Juan Pascual Llobell
María Dolores Frías Navarro
José Fernando García Pérez
Universitat de València

Manual de psicología experimental

Metodología de investigación

Ariel Psicología, 1996

Capítulo III

Diseño de medidas repetidas

CUANDO administramos los tratamientos objeto de nuestra investigación a los mismos sujetos y, en consecuencia, estos reciben más de un tratamiento experimental, disponiendo al menos de una observación por tratamiento y sujeto, decimos que estamos en presencia de un diseño *intra-sujetos* o de *medidas repetidas.* Estos diseños constituyen una alternativa válida y en muchos casos más adecuada que los diseños *entre-sujetos* —o *entre-grupos*— analizados en el capítulo anterior. Los diseños de medidas repetidas constituyen en la actualidad uno de los procedimientos analíticos más ampliamente utilizados en Psicología y en Ciencias Sociales y de la Salud.

En este capítulo expondremos los supuestos sobre los que se sustentan estos diseños, las razones de su mayor potencia, las restricciones que imponen a los datos para su correcta aplicación así como los modelos básicos que se pueden formular con el objetivo último de que el lector aprenda no sólo a interpretarlos correctamente sino también a planificar la investigación de acuerdo con el esquema formal que les es propio.

La situación más simple de un diseño de *medidas repetidas* se caracteriza por incluir varias observaciones por unidad experimental (sujetos), obtenidas cada una de ellas bajo una condición experimental diferente (tratamiento). Nos encontramos entonces con un cruce entre la variable sujetos y la variable tratamiento que configura un diseño de *medidas totalmente repetidas* (Lee, 1975), diseño *intra* (Keppel, 1982) o diseño de *sujetos* $\times$ *tratamientos*. Puede ocurrir, sin embargo, que además del factor de medidas repetidas (factor de tratamiento), el diseño incluya una o más variables independientes de naturaleza *entre-*

sujetos, planteándose entonces un diseño de *medidas parcialmente repetidas* (Lee, 1975), también conocido como diseño *factorial mixto* (Winer, 1971), diseño *split – plot* (Kirk, 1982), diseño de *grupos × sujetos intra-grupos × tratamiento* (Arnau, 1990a), diseño *Lindquist Tipo I* (Lindquist, 1953) o diseño de *medidas repetidas de muestra dividida* (Vallejo, 1991). Serán diseños bastante más complejos pero que a todas luces participan de las mismas condiciones generales de uso que el primero.

Los diseños *intra-sujetos* o de *medidas repetidas* se pueden utilizar en diferentes tipos de situaciones (Maxwell y Delaney, 1990):

a) evaluación longitudinal del cambio a lo largo del tiempo
b) evaluación de la actuación de los sujetos bajo diferentes condiciones de tratamiento en estudios transversales y,
c) en aquellas otras situaciones en las que deseamos comparar las puntuaciones de los mismos sujetos obtenidas en diferentes pruebas psicométricas.

Esta tercera situación es conocida como *análisis de perfiles* cuyo objetivo es medir el perfil medio de las puntuaciones de los sujetos en los *tests*, comparando los resultados entre dos o más grupos de sujetos (Stevens, 1992). La situación primera alude directamente a condiciones en las que la variable tiempo tiene naturaleza propia (*diseños de series temporales*). En este capítulo trataremos aquellos casos que tienen que ver casi exclusivamente con la situación segunda. Imagine el siguiente ejemplo: queremos comprobar si el recuerdo de los sujetos depende del tipo de procesamiento (niveles de procesamiento) mediante el que se analiza la información. Para ello hacemos leer a nuestros sujetos experimentales un conjunto de palabras que llevan asociadas unas preguntas (una sola pregunta por palabra) que canalizan la atención sobre ciertos atributos de las palabras como ¿está escrita en mayúsculas? ¿es sinónimo de…? ¿te la puedes imaginar? ¿sabes lo que significa? La pregunta que se plantea el investigador es clara: *¿qué palabras se recordarán mejor?* ¿ El recuerdo será igual para todas? La variable independiente de interés es el tipo de procesamiento (operacionalizada como tipo de pregunta) y cada uno de los sujetos lee las 40 palabras que 10 a 10 llevan asociadas un tipo particular de pregunta.

El ejemplo anterior constituye una situación típica de lo que entendemos por diseño *intra-sujetos* o diseño de *medidas repetidas*. En el contexto de este libro vamos a utilizar de forma intercambiable ambos términos.

Ventajas e inconvenientes

Las ventajas asociadas a esta categoría de diseños son fundamentalmente tres: el menor costo en número de sujetos, el control de las diferencias individuales y el incremento de la potencia del diseño. Las tres están intrínsecamente relacionadas y son interdependientes entre sí.

La reducción del número de sujetos es una ventaja importante bajo múltiples punto de vista, incluidos los ahorro de tiempo y costos. Teniendo en cuenta que el mismo grupo de sujetos recibe todos los tratamientos, el número de sujetos que se necesitan para completar la investigación es menor en comparación con los diseños *entre-sujetos*. Por ejemplo, si un investigador plantea un estudio con dos variables independientes, cada una de ellas con tres niveles, (3×3), una variable dependiente y diez observaciones por condición, necesitará 90 sujetos si aplica un diseño *entre-sujetos* mientras que sólo serán necesarios 10 sujetos si el diseño es de *medidas repetidas*. Evidentemente, cada uno de los diez sujetos tendrá que pasar por las nueve condiciones experimentales. Conviene anotar que en los diseños *entre-sujetos* el número de observaciones se corresponde con el número total de sujetos incluidos en el estudio dado que cada sujeto es medido en una única ocasión. En los diseños de *medidas repetidas* el número de observaciones no coincide con el total de sujetos ya que cada uno de ellos es medido en más de una condición experimental, generando los n sujetos $n \cdot a$ observaciones. Los diez sujetos del diseño anterior de *medidas repetidas* dan lugar a noventa observaciones ($10 \cdot (3 \cdot 3) = 90$) mientras que si el diseño fuese *entre-sujetos,* esos diez sujetos únicamente producirían diez observaciones, siendo necesario contar con noventa sujetos ($n \cdot a$ sujetos) para completar una investigación con noventa observaciones.

La equivalencia previa de los sujetos es un elemento clave para garantizar la validez de nuestras conclusiones. Técnicas como la asignación aleatoria de los sujetos a los grupos, el apareamiento en alguna variable relevante o su bloqueo, tienen como principal propósito asegurar esa equivalencia, proporcionando grupos homogéneos y controlando así diferencias anteriores. Que la administración de los tratamientos sea *intra-sujetos* y no *entre-sujetos* es una garantía de equivalencia y de reducción de la varianza de error. Hemos de recordar que el diseño experimental que garantiza por definición la aleatorización de los sujetos a las condiciones también garantiza por la misma razón, la igualdad o equivalencia (homogeneidad de varianzas) de los grupos de tratamiento. No obstante la aleatorización a veces no es suficiente por lo que otras técnicas, tales como las enumeradas anteriormente, son necesarias.

En los diseños de *medidas repetidas*, la equivalencia previa de los sujetos no constituye un problema relevante dado que, al ser medidos los sujetos bajo todos los niveles o condiciones experimentales, el sujeto actúa como su propio control, no afectando a la estimación de los efectos las posibles diferencias individuales. En estos diseños, por el contrario, la técnica del bloqueo llega a su punto máximo ya que cada sujeto constituye un bloque, eliminando del componente de error la variabilidad atribuida a las diferencias individuales (Winer, 1971; Riba, 1990; Arnau, 1995). De ahí el nombre de diseño de *bloques aleatorizados* ya que cada sujeto, seleccionado aleatoriamente, forma un ‘bloque’ y, dentro de cada ‘bloque’ se producen *a* replicas del mismo sujeto, al menos una por tratamiento.

Otra ventaja, relacionada con la anterior, radica en la mayor potencia de los diseños de *medidas repetidas* para detectar efectos experimentales que realmente existen (Stevens, 1992). Como ya se ha comentado, la varianza del término de error en estos diseños está despojada de la variabilidad atribuida a las diferencias individuales, codificadas en el diseño de *medidas repetidas* como un ‘factor sujeto’ que adopta los supuestos de un modelo de efectos aleatorios. Es decir, dado que las diferencias entre los tratamientos son valoradas *dentro* de cada sujeto y no *entre* los grupos, la variabilidad atribuida a las diferencias individuales entre los sujetos, (efecto principal del factor sujetos), es eliminada del término de error, aumentando con ello la precisión y potencia de la prueba estadística y acentuando así la validez de *conclusión estadística* (Maxwell y Delaney, 1990; Vallejo, 1991).

Pese a sus evidentes ventajas, los diseños de *medidas repetidas* presentan también una serie de inconvenientes que han de ser tenidos en cuenta para evitar su uso incorrecto. El fenómeno de atrición o los efectos de orden o la secuencia de administración de los tratamientos pueden sesgar los resultados. El efecto de práctica, los efectos *'carry-over'* o de arrastre, provocados por los residuales de anteriores tratamientos, o la propia fatiga, constituyen posibles secuelas tras la administración del tratamiento que deberán ser neutralizadas, evitando inferencias incorrectas al atribuir los efectos a causas que realmente no lo son.

Las técnicas para neutralizar dichos efectos son diversas y conocidas. Los efectos de atrición pueden ser debilitados con un cuidadoso análisis de los datos faltantes o *'missing'*. Respecto a los efectos provocados por la secuencia de aplicación de la variable independiente, gran parte de la acción residual puede ser parcialmente controlada utilizando *técnicas de contrabalanceo* (Underwood y Shaughnessy, 1978), equilibrando el orden de administración de los tratamientos o asignando aleatoriamente a cada sujeto el orden de presentación de los mismos (Pascual, García y Frías, 1995). El diseño de espacios temporales amplios entre la aplicación de los tratamientos minimiza la presencia de efectos de arrastre, de la práctica e incluso impedir la fatiga.

Los diseños *intra* tienen otro inconveniente que a nuestro entender es el que tiene mayor importancia y por tanto le concederemos un especial tratamiento. Nos referimos a la posible violación de algún supuesto de naturaleza estadística. Puesto que son los mismos sujetos los que reciben cada una de las condiciones experimentales es normal que en estos diseños aparezca un efecto *sistemático* en las respuestas, dando lugar a la aparición de correlación o dependencia entre los errores, algo por lo demás muy habitual en estos diseños (O'Brien y Kaiser, 1985). La violación de este supuesto repercute gravemente en los resultados obtenidos, ya que la prueba F no es robusta ante observaciones correlacionadas o dependientes. Los intentos de solución a este problema han llevado a plantearse dos estrategias de análisis distintas: la clásica o tradicional, con todos los ajustes necesarios para el correcto control del error de *Tipo I* y la perspectiva multivariada.

Técnicas de análisis

El adecuado análisis de los diseños de *medidas repetidas* requiere el cumplimiento de ciertos supuestos que aseguren la validez del contraste de hipótesis mediante la prueba *F* con una distribución con $a - 1$ y $(a - 1) \cdot (n - 1)$ grados de libertad para las fuentes de varianza del tratamiento y del error respectivamente. Asumimos la *homogeneidad de las varianzas intra- tratamiento*, también la *independencia de las observaciones* y se asume que las observaciones en los diferentes niveles del tratamiento siguen una *distribución normal* conjunta multivariada, si consideramos cada medición por tratamiento como una variable dependiente. La violación de alguno de estos supuestos afectará al error de *Tipo I* y a la potencia de la prueba *F*; de entre ellos es especialmente grave la existencia de correlación entre los errores (Winer, 1971; Kirk, 1982).

Teniendo en cuenta que son los mismos sujetos los que reciben las diferentes condiciones experimentales, no es extraño que los errores puedan estar correlacionados. En todos los casos, la dependencia serial entre las observaciones tomadas a una misma unidad de análisis (sujeto) en distintos momentos puede llegar a ser de considerable importancia. Como señala Vallejo (1986), las observaciones registradas a un mismo sujeto, además de estar positivamente correlacionadas, presentan una matriz de varianzas-covarianzas entre las medidas repetidas que tiene una estructura *Toeplitz* (las puntuaciones más próximas presentan una correlación más elevada). Cuanto mayor sea el patrón de correlaciones, menor será la varianza residual, aumentando la probabilidad de que los resultados de la prueba *F* resulten estadísticamente significativos (Girden, 1992, p.14). La correlación sesga pues, positivamente la significación de la prueba *F*.

Análisis de supuestos: homogeneidad, esfericidad y simetría compuesta.

El principal supuesto que ha de tener en cuenta el investigador al analizar los diseños de *medidas repetidas* se refiere al de homogeneidad o igualdad de las varianzas de las diferencias entre pares de puntuaciones. Huynt y Feldt (1970) y Rouanet y Lépine (1970) demostraron que la homogeneidad de las varianzas de las diferencias equivale a asumir que la matriz de covarianza tiene determinada forma, denominada *esfericidad o circularidad*, siendo la *simetría compuesta* o *combinada* un caso particular de la esfericidad.

El supuesto de *simetría compuesta* se cumple si y sólo si todas las varianzas son iguales y también lo son las covarianzas.

…………………………………………….

Existen diferentes procedimientos que permiten comprobar la *hipótesis nula* de la igualdad u homogeneidad de las varianzas poblacionales de las puntuaciones de diferencia. Una de ellas es la prueba *W* de Mauchly (1940). Aunque algunos autores como Keselman et al. (1980) le conceden poco valor, creemos conveniente presentar su desarrollo. Analicemos el proceso de cálculo y sus características más destacadas dado que es el procedimiento desarrollado en los paquetes estadísticos como técnica de análisis del grado de esfericidad de la matriz de varianzas-covarianzas.

Prueba de esfericidad de Mauchly

Existen varias pruebas que permiten comprobar el nivel de esfericidad de la matriz de varianzas-covarianzas; el *test W* de Mauchly (1940) es una de ellas (Kirk, 1982, p. 259; Vallejo, 1991, p. 31). Esta prueba parte de la estimación muestral de la matriz de varianzas-covarianzas poblacional, Σ, con el propósito de probar si $\mathbf{C^{*\prime}} \Sigma \mathbf{C^{*}}$ cumple el supuesto de *esfericidad* (Huynh y Mandeville, 1979). En los trabajos de Consul (1967; 1969) y Nagarsenker y Pillai (1973) puede consultarse la distribución exacta de *W* así como las tablas de valores críticos.

En el caso de que no se pudiese mantener el supuesto de *esfericidad*, la razón *F* del análisis de la varianza estaría positivamente sesgada, tal y como demostró Box (1954a).

Los estudios de Monte Carlo no recomiendan la utilización de la prueba W de Mauchly como técnica preliminar de análisis de la *esfericidad* (Keselman, Rogan, Mendoza y Breen, 1980; Rogan, Keselman y Mendoza, 1979) ya que es una prueba muy sensible a la desviación de la normalidad multivariada (Maxwell y Delaney, 1990; Stevens, 1992).

Cuando la matriz de varianzas-covarianzas se aleja, significativamente, del supuesto de *esfericidad*, la prueba *F* presenta un sesgo positivo, aumentando la tasa de error de *Tipo I* lo que invalida las conclusiones derivadas del análisis univariado. El investigador tiene entonces que optar por ajustar los grados de libertad o aplicar un análisis multivariado de la varianza basándose en $a - 1$ contrastes independientes. Analicemos en primer lugar el proceso de cálculo bajo la perspectiva univariada tradicional sin ajustar y ajustando los grados de libertad bajo diferentes opciones.

Aproximación univariada

El modelo de predicción de un diseño de *medidas repetidas* con un factor de tratamiento se ajusta plenamente a la estructura de un diseño factorial, ya conocido y analizado en el capítulo anterior, con la diferencia que ahora sólo existe un factor de tratamiento siendo el segundo el denominado ‘factor sujetos’. La variable ‘sujetos’ se define como un factor cruzado dentro de todos los niveles del factor tratamiento o, si el diseño es factorial, dentro de todas las celdillas de combinación de niveles de dos o más factores de medidas repetidas.

Desde el punto de vista tradicional, el método de análisis de los diseños de *medidas repetidas* se basa en el modelo estructural del ANOVA con valores esperados mixtos dado que el denominado factor *Sujetos* se considera que asume un modelo de efectos aleatorios puesto que los sujetos o unidades experimentales han sido seleccionados aleatoriamente de entre la población de sujetos, mientras que el factor *Tratamiento* normalmente asume un modelo de efectos fijos (McCall y Appelbaum, 1973) ya que sus niveles son seleccionados arbitrariamente por el investigador.

Desde la aproximación univariada, la ecuación del modelo estructural de un diseño de *medidas repetidas* con un factor de tratamiento y un factor sujetos es:

$$H_1 \equiv \mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{S} + \mathbf{AS} + \mathbf{E} \qquad [1]$$

En este caso, la interacción **AS** está formada por una sola observación (aunque no necesariamente) y su variabilidad coincide con la varianza *no explicada* o varianza *residual*. Dado que se está trabajando con un modelo *mixto*, asumiendo el efecto del tratamiento o efecto principal un modelo de *efectos fijos*, mientras que el factor sujetos asume un modelo de *efectos aleatorios*, el término de error adecuado o el denominador de la prueba *F* que debemos utilizar al realizar el contraste de hipótesis, es la interacción entre el factor de efectos fijos y el factor de efectos aleatorios. En el caso particular de una sola observación por condición y sujeto no puede haber estimación independiente del término de error y de la interacción; la variabilidad atribuible a dicha interacción la consideramos como varianza residual no explicada quedando entonces definida la ecuación estructural como:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{S} + \mathbf{AS} \qquad [2]$$

Estimemos los valores de cada una de las fuentes de variación y las *Sumas de Cuadrados* correspondientes por el mismo procedimiento que hemos utilizado hasta ahora.

$$\mathbf{A} = \mathbf{M}_{\mathrm{A}} - \mathbf{M} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \end{pmatrix} - \begin{pmatrix} 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \end{pmatrix} = \begin{pmatrix} -5 \\ -5 \\ -5 \\ -5 \\ -5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 5 \\ 5 \\ 5 \\ 5 \\ 5 \end{pmatrix}$$

$$\mathbf{SC}_{\mathbf{intra}\ \mathrm{A}} = \mathbf{A^2} = \mathbf{A'\,A} = \boxed{250}$$

$$\mathbf{S} = \mathbf{M}_{\mathrm{S}} - \mathbf{M} = \begin{pmatrix} 9 \\ 20 \\ 21 \\ 11 \\ 14 \\ 9 \\ 20 \\ 21 \\ 11 \\ 14 \\ 9 \\ 20 \\ 21 \\ 11 \\ 14 \end{pmatrix} - \begin{pmatrix} 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \end{pmatrix} = \begin{pmatrix} -6 \\ 5 \\ 6 \\ -4 \\ -1 \\ -6 \\ 5 \\ 6 \\ -4 \\ -1 \\ -6 \\ 5 \\ 6 \\ -4 \\ -1 \end{pmatrix}$$

$$\mathbf{SC}_{\mathbf{entre}\ \mathrm{S}} = \mathbf{S^2} = \mathbf{S'\,S} = \boxed{342}$$

El cálculo de la fuente de varianza de interacción se realiza también de la forma usual, restando a las medias de interacción $\mathbf{S} \times \mathbf{A}$ (en este caso sobre las puntuaciones directas) la media general y los efectos implicados en la interacción objeto de análisis.

Continuemos con el cálculo del término de error,

$$\mathbf{AS} = \mathbf{M}_{\mathrm{AS}} - \mathbf{M} - \mathbf{A} - \mathbf{S} = \begin{pmatrix} 4 \\ 17 \\ 12 \\ 7 \\ 10 \\ 9 \\ 12 \\ 27 \\ 12 \\ 15 \\ 14 \\ 31 \\ 24 \\ 14 \\ 17 \end{pmatrix} - \begin{pmatrix} 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \end{pmatrix} - \begin{pmatrix} -5 \\ -5 \\ -5 \\ -5 \\ -5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 5 \\ 5 \\ 5 \\ 5 \\ 5 \end{pmatrix} - \begin{pmatrix} -6 \\ 5 \\ 6 \\ -4 \\ -1 \\ -6 \\ 5 \\ 6 \\ -4 \\ -1 \\ -6 \\ 5 \\ 6 \\ -4 \\ -1 \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 2 \\ -4 \\ 1 \\ 1 \\ 0 \\ -8 \\ 6 \\ 1 \\ 1 \\ 0 \\ 6 \\ -2 \\ -2 \\ -2 \end{pmatrix}$$

$$\mathbf{SC}_{\mathbf{residual\ AS}} = \mathbf{AS^2} = \mathbf{AS^2} = \boxed{172}$$

Aplicando la descomposición de la ecuación estructural a las puntuaciones de los sujetos del supuesto obtenemos la siguiente predicción:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{S} + \mathbf{AS} = \begin{pmatrix} 4 \\ 10 \\ 5 \\ 4 \\ 7 \\ 6 \\ 2 \\ 17 \\ 6 \\ 9 \\ 5 \\ 15 \\ 8 \\ 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \\ 15 \end{pmatrix} + \begin{pmatrix} -5 \\ -5 \\ -5 \\ -5 \\ -5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 5 \\ 5 \\ 5 \\ 5 \\ 5 \end{pmatrix} + \begin{pmatrix} -6 \\ 5 \\ 6 \\ -4 \\ -1 \\ -6 \\ 5 \\ 6 \\ -4 \\ -1 \\ -6 \\ 5 \\ 6 \\ -4 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ 2 \\ -4 \\ 1 \\ 1 \\ 0 \\ -8 \\ 6 \\ 1 \\ 1 \\ 0 \\ 6 \\ -2 \\ -2 \\ -2 \end{pmatrix}$$

Como señala Riba (1990) no suele ser usual realizar una estimación de los parámetros asociados al factor sujeto (factor considerado como *entre*) ya que por regla general el investigador no se interesa por conocer los efectos individualizados de los sujetos. Efectivamente, en el contexto de la psicología experimental básica todo lo que sea variabilidad asociada a las diferencias individuales se tiende a considerarlo como varianza de error, dado que lo que le interesa a este investigador básico es detectar leyes o reglas generales de conducta. No obstante, desde una óptica diferencialista sería igualmente legítimo el uso de este tipo de diseños, en cuyo caso lo más importante sería interpretar el factor sujetos y sus posibles interacciones con los tratamientos.

A continuación sólo resta aplicar la prueba de la hipótesis a los resultados obtenidos (véase ***Tabla 1***).

Tabla 1 *Diseño con medidas repetidas en a = 3*

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*
Sujetos	342.000	4	85.500	3.977	0.046
A	250.000	2	125.000	5.814	0.028
A × *Sujetos*	172.000	8	21.500		
Total	764.000	14			

Los resultados señalan la existencia de diferencias estadísticamente significativas en el nivel de recuerdo de los sujetos en función del uso de las palabras ($F_{(2, 8)} = 5.814$, $p < 0.05$). Por tanto hemos de rechazar la *hipótesis de nulidad* del efecto de tratamiento, $H_0 \equiv \mu_{a_1} = \mu_{a_2} = \mu_{a_3}$. La prueba de la hipótesis para la fuente de varianza sujetos, aunque para nosotros no tenga interés, también resulta significativa ($F_{(4, 8)} = 3.977$, $p < 0.05$).

Ajuste de los grados de libertad

La técnica tradicional de ajuste que acompaña al análisis de varianza univariado, como ya señalamos anteriormente, y que tiene como objetivo el control de la probabilidad de error de *Tipo I*, se basa en el ajuste de los grados de libertad. Esta estrategia tiene como propósito ajustar los grados de libertad tanto de la variabilidad explicada por el tratamiento como la atribuida al término de error a través de la *reducción* de los mismos, tratando de compensar el sesgo positivo de la prueba *F* cuando no se cumple la asunción de homogeneidad de las varianzas de las diferencias entre pares de observaciones. El investigador estará trabajando entonces con una prueba de la hipótesis bastante más conservadora de lo habitual.

Box (1954a, b) demostró que el error de *Tipo I* supera al establecido a priori si la matriz de varianzas y covarianzas del diseño no cumple el supuesto de homogeneidad. Por ejemplo, si el investigador establece a priori como riesgo de error de *Tipo I* un $\alpha = 0.05$, en realidad se podría estar rechazando falsamente la *hipótesis nula* en el 8% o 10% de las ocasiones, inclusive bastante más.

Para controlar este sesgo, bajo la *hipótesis de nulidad* de no diferencias entre las medidas repetidas, la distribución del estadístico F se puede aproximar a una distribución F_u con menos grados de libertad. El nivel de reducción o ajuste está en dependencia directa del patrón de correlación previsto. El mismo autor derivó un parámetro ε —épsilon— que refleja la desviación de la matriz de varianzas-covarianzas poblacional de la homogeneidad.

El grado, pues, en que la matriz de varianzas-covarianzas se aleja de la *esfericidad* está indicado por el parámetro ε. El rango de valores que puede adoptar el factor correctivo ε oscila entre 1.00 y 0.00 de tal modo que existirá homogeneidad de la matriz de varianzas-covarianzas cuando $\varepsilon = 1$, no modificándose los grados de libertad originales del diseño.

Los trabajos de Box (1954a) demostraron que cuando el supuesto de homogeneidad de las varianzas de las puntuaciones de diferencias no se cumple, la razón F, (MC_A / MC_{AS}), sigue aproximadamente una distribución F_u con los grados de libertad ajustados de la siguiente manera:

$$\textit{Grados de libertad}_{\textbf{NUMERADOR}}: \varepsilon \cdot (a - 1) \qquad [3]$$

$$\textit{Grados de libertad}_{\textbf{DENOMINADOR}}: \varepsilon \cdot (n - 1) \cdot (a - 1) \qquad [4]$$

Una vez conocido el valor del factor correctivo ε, simplemente hay que multiplicarlo por los grados de libertad originales del numerador y del denominador de la razón F y ajustar con ello los resultados para contrastar con las tablas de la distribución teórica de F. Puede ocurrir que aparezcan grados de libertad con decimales en cuyo caso el investigador puede o bien redondear el valor, o bien interpolar o bien encontrar el valor crítico de F para los grados de libertad decimales utilizando un programa de ordenador.

El factor ε constituye un parámetro poblacional desconocido, que puede ser estimado a través de tres alternativas (Maxwell y Delaney, 1990):

- *Ajuste del límite inferior de* Geisser-Greenhouse
- *Ajuste* $\hat{\varepsilon}$ *de* Box
- *Ajuste* $\tilde{\varepsilon}$ *de* Huynh-Feldt

El cálculo del factor correctivo es tedioso pero hoy en día prácticamente todos los paquetes estadísticos, (BMDP, SAS y SPSS), incorporan estos procedimientos. Es conveniente anotar que el ajuste de $\hat{\varepsilon}$ se describe como el procedimiento *Epsilon de* Greenhouse-Geisser seguramente motivado por el hecho de que fueron estos autores los que difundieron el ajuste $\hat{\varepsilon}$ a diseños más complejos (Maxwell y Delaney, 1990).

Ajuste del límite inferior de Geisser y Greenhouse

Geisser y Greenhouse (1958) derivaron el *límite inferior* de los grados de libertad, definiendo su máxima reducción. Dado un determinado diseño de *medidas repetidas*, el mínimo valor posible del factor correctivo épsilon es igual a 1 / (a – 1), considerado su *límite inferior*; valor indicativo de la máxima violación posible del supuesto de *esfericidad*. Cuanto menor sea el valor de ε, más se alejará la matriz del supuesto de *esfericidad* y, como consecuencia, mayor ajuste (reducción) se realizará de los grados de libertad. En el paquete estadístico SPSS el valor del límite inferior desarrollado por Geisser y Greenhouse aparece rotulado como *límite inferior de Epsilon*.

Cuando se lleva a cabo la *máxima reducción* de los grados de libertad (ajuste del *límite inferior*), el valor de la F crítica tenderá a ser superior en comparación al valor que se observaba en las tablas previamente a la multiplicación de los grados de libertad por el valor del factor correctivo. Los grados de libertad de referencia con los que contrastar en tablas habrán disminuido. Así, la prueba estará sesgada en sentido conservador.

Los autores elaboraron una regla de cómputo de los grados de libertad bajo una perspectiva altamente conservadora. Se trata de *fingir* que todos los factores de tratamiento del diseño tienen únicamente dos niveles y aplicar el cálculo de los grados de libertad de la forma usual. Así, cuando el diseño tiene solamente factores de tratamiento de *medidas repetidas*, los grados de libertad más conservadores de todas las pruebas F adoptarán los nuevos valores de 1 para el numerador y de n – 1 para el denominador.

$$\textit{Grados de libertad}_{\text{NUMERADOR}}: \frac{1}{a-1} \cdot (a-1) = \boxed{1} \qquad [5]$$

$$\textit{Grados de libertad}_{\text{DENOMINADOR}}: \frac{1}{a-1} \cdot (n-1) \cdot (a-1) = \boxed{(n-1)} \qquad [6]$$

Si trabajamos con un diseño de medidas repetidas con un factor de tratamiento que tiene dos niveles, la corrección será igual a uno y como consecuencia los grados de libertad originales no se verán afectados (valor del *límite inferior* de 1 / (2 – 1) = 1). En cambio, si el diseño tuviese un factor de tratamiento con cuatro niveles, la corrección no podría ser superior a 0.33 ya que 1 / (4 – 1) = 0.33.

En nuestro ejemplo, un diseño de *medidas repetidas* con un factor de tratamiento con tres niveles, el valor del límite inferior de la corrección es de 0.5 (1 / 3 – 1), siendo los grados de libertad corregidos con la *prueba F conservadora* de 1 para la fuente de varianza explicada o del tratamiento y de 4 para la fuente de varianza no explicada o del error:

$$\textit{Grados de libertad}_{\text{NUMERADOR}}: \frac{1}{2-1} \cdot (2-1) = 1$$

$$\textit{Grados de libertad}_{\text{DENOMINADOR}}: \frac{1}{2-1} (5-1)(2-1) = 4$$

Si aplicamos la prueba *F Conservadora*, reduciendo los grados de libertad como si se hubiese violado la *esfericidad* en su punto máximo, el valor crítico o teórico que utilizamos para nuestro contraste es de $F_{0.05,1,4}$ = de 7.709; valor superior al obtenido cuando trabajábamos con los grados de libertad sin ajustar ($F_{0.05,\ 2,\ 8}$ = 4.459). La interpretación de los resultados nos lleva ahora a aceptar el modelo de la *hipótesis nula* del efecto del tratamiento ($F_e = 5.814 < F_t = 7.709$), concluyendo que las diferencias en el nivel de recuerdo de los sujetos en función del uso de las palabras no son estadísticamente significativas

La prueba conservadora de Geisser y Greenhouse (1958) tiene la ventaja de ser muy simple al utilizar sistemáticamente 1 y $n - 1$ como grados de libertad de la fuente de varianza explicada y no explicada respectivamente. Por ello mismo su principal inconveniente deriva de la condición de ser una prueba altamente conservadora, sacrificando en parte la potencia que los diseños de *medidas repetidas* aportan. Cuando la utilización de las computadoras no era común, el ajuste de los grados de libertad mediante 1 y $n - 1$ era muy habitual. En la actualidad, la presentación de los valores de épsilon en los resultados de los paquetes estadísticos ha hecho que se trabaje más con los grados de libertad ajustados tras la multiplicación por el factor correctivo debido a su naturaleza menos conservadora. Se controla el *Error de Tipo I* y se aumenta la potencia de la prueba, en comparación con el ajuste del *límite inferior* (Collier, Baker, Mandeville y Hayes, 1967).

Ajuste $\hat{\varepsilon}$ de Box

El ajuste $\hat{\varepsilon}$ de Box permite estimar el valor poblacional de ε en función de los datos observados a través de la siguiente expresión

$$\hat{\varepsilon} = \frac{a^2 (\bar{S}_{jj} - \bar{S}_{..})^2}{(a-1)(\sum\sum S_{jj}^2 - (2a \sum \bar{S}_j^2) + (a^2 \bar{S}_{..}^2))} \qquad [7]$$

donde:

- a ≡ número de niveles del *factor de tratamiento* de medidas repetidas (a, b, ab, …)
- $\bar{S}_{jj}$ ≡ media de los elementos de la diagonal principal de la matriz de varianzas-covarianzas (es decir, media de las varianzas)
- $\bar{S}_{.}$ ≡ media de todos los elementos de la matriz de varianzas-covarianzas
- $\sum\sum S_{jj}^2$ ≡ suma al cuadrado de cada elemento de la matriz de varianzas-covarianzas
- $\sum \bar{S}_j^2$ ≡ suma al cuadrado de la media de los elementos de cada fila de la matriz de varianzas-covarianzas

En nuestro ejemplo, (véase datos de la **¡Error! No se encuentra el origen de la referencia.**), el resultado del ajuste $\hat{\varepsilon}$ de Box (etiquetado como *épsilon de Greenhouse-Geisser* en el paquete estadístico SPSS) es el siguiente:

$$\hat{\varepsilon} = \frac{3^2\,(43 - 28.667)^2}{(3-1)\,(13941 - ((2)(3)(3298)) + (3^2\,(28.667^2)))} = \underline{0.59684}$$

....

En el ejemplo el valor de $\hat{\varepsilon}$ es de 0.59684. El contraste de hipótesis tendrá mayor validez si reducimos los grados de libertad del numerador y del denominador al multiplicarlos por el valor del factor correctivo 0.59684. Los grados de libertad originales asociados a la prueba *F*, 2 para el tratamiento y 8 para el error, serán de 1 (2 x 0.59584) y 5 (8 x 0.59584) respectivamente.

El valor crítico utilizado sin ajuste era de $F_{0.05,\,2,\,8} = 4.459$, pasando a ser después de efectuar la corrección de los grados de libertad mediante $\hat{\varepsilon}$ de $F_{0.05,\,1,\,5} = 6.608$, aceptándose el modelo restringido vinculado a la *hipótesis nula* ya que $F_e = 5.814 < F_t = 6.608$.

Ajuste $\hat{\varepsilon}$ de Huynh y Feldt

Posteriormente, Huynh y Feldt (1976) presentaron una nueva estimación del ajuste $\hat{\varepsilon}$ por considerarlo un estimador sesgado ya que bajo la *hipótesis de nulidad*, pequeñisimas desigualdades en las varianzas de diferencias muestrales se traducen siempre en valores del factor correctivo inferiores a la unidad.

Si observamos la formula del ajuste $\hat{\varepsilon}$ de Box, el valor de $\hat{\varepsilon}$ será igual a 1.0 sólo si todas las varianzas de diferencias de los datos son exactamente iguales; en caso contrario, $\hat{\varepsilon}$ será menor a uno. Puede ocurrir que exista homogeneidad, es decir $\varepsilon = 1$, de manera que las varianzas de diferencias de la *población* sean estadísticamente iguales y, sin embargo, las varianzas de diferencias de la *muestra* mostrar algunas diferencias que conducirán a un valor de $\hat{\varepsilon}$ menor a la unidad. Huynh y Feldt (1976) trataron de corregir este problema desarrollando otro estimador del parámetro ε, el ajuste conocido como $\tilde{\varepsilon}$ y denominado por los autores como *estimador de máxima verosimilitud*. De nuevo, el valor poblacional de ε es estimado a partir de los datos muestrales recogidos.

La corrección de los grados de libertad a través del parámetro $\tilde{\varepsilon}$ siempre dará lugar a valores mayores que los obtenidos con la corrección de $\hat{\varepsilon}$, $(\tilde{\varepsilon} \geq \hat{\varepsilon})$, de tal manera que los valores críticos de la *prueba F* con los que comparar los resultados observados o empíricos serán inferiores, excepto cuando $\hat{\varepsilon} = \frac{1}{a-1}$, en cuyo caso $\hat{\varepsilon} = \tilde{\varepsilon}$. Se postula así la naturaleza menos conservadora del parámetro $\tilde{\varepsilon}$ de Huynh y Feldt.

Si se conoce el valor de $\hat{\varepsilon}$ entonces el parámetro de Huynh y Feldt se puede obtener fácilmente. Los autores demostraron que:

$$\tilde{\varepsilon} = \frac{n(a-1)\,\hat{\varepsilon} - 2}{(a-1)\,((n-1-(a-1)\,\hat{\varepsilon}))} \qquad [8]$$

El cómputo de la corrección $\tilde{\varepsilon}$ para los datos del ejercicio que venimos teniendo como referente es el siguiente:

$$\tilde{\varepsilon} = \frac{5(3-1)\ 0.59684 - 2}{(3-1)\ (5-1-((3-1)\ 0.59684))} = \frac{3.9684}{5.61264} = \underline{0.70704}$$

En el ejemplo el valor de $\tilde{\varepsilon}$ es de 0.70704. Ajustando los grados de libertad con el estimador de Huynh y Feldt, los nuevos grados de libertad son de 1 (2×0.70704) para el numerador y 6 (8×0.70704) para el denominador. El valor crítico utilizado ahora es de $F_{0.05,\,1,\,6} = 5.987$, superior al valor tabular sin ajuste ($F_{0.05,\,2,\,8} = 4.459$), pero inferior al valor tabular que utilizamos al trabajar con el ajuste de Box ($F_{0.05,\,1,\,5} = 6.608$). Con este nuevo ajuste y redondeando los grados de libertad, tampoco podemos rechazar la hipótesis de nulidad del efecto del tratamiento ($F_e = 5.814 < F_t = 5.987$) aunque observará el lector la naturaleza menos restrictiva de este ajuste en comparación con el ajuste de Box y, especialmente con los resultados obtenidos con el ajuste del *límite inferior*. Las interpretaciones realizadas ajustando y no ajustando los grados de libertad son dispares. Utilizando los grados de libertad exactos rechazaríamos la hipótesis nula.

Como resumen de la aproximación univariada tradicional presentamos en la ***Tabla 2*** los resultados de las cuatro técnicas analizadas hasta el momento, utilizando los datos del supuesto propuesto anteriormente, ordenadas en términos de la probabilidad de rechazar la *hipótesis nula* (de mayor a menor probabilidad): sin ajuste, ajuste $\tilde{\varepsilon}$ de Huynh-Feldt, ajuste $\hat{\varepsilon}$ de Box y ajuste del *límite inferior* de Geisser-Greenhouse. Todas utilizan la misma prueba estadística, distinguiéndose sólo por los grados de libertad ajustados a la varianza *explicada* y *no explicada* que utilizan para encontrar los valores teóricos de la distribución de la prueba y así comparar con los resultados empíricos. El valor de *p* es exacto y ha sido obtenido con un programa de ordenador.

Tabla 2 *Resumen de las aproximaciones basadas en un modelo de efectos mixtos*

Técnica	Test estadístico	$F_{\text{empírico}}$	Grados de libertad		p
			entre	*error*	
Sin ajuste	$F = \frac{MC_A}{MA_{AS}}$	5.814	2.00	8.00	**.028**
$\tilde{\varepsilon}$ de Huynh-Feldt	$F = \frac{MC_A}{MA_{AS}}$	5.814	1.41	5.66	**.049**
$\hat{\varepsilon}$ de Box	$F = \frac{MC_A}{MA_{AS}}$	5.814	1.19	4.77	**.060**
Límite inferior de Geisser-Greenhouse	$F = \frac{MC_A}{MA_{AS}}$	5.814	1.00	4.00	**.073**

Elección del procedimiento

¿Qué aproximación utilizar cuando aplicamos diseños de *medidas repetidas*? Por supuesto, la técnica tradicional sin ajuste no es adecuada ya que es muy sensible a la violación del supuesto de *esfericidad*. En este preciso momento y lugar nos sentimos en la obligación de hacer una llamada a los investigadores, pues a pesar del riesgo de cometer más error de *Tipo I* del deseado son todavía muchas las publicaciones que utilizan estos diseños y análisis sin realizar ningún tipo de ajuste, descalificando de este modo las conclusiones que a veces tan arduamente se defienden.

Cuando se trata de ajustar los grados de libertad, la elección entre los diversos procedimientos que hemos expuesto se tiene que hacer dependiendo del riesgo que estamos dispuestos a asumir. El ajuste del *límite inferior* de Geisser-Greenhouse tiene la ventaja de la sencillez de su cálculo, siendo su valor independiente de las varianzas y covarianzas de la matriz poblacional; sin embargo, es altamente conservador, constituyendo los ajustes $\hat{\varepsilon}$ y $\tilde{\varepsilon}$ las correcciones más eficaces dado que aumentan la potencia y controlan al mismo tiempo el error de *Tipo I* fijado a priori por el investigador.

Barcikowsky y Robey (1984) únicamente recomiendan el ajuste del *límite inferior* en estudios exploratorios. Es conveniente, pues, analizar el comportamiento de las aproximaciones $\hat{\varepsilon}$ y $\tilde{\varepsilon}$ en relación al riesgo de error de *Tipo I* y de la potencia de la prueba *F*.

Los resultados del trabajo de simulación de Maxwell y Arvey (1982), realizados con un amplio número de medidas repetidas y un tamaño muestral relativamente pequeño en un diseño de *medidas parcialmente repetidas* o diseño *mixto*, aportan información especialmente útil. Como se observa en la ***Tabla 3***, los autores analizan el comportamiento del ajuste de Box y el ajuste de Huynh y Feldt, variando el número de niveles del factor entre-sujetos, el tamaño de la muestra y la matriz de covarianza, manteniendo constante en 13 niveles el factor de medidas repetidas.

Tabla 3 *Puntuaciones medias de* $\hat{\varepsilon}$ *y* $\tilde{\varepsilon}$

		ε = 1.00		ε = 0.74		ε = 0.41	
Factor entre-sujetos		$\hat{\varepsilon}$	$\tilde{\varepsilon}$	$\hat{\varepsilon}$	$\tilde{\varepsilon}$	$\hat{\varepsilon}$	$\tilde{\varepsilon}$
$a = 2$	$n = 15$	0.688	0.986	0.558	0.780	0.366	0.462
	$n = 10$	0.590	0.983	0.492	0.804	0.342	0.487
	$n = 5$	0.393	0.964	0.350	0.847	0.276	0.579
$a = 6$	$n = 8$	0.768	0.999	0.609	0.838	0.379	0.483
	$n = 2$	0.326	0.997	0.298	0.983	0.248	0.833

A tenor de los datos reflejados en la tabla podemos concluir con Maxwell y Arvey (1982) que el ajuste $\tilde{\varepsilon}$ tiende a sobrestimar el valor poblacional de ε, especialmente si la muestra es pequeña. Cuando n decrece, su valor aumenta incrementando el error de *Tipo I*. De hecho, $\tilde{\varepsilon}$ puede alcanzar valores superiores a la unidad. Si esto ocurre, en realidad $\tilde{\varepsilon}$ es igual a 1.0 ya que el parámetro poblacional de épsilon no puede ser superior a la unidad. La validez de su ajuste está mediatizada por el tamaño muestral, aumentando su sesgo positivo con matrices especialmente desviadas de la esfericidad ($\varepsilon \leq 0.85$). Por otro lado, el ajuste $\hat{\varepsilon}$ de Box tiende a infraestimar el valor poblacional de ε, actuando de forma más conservadora, especialmente cuando el tamaño de la muestra es pequeño. El sesgo negativo de $\hat{\varepsilon}$ se acentúa cuando el valor de ε se encuentra próximo a uno ($0.75 < \varepsilon < 1$). Por lo tanto, cuando la matriz de covarianzas está próxima a la esfericidad, el ajuste $\tilde{\varepsilon}$ de Huynh y Feldt mantiene el α en niveles más cercanos al establecido a priori por el investigador que el ajuste $\hat{\varepsilon}$ de Box, cuya distribución es bastante exacta cuando las desviaciones son de moderadas a severas.

Como recomendaciones generales, Maxwell y Delaney (1990) prefieren utilizar el ajuste $\hat{\varepsilon}$ de Box dado que en ocasiones el parámetro $\tilde{\varepsilon}$ fracasa en el control del error de *Tipo I*, presentando un sesgo positivo en la estimación de ε. Trabajos como los de Keselman et al. (1980) y el de Muller y Barton (1989) también recomiendan el ajuste de Box como estimador del parámetro poblacional épsilon.

Conocido el sesgo negativo del ajuste $\hat{\varepsilon}$, Greenhouse y Geisser (1959) proponen una aproximación basada en tres pasos (véase ***Figura 1***). En un primer momento se trata de aplicar la prueba F sin realizar ningún ajuste. Si se mantiene la *hipótesis de nulidad* de no diferencias entre las medias se concluye directamente que las diferencias no son estadísticamente significativas y ya no es necesario pasar al ajuste de los grados de libertad dado que no se modificarán las conclusiones de significatividad estadística. Si se rechaza la *hipótesis nula* de no diferencias, aplicamos directamente la prueba *F conservadora* ajustando al máximo los grados de libertad. Si rechazamos la *hipótesis* *nula*, podemos concluir que existen diferencias estadísticamente significativas y detener aquí el proceso ya que cualquier otro ajuste no modificaría las conclusiones. En cambio, si tras ajustar los grados de libertad mediante la técnica del *límite inferior* mantenemos la *hipótesis nula* es necesario estimar el valor de épsilon mediante $\hat{\varepsilon}$ y ajustar los grados de libertad para comparar adecuadamente con el valor teórico correspondiente.

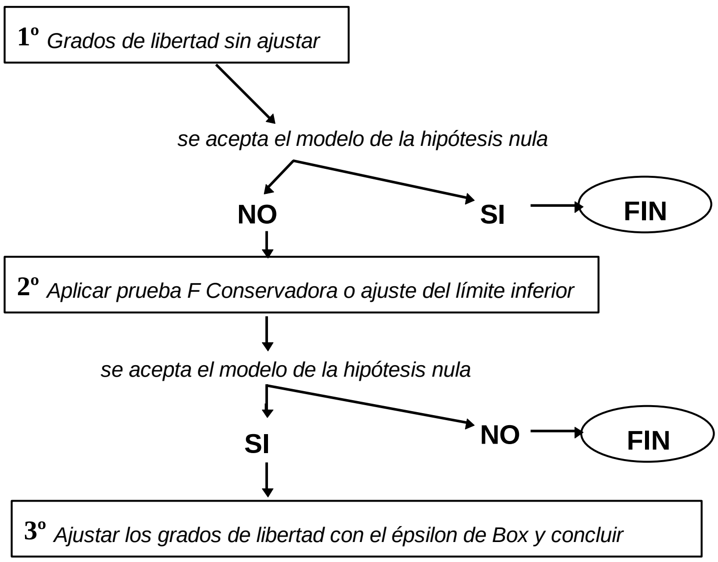

Figura 1 *Procedimiento para el ajuste de los grados de libertad en tres pasos*

Por su parte, Stevens (1992) recomienda realizar una media de los dos estimadores dada la tendencia a infraestimar y sobrestimar que los ajustes $\hat{\varepsilon}$ y $\tilde{\varepsilon}$ realizan del parámetro poblacional ε, respectivamente. Recomendación ya sugerida anteriormente por Maxwell y Arvey (1982).

$$\bar{\varepsilon} = \frac{\hat{\varepsilon} + \tilde{\varepsilon}}{2} \qquad [9]$$

Como comentamos anteriormente, las conclusiones del trabajo de simulación de Keselman et al. (1980) invalidaban la aplicación de la prueba *W* de Mauchly como método de análisis del nivel de esfericidad de la matriz de covarianzas dada su alta sensibilidad a la normalidad. Sin embargo, estos mismos autores también desaconsejan utilizar $\hat{\varepsilon}$ o $\tilde{\varepsilon}$, recomendando la estrategia de los tres pasos de Greenhouse y Geisser (1959) ya que si se mantiene la *hipótesis nula* de no diferencias entre las medias mediante la prueba *F* sin ajuste no sería ya necesario pasar al cómputo de $\hat{\varepsilon}$ o $\tilde{\varepsilon}$.

En resumen, dado que la matriz de covarianzas poblacional es desconocida y que tenemos que estimar el valor poblacional de épsilon, la elección del procedimiento de ajuste de los grados de libertad depende especialmente del nivel de riesgo que estamos dispuestos a asumir. Afortunadamente el comportamiento de los ajuste de $\hat{\varepsilon}$ y $\tilde{\varepsilon}$ no son tan dispares. A medida que *n* aumenta, sus valores se aproximan hasta ser casi idénticos y por ello, las conclusiones que se obtienen suelen ser similares excepto si el tamaño muestral es muy pequeño.

Otra aproximación para el análisis de los diseños de *medidas repetidas* consiste en la utilización de la perspectiva multivariada. El modelo multivariado tiene la ventaja de que no requiere el cumplimiento de la presencia de *esfericidad*. A continuación presentaremos las razones de por qué esto es así junto con los desarrollos analíticos para la obtención de la prueba de la hipótesis a partir del enfoque multivariado.

Diseño de medidas parcialmente repetidas

Los diseños denominados *'Split-Plot'* (*división de los sujetos en grupos*) son ampliamente aplicados en la investigación psicológica configurándose como diseños que combinan el diseño de *medidas repetidas* y el diseño *entre-sujetos*. Suelen nombrarse como *diseños mixtos* o *diseños de medidas parcialmente repetidas* dado que incorporan conjuntamente una variable *entre-sujetos* con al menos otra *intra* o de mediciones repetidas. También suelen ser denominados diseños *factoriales anidados* ya que el 'factor sujeto' se encuentra cruzado con el o los factores *intra* mientras que está anidado dentro del factor o factores *entre*. (El término *diseño mixto* crea cierta confusión y ambigüedad ya que la estadística suele reservar dicho término para aquellos diseños que incluyen factores de efectos aleatorios y factores de efectos fijos).

La ecuación estructural del diseño de *medidas parcialmente repetidas* con un factor *A entre* y un factor *B intra* o de medidas repetidas es la siguiente:

$$H_1 \equiv \mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{S/A} + \mathbf{AB} + \mathbf{BS/A} + \mathbf{E} \qquad [10]$$

La interacción de segundo orden **BS/A** coincide con la *varianza no explicada* o *varianza residual* quedando entonces definida la ecuación estructural como:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{S/A} + \mathbf{AB} + \mathbf{BS/A} \qquad [11]$$

Aunque se consideran tres factores no es posible estimar todas las interacciones existentes entre ellos. La razón es obvia, el factor *A* no está cruzado factorialmente con el denominado factor sujetos, pues los primeros cuatro sujetos pertenecen a a_1 y los cuatro segundos a a_2. De cruzarse ambas variables, siendo así que hay ocho sujetos y *A* tiene dos valores tendríamos 16 observaciones por cada nivel de *B*. Y no es este el caso, disponiendo solamente de 8 observaciones

El factor denominado sujetos *–S/A–* constituye el término de error de la variabilidad asociado al factor *entre-sujetos A*, mientras que la interacción del factor sujeto, que está anidado en el factor *A*, por el factor de medidas repetidas *–BS/A–* es el error del la variabilidad denominada *intra*, a saber, *B* y *AB*. Dado que el *factor sujeto* está anidado en el factor *entre A*, no es posible calcular la interacción entre ambos. Obsérvese que la interacción del factor de medidas repetidas y el factor de medidas no repetidas tiene como término de error o denominador de la prueba *F* la fuente de varianza *BS/A*.

El cálculo de los grados de libertad para cada fuente de varianza sigue las mismas pautas que hasta aquí se han desarrollado. Para los factores principales es el número de niveles del factor menos 1, para la interacción la multiplicación entre los grados de libertad de los factores implicados y para el término de error de la fuente de varianza *entre*, *S/A*, será $a\,(n - 1)$ mientras que para el error de la fuente de varianza *intra* los grados de libertad se obtienen como $a\,(n - 1)\,(b - 1)$. Trabajemos con un nuevo ejemplo.

Supuesto 1

Ocho estudiantes de primer ciclo de Universidad participan voluntariamente en una investigación sobre el estudio del efecto que la longitud de las palabras tiene sobre su recuerdo. Cada estudiante recibió dos listas de palabras con longitud baja y alta que tuvieron que leer para medir su nivel de recuerdo posterior. Además, la mitad del grupo fue sometido a una sesión auditiva de ambas listas de palabras mientras que la otra mitad únicamente realizó su lectura. Los resultados fueron los siguientes.

Tabla 4 *Matriz de resultados y medias marginales*

		b_1 *baja*	b_2 *alta*	$\mathbf{M}_{S/A}$	$\mathbf{M}_A$
	sujeto 1	18	22	*20*	
	sujeto 2	14	18	*16*	
a_1 audición	*sujeto 3*	6	22	*14*	
	sujeto 4	2	26	*14*	*16*
	sujeto 5	22	18	*20*	
a_2 No audición	*sujeto 6*	22	22	*22*	
	sujeto 7	22	30	*26*	
	sujeto 8	22	34	*28*	*24*
	$\mathbf{M}_B$	*16*	*24*		$\mathbf{M}$ = *20*

Estimemos los efectos según la predicción del *modelo completo*. La estimación de los efectos principales sigue las mismas pautas: media de la condición menos la media total.

$$\mathbf{A} = \mathbf{M}_A - \mathbf{M} = \begin{pmatrix} 16\\16\\16\\16\\24\\24\\24\\24\\16\\16\\16\\16\\24\\24\\24\\24 \end{pmatrix} - \begin{pmatrix} 20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20 \end{pmatrix} = \begin{pmatrix} -4\\-4\\-4\\-4\\4\\4\\4\\4\\-4\\-4\\-4\\-4\\4\\4\\4\\4 \end{pmatrix}$$

$$\mathbf{SC_{entre}}\ _A = \mathbf{A}^2 = \mathbf{A}'\mathbf{A} = \boxed{256}$$

El cálculo del efecto del factor sujetos que está anidado en el factor *entre-sujetos* se resuelve mediante la expresión:

$$\mathbf{S/A} = \mathbf{M}_{S/A} - \mathbf{M} - \mathbf{A} = \begin{pmatrix} 20\\16\\14\\14\\20\\22\\26\\28\\20\\16\\14\\14\\20\\22\\26\\28 \end{pmatrix} - \begin{pmatrix} 20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20 \end{pmatrix} - \begin{pmatrix} -4\\-4\\-4\\-4\\4\\4\\4\\4\\-4\\-4\\-4\\-4\\4\\4\\4\\4 \end{pmatrix} = \begin{pmatrix} 4\\0\\-2\\-2\\-4\\-2\\2\\4\\4\\0\\-2\\-2\\-4\\-2\\2\\4 \end{pmatrix}$$

$$\mathbf{SC_{entre}}\ _{S/A} = \mathbf{S/A}^2 = \mathbf{S/A}'\,\mathbf{S/A} = \boxed{224}$$

El efecto del factor de medidas repetidas *B* también es calculado del mismo modo:

$$\mathbf{B} = \mathbf{M}_{\mathrm{B}} - \mathbf{M} = \begin{pmatrix} 16 \\ 16 \\ 16 \\ 16 \\ 16 \\ 16 \\ 16 \\ 16 \\ 24 \\ 24 \\ 24 \\ 24 \\ 24 \\ 24 \\ 24 \\ 24 \end{pmatrix} - \begin{pmatrix} 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \end{pmatrix} = \begin{pmatrix} -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \end{pmatrix}$$

$$\mathbf{SC}_{\mathbf{intra}\ \mathrm{B}} = \mathbf{B^2} = \mathbf{B'}\,\mathbf{B} = \boxed{256}$$

El efecto de interacción entre el factor de medidas repetidas y el factor de medidas no repetidas supone eliminar de la media de interacción la media total y los efectos estimados de los dos factores.

$$\mathbf{AB} = \mathbf{M}_{\mathrm{AB}} - \mathbf{M} - \mathbf{A} - \mathbf{B} =$$

$$= \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 22 \\ 22 \\ 22 \\ 22 \\ 22 \\ 22 \\ 22 \\ 22 \\ 26 \\ 26 \\ 26 \\ 26 \end{pmatrix} - \begin{pmatrix} 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \end{pmatrix} - \begin{pmatrix} -4 \\ -4 \\ -4 \\ -4 \\ 4 \\ 4 \\ 4 \\ 4 \\ -4 \\ -4 \\ -4 \\ -4 \\ 4 \\ 4 \\ 4 \\ 4 \end{pmatrix} - \begin{pmatrix} -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \end{pmatrix} = \begin{pmatrix} -2 \\ -2 \\ -2 \\ -2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ -2 \\ -2 \\ -2 \\ -2 \end{pmatrix}$$

$$\mathbf{SC}_{\mathbf{intra}\ \mathrm{AB}} = \mathbf{AB^2} = \mathbf{AB'}\,\mathbf{AB} = \boxed{64}$$

Como se ha comentado anteriormente, en estos diseños no es posible calcular un efecto de interacción entre el factor sujetos y el factor de medidas no repetidas ya que los sujetos están anidados y no cruzados con dicho factor. Sin embargo, sí es posible obtener la interacción del factor sujetos con el factor de medidas repetidas ya que ambos están cruzados. La estimación de este efecto de interacción implica eliminar de la media de interacción la media total, el efecto del factor de medidas repetidas, el efecto del factor sujetos y también el efecto de interacción entre los factores de tratamiento *AB*.

$$\mathbf{BS/A} = \mathbf{M}_{BS/A} - \mathbf{M} - \mathbf{A} - \mathbf{S/A} - \mathbf{B} - \mathbf{AB} =$$

$$\begin{pmatrix} 18\\14\\6\\2\\22\\22\\22\\22\\22\\18\\22\\26\\18\\22\\30\\34 \end{pmatrix} - \begin{pmatrix} 20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20\\20 \end{pmatrix} - \begin{pmatrix} -4\\-4\\-4\\-4\\4\\4\\4\\4\\-4\\-4\\-4\\-4\\4\\4\\4\\4 \end{pmatrix} - \begin{pmatrix} 4\\0\\-2\\-2\\-4\\-2\\2\\4\\4\\0\\-2\\-2\\-4\\-2\\2\\4 \end{pmatrix} - \begin{pmatrix} -4\\-4\\-4\\-4\\-4\\-4\\-4\\-4\\4\\4\\4\\4\\4\\4\\4\\4 \end{pmatrix} - \begin{pmatrix} -2\\-2\\-2\\-2\\2\\2\\2\\2\\2\\2\\2\\2\\-2\\-2\\-2\\-2 \end{pmatrix} = \begin{pmatrix} 4\\4\\-2\\-6\\4\\2\\-2\\-4\\-4\\-4\\2\\6\\-4\\-2\\2\\2 \end{pmatrix}$$

$$\mathbf{SC_{intra}}\ _{BS/A} = \mathbf{BS/A}^2 = \mathbf{BS/A}'\,\mathbf{BS/A} = \boxed{128}$$

La suma de cada una de estos efectos al cuadrado permite resolver el análisis de la varianza (véase ***Tabla 5***)

Tabla 5 *Análisis de la varianza (factor A 'entre' / factor B 'intra')*

Fuente	*SC*	*GL*	*MC*	*Razón F*	*p*
A	256	1	256.000	6.857	< 0.05
Sujetos/A	224	6	37.333		
B	256	1	256.000	12.000	< 0.05
$A \times B$	64	1	64.000	3.000	> 0.05
$B \times S/A$	128	6	21.333		
Total	928	15			

Los resultados del análisis de la varianza realizado señalan la ausencia de un efecto de interacción entre las dos variables estadísticamente significativo, pasando por lo tanto a observar a continuación el comportamiento de los factores principales. Respecto a ellos, se rechaza la hipótesis de la nulidad del efecto del tratamiento, concluyendo que tanto la longitud de las palabras como la naturaleza de la presentación de las mismas afectan de forma significativa al nivel de recuerdo, no detectándose un patrón interaccional entre ambas variables. En concreto, los sujetos que reciben las palabras con una presentación sólo auditiva, ($\overline{a}_1 = 16$), tienen un nivel de recuerdo significativamente menor que aquellos que reciben la estimulación tanto visual como auditiva ($\overline{a}_2 = 24$). Además, las palabras de menor longitud también son recordadas significativamente mejor ($\overline{b}_1 = 16$), que las palabras de mayor longitud ($\overline{b}_2 = 24$).

Supuestos del diseño de medidas parcialmente repetidas

En el diseño de *medidas parcialmente repetidas* se combinan aspectos de los diseños *entre-sujetos* y elementos de los diseños *intra-sujetos*, requiriendo la presencia de los supuestos mantenidos por ambos tipos de diseños tales como normalidad, homogeneidad de las varianzas e independencia de las observaciones. Además, las fuentes de varianza *intra* también deben sustentar la asunción de homogeneidad de las varianzas de las puntuaciones de diferencia.

El ajuste de los grados de libertad de las fuentes de varianza *intra* es necesario para compensar el sesgo positivo que se produce al aplicar la prueba estadística F por la razones ya expuestas anteriormente. El ajuste se realiza únicamente sobre las fuentes de varianza de medidas repetidas, no siendo necesario en las fuentes de varianza *entre*.

Supongamos que tenemos un diseño de *medidas parcialmente repetidas* con dos factores de tratamiento, el factor A de medidas repetidas con dos niveles y el factor B de medidas no repetidas con tres niveles y 25 observaciones. Si aplicamos el ajuste del *límite inferior*, los grados de libertad del efecto principal *entre* B no sufrirán ningún tipo de ajuste, siendo el original de 2, $(a - 1)$, para la varianza explicada, manteniendo 72 grados del libertad para su término de error $(b\,(n - 1))$. Respecto a la interacción de un factor *entre* - *sujetos* con un factor intra - *sujetos*, si el factor de medidas repetidas tiene únicamente dos niveles, no se produce ningún cambio o ajuste. Así la interacción $A \times B$ tiene asociado 2 grados de libertad para el numerador $((a-1)(b-1))$ y 72 para el denominador $(b(a-1)(n-1))$. El ajuste de los grados de libertad se realiza en los factores de medidas repetidas que tengan más de dos niveles y en aquellas interacciones del factor *entre* - *sujetos* con factores de medidas repetidas que originariamente tienen más de dos niveles.

El ajuste de los grados de libertad puede realizarse a través de las tres técnicas ya conocidas: el ajuste del *límite inferior*, el ajuste de $\hat{\varepsilon}$ Box (también conocido como ajuste $\hat{\varepsilon}$ de Geisser-Greenhouse) y el ajuste $\tilde{\varepsilon}$ de Huynh y Feldt. Bajo la perspectiva del *límite inferior* simplemente se trata de *fingir* que el factor de medidas repetidas tiene sólo dos niveles y, a partir de ahí calcular los grados de libertad ajustados para las fuentes de varianza *intra*.

Es importante tener en cuenta que en los diseños de *medidas parcialmente repetidas* sólo existirá un valor de $\hat{\varepsilon}$ y un valor de $\tilde{\varepsilon}$ ya que todas las fuentes de varianza *intra* utilizan el mismo término de error, a saber, la Media Cuadrática $_{B \times S/A}$, no necesitando un ajuste diferente tal y como se realiza en los diseños donde todos los factores de tratamiento son de medidas repetidas y se trabaja con diferentes términos de error. Es decir, la fuente de varianza relacionada con el efecto del factor principal de medidas repetidas y la de la interacción entre dicho factor y el de medidas no repetidas tendrán un único valor de $\hat{\varepsilon}$ y un único valor de $\tilde{\varepsilon}$ ya que sólo existe un término de error para las fuentes de varianza *intra*.