0.077	OK	A
FRENTE	0.0297	0.0648	0.2262	0.2865	0.783	OK	A
LLAVE	0.0419	0.0671	0.2153	0.3117	2.137	OK	A
CRUZ	0.204	-0.1284	0.2297	-0.559	3.844	OK	A
FUEGO	0.0059	-0.0092	0.2314	-0.0398	3.564	OK	A
PENA	0.1682	-0.1198	0.2298	-0.5213	4.006	OK	A
MODELO	5.4253	-0.6014	0.2491	-2.4143	0.938	Flag	B
OIDO	0.0383	-0.0651	0.2148	-0.3031	0.167	OK	A

En el Ensayo 2 (Tabla 66) el ítem DADO con DIF elevado “C”, a favor de la muestra de referencia (mexicana).

Tabla 66

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	16.5086	0.9196	0.2225	4.133	0.117	Flag	C
CINTA	0.3228	-0.1442	0.2137	-0.6748	0.104	OK	A
NORTE	4.631	-0.5131	0.2268	-2.2623	0.079	OK	B
JARRO	0.3007	-0.1441	0.2187	-0.6589	0.173	OK	A
POLLO	1.5955	0.2887	0.2108	1.3695	0.851	OK	A
FRENTE	1.1779	0.2509	0.2102	1.1936	0.623	OK	A
LLAVE	1.43	-0.2798	0.2152	-1.3002	0.113	OK	A
CRUZ	0.7265	-0.2015	0.211	-0.955	1.489	OK	A
FUEGO	2.1199	-0.3449	0.2201	-1.567	0.057	OK	A
PENA	0.0078	-0.0038	0.213	-0.0178	2.037	OK	A
MODELO	0.4012	-0.1565	0.2114	-0.7403	0.864	OK	A
OIDO	1.8216	-0.3039	0.2088	-1.4555	2.194	OK	A

En el ensayo 3, 5 y 6 ningún ítem fue marcado “flag” (Ver Tablas 67, 69 y 70). Mientras que en el Ensayo 4 el ítem FUEGO con DIF no homogéneo y bajo al observar el estadístico BD. El ítem OIDO también en el mismo ensayo con DIF moderado a favor del grupo focal E.U.A. (Tabla 68).

Tabla 67

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	2.0375	-0.454	0.2892	-1.5698	0.014	OK	A
CINTA	1.4981	0.2963	0.2218	1.3359	0.449	OK	A
NORTE	0.1154	0.0993	0.2205	0.4503	3.719	OK	A
JARRO	0.2451	-0.1284	0.2131	-0.6025	0.573	OK	A
POLLO	0.939	0.2408	0.2221	1.0842	0.419	OK	A
FRENTE	3.363	-0.4342	0.2233	-1.9445	0.072	OK	A
LLAVE	0.9743	-0.241	0.2191	-1.1	0.274	OK	A
CRUZ	2.2314	0.3452	0.215	1.6056	0.027	OK	A
FUEGO	0.6957	0.2085	0.2199	0.9482	0.42	OK	A
PENA	0.0086	-0.043	0.2147	-0.2003	1.302	OK	A
MODELO	0.3806	-0.1515	0.2097	-0.7225	0.382	OK	A
OIDO	0.145	-0.0994	0.2054	-0.4839	1.755	OK	A

Tabla 68

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.0024	0.2747	0.2432	1.1295	1.511	OK	A
CINTA	0.0934	-0.1057	0.2462	-0.4293	0.003	OK	A
NORTE	0.2544	-0.142	0.2294	-0.619	0.358	OK	A
JARRO	0.3667	-0.1803	0.2476	-0.7282	1.045	OK	A
POLLO	0.4118	0.1749	0.2306	0.7585	0.063	OK	A
FRENTE	2.1377	0.3529	0.2237	1.5776	0.332	OK	A
LLAVE	1.8019	0.3263	0.2239	1.4573	0.132	OK	A
CRUZ	0.0728	0.0864	0.2252	0.3837	4.033	OK	A
FUEGO	0.104	0.093	0.216	0.4306	6.074	Flag	A
PENA	0.2697	-0.1396	0.221	-0.6317	0.267	OK	A
MODELO	3.4828	-0.4595	0.2318	-1.9823	0.029	OK	A
OIDO	6.8825	-0.5796	0.213	-2.7211	0.478	Flag	B

Tabla 69

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.0351	0.0978	0.2941	0.3325	1.303	OK	A
CINTA	0.0439	0.078	0.2376	0.3283	0.602	OK	A
NORTE	0.1119	-0.1044	0.2322	-0.4496	3.133	OK	A
JARRO	1.2694	0.2813	0.2273	1.2376	1.196	OK	A
POLLO	4.1191	0.5522	0.2591	2.1312	0.629	OK	B
FRENTE	0.0955	0.1042	0.2424	0.4299	1.143	OK	A
LLAVE	1.573	0.3183	0.2319	1.3726	0.88	OK	A
CRUZ	1.1082	0.2639	0.2256	1.1698	0.134	OK	A
FUEGO	1.0847	0.2675	0.2316	1.155	1.792	OK	A
PENA	0.0001	-0.0269	0.2204	-0.1221	1.635	OK	A
MODELO	0.1146	0.1009	0.2233	0.4519	0.441	OK	A
OIDO	0.1022	-0.093	0.2168	-0.429	0.005	OK	A

Tabla 70

Análisis del DIF mediante Mantel-Haenszel entre población mexicana y estadounidense posterior a purificación Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.7722	0.2759	0.2705	1.02	2.452	OK	A
CINTA	0.4453	0.2143	0.2668	0.8032	0.09	OK	A
NORTE	0.4619	0.2016	0.2494	0.8083	3.302	OK	A
JARRO	0.8254	0.2888	0.2802	1.0307	4.108	OK	A
POLLO	0.1222	-0.1282	0.2651	-0.4836	1.034	OK	A
FRENTE	2.4365	0.3989	0.2383	1.6739	0.97	OK	A
LLAVE	0.2683	0.1538	0.2402	0.6403	0.154	OK	A
CRUZ	0.0735	-0.1007	0.2523	-0.3991	3.395	OK	A
FUEGO	0.004	-0.0112	0.2262	-0.0495	0.806	OK	A
PENA	0.2528	-0.1411	0.2282	-0.6183	0.163	OK	A
MODELO	0.3852	-0.1783	0.2415	-0.7383	1.586	OK	A
OIDO	3.6616	-0.4684	0.232	-2.019	0.72	OK	A

QUINTA ETAPA

5.6 Análisis del DIF en las muestras focales fusionadas

Al observar con base en los análisis anteriores, que las muestras focales parecen tener un funcionamiento diferencial en cuanto al vSRT mayor con respecto a todos los análisis utilizando la muestra española como grupo de referencia. Por lo anterior, se decidió incluir esta quinta etapa en la que se realiza un nuevo análisis del DIF con las muestras focales fusionadas: mexicana y estadounidense, para su comparación con el grupo de referencia español.

Se realizó una recategorización de los niveles de habilidad, quedando distribuidos los participantes por grupo de referencia y focal como se presenta en la Tabla 71.

Tabla 71

Tabla de contingencia: (n) participantes x Niveles de habilidad (θ) por población de referencia España y focales fusionadas.

Habilidad (θ)	Grupo (G)		Total (n)
	Referencia (n)	Focal (n)	
1 (<42)	50	197	247
2 (42-49)	73	95	168
3 (>49)	88	114	202
Total	211	406	617

Se volvieron a correr los análisis del DIF, utilizando el procedimiento de Mantel-Haenszel mediante el paquete DIFAS (Ver Anexo 18). Los principales estadísticos para cada ítem se presentan divididos por Ensayos en cada una de las Tablas 72 a la 77.

Como se aprecia en la Tabla 72 correspondiente al ensayo 1, es donde se presentaron mayor cantidad de ítems marcados “flag”, 6 de acuerdo a la CDR. Siendo éstos: DADO, MODELO con DIF fuerte y PENA con DIF moderado a favor del grupo de referencia. Mientras que los ítems NORTE y POLLO, con DIF moderado a favor del grupo focal.

Tabla 72

Análisis del DIF mediante Mantel-Haenszel entre población española y muestras focales fusionadas Ensayo 1.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	8.0577	1.4166	0.4975	2.8474	2.066	Flag	C
CINTA	1.2429	0.2201	0.1827	1.2047	0.83	OK	A
NORTE	5.3024	-0.5045	0.2113	-2.3876	0.309	Flag	B
JARRO	0.3643	-0.1414	0.2007	-0.7045	0.078	OK	A
POLLO	7.4648	-0.5335	0.1906	-2.7991	0.306	Flag	B
FRENTE	0.0025	0.0259	0.1833	0.1413	3.221	OK	A
LLAVE	1.2329	-0.2317	0.1917	-1.2087	0.165	OK	A
CRUZ	3.191	0.3476	0.1862	1.8668	0.168	OK	A
FUEGO	0.4633	-0.1507	0.1958	-0.7697	2.501	OK	A
PENA	13.7377	0.7059	0.1857	3.8013	0.198	Flag	B
MODELO	1.6328	0.2747	0.1979	1.3881	0.221	OK	A
OIDO	34.2809	1.0683	0.1815	5.886	0.353	Flag	C

En el ensayo 2, solamente el primer ítem: DADO con DIF fuerte y CRUZ con DIF moderado ambos a favor del grupo de referencia español; aunque el ítem CRUZ, también presenta elevación en el estadístico Breslow-Day, por lo que podríamos sospechar de un DIF no uniforme o mixto.

Tabla 73

Análisis del DIF mediante Mantel-Haenszel entre población española y muestras focales Ensayo 2.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	30.1217	1.2005	0.2244	5.3498	3.165	Flag	C
CINTA	1.9987	0.2855	0.1895	1.5066	0.32	OK	A
NORTE	0.3202	0.125	0.1899	0.6582	0.82	OK	A
JARRO	0.0018	0.0104	0.1922	0.0541	0.066	OK	A
POLLO	2.6446	0.3179	0.1853	1.7156	0.321	OK	A
FRENTE	0.0001	-0.015	0.1857	-0.0808	1.05	OK	A
LLAVE	3.7866	-0.383	0.1868	-2.0503	3.484	OK	A
CRUZ	5.1951	0.4446	0.1862	2.3878	5.321	Flag	B
FUEGO	0.4875	0.143	0.1814	0.7883	0.519	OK	A
PENA	1.7756	0.2596	0.1814	1.4311	0.289	OK	A
MODELO	3.2213	0.3494	0.1854	1.8846	0.217	OK	A
OIDO	3.8153	-0.3732	0.1825	-2.0449	0.112	OK	A

En el ensayo 3, los ítems identificados son los dos últimos: MODELO y OÍDO con DIF moderado a favor del grupo focal. Mientras que con posible DIF no uniforme y bajo el ítem FUEGO, según lo indica el estadístico Breslow-Day (Referirse a Tabla 74).

Tabla 74

Análisis del DIF mediante Mantel-Haenszel entre población española y muestras focales Ensayo 3.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.8164	-0.2551	0.2467	-1.034	0.405	OK	A
CINTA	1.9775	0.3009	0.2004	1.5015	0.926	OK	A
NORTE	0.4866	-0.1538	0.1934	-0.7952	0.004	OK	A
JARRO	0.0141	-0.0383	0.1824	-0.21	0.637	OK	A
POLLO	0.1806	-0.0994	0.1914	-0.5193	2.307	OK	A
FRENTE	0.3594	0.136	0.1949	0.6978	0	OK	A
LLAVE	1.1486	0.2218	0.1909	1.1619	2.676	OK	A
CRUZ	2.5143	0.3222	0.193	1.6694	1.218	OK	A
FUEGO	1.7999	-0.2621	0.1854	-1.4137	7.452	Flag	A
PENA	0.3869	0.1314	0.1843	0.713	0.747	OK	A
MODELO	5.7108	0.4564	0.1836	2.4858	2.444	Flag	B
OIDO	10.5567	0.6116	0.1841	3.3221	0.194	Flag	B

En el ensayo 4, nuevamente se identifican posibles DIF no uniformes de nivel bajo en los ítems JARRO y LLAVE (Ver Tabla 75).

Tabla 75

Análisis del DIF mediante Mantel-Haenszel entre población española y muestras focales Ensayo 4.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	1.5207	0.3074	0.2287	1.3441	0.426	OK	A
CINTA	0.3193	-0.1451	0.2154	-0.6736	0.833	OK	A
NORTE	0.7436	-0.1898	0.1983	-0.9571	0.432	OK	A
JARRO	0.2622	-0.1264	0.206	-0.6136	6.446	Flag	A
POLLO	0.0031	-0.0322	0.2043	-0.1576	0.055	OK	A
FRENTE	0.0318	0.0566	0.2023	0.2798	0.319	OK	A
LLAVE	0.1966	-0.1008	0.1888	-0.5339	5.129	Flag	A
CRUZ	0	0.021	0.2029	0.1035	2.502	OK	A
FUEGO	0.2718	0.1165	0.1891	0.6161	0.21	OK	A
PENA	0.0019	-0.0101	0.1929	-0.0524	0.022	OK	A
MODELO	0.2168	0.1138	0.2016	0.5645	0.311	OK	A
OIDO	1.8775	-0.2654	0.1816	-1.4615	0.012	OK	A

En el Ensayo 5 presentado en la Tabla 76, nuevamente un posible DIF no uniforme en el ítem POLLO; DIF moderado a favor del grupo focal en FUEGO, mientras que para el ítem OÍDO a favor del grupo de referencia.

Tabla 76

Análisis del DIF mediante Mantel-Haenszel entre población española y muestras focales Ensayo 5.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.1749	-0.1463	0.2634	-0.5554	1.084	OK	A
CINTA	2.1256	0.354	0.2267	1.5615	0.257	OK	A
NORTE	0.2835	0.1314	0.2066	0.636	0.039	OK	A
JARRO	0.0715	0.0781	0.2095	0.3728	0.002	OK	A
POLLO	1.9206	-0.326	0.2179	-1.4961	8.192	Flag	A
FRENTE	0.0113	-0.047	0.2178	-0.2158	0.8	OK	A
LLAVE	0.8318	-0.2022	0.1998	-1.012	0.268	OK	A
CRUZ	0.943	0.2271	0.2111	1.0758	1.601	OK	A
FUEGO	11.5019	-0.6693	0.1947	-3.4376	3.12	Flag	B
PENA	0.2216	0.112	0.1965	0.57	0.01	OK	A
MODELO	0.8931	0.214	0.204	1.049	0.912	OK	A
OIDO	5.2154	0.4734	0.2001	2.3658	1.837	Flag	B

En el último Ensayo (6), solamente dos ítems con DIF moderado a favor del grupo focal: JARRO y LLAVE (Tabla 77).

Tabla 77

Análisis del DIF mediante Mantel-Haenszel entre población española y muestras focales Ensayo 6.

Variable	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
DADO	0.2951	0.1733	0.2579	0.672	1.258	OK	A
CINTA	0.2636	-0.1491	0.2372	-0.6286	2.08	OK	A
NORTE	1.4826	-0.2878	0.2169	-1.3269	0.821	OK	A
JARRO	9.355	-0.7232	0.2291	-3.1567	2.124	Flag	B
POLLO	0.0871	0.0992	0.2389	0.4152	0.724	OK	A
FRENTE	0.3933	0.1623	0.2202	0.7371	0.422	OK	A
LLAVE	5.5641	-0.501	0.2043	-2.4523	0.088	Flag	B
CRUZ	0.1059	0.1011	0.2297	0.4401	0.007	OK	A
FUEGO	0.3544	0.1442	0.2064	0.6986	0.163	OK	A
PENA	0.0012	-0.0137	0.2035	-0.0673	0.078	OK	A
MODELO	0.9359	0.2435	0.2266	1.0746	0.231	OK	A
OIDO	3.7345	-0.3995	0.1973	-2.0248	1.446	OK	A

SEXTA ETAPA

5.7 Análisis del Funcionamiento Diferencial del Test (DTF)

Una ventaja agregada del paquete estadístico DIFAS, es que además del análisis del funcionamiento diferencial de cada uno de los ítems, permite evaluar el Funcionamiento Diferencial del Test, que como ya vimos es un indicador global del nivel de DIF, presente en una prueba. Los resultados de este procedimiento se presentan en Tablas en este apartado.

La varianza del efecto del DIF a través del conjunto de ítems de un test o escala ha sido propuesta como un medida de DTF para conjuntos de ítems dicotómicos (Penfield, 2005). La varianza del efecto del DIF entre ítems dicotómicos se simboliza como τ^2(tau-squared) (Longford, Holland, & Thayer, 1993), o la versión ligera τ_ω^2 (weighted tau-squared), propuesta por Camilli & Penfield (1997). Ambos estadísticos son calculados en el paquete estadístico DIFAS y su interpretación puede servir como indicador para determinar el funcionamiento diferencial del test en su conjunto (C. R. Rao & Sinharay, 2007).

Para efectos del presente estudio y considerando las características de la prueba se utilizó el procedimiento de análisis del Funcionamiento Diferencial del

Test (DTF) no paramétrico para ítems dicotómicos, en cada par de comparaciones.

Para la interpretación de los estadísticos se ha seguido la interpretación del estimador τ^2(tau-squared). Siguiendo el procedimiento descrito por Brown (2010).

Efecto de la varianza del DIF pequeño, 0.07< τ^2<0.14 se considera DTF pequeño o inexistente. (Aproximadamente menos del 10% de los ítems son identificados con valores de MH-LOR < ±0.43).

Efecto de la varianza del DIF medio, 0.07< τ^2<0.14, se considera DTF mediano. (Aproximadamente menos del 25% de los ítems son identificados con valores de MH-LOR> ±0.43).

Efecto de la varianza del DIF alto, valores de τ^2> 0.14 se considera DTF alto. (Aproximadamente más del 25% de los ítems que conforman el test son identificados con valores de MH-LOR> ±0.43).

5.7.1 Análisis del DTF entre población española y mexicana

En el análisis del DTF entre las muestras de España y México, correspondiente al análisis presentado en 5.3.2 (Anexo 7); podemos observar que este presenta un nivel medio τ^2= 0.10>0.07 (Ver Tabla 78).

Tabla 78

Análisis del DTF entre población española y mexicana

Estadístico	Valor	SE	Z
Tau^2	0.102	0.027	3.778
Weighted Tau^2	0.082	0.022	3.727

Posterior a la purificación en la comparación España y México, información de DIF en los puntos 5.5.2.1 y 5.5.2.2 (Anexo 15) el valor de DTF es el presentado en la Tabla 79 bajando ligeramente, pero se mantiene en un nivel medio τ^2= 0.093>0.07.

Tabla 79

Análisis del DTF entre población española y mexicana posterior a purificación

Estadístico	Valor	SE	Z
Tau^2	0.093	0.025	3.72
Weighted Tau^2	0.079	0.022	3.591

5.7.2 Análisis del DTF entre población española y estadounidense

En cuanto al primer análisis del DIF mediante MH en la comparación población española y estadounidense como se presentó en el punto 5.3.1 de este documento (Anexo 6), el Funcionamiento Diferencial del Test fue medio τ^2= 0.11>0.07, como se puede observar en la Tabla 80.

Tabla 80

Análisis del DTF entre población española y estadounidense

Estadístico	Valor	SE	Z
Tau^2	0.11	0.028	3.929
Weighted Tau^2	0.096	0.025	3.84

Posterior al proceso de purificación con el primer análisis de MH (para revisar en detalle, ver tema 5.5.2.3 de esta tesis), el DTF en esta comparación aumentó ligeramente; no obstante, se mantiene en un nivel medio τ^2= 0.116>0.07, como se puede observar en la Tabla 81 (Anexo 16).

Tabla 81

Análisis del DTF entre población española y estadounidense posterior a purificación

Estadístico	Valor	SE	Z
Tau^2	0.116	0.025	4
Weighted Tau^2	0.097	0.025	3.88

5.7.3 Análisis del DTF entre población mexicana y estadounidense

Al analizar el funcionamiento Diferencial del Test en su conjunto en la comparación realizada entre la muestra mexicana y estadounidense (Ver Tema 5.3.3 del presente documento), se observa que este es bajo o inexistente τ^2= 0.25 <0.07, como se puede apreciar en la Tabla 82 (Anexo 8).

Tabla 82

Análisis del DTF entre población mexicana y estadounidense

Estadístico	Valor	SE	Z
Tau^2	0.025	0.013	1.923
Weighted Tau^2	0.025	0.013	1.923

Al realizar la purificación presentada en el punto 5.5.2.5 de esta tesis (Anexo 17), los valores del DTF en esta comparación disminuyeron ligeramente sin embargo se siguen manteniendo en nivel bajo o inexistente τ^2= 0.022 <0.07 (Ver Tabla 83).

Tabla 83

Análisis del DTF entre población mexicana y estadounidense posterior a purificación

Estadístico	Valor	SE	Z
Tau^2	0.022	0.013	1.692
Weighted Tau^2	0.023	0.012	1.917

CAPITULO VI

DISCUSION Y CONCLUSIONES

6.1 Introducción

En el presente capítulo se plasman las conclusiones obtenidas de nuestra investigación, principales limitantes a las cuales nos enfrentamos y recomendaciones para futuras investigaciones.

Cómo se puede observar, los resultados presentados por ambas técnicas MH y Regresión logística son muy similares, y que el proceso de purificación parece disminuir en cierta medida la cantidad de ítems detectados con DIF; sin embargo, no cambia de manera global el Funcionamiento Diferencial del Test entre las diferentes comparaciones.

Vemos que existe una cantidad considerable de ítems identificados con DIF, principalmente al comparar a las muestras mexicanas y estadounidenses contra los españoles, sin embargo esto es una situación que no disminuye considerablemente al realizar los procesos de purificación. Esto nos daría indicios

de las causas del funcionamiento diferencial de algunos ítems y resultados globales de la prueba distintos. Al observar el DIF en los diferentes ensayos y corroborados mediante ambas técnicas Mantel Haenszel y Regresión Logística, podemos ver como estos se van disminuyendo posiblemente como producto del aprendizaje. También observamos que no son constantes, sino que en algunos ensayos aparecen a favor del grupo de referencia y en otros a favor del focal, es decir el DIF es bidireccional.

Esto nos lleva a pensar que el funcionamiento diferencial encontrado, no es debido a un *sesgo* en los ítems sino a un *impacto* del ítem, y que las diferencias en la memorización de la lista de palabras están relacionadas con los efectos de primicia y recencia, que sabemos se presentan en el proceso de aprendizaje y codificación de la memoria, que por efecto de la repetición se ven disminuidos; por ello, al finalizar los seis ensayos, prácticamente no se presenta funcionamiento diferencial e incluso, el DIF cambia a favor del grupo focal

Lo anterior se podría explicar como una consecuencia de la falta del recordatorio para la muestra de referencia que ha tenido un funcionamiento más alto en los primeros ensayos, los ítems recordados en un principio, empiezan a ser olvidados, como resultado de la interferencia anterógrada, mientras que los sujetos que en los primeros ensayos tuvieron fallos, recibieron mayor cantidad de recordatorios, que les permite al finalizar los ensayos mejorar su tasa de recuerdo.

Por tanto, se podría considerar que el DIF detectado, se debe a una diferencia real en la habilidad (memoria) de las poblaciones evaluadas y no al sesgo del ítem.

6.2 Discusión

En México, como en la mayoría de los países en vías de desarrollo, el uso de pruebas psicométricas desarrolladas para otros países y/o culturas, es una práctica común, sin embargo como se pudo observar en capítulos anteriores, esto conlleva una disminución considerable e incluso la anulación de su confiabilidad y validez. El campo de la evaluación neuropsicológica de la memoria no es la excepción en nuestro país; motivo por el cual, con la presente investigación se pretende hacer extensivo el uso del Test de Recuerdo Verbal Selectivo en su versión en español (Campo, Morales y Juan-Malpartida, 2000), al evaluar el Funcionamiento Diferencial del Test de Recuerdo Verbal Selectivo entre población española (población de referencia), mexicana e hispana en Estados Unidos de Norteamérica (poblaciones focales). Esto debido a que la prueba cuenta con datos normativos (Campo y Morales, 2004), de confiabilidad (Campo et al., 2000) y validez (Campo, Morales y Martínez-Castillo, 2003) para población española, y de no existir DIF o DFT, la prueba podría ser utilizada en nuestras poblaciones focales con los mismos criterios de confiabilidad, validez y normatividad que los obtenidos en la población de referencia.

Con base en el análisis estadístico realizado, podemos determinar que en su conjunto, el tRVS presentó funcionamiento diferencial mayor al comparar las muestras españolas contra las mexicanas y estadounidenses.

En cambio, al realizar las comparaciones de DIF entre las muestras mexicana (referencia) y estadounidense (focal), los ítems y el grado de DIF fue significativamente menor al presentado en las comparaciones contra la muestra española. A pesar de que se detectaron DIF moderados y no homogéneos en algunos de los ítems en los diversos ensayos, estos no llegaron a ser significativamente favorecedores para ninguna de las dos poblaciones estudiadas.

Además, al evaluar el Funcionamiento Diferencial del Test en su conjunto entre estas dos poblaciones, el grado de DTF, fue bajo o nulo, incluso sin necesidad de realizar purificación en la puntuación global. Esto podría ser indicio de una mayor similitud en aspectos socio-culturales que pudieran influir en el desempeño en el vSRT y en otras pruebas neuropsicológicas, entre estas dos poblaciones que con respecto a la población española.

Por lo anterior, podemos concluir que no se recomienda utilizar esta prueba en población mexicana e hispana en EUA, con los datos normativos, confiabilidad

y validez establecidos para población española, sino que es necesario realizar un nuevo procedimiento de estandarización para estas poblaciones.

Sin embargo a la luz de estos resultados, podríamos considerar la posibilidad de realizar un proceso de estandarización con los resultados de las muestras focales fusionadas; es decir, las muestras mexicanas y estadounidenses, ya que como se observó en los análisis previos, estos grupos parecen tener un funcionamiento muy similar en el tRVS; por lo que la línea de investigación futura y más próxima sería ésta.

6.3 Conclusiones generales

Cabe hacer mención, que estos resultados son aplicables solamente a la Forma 1 del tRVS adaptada por Campo et al. (2000) y que fue estandarizado para población española por Campo y Morales (2004), aunque en un estudio complementario realizado en el 2004, estos investigadores presentaron datos que confirmaron la equivalencia de las dos versiones desarrolladas de esta prueba (para una revisión a profundidad de estos resultados se puede revisar a Campo y Morales, 2004).

También sería conveniente tener en mente las características de las muestras focales utilizadas para este estudio. Con lo que respecta a la muestra focal mexicana, ésta proviene de una población del noreste de la república mexicana: la ciudad de H. Matamoros, Tamaulipas, la cual por ser una ciudad fronteriza con el sureste de los Estados Unidos de Norteamérica presenta características sociodemográficas muy particulares. Como cualquier población fronteriza, ésta se encuentra marcada por la migración de personas de otros estados de la República Mexicana e incluso de otros países de Sudamérica, que arriban a la ciudad, con la idea de emigrar a los EUA o como resultado de la deportación de aquél país. De esta forma, algunas de ellas se establecen por tiempo indefinido en este territorio, siendo usuarios regulares de los servicios públicos de la región. Por ello los resultados de esta investigación no son aplicables a esta "población flotante" y sería recomendable realizar estudios adicionales para evaluar el funcionamiento diferencial de la prueba en estos sub-grupos culturales.

Por otra parte, la muestra estadounidense, recordemos que está formada por población hispana en la ciudad de Brownsville, Texas, siendo esta también una ciudad fronteriza, la etnicidad está fuertemente marcada por la migración de personas principalmente de México; sin embargo, existen grupos hispanos también de otras partes de Sudamérica: Guatemala, Nicaragua, Ecuador, que aunque son menos representativos de la población, también están presentes y suelen hacer uso de los mismos servicios de salud y educativos públicos.

Por lo anterior expuesto; sería conveniente extender esta investigación realizando estudios comparativos e interculturales con otros grupos poblacionales de México e hispanos en EUA, de diversa procedencia cultural. Recordemos lo visto en los capítulos previos acerca del término hispano, utilizado independientemente del país de origen, e incluso para aquellos que son nacidos en EUA, segundas o terceras generaciones, hijos, nietos de personas inmigrantes, que no solamente difieren en los usos del lenguaje (generalmente bilingües) sino que también van formando una nueva cultura producto del mestizaje de usos y costumbres.

También sería recomendable ampliar las investigaciones en otros países de América Latina a fin de brindar a los profesionistas encargados de evaluar la memoria en estos países una herramienta que ha comprobado repetidamente su eficiencia en la evaluación neuropsicológica de la memoria.

Adicionalmente, se puede ampliar el conocimiento de la prueba en sujetos con niveles educativos más bajos e incluso iletrados así como con sujetos que presentan alteraciones graves y moderadas de la memoria, en estos grupos culturales. Como vemos el campo de estudio es amplio, en nuestras poblaciones. Esperamos con esta investigación contribuir al surgimiento de un interés por este campo de estudio.

BIBLIOGRAFIA

Abedalaziz, N. (2010). Detecting Gender Related DIF using Logistic Regression and Mantel-Haenszel Approaches. *Procedia - Social and Behavioral Sciences*, *7*, 406–413. http://doi.org/10.1016/j.sbspro.2010.10.055

Abrisqueta-Gomez, J., Ostrosky-Solis, F., Bertolucci, P. H. F., & Bueno, O. F. A. (2008). Applicability of the abbreviated neuropsychologic battery (NEUROPSI) in Alzheimer disease patients. *Alzheimer Disease & Associated Disorders*, *22*(1), 72–78.

Abwender, D. A., & Sfikouris, S. A. (2005). Validity of the NEUROPSI among Spanish speakers in the United States. *Clinical Neuropsychologist*, *19*(3-4), 554–554.

Ackerman, T. A. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. *Journal of Educational Measurement*, *29*(1), 67–91. http://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Alderman, D. L. (1982). Language proficiency as a moderator variable in testing academic aptitude. *Journal of Educational Psychology*, *74*(4), 580–587. http://doi.org/10.1037/0022-0663.74.4.580

American Educational Research Association, American Psychological Association, National Council on Meassurement in Education, & Joint Committee on

Standards for Educational & Psychological Testing. Standards for educational and psychological testing (1999).

Andrewes, D. (2013). *Neuropsychology: From Theory to Practice*. Psychology Press.

Angoff, W. H. (1972). A Technique for the Investigation of Cultural Differences. Retrieved from http://eric.ed.gov/?id=ED069686

APA Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations. (2013). Retrieved June 28, 2013, from http://www.apa.org/pi/oema/resources/policy/provider-guidelines.aspx

Ardila, A. (1995). Directions of research in cross-cultural neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*(1), 143–150. http://doi.org/10.1080/13803399508406589

Ardila, A. (2013). The impact of culture in neuropsychological test performance. In B. P. Uzzell, M. Ponton, & A. Ardila, *International Handbook of Cross-Cultural Neuropsychology*. Psychology Press.

Ardila, A., & Ostrosky Solis, F. (1991). *Diagnóstico del daño cerebral: enfoque neuropsicológico*. Trillas.

Ardila, A., Ostrosky-Solis, F., Rosselli, M., & Gomez, C. (2000). Age-related cognitive decline during normal aging: The complex effect of education. *Archives of Clinical Neuropsychology*, *15*(6), 495–513. http://doi.org/10.1016/S0887-6177(99)00040-2

Ardila, A., & Ramos, E. (2007). *Speech and languaje disorders in hispanics*. Nova Science Publishers.

Ardila, A., & Rosselli, M. (1994). Development of language, memory, and visuospatial abilities in 5- to 12-year-old children using a neuropsychological battery. *Developmental Neuropsychology*, *10*(2), 97–120. http://doi.org/10.1080/87565649409540571

Ardila, A., Rosselli, M., & Ostrosky-Solis, F. (1992). Socioeducational. In A. E. Puente & R. J. McCaffrey (Eds.), *Handbook of Neuropsychological Assessment* (pp. 181–192). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4899-0682-3_7

Arffman, I. (2013). Problems and Issues in Translating International Educational Achievement Tests. *Educational Measurement: Issues and Practice*, *32*(2), 2–14. http://doi.org/10.1111/emip.12007

Arnold, B. R., & Matus, Y. E. (2000). Test translation and cultural equivalence methodologies for use with diverse populations. In *Handbook of multicultural mental health: Assessment and treatment of diverse populations* (pp. 121–136). U.S.A.: Academic Press. Retrieved from http://books.google.com.mx/books?hl=es&lr=&id=AjpW3jNYez0C&oi=fnd&pg=PA121&dq=Hambleton+1996&ots=T_8i3CACvX&sig=Nhl41TqA8EMgbDotAcmPiUwqLAQ

Arnold, B. R., Montgomery, G. T., Castañeda, I., & Longoria, R. (1994). Acculturation and Performance of Hispanics on Selected Halstead-Reitan

Neuropsychological Tests. *Assessment*, *1*(3), 239–248. http://doi.org/10.1177/107319119400100303

Artiola I Fortuny, L., Heaton, R. K., & Hermosillo, D. (1998). Neuropsychological comparisons of Spanish-speaking participants from the U.S.-Mexico border region versus Spain. *Journal of the International Neuropsychological Society*, *4*(4), 363–379.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In Kenneth W. Spence and Janet Taylor Spence (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 2, pp. 89–195). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742108604223

Baddeley, A. D. (1999). *Memoria humana: teoría y práctica*. McGraw-Hill.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement*, 59–62.

Bandeira Andriola, W., & Gaviria Soto, J. L. (2002). *Detección del funcionamiento diferencial del item (DIF) en tests de rendimiento: aportaciones teóricas y metodológicas*. Universidad Complutense de Madrid. Retrieved from http://biblioteca.ucm.es/tesis/edu/ucm-t26457.pdf

Barbero García, M. I., Prieto, P., & San Luis, C. (2000). Procedimiento para la detección del FDI tanto en ítems politómicos como dicotómicos. *Psicothema*, *12*(Suplemento), 69–73.

Barbero García, M. I., Vila Abad, E., & Holgado Tello, F. P. (2008). La adaptación de los test en estudios comparativos interculturales [Tests Adaptation in cross-cultural comparative studies]. *Acción Psicológica*, *5*(2), 7–16.

Bartok, J. A., Wilson, C. S., Giordani, B., Keys, B. A., Persad, C. C., Foster, N. L., & Berent, S. (1997). Varying patterns of verbal recall, recognition, and response bias with progression of alzheimer's disease. *Aging, Neuropsychology, and Cognition*, *4*(4), 266–272. http://doi.org/10.1080/13825589708256651

Bauer, R. M., Tobias, B., Valenstein, E., & Heilman, K. M. (1993). Clinical neuropsychology.

Beatty, W. W., Krull, K. R., Wilbanks, S. L., Blanco, C. R., Hames, K. A., & Paul, R. H. (1996). Further Validation of Constructs from the Selective Reminding Test. *Journal of Clinical and Experimental Neuropsychology*, *18*(1), 52–55. http://doi.org/10.1080/01688639608408261

Beatty, W. W., Wilbanks, S. L., Blanco, C. R., Hames, K. A., Tivis, R., & Paul, R. H. (1996). Memory Disturbance in Multiple Sclerosis: Reconsideration of Patterns of Performance on the Selective Reminding Test. *Journal of Clinical and Experimental Neuropsychology*, *18*(1), 56–62. http://doi.org/10.1080/01688639608408262

Behling, O., & Law, K. S. (2000). *Translating Questionnaires and Other Research Instruments: Problems and Solutions*. SAGE.

Bell, B. D., Fine, J., Dow, C., Seidenberg, M., & Hermann, B. P. (2005). Temporal Lobe Epilepsy and the Selective Reminding Test: The Conventional 30-Minute Delay Suffices. *Psychological Assessment*, *17*(1), 103–109. http://doi.org/10.1037/1040-3590.17.1.103

Belli, R. F. (2011). *True and False Recovered Memories: Toward a Reconciliation of the Debate*. Springer Science & Business Media.

Benedet, M. J., & Alejandre, M. Á. (1998). *TAVEC: test de aprendizaje verbal España-Complutense: manual*. TEA ediciones. Retrieved from http://dialnet.unirioja.es/servlet/libro?codigo=212146

Berry, J. W. (1979). A Cultural Ecology of Social Behavior. In Leonard Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. Volume 12, pp. 177–206). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0065260108602622

Berry, J. W. (2005). Acculturation: Living successfully in two cultures. *International Journal of Intercultural Relations*, *29*(6), 697–712. http://doi.org/10.1016/j.ijintrel.2005.07.013

Binet, A., & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'annee Psychologique*, *12*, 191–244.

Binet, A., & Simon, T. (1948). The development of the Binet-Simon Scale, 1905-1908. In *Readings in the history of psychology* (pp. 412–424). East Norwalk, CT, US: Appleton-Century-Crofts.

Boeve, B., McCormick, J., Smith, G., Ferman, T., Rummans, T., Carpenter, T., … Petersen, R. (2003). Mild cognitive impairment in the oldest old. *Neurology*, *60*(3), 477–480. http://doi.org/10.1212/WNL.60.3.477

Boone, K., Victor, T., Wen, J., Razani, J., & Ponton, M. (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology*, *22*(3), 355–365. http://doi.org/10.1016/j.acn.2007.01.010

Boringa, J. B., Lazeron, R. H., Reuling, I. E., Adèr, H. J., Pfennings, L. E., Lindeboom, J., … Polman, C. H. (2001). The Brief Repeatable Battery of Neuropsychological Tests: normative values allow application in multiple sclerosis clinical practice. *Multiple Sclerosis*, *7*(4), 263–267. http://doi.org/10.1177/135245850100700409

Brandt, J. (1991). The hopkins verbal learning test: Development of a new memory test with six equivalent forms. *Clinical Neuropsychologist*, *5*(2), 125–142. http://doi.org/10.1080/13854049108403297

Brandt, J., & Benedict, R. H. (2001). *Hopkins verbal learning test, revised: professional manual*. Psychological Assessment Resources.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216.

Brislin, R. W. (1986). The wording and translation of research instruments. Retrieved from http://psycnet.apa.org/psycinfo/1987-97046-005

Brown, A. (2010). Introducing concepts of measurement invariance. Investigating Differential Item Functioning (DIF) using various approaches (Mantel-Haenszel and Confirmatory Factor Analysis (CFA) with covariates). Retrieved from http://www.psychometrics.cam.ac.uk/uploads/documents/ESRC_RDI_Sep_10/summer-school-2-day4.pdf

Buschke, H. (1973). Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior*, *12*(5), 543–550. http://doi.org/10.1016/S0022-5371(73)80034-9

Buschke, H. (1984). Cued recall in Amnesia. *Journal of Clinical Neuropsychology*, *6*(4), 433–440. http://doi.org/10.1080/01688638408401233

Buschke, H., & Fuld, P. A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology*, *24*(11), 1019–1019. http://doi.org/10.1212/WNL.24.11.1019

Butman, T. (2001). Designing an instrument of early diagnosis of dementia in primary care. *Acta Psyiquiatr Psicol Am Lat*, *47*(1), 79–87.

Byrne, B. M., & Campbell, T. L. (1999). Cross-Cultural Comparisons and the Presumption of Equivalent Measurement and Theoretical Structure A Look Beneath the Surface. *Journal of Cross-Cultural Psychology*, *30*(5), 555–574. http://doi.org/10.1177/0022022199030005001

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology*, *3*(2), 94–105. http://doi.org/10.1037/a0014516

Cabassa, L. J. (2003). Measuring Acculturation: Where We Are and Where We Need to Go. *Hispanic Journal of Behavioral Sciences*, *25*(2), 127–146. http://doi.org/10.1177/0739986303025002001

Camilli, G. (1979). *A critique of the chi-square method for assessing item bias*. Documento no publicado, E.U.A.

Camilli, G., & Hopkins, K. D. (1979). Testing for association in 2 x 2 contingency tables with very small sample sizes. *Psychological Bulletin*, *86*(5), 1011.

Camilli, G., & Penfield, D. A. (1997). Variance Estimation for Differential Test Functioning Based on Mantel-Haenszel Statistics. *Journal of Educational Measurement*, *34*(2), 123–139.

Camilli, G., & Shepard, L. A. (1987). The Inadequacy of ANOVA for Detecting Test Bias. *Journal of Educational and Behavioral Statistics*, *12*(1), 87–99. http://doi.org/10.3102/10769986012001087

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage publications London.

Campo, P., & Morales, M. (2004). Normative data and reliability for a Spanish version of the verbal Selective Reminding Test. *Archives of Clinical Neuropsychology*, *19*(3), 421–435. http://doi.org/10.1016/S0887-6177(03)00075-1

Campo, P., Morales, M., & Juan-Malpartida, M. (2000a). Development of two Spanish versions of the verbal selective reminding test. *Journal of Clinical and Experimental Neuropsychology*, *22*(2), 279–285.

Campo, P., Morales, M., & Juan-Malpartida, M. (2000b). Versiones españolas del test de Recuerdo Verbal Selectivo. *Psicothema*, *12*(Suplemento), 108–110.

Campo, P., Morales, M., & Martínez-Castillo, E. (2003). Discrimination of Normal from Demented Elderly on a Spanish Version of the Verbal Selective Reminding Test. *Journal of Clinical and Experimental Neuropsychology*, *25*(7), 991–999. http://doi.org/10.1076/jcen.25.7.991.16492

Cardall, C., & Coffman, W. E. (1964). A method for comparing the performance of different groups on the items in a test. *Research and Development Reports*, *9*, 64–65.

Cattell, J. M. (1890). V.—MENTAL TESTS AND MEASUREMENTS. *Mind*, *os-XV*(59), 373–381. http://doi.org/10.1093/mind/os-XV.59.373

Chang, A. M., Chau, J. P., & Holroyd, E. (1999). Translation of questionnaires and issues of equivalence. *Journal of Advanced Nursing*, *29*(2), 316–322.

Cheung, G. W., & Rensvold, R. B. (1999). Testing Factorial Invariance across Groups: A Reconceptualization and Proposed New Method. *Journal of Management*, *25*(1), 1–27. http://doi.org/10.1177/014920639902500101

Chiaravalloti, N. D., DeLuca, J., Moore, N. B., & Ricker, J. H. (2005). Treating learning impairments improves memory performance in multiple sclerosis: a randomized clinical trial. *Multiple Sclerosis*, *11*(1), 58–68. http://doi.org/10.1191/1352458505ms1118oa

Chiaravalloti, N. D., Demaree, H., Gaudino, E. A., & DeLuca, J. (2003). Can the repetition effect maximize learning in multiple sclerosis? *Clinical Rehabilitation*, *17*(1), 58–68. http://doi.org/10.1191/0269215503cr586oa

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The Effects of Purification of Matching Criterion on the Identification of DIF Using the Mantel-Haenszel Procedure. *Applied Measurement in Education*, *6*(4), 269–279. http://doi.org/10.1207/s15324818ame0604_2

Cleary, T. A., & Hilton, T. L. (1968). An Investigation of Item Bias. *Educational and Psychological Measurement*, *28*(1), 61–75. http://doi.org/10.1177/001316446802800106

Coffey, D., Marmol, L., Schock, L., & Adams, W. (2005). The influence of acculturation on the Wisconsin Card Sorting Test by Mexican Americans. *Archives of Clinical Neuropsychology*, *20*(6), 795–803. http://doi.org/10.1016/j.acn.2005.04.009

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*(4), 335–350.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., … van Belle, G. (2008). Item response theory facilitated co-calibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, *61*(10), 1018–27.e9. http://doi.org/10.1016/j.jclinepi.2007.11.011

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Pedraza, O., Mehta, K. M., Tang, Y., … Mungas, D. M. (2008). Composite scores for executive function items: Demographic heterogeneity and relationships with quantitative magnetic resonance imaging. *Journal of the International Neuropsychological Society*, *14*(05), 746–759. http://doi.org/10.1017/S1355617708081162

Dana, R. H. (1996). Culturally Competent Assessment Practice in the United States. *Journal of Personality Assessment*, *66*(3), 472–487. http://doi.org/10.1207/s15327752jpa6603_2

Degenszajn, J., Caramelli, P., Caixeta, L., & Nitrini, R. (2001). Encoding process in delayed recall impairment and rate of forgetting in Alzheimer's disease. *Arquivos de Neuro-Psiquiatria*, *59*(2A), 171–174. http://doi.org/10.1590/S0004-282X2001000200003

Delis, D. C., & Kramer, J. H. (2000). *California Verbal Learning Test: CvLT-II; Adult Version; Manual*.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). California verbal learning test, research edition. *New York: The Psychological Corporation*.

Devanand, D. P., Pradhaban, G., Liu, X., Khandji, A., Santi, S. D., Segal, S., … Leon, M. J. de. (2007). Hippocampal and entorhinal atrophy in mild cognitive impairment Prediction of Alzheimer disease. *Neurology*, *68*(11), 828–836. http://doi.org/10.1212/01.wnl.0000256697.20968.d7

Dolnicar, S., & Grün, B. (2013). "Translating" between survey answer formats. *Journal of Business Research*, *66*(9), 1298–1306. http://doi.org/10.1016/j.jbusres.2012.02.029

Dorans, N. J. (2013). *ETS Contributions to the Quantitative Assessment of Item, Test, and Score Fairness*. Educational Test Commission. Retrieved from http://www.ets.org/Media/Research/pdf/RR-13-27.pdf

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. Retrieved from http://eric.ed.gov/?id=ED387526

Drane, D. L., Loring, D. W., Lee, G. P., & Meador, K. J. (1998). Trial-Length Sensitivity of the Verbal Selective Reminding Test to Lateralized Temporal Lobe Impairment. *The Clinical Neuropsychologist*, *12*(1), 68–73. http://doi.org/10.1076/clin.12.1.68.1728

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting Inappropriate Test Scores with Optimal and Practical Appropriateness Indices. *Applied Psychological Measurement*, *11*(1), 59–79. http://doi.org/10.1177/014662168701100105

Dunckley, M., Hughes, R., Addington-Hall, J. M., & Higginson, I. J. (2003). Translating clinical tools in nursing practice. *Journal of Advanced Nursing*, *44*(4), 420–426.

Ebbinghaus, H. (2014). Memory: A Contribution to Experimental Psychology. *Annals of Neurosciences*, *20*(4). http://doi.org/10.5214/ans.0972.7531.200408

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). *Intelligence and cultural differences: A study of cultural learning and problem-solving*. University of Chicago Press.

Elbulok-Charcape, M. M., Rabin, L. A., Spadaccini, A. T., & Barr, W. B. (2014). Trends in the neuropsychological assessment of ethnic/racial minorities: A survey of clinical neuropsychologists in the United States and Canada. *Cultural Diversity and Ethnic Minority Psychology*, *20*(3), 353–361. http://doi.org/10.1037/a0035023

Elosua, P., & López-Jáuregui, A. (2007). Aplicación de cuatro procedimientos de detección del funcionamiento diferencial sobre ítems politómicos. *Psicothema*, *19*(2), 329–336.

Elosua, P., & López-Jáuregui, A. (2008). Equating Between Linguistically Different Tests: Consequences for Assessment. *The Journal of Experimental Education*, *76*(4), 387–402. http://doi.org/10.3200/JEXE.76.4.387-402

Elosua, P., & Wells, C. S. (2013). Detecting DIF in Polytomous Items Using MACS, IRT and Ordinal Logistic Regression. *Psicologica*, *34*(2), 327–342.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341–349. http://doi.org/10.1037/1040-3590.8.4.341

Ferreres Traver, D., Fidalgo Aliste, Á. M., & Muñiz, J. (2000). Detección del funcionamiento diferencial de los ítems no uniforme: comparación de los métodos Mantel-Haenszel y regresión logística. *Psicothema*, *12*(Suplemento), 220–225.

Ferreres Traver, D., González Romá, V., & Gómez Benito, J. (2000). Comparación del estadístico Mantel-Haenszel y la regresión logística en el

funcionamiento diferencial de los ítems en dos pruebas de aptitud intelectual en un contexto bilingüe. *Psicothema*, *12*(Suplemento), 214–219.

Fidalgo Aliste, Á. M. (1996). *Funcionamiento diferencial de los ítems. Procedimiento mantel-haenszel y modelos loglineales*. Universidad de Oviedo. Retrieved from http://dspace.sheol.uniovi.es/dspace/handle/10651/16659

Fidalgo Aliste, Á. M. (2011). A New Approach for Differential Item Functioning Detection Using Mantel-Haenszel Methods. The GMHDIF Program. *The Spanish Journal of Psychology*, *14*(02), 1018–1022. http://doi.org/10.5209/rev_SJOP.2011.v14.n2.47

Fidalgo Aliste, Á. M., Alavi, S. M., & Amirian, S. M. R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing*, *31*(4), 433–451. http://doi.org/10.1177/0265532214526748

Fidalgo Aliste, Á. M., Ferreres Traver, D., & Muñiz, J. (2004). Liberal and Conservative Differential Item Functioning Detection Using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II Error Rates. *The Journal of Experimental Education*, *73*(1), 23–39. http://doi.org/10.3200/JEXE.71.1.23-40

Fidalgo Aliste, Á. M., Mellenberg, G. J., & Muñiz, J. (2000). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure and the iterative logit method. *REMA*, *5*(1), 13–24.

Fidalgo Aliste, Á. M., Mellenbergh, G. J., & Muñiz, J. (1999). Aplicación en una etapa, dos etapas e iterativamente de los estadísticos de Mantel-Haenszel. *Psicológica*, *20*, 227–242.

Fidalgo Aliste, Á. M., & Paz, M. D. (1995). Modelos lineales logarítmicos y funcionamiento diferencial de los ítems. *Anuario de Psicología*, *64*, 57–66.

Finch, W. H., & French, B. F. (2007). Detection of Crossing Differential Item Functioning: A Comparison of Four Methods. *Educational and Psychological Measurement*, *67*(4), 565–582. http://doi.org/10.1177/0013164406296975

Franzen, M. D. (2000). *Reliability and Validity in Neuropsychological Assessment*. Springer Science & Business Media.

French, A. W., & Miller, T. R. (1996). Logistic Regression and Its Use in Detecting Differential Item Functioning in Polytomous Items. *Journal of Educational Measurement*, *33*(3), 315–332. http://doi.org/10.1111/j.1745-3984.1996.tb00495.x

Fuster, J. M. (2001). The prefrontal cortex-An update-Time is of the essence. *Neuron*, *30*(2), 319–333.

Fuster, J. M. (2003). *Cortex and Mind: Unifying Cognition*. Oxford University Press.

Fuster, J. M. (2005). *Cortex and Mind:Unifying Cognition: Unifying Cognition*. Oxford University Press, USA.

Ganor-Stern, D., Seamon, J. G., & Carrasco, M. (1998). The role of attention and study time in explicit and implicit memory for unfamiliar visual stimuli. *Memory & Cognition*, *26*(6), 1187–1195.

Gasquoine, P. G. (1999). Variables Moderating Cultural and Ethnic Differences in Neuropsychological Assessment: The Case of Hispanic Americans. *The Clinical Neuropsychologist*, *13*(3), 376–383. http://doi.org/10.1076/clin.13.3.376.1735

Gasquoine, P. G. (2001). Research in clinical neuropsychology with Hispanic American participants: A review. *Clinical Neuropsychologist*, *15*(1), 2–12. http://doi.org/10.1076/clin.15.1.2.1915

Gasquoine, P. G., Cavazos, A., Cantu, J., & Weimer, A. A. (2010). Bilingualism and hispanic American intelligence test scores. In *Bilinguals: Cognition, Education and Language Processing* (pp. 181–199).

Gasquoine, P. G., Croyle, K. L., Cavazos-Gonzalez, C., & Sandoval, O. (2007). Language of administration and neuropsychological test performance in neurologically intact Hispanic American bilingual adults. *Archives of Clinical Neuropsychology*, *22*(8), 991–1001. http://doi.org/10.1016/j.acn.2007.08.003

Gigi, A., Michal Schnaider-Beeri, B. A., Davidson, M., & Prohovnik, I. (1999). Validation of a hebrew selective reminding test. *Israel Journal of Psychiatry and Related Sciences*, *36*(1), 11–17.

Goldstein, G. (1992). Historical Perspectives. In A. E. Puente & R. J. McCaffrey (Eds.), *Handbook of Neuropsychological Assessment* (pp. 1–10). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4899-0682-3_1

Goldstein, K. (1939). The organism: A holistic approach to biology derived from pathological data in man. Retrieved from http://psycnet.apa.org/psycinfo/2004-16223-000

Gómez-Benito, J., Hidalgo, M. D., & Guilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles Del Psicólogo*, *31*(1), 75–84.

Gómez Benito, J., & Hidalgo Montesinos, M. D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: una revisión metodológica. *Anuario de Psicología/The UB Journal of Psychology*, (74), 3–32.

Gonzalez da Silva, C., Petersson, K. M., Faísca, L., Ingvar, M., & Reis, A. (2004). The Effects of Literacy and Education on the Quantitative and Qualitative Aspects of Semantic Verbal Fluency. *Journal of Clinical and Experimental Neuropsychology*, *26*(2), 266–277. http://doi.org/10.1076/jcen.26.2.266.28089

Gonzalez, H., Mungas, D., & Haan, M. (2005). A semantic verbal fluency test for English- and Spanish-speaking older Mexican-Americans. *Archives of Clinical Neuropsychology*, *20*(2), 199–208. http://doi.org/10.1016/j.acn.2004.06.001

Grasso, L., & Peraita, H. (2011a). Evaluation battery for semantic memory deterioration in dementia of the alzheimer type (EMSDA): Item's Adjustment To The Population Of Buenos Aires City. *Interdisciplinaria*, *28*(1), 37–56.

Grasso, L., & Peraita, H. (2011b). Semantic memory deficits assessment in dementia of the alzheimer type. In *Perspectives on Alzheimer's Disease* (pp. 51–75).

Grigorenko, E. L. (2009). *Multicultural Psychoeducational Assessment*. Springer Publishing Company.

Grober, E., Dickson, D., Sliwinski, M. J., Buschke, H., Katz, M., Crystal, H., & Lipton, R. B. (1999). Memory and mental status correlates of modified Braak staging. *Neurobiology of Aging*, *20*(6), 573–579. http://doi.org/10.1016/S0197-4580(99)00063-9

Grober, E., Ehrlich, A. R., Troche, Y., Hahn, S., & Lipton, R. B. (2014). Screening Older Latinos for Dementia in the Primary Care Setting. *Journal of the International Neuropsychological Society*, *20*(08), 848–855. http://doi.org/10.1017/S1355617714000708

Grober, E., Hall, C. B., Lipton, R. B., Zonderman, A. B., Resnick, S. M., & Kawas, C. (2008). Memory impairment, executive dysfunction, and intellectual decline in preclinical Alzheimer's disease. *Journal of the International Neuropsychological Society*, *14*(02), 266–278. http://doi.org/10.1017/S1355617708080302

Guilera, G., Gómez Benito, J., Hidalgo Montesinos, M. D., & Sánchez Meca, J. (2007). Un meta-análisis del procedimiento Mantel-Haenszel en la detección del DIF en ítems dicotómicos. *Anuario de psicología / The UB Journal of psychology*, *38*(3), 431–442.

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-Cultural Adaptation of Health-Related Quality-of-Life Measures - Literature-Review and Proposed Guidelines. *Journal of Clinical Epidemiology*, *46*(12), 1417–1432. http://doi.org/10.1016/0895-4356(93)90142-N

Guinn, R., Vincent, V., Wang, L., & Villas, P. (2011). Acculturation Tendencies in a Border Latino Population. *Hispanic Journal of Behavioral Sciences*, *33*(2), 170–183. http://doi.org/10.1177/0739986311398209

Gunnarsson, B. (1978). *A look at the content similarities between intelligence, achievement, personality, and language tests*. Newbury House Publishers.

Hambleton, R. K. (1996). Guidelines for Adapting Educational and Psychological Tests. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED399291

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, *17*(3), 164.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2004). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Psychology Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting Educational and Psychological Tests for Cross-cultural Assessment*. Taylor & Francis Group.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Association of Test Publishers*, *1*(1), 1–13.

Hannay, H. J. (1998). Proceedings of the Houston Conference on specialty education and training in clinical neuropsychology, september 3–7, 1997, University of Houston Hilton and Conference Center. *Archives of Clinical Neuropsychology*, *13*(2), 157–158. http://doi.org/10.1093/arclin/13.2.157

Hannay, H. J., & Levin, H. S. (1985). Selective reminding test: An examination of the equivalence of four forms. *Journal of Clinical and Experimental Neuropsychology*, *7*(3), 251–263. http://doi.org/10.1080/01688638508401258

Harkness, J. A. (2006). 4. Measurement and Comparability in Cross-National Health Surveys Used to Inform Policy Decisions. *5. The Impact of Social Science Research on Social Policy: Governance and Management*, 27.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., … Smith, T. W. (2010). *Survey Methods in Multicultural, Multinational, and Multiregional Contexts*. John Wiley & Sons.

Harris, J. G., Cullum, C. M., & Puente, A. E. (1995). Effects of bilingualism on verbal learning and memory in Hispanic adults. *Journal of the International Neuropsychological Society*, *1*(01), 10–16. http://doi.org/10.1017/S1355617700000059

Hebben, N., & Milberg, W. (2011). *Fundamentos para la evaluación neuropsicológica*. El Manual Moderno.

Hécaen, H., & Albert, M. L. (1986). *Human neuropsychology*. RE Krieger Publishing Company.

Herdman, M., Fox-Rushby, J., & Badia, X. (1997). "Equivalence" and the translation and adaptation of health-related quality of life questionnaires. *Quality of Life Research*, *6*(3), 0–0. http://doi.org/10.1023/A:1026410721664

Hidalgo Montesinos, M. D., Gómez Benito, J., & Padilla García, J. L. (2005). Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial del ítem. *Psicothema*, *17*(3), 509–515.

Hidalgo Montesinos, M. D., & Lopez Pina, J. A. (2004). Differential Item Functioning Detection and Effect Size: A Comparison between Logistic Regression and Mantel-Haenszel Procedures. *Educational and*

Psychological Measurement, *64*(6), 903–915. http://doi.org/10.1177/0013164403261769

Hilton, A., & Skrutkowski, M. (2002). Translating instruments into other languages: development and testing processes. *Cancer Nursing*, *25*(1), 1–7.

Holland, P. W. (1985). On the study of differential item performance without IRT. In *Proceedings of the Military Testing Association*.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test Validity*, 129–145.

Holland, P. W., & Wainer, H. (2012). *Differential Item Functioning*. Routledge.

International Standards Organization. ISO 20252 Market Research (2006). Retrieved from http://www.standards.org/standards/listing/iso_20252

International Test Commission. (2010). Guidelines for translating and adapting tests, *1*, 2012.

Ivnik, R. J., Smith, G. E., Lucas, J. A., Tangalos, E. G., Kokmen, E., & Petersen, R. C. (1997). Free and cued selective reminding test: Moans norms. *Journal of Clinical and Experimental Neuropsychology*, *19*(5), 676–691. http://doi.org/10.1080/01688639708403753

Jacobs, D. M., Marder, K., Cote, L. J., Sano, M., Stern, Y., & Mayeux, R. (1995). Neuropsychological characteristics of preclinical dementia in Parkinson's disease. *Neurology*, *45*(9), 1691–1696. http://doi.org/10.1212/WNL.45.9.1691

Jacobs, D. M., Winston, T. D., & Polanco, C. L. (1997). Assessment of verbal memory in spanish-speaking elders: Development of two frequency-matched list learning tests. *Journal of Clinical and Experimental Neuropsychology*, *19*(1), 119–125. http://doi.org/10.1080/01688639708403841

Johnson-Markve, B. L., Lee, G. P., Loring, D. W., & Viner, K. M. (2011). Usefulness of verbal selective reminding in distinguishing frontal lobe memory disorders in epilepsy. *Epilepsy & Behavior*, *22*(2), 313–317. http://doi.org/10.1016/j.yebeh.2011.06.039

Johnson, M. K., Raye, C. L., Mitchell, K. J., & Ankudowich, E. (2012). The Cognitive Neuroscience of True and False Memories. In R. F. Belli (Ed.), *True and False Recovered Memories* (Vol. 58, pp. 15–52). New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-1195-6_2

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. *ZUMA-Nachrichten Spezial*, *3*, 1–40.

Judd, T., Capetillo, D., Carrion-Baralt, J., Marmol, L. M., Miguel-Montes, L. S., Navarrete, M. G., … Valdes, J. (2009). Professional Considerations for Improving the Neuropsychological Evaluation of Hispanics: A National Academy of Neuropsychology Education Paper. *Archives of Clinical Neuropsychology*, *24*(2), 127–135. http://doi.org/10.1093/arclin/acp016

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika*, *54*(4), 681–697.

Khalid, M. N., & Glas, C. A. W. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186–197. http://doi.org/10.1016/j.measurement.2013.12.019

Kim, S.-H., & Cohen, A. S. (1995). A Comparison of Lord's Chi-Square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning. *Applied Measurement in Education*, *8*(4), 291–312. http://doi.org/10.1207/s15324818ame0804_2

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of Differential Item Functioning in Multiple Groups. *Journal of Educational Measurement*, *32*(3), 261–276. http://doi.org/10.1111/j.1745-3984.1995.tb00466.x

Kunnan, A. J. (2007). Test Fairness, Test Bias, and DIF. *Language Assessment Quarterly*, *4*(2), 109–112. http://doi.org/10.1080/15434300701375865

Larrabee, G. J., Trahan, D. E., Curtiss, G., & Levin, H. S. (1988). Normative data for the Verbal Selective Reminding Test. *Neuropsychology*, *2*(3-4), 173–182. http://doi.org/10.1037/h0091731

Lei, P.-W., & Li, H. (2013). Small-Sample DIF Estimation Using SIBTEST, Cochran's Z, and Log-Linear Smoothing. *Applied Psychological Measurement*, *37*(5), 397–416. http://doi.org/10.1177/0146621613478150

Levy, G., Jacobs, D. M., Tang, M.-X., Côté, L. J., Louis, E. D., Alfaro, B., … Marder, K. (2002). Memory and executive function impairment predict dementia in Parkinson's disease. *Movement Disorders*, *17*(6), 1221–1226. http://doi.org/10.1002/mds.10280

Lezak, M. D. (2004). *Neuropsychological Assessment*. Oxford University Press.

Li, H., & Stout, W. F. (1993). A new procedure for detection of crossing DIF/bias. Presented at the Annual meeting of the American Educational Research Association, Atlanta.

Llorente, A. M. (2008). Chapter 2. In A. M. Llorente (Ed.), *Principles of Neuropsychological Assessment with Hispanics: Theoretical Foundations and Clinical Practice*. Springer.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. Retrieved from http://psycnet.apa.org/?fa=main.doiLanding&uid=1993-97193-009

López de Ibáñez, M. (1998). *Evaluación neuropsicológica: principios y métodos*. CDCH UCV.

Magis, D., & de Boeck, P. (2014). Type I Error Inflation in DIF Identification With Mantel-Haenszel: An Explanation and a Solution. *Educational and Psychological Measurement*, *74*(4), 713–728. http://doi.org/10.1177/0013164413516855

Magis, D., & Facon, B. (2013). Item Purification Does Not Always Improve DIF Detection A Counterexample With Angoff's Delta Plot. *Educational and Psychological Measurement*, *73*(2), 293–311. http://doi.org/10.1177/0013164412451903

Maneesriwongul, W., & Dixon, J. K. (2004). Instrument translation process: a methods review. *Journal of Advanced Nursing*, *48*(2), 175–186.

Manly, J. J. (2008). Critical Issues in Cultural Neuropsychology: Profit from Diversity. *Neuropsychology Review*, *18*(3), 179–183. http://doi.org/10.1007/s11065-008-9068-8

Manly, J. J., Byrd, D. A., Touradji, P., & Stern, Y. (2004). Acculturation, Reading Level, and Neuropsychological Test Performance Among African American Elders. *Applied Neuropsychology*, *11*(1), 37–46. http://doi.org/10.1207/s15324826an1101_5

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Retrieved from http://arch.neicon.ru/xmlui/handle/123456789/3913344

Masur, D. M., Fuld, P. A., Blau, A. D., Crystal, H., & Aronson, M. K. (1990). Predicting development of dementia in the elderly with the Selective Reminding Test. *Journal of Clinical and Experimental Neuropsychology*, *12*(4), 529–538. http://doi.org/10.1080/01688639008400999

Matsumoto, D., & Van de Vijver, F. J. R. (2010). *Cross-Cultural Research Methods in Psychology*. Cambridge University Press.

Mayes, A. R. (1986). Learning and memory disorders and their assessment. *Neuropsychologia*, *24*(1), 25–39.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of Nonuniform Differential Item Functioning Using a Variation of the Mantel-Haenszel Procedure. *Educational and Psychological Measurement*, *54*(2), 284–291. http://doi.org/10.1177/0013164494054002003

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105–118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. http://doi.org/10.1016/0883-0355(89)90002-5

Millsap, R. E., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement*, *17*(4), 297–334. http://doi.org/10.1177/014662169301700401

Mitrushina, M. (2005). *Handbook of Normative Data for Neuropsychological Assessment*. Oxford University Press.

Mokri, H., Alberto Avila-Funes, J., Meillon, C., Gutierrez Robledo, L. M., & Amieva, H. (2013). Normative data for the Mini-Mental State Examination, the Free and Cued Selective Reminding Test and the Isaacs Set Test for an older

adult Mexican population: The Coyoacan Cohort Study. *Clinical Neuropsychologist*, *27*(6), 1004–1018. http://doi.org/10.1080/13854046.2013.809793

Morales, M., Campo, P., Fernandez, A., Moreno, D., Yanez, J., & Sanudo, I. (2010). Normative Data for a Six-Trial Administration of a Spanish Version of the Verbal Selective Reminding Test. *Archives of Clinical Neuropsychology*, *25*(8), 745–761. http://doi.org/10.1093/arclin/acq076

Moran, A., Diez Roux, A. V., Jackson, S. A., Kramer, H., Manolio, T. A., Shrager, S., & Shea, S. (2007). Acculturation is associated with hypertension in a multiethnic sample. *American Journal of Hypertension*, *20*(4), 354–363. http://doi.org/10.1016/j.amjhyper.2006.09.025

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema*, *25*(2), 151–157.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, *18*(4), 315–328. http://doi.org/10.1177/014662169401800403

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257–274. http://doi.org/10.1177/014662169602000306

Noe, E., Marder, K., Bell, K. L., Jacobs, D. M., Manly, J. J., & Stern, Y. (2004). Comparison of dementia with Lewy bodies to Alzheimer's disease and Parkinson's disease with dementia. *Movement Disorders*, *19*(1), 60–67. http://doi.org/10.1002/mds.10633

Núñez Núñez, R. M., Hidalgo Montesinos, M. D., & López Pina, J. A. (2000). Influencia de la igualación iterativa en la detección del funcionamiento diferencial del ítem mediante las medidas de área de Raju y el estadístico de Lord. *Psicothema*, *12*(3), 495–502.

O'Connell, M. E., & Tuokko, H. (2002). The 12-Item Buschke Memory Test: Appropriate for Use Across Levels of Impairment. *Applied Neuropsychology*, *9*(4), 226–233. http://doi.org/10.1207/S15324826AN0904_5

Oller, J. W., & Perkins, K. (1978). Intelligence and Language Proficiency as Sources of Variance in Self-Reported Affective Variables1. *Language Learning*, *28*(1), 85–97. http://doi.org/10.1111/j.1467-1770.1978.tb00306.x

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: detection and evaluation of impact. *Psychological Assessment*, *14*(1), 50–59.

Ostrosky-Solis, F., Ardila, A., & Rosselli, M. (1999). NEUROPSI: A brief neuropsychological test battery in Spanish with norms by age and educational level. *Journal of the International Neuropsychological Society*, *5*(5), 413–433. http://doi.org/10.1017/S1355617799555045

Ostrosky-Solis, F., Gomez-Perez, M. E., Matute, E., Rosselli, M., Ardila, A., & Pineda, D. (2007). NEUROPSI ATTENTION AND MEMORY: A neuropsychological test battery in Spanish with norms by age and educatioal level. *Applied Neuropsychology*, *14*(3), 156–170.

Ostrosky-Solis, F., Gutierrez, A. L., Flores, M. R., & Ardila, A. (2007). Same or different? Semantic verbal fluency across Spanish-speakers from different countries. *Archives of Clinical Neuropsychology*, *22*(3), 367–377. http://doi.org/10.1016/j.acn.2007.01.011

Palomo, R., Casals-Coll, M., Sánchez-Benavides, G., Quintana, M., Manero, R. M., Rognoni, T., … Peña-Casanova, J. (2013). Spanish normative studies in young adults (NEURONORMA young adults project): Norms for the Rey–Osterrieth Complex Figure (copy and memory) and Free and Cued Selective Reminding Test. *Neurología (English Edition)*, *28*(4), 226–235. http://doi.org/10.1016/j.nrleng.2012.03.017

Pedraza, O., & Mungas, D. (2008). Measurement in Cross-Cultural Neuropsychology. *Neuropsychology Review*, *18*(3), 184–193. http://doi.org/10.1007/s11065-008-9067-9

Peña-Casanova, J., Blesa, R., Aguilar, M., Gramunt-Fombuena, N., Gómez-Ansón, B., Oliva, R., … Sol, J. M. (2009). Spanish multicenter normative studies (NEURONORMA project): Methods and sample characteristics. *Archives of Clinical Neuropsychology*, *24*(4), 307–319. http://doi.org/10.1093/arclin/acp027

Peña-Casanova, J., Gramunt-Fombuena, N., Quiñones-Úbeda, S., Sánchez-Benavides, G., Aguilar, M., Badenes, D., … Blesa, R. (2009). Spanish Multicenter Normative Studies (NEURONORMA Project): Norms for the Rey–Osterrieth Complex Figure (Copy and Memory), and Free and Cued Selective Reminding Test. *Archives of Clinical Neuropsychology*, acp041. http://doi.org/10.1093/arclin/acp041

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement*, *29*(2), 150–151. http://doi.org/10.1177/0146621603260686

Penfield, R. D. (2007). An Approach for Categorizing DIF in Polytomous Items. *Applied Measurement in Education*, *20*(3), 335–355. http://doi.org/10.1080/08957340701431435

Peraita, H., Diaz, C., & Gonzalez-Labra, M. J. (2000). Evaluation of semantic/categorical deterioration in Alzheimer disease. *International Journal of Psychology*, *35*(3-4), 147–147.

Peters, M., & Passchier, J. (2006). Translating Instruments for Cross-Cultural Studies in Headache Research. *Headache: The Journal of Head and Face Pain*, *46*(1), 82–91. http://doi.org/10.1111/j.1526-4610.2006.00298.x

Plenger, P. M., Breier, J. I., Wheless, J. W., Papanicolaou, A. C., Brookshire, B., Thomas, A., … Willmore, L. J. (1996). Nonverbal selective reminding test: Efficacy in the assessment of adults with temporal lobe epilepsy. *Journal of Epilepsy*, *9*(1), 65–69. http://doi.org/10.1016/0896-6974(95)00059-3

Pontón, M. O., & Carrión, J. L. (2001). *Neuropsychology and the Hispanic Patient: A Clinical Handbook*. Routledge.

Poreh, A., & Sultan, A. (2009). Neurocognitive Testing of Minorities in Mental Health Settings. In S. Loue & M. Sajatovic (Eds.), *Determinants of Minority Mental Health and Wellness* (pp. 1–14). Springer New York. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-75659-2_16

Prieto, G., Delgado, A. R., Perea, M. V., & Ladera, V. (2011). Funcionamiento diferencial de los ítems del test Mini-mental en función de la patología. *Neurología*, *26*(8), 474–480. http://doi.org/10.1016/j.nrl.2011.01.013

Puente, A. E. (1993). Ethics in Neuropsychology - a Cross-Cultural-Perspective. *Journal of Clinical and Experimental Neuropsychology*, *15*(1), 19–19.

Puente, A. E., & Ardila, A. (2000). Neuropsychological assessment of hispanics. In *Handbook of Cross-Cultural Neuropsychology* (pp. 87–104). Springer Science & Business Media.

Puente, A. E., & McCaffrey, R. J. (1992). *Handbook of Neuropsychological Assessment*. Springer US.

Quaranta, M. (2013). Measuring political protest in Western Europe: Assessing cross-national equivalence. *European Political Science Review*, *5*(3), 457–482. http://doi.org/10.1017/S1755773912000203

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Rao, C. R., & Sinharay, S. (2007). *Psychometrics*. Elsevier.

Rao, S. L., & Andrade, C. (1998). Selective Reminding Test to measure verbal and visual memory. *Indian J Clin Psychol*, *25*(2), 149–153.

Rao, S. M., Leo, G. J., & Aubin-Faubert, P. S. (1989). On the nature of memory disturbance in multiple sclerosis. *Journal of Clinical and Experimental Neuropsychology*, *11*(5), 699–712. http://doi.org/10.1080/01688638908400926

Reis, A., Guerreiro, M., & Castro-Caldas, A. (1994). Influence of educational level of non brain-damaged subjects on visual naming capacities. *Journal of Clinical and Experimental Neuropsychology*, *16*(6), 939–942.

Reis, A., Petersson, K. M., Castro-Caldas, A., & Ingvar, M. (2001). Formal Schooling Influences Two- but Not Three-Dimensional Naming Skills. *Brain and Cognition*, *47*(3), 397–411. http://doi.org/10.1006/brcg.2001.1316

Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation*. Neuropsychology Press Tucson, AZ.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique [The Rey Auditory Verbal Learning Test. *RVLMT]. Arch. Psy-Chol. (Geneve)*, *28*, 21.

Rey, A. (1958). L'examen clinique en psychologie. Retrieved from http://psycnet.apa.org/psycinfo/1959-03776-000

Reynolds, C. R., & Voress, J. (2007). Test of Memory and Learning (TOMAL-2). *TX: PRO-ED*. Retrieved from http://www.v-psyche.com/doc/Clinical%20Test/Test%20of%20Memory%20%26%20Learning(TOMAL2).doc

Riedel-Heller, S. G., Matschinger, H., Schork, A., & Angermeyer, M. C. (1999). Do memory complaints indicate the presence of cognitive impairment? - Results of a field study. *European Archives of Psychiatry & Clinical Neuroscience*, *249*(4), 197.

Robinson, J. P. (2010). The Effects of Test Translation on Young English Learners' Mathematics Performance. *Educational Researcher*, *39*(8), 582–590. http://doi.org/10.3102/0013189X10389811

Rogers, H. J., & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, *17*(2), 105–116. http://doi.org/10.1177/014662169301700201

Rosselli, M., Ardila, A., Marquez, M., Matos, L., Salvatierra, J. L., Weekes, V. A., & Ostrosky, F. (1999). Linguistic organization in verbal fluency tests among English and Spanish speakers and Spanish-English bilinguals. *Archives of Clinical Neuropsychology*, *14*(8), 714–714.

Rosselli, N., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition*, *52*(3), 326–333. http://doi.org/10.1016/S0278-2626(03)00170-2

Roussos, L. A., & Stout, W. F. (1996). Simulation Studies of the Effects of Small Sample Size and Studied Item Parameters on SIBTEST and Mantel-Haenszel Type I Error Performance. *Journal of Educational Measurement*, *33*(2), 215–230. http://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Ruff, R. M., Light, R. H., & Quayhagen, M. (1989). Selective reminding tests: A normative study of verbal learning in adults. *Journal of Clinical and Experimental Neuropsychology*, *11*(4), 539–550. http://doi.org/10.1080/01688638908400912

Schacter, D. L. (1987). Memory, amnesia, and frontal lobe dysfunction. *Psychobiology*, *15*(1), 21–36.

Schacter, D. L., Chiu, C.-Y. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual Review of Neuroscience*, *16*(1), 159–182.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*(3), 143–152.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., … Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*(3), 288–295. http://doi.org/10.1016/j.jclinepi.2008.06.003

Scoville, W. B., & Milner, B. (1957). LOSS OF RECENT MEMORY AFTER BILATERAL HIPPOCAMPAL LESIONS. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. http://doi.org/10.1007/BF02294572

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational and Behavioral Statistics*, *6*(4), 317–375.

Siedlecki, K. L., Manly, J. J., Brickman, A. M., Schupf, N., Tang, M.-X., & Stern, Y. (2010). Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? *Neuropsychology*, *24*(3), 402–411. http://doi.org/10.1037/a0017515

Silva, C., Faísca, L., Ingvar, M., Petersson, K. M., & Reis, A. (2012). Literacy: Exploring working memory systems. *Journal of Clinical and Experimental Neuropsychology*, *34*(4), 369–377. http://doi.org/10.1080/13803395.2011.645017

Sireci, S. G. (1997). Problems and Issues in Linking Assessments Across Languages. *Educational Measurement: Issues and Practice*, *16*(1), 12–19. http://doi.org/10.1111/j.1745-3992.1997.tb00581.x

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, *20*(2), 148–166. http://doi.org/10.1191/0265532203lt249oa

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating Guidelines For Test Adaptations A Methodological Analysis of Translation Quality. *Journal of Cross-Cultural Psychology*, *37*(5), 557–567. http://doi.org/10.1177/0022022106290478

Smith, R. L., Goode, K. T., La Marche, J. A., & Boll, T. J. (1995). Selective Reminding Test short form administration: A comparison of two through twelve trials. *Psychological Assessment*, *7*(2), 177–182. http://doi.org/10.1037/1040-3590.7.2.177

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. http://doi.org/10.1037/0033-295X.99.2.195

Squire, L. R., & Shimamura, A. (1996). The neuropsychology of memory dysfunction and its assessment. *Neuropsychological Assessment of Neuropsychiatric Disorders*, 232–262.

Squires, A., Finlayson, C., Gerchow, L., Cimiotti, J. P., Matthews, A., Schwendimann, R., … Sermeus, W. (2014). Methodological considerations when translating “burnout.” *Burnout Research*, *1*(2), 59–68. http://doi.org/10.1016/j.burn.2014.07.001

Stanczak, D. E., Stanczak, E. M., & Awadalla, A. W. (2001). Development and initial validation of an Arabic version of the Expanded Trail Making Test: Implications for cross-cultural assessment. *Archives of Clinical Neuropsychology*, *16*(2), 141–149.

Stansfield, C. W. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, *20*(2), 189–207. http://doi.org/10.1191/0265532203lt252oa

Stedman, L. C. (1994). Incomplete Explanations: The Case of U.S. Performance in the International Assessments of Education. *Educational Researcher*, *23*(7), 24–32. http://doi.org/10.3102/0013189X023007024

Stern, Y., Albert, S., Tang, M.-X., & Tsai, W.-Y. (1999). Rate of memory decline in AD is related to education and occupation Cognitive reserve? *Neurology*, *53*(9), 1942–1942. http://doi.org/10.1212/WNL.53.9.1942

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*(4), 361–370. http://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tabert, M. H., Manly, J. J., Liu, X., Pelton, G. H., Rosenblum, S., Jacobs, M., … Devanand, D. P. (2006). Neuropsychological prediction of conversion to alzheimer disease in patients with mild cognitive impairment. *Archives of General Psychiatry*, *63*(8), 916–924. http://doi.org/10.1001/archpsyc.63.8.916

Tamayo, F., Casals-Coll, M., Sánchez-Benavides, G., Quintana, M., Manero, R. M., Rognoni, T., … Peña-Casanova, J. (2012). Spanish normative studies in a young adult population (NEURONORMA young adults project): Guidelines for the span verbal, span visuo-spatial, Letter-Number Sequencing, Trail Making Test and Symbol Digit Modalities Test. *Neurología (English Edition)*, *27*(6), 319–329. http://doi.org/10.1016/j.nrleng.2012.07.008

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118–128. http://doi.org/10.1037/0033-2909.99.1.118

Torres, I. J., Mundt, A. J., Sweeney, P. J., Llanes–Macy, S., Dunaway, L., Castillo, M., & Macdonald, R. L. (2003). A longitudinal neuropsychological study of partial brain radiation in adults with brain tumors. *Neurology*, *60*(7), 1113–1118. http://doi.org/10.1212/01.WNL.0000055862.20003.4A

Touradji, P., Manly, J. J., Jacobs, D. M., & Stern, Y. (2001). Neuropsychological Test Performance: A Study of Non-Hispanic White Elderly*. *Journal of Clinical and Experimental Neuropsychology*, *23*(5), 643–649. http://doi.org/10.1076/jcen.23.5.643.1246

Trahan, D. E., & Larrabee, G. J. (1993). Clinical and methodological issues in measuring rate of forgetting with the verbal selective reminding test. *Psychological Assessment*, *5*(1), 67–71. http://doi.org/10.1037/1040-3590.5.1.67

Traykov, L., Baudic, S., Raoux, N., Latour, F., Rieu, D., Smagghe, A., & Rigaud, A.-S. (2005). Patterns of memory impairment and perseverative behavior discriminate early Alzheimer's disease from subcortical vascular dementia. *Journal of the Neurological Sciences*, *229–230*, 75–79. http://doi.org/10.1016/j.jns.2004.11.006

Umfleet, L. G., Janecek, J. K., Quasney, E., Sabsevitz, D. S., Ryan, J. J., Binder, J. R., & Swanson, S. J. (2014). Sensitivity and Specificity of Memory and Naming Tests for Identifying Left Temporal-Lobe Epilepsy. *Applied Neuropsychology: Adult*, *0*(0), 1–8. http://doi.org/10.1080/23279095.2014.895366

U. S. Census Bureau, D. I. S. (2005). 2005 Interim State Population Projections. Retrieved June 26, 2013, from http://www.census.gov/population/projections/data/state/projectionsagesex.html

U. S. Census Bureau, P. (2010). *U. S. Census Bureau 2010*. U. S.Census Bureau. Retrieved from http://www.census.gov/2010census/

Van Der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An Iterative Item Bias Detection Method. *Journal of Educational Measurement*, *21*(2), 131–145. http://doi.org/10.1111/j.1745-3984.1984.tb00225.x

Van De Vijver, F. J. R. (2013). Contributions of internationalization to psychology: Toward a global and inclusive discipline. *American Psychologist*, *68*(8), 761–770. http://doi.org/10.1037/a0033762

Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing Across Cultures. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in Educational and Psychological Testing: Theory and Applications* (pp. 277–308). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-94-009-2195-5_10

Van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural Equivalence in Multilevel Research. *Journal of Cross-Cultural Psychology*, *33*(2), 141–156. http://doi.org/10.1177/0022022102033002002

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119–135. http://doi.org/10.1016/j.erap.2003.12.004

Varela Ruíz, M., Fortoul, V., Ávila Costa, Ávila Acosta, M. R., & Fortoul van derGoes, T. I. (2005). *La Memoria*. Ed. Médica Panamericana.

Walker, C. M., Beretvas, S. N., & Ackerman, T. (2001). An Examination of Conditioning Variables Used in Computer Adaptive Testing for DIF Analyses. *Applied Measurement in Education*, *14*(1), 3–16. http://doi.org/10.1207/S15324818AME1401_02

Wang, W.-L., Lee, H.-L., & Fetzer, S. J. (2006). Challenges and strategies of instrument translation. *Western Journal of Nursing Research*, *28*(3), 310–321.

Warrington, E. K., & Weiskrantz, L. (1968). A study of learning and retention in amnesic patients. *Neuropsychologia*, *6*(3), 283–291.

Warrington, E. K., & Weiskrantz, L. (1970). Amnesic syndrome: Consolidation or retrieval? *Nature*. Retrieved from http://doi.apa.org/psycinfo/1971-23574-001

Wechsler, D. (1945). Wechsler memory scale. Retrieved from http://psycnet.apa.org/psycinfo/1946-00348-000

Wechsler, D. (1955). *MANUAL FOR THE WECHSLER ADULT INTELLIGENCE SCALE* (Vol. vi). Oxford, England: Psychological Corp.

Wechsler, D. (2009). *WMS-IV.: Wechsler Memory Scale-Administration and Scoring Manual*. Psychological Corporation.

Westbury, I. (1992). Comparing American and Japanese Achievement: Is the United States Really a Low Achiever? *Educational Researcher*, *21*(5), 18–24. http://doi.org/10.3102/0013189X021005018

Westerveld, M., Sass, K. J., Sass, A., & Henry, H. G. (1994). Assessment of verbal memory in temporal lobe epilepsy using the selective reminding test: Equivalence and reliability of alternate forms. *Journal of Epilepsy*, *7*(1), 57–63. http://doi.org/10.1016/0896-6974(94)90122-8

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO)

Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health*, *8*(2), 94–104. http://doi.org/10.1111/j.1524-4733.2005.04054.x

Williams, V. S. L. (1997). The “Unbiased” Anchor: Bridging the Gap Between DIF and Item Bias. *Applied Measurement in Education*, *10*(3), 253–267. http://doi.org/10.1207/s15324818ame1003_4

Wong, T. M., & Fujii, D. E. (2004). Neuropsychological Assessment of Asian Americans: Demographic Factors, Cultural Diversity, and Practical Guidelines. *Applied Neuropsychology*, *11*(1), 23–36. http://doi.org/10.1207/s15324826an1101_4

Wong, T. M., Strickland, T. L., Fletcher-Janzen, E., Ardila, A., & Reynolds, C. R. (2000). Theoretical and practical issues in the neuropsychological assessment and treatment of culturally dissimilar patients. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology* (pp. 3–18). Retrieved from http://www.springerlink.com/index/Q6056RK065478Q1H.pdf

Xavier, F. M. F., Ferraz, M. P. T., Trentini, C. M., Freitas, N. K., & Moriguchi, E. H. (2002). Bereavement-related cognitive impairment in an oldest-old community-dwelling Brazilian sample. *Journal of Clinical and Experimental Neuropsychology*, *24*(3), 294–301. http://doi.org/10.1076/jcen.24.3.294.983

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. http://doi.org/10.1177/0265532209349465

Zec, R. F., Zellers, D., Belman, J., Miller, J., Matthews, J., Ferneau-Belman, D., & Robbs, R. (2001). Long-Term Consequences of Severe Closed Head Injury on Episodic Memory. *Journal of Clinical and Experimental Neuropsychology*, *23*(5), 671–691. http://doi.org/10.1076/jcen.23.5.671.1247

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection Of Differential Item Functioning In Large-Scale State Assessments: A Study Evaluating A Two-Stage Approach. *Educational and Psychological Measurement*, *63*(1), 51–64. http://doi.org/10.1177/0013164402239316

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233. http://doi.org/10.1080/15434300701375832

Zwick, R. (2012). A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. *ETS Research Report Series*, *2012*(1), i–30. http://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Zwick, R., Donogue, J., Grima, A., Holland, P. W., Thayer, D., Thomas, N., & Wingersky, M. (1992). Differential item functioning analysis for new models of assesment. Presented at the Annual Convention of the National Council on Measurement in Education, San Francisco.

Zwick, R., & Thayer, D. T. (1994). Evaluation of the Magnitude of Differential Item Functioning in Polytomous Items. *ETS Research Report Series*, *1994*(1), i–25. http://doi.org/10.1002/j.2333-8504.1994.tb01586.x