© International Journal of Clinical and Health Psychology

ISSN 1697-2600
2007, Vol. 7, Nº 1, pp. 153-175

Cómo elegir la mejor prueba estadística para analizar un diseño de medidas repetidas[1]

Paula Fernández[2] *(Universidad de Oviedo, España)*,
Pablo Livacic-Rojas (*Universidad de Santiago de Chile, Chile)* y
Guillermo Vallejo *(Universidad de Oviedo, España)*

(Recibido 30 de mayo 2005/ Received May 30, 2005)
(Aceptado 8 de abril 2006/ Accepted April 8, 2006)

RESUMEN. La potencia y robustez de los procedimientos estadísticos para analizar los efectos en los diseños de medidas repetidas están en función de la satisfacción de los supuestos asociados al análisis, en especial, el supuesto de esfericidad y de homogeneidad de las matrices de covarianza. Desafortunadamente, la violación de estos supuestos es habitual en los datos de las investigaciones aplicadas educativas y psicológicas. En este artículo teórico revisamos las competencias de varios estadísticos con respecto al error de Tipo I y la potencia obtenidos por diferentes autores mediante estudios de simulación Monte Carlo. También realizamos una investigación Monte Carlo *ad hoc* para ejemplificar la cuantía del error de Tipo I en los efectos intra-sujeto en un diseño split-plot de medidas repetidas. Examinando todo lo anterior advertimos que diferentes métodos de análisis son apropiados en diferentes situaciones. Concluimos aportando recomendaciones para el análisis de estos diseños en función de la violación o no de las asunciones subyacentes.

PALABRAS CLAVE. Medidas repetidas. Investigación aplicada. Procedimientos estadísticos. Violación de asunciones. Recomendaciones. Estudio teórico.

[1] Este trabajo ha sido realizado con la ayuda concedida por el MEC-SEJ2005-01883.
[2] Correspondencia: Universidad de Oviedo. Departamento de Psicología. Plaza Feijóo, s/n. 33003 Oviedo (España). E-Mail: paula@uniovi.es

ABSTRACT. The power and robustness of the statistical procedures for the analysis of the repeated measures designs' effects are sensitive to the not satisfaction of the analysis assumptions, especially, the assumption of sphericity and homogeneity of the covariance matrices. Unfortunately, the violation of these assumptions is habitual in the data of the educational and psychological applied research. In this theoretical study we examined the performance of the several statistics procedures with regard to the Type I error and the power obtained for several authors through Monte Carlo methods. As well a Monte Carlo investigation was conduced *ad hoc* for exemplify the quantity of the Type I error in the within-subject effects in a split-plot repeated measures design. Examinee all previous we advise that different methods of analysis are appropriate in different situations and we conclude bringing recommendations for the analysis of these designs in order of the violation or not of the derivational assumptions.

KEYWORDS. Repeated measures. Applied research. Statistical procedures. Violation of assumptions. Recommendations. Theoretical study.

RESUMO. A potência e robustez dos procedimentos estatísticos para analisar os efeitos em planos de medidas repetidas estão em função da satisfação dos pressupostos associados da análise, em especial, o pressuposto da esfericidade e da homogeneidade das matrizes de covariância. Infelizmente, a violação destes pressupostos é habitual nos dados das investigações educativas e psicológicas aplicadas. Neste artigo teórico revemos a realização de vários procedimentos estatisticos em relação ao erro de Tipo I e a potência obtidos por diferentes autores através de estudos de simulação Monte Carlo. Também realizamos uma investigação Monte Carlo ad hoc para exemplificar a quantidade do erro Tipo I nos efeitos intra-sujeito num plano split-plot de medidas repetidas. Analisando tudo isto, advertimos que métodos de análise são apropriados em diferentes situações. Concluimos contribuindo com recomendações para a análise destes planos em função da violação ou não das situações adjacentes.

PALAVRAS CHAVE. Medidas repetidas. Investigação aplicada. Procedimentos estatísticos. Violação de pressupostos. Recomendações. Estudo teórico.

Introducción

La popularidad e importancia de los diseños de investigación en los cuales se efectúan mediciones repetidas de cada uno de los sujetos se justifica por ser los únicos que permiten describir, pronosticar y explicar los procesos (biológicos, psicológicos, sociales, etc) que se producen como consecuencia del paso del tiempo. Este artículo viene motivado precisamente porque esta apreciación es tan poco original y aislada como sólidamente confirmada (Keselman, Algina y Kowalchuk, 2001). Así pues, atendiendo al vigor que estos diseños ostentan, nuestra intención no es dirigirnos a los estudiosos del diseño y del método sino a los consumidores del mismo en la investigación aplicada.

El diseño es un modo de recogida de datos. Estos que aquí nos ocupan, en función de la naturaleza de la/s variable/s intra-sujeto (variables independientes-diferentes tra-

tamientos-, variables independientes asignadas -tiempo, edad-), de la muestra (sea uno o varios grupos definidos por los niveles de la/s variable/s entre-sujetos -variables independientes o de clasificación-), de la regla de asignación (si los niveles de la/s variable/s intra y entre-sujetos han sido elegidos y asignados de modo aleatorio o no) y de la cantidad de variables dependientes, se despliegan en un amplio abanico de diseños particulares. No obstante, la estructura más habitualmente utilizada es aquella que tiene un único factor intra-sujeto y un único factor entre-sujetos. Los sujetos (i = 1,..., n_j, $\Sigma\ n_j = N$), clasificados en función de, o asignados aleatoriamente a los niveles del factor entre-sujetos (j = 1,..., p) son observados y medidos en todos los niveles del factor intra-sujeto (k = 1,..., q). Dos ejemplos son los que siguen:

Palmero, Breva, Diago, Díez y García (2002) realizan un experimento para explorar el papel que juegan los patrones emocionales Tipo A y Tipo B (variable entre-sujetos de clasificación) en la activación,variabilidad y recuperación cardiaca (variables dependientes), en tres situaciones que representa la variable independiente intra-sujeto (habituación, tarea –presentación de estímulos estresantes– y recuperación).

Méndez, Orgiles y Espada (2004) llevan a cabo otro experimento para probar la eficiencia del programa de escenificaciones emotivas para el tratamiento de la fobia a la oscuridad. 45 niños fueron asignados aleatoriamente a dos condiciones experimentales -grupo de tratamiento mediante escenificaciones emotivas o grupo de control en lista de espera- (variable entre-sujetos independiente). Sendos grupos fueron observados en cuatro momentos temporales –pretest, postest, seguimiento a los 3 meses y seguimiento a los 6 meses– (variable independiente intra-sujeto asignada).

Estos diseños tienen ventajas de considerable valor práctico, estadístico y sustantivo. A saber:

a) Valor práctico: suponen una alternativa a los diseños de grupos cuando se dispone de pocas unidades muestrales para realizar la investigación porque todos los sujetos son observados en todas las condiciones experimentales.
b) Valor estadístico: una de las exigencias en los diseños trasversales es que los grupos sean homogéneos. En principio, la homogeneidad quedaría garantizada configurando los grupos al azar. Sin embargo, si la muestra es pequeña (es habitual que cuanto menor sea el tamaño de la muestra se advierta una mayor variabilidad) o aun siendo grande es demasiado heterogénea, la aleatorización no va a resultar efectiva generándose entonces una gran varianza del error. Esta circunstancia supone una fuerte amenaza a la validez de la conclusión estadística de la investigación. Por su parte, los diseños de medidas repetidas permiten eliminar mucha variabilidad del error experimental al servir cada sujeto como su propio control; de ahí que sean más potentes que los diseños transversales con el mismo número de mediciones.
c) Valor sustantivo: anteriormente hemos señalado que no siempre las respuestas de cada uno de los sujetos del estudio son obtenidas bajo condiciones experimentales, sino que cabe también la posibilidad de que se tomen registros en función de múltiples puntos temporales. Cuando los sujetos no reciben un tratamiento diferente en cada punto del tiempo, estos diseños son los únicos

que nos permiten obtener información concerniente a los patrones de cambio individual.

Los datos así recogidos se analizan habitualmente utilizando el AVAR (Análisis Univariado de la Varianza) o el AMVAR (Análisis Multivariado de la Varianza). Ambas son alternativas acertadas si se satisfacen los supuestos del modelo elegido. En común comparten los siguientes: normalidad conjunta multivariada, independencia entre los vectores de observaciones de las diferentes unidades experimentales y homogeneidad de las matrices de covarianza o de dispersión (Σ). De modo particular, el modelo Univariado (AVAR) también requiere que la estructura de la matriz Σ sea esférica. ¿Qué es esfericidad? Existe esfericidad cuando las varianzas correspondientes a las diferencias entre las distintas ocasiones de medida son iguales o bien cuando tenemos varianzas iguales y covarianzas iguales. A las matrices que representan la primera condición se las llama Tipo H o de Huynh-Feldt (Huynh y Feldt, 1970). A las que representan la segunda condición se las denomina Tipo S o de Simetría Combinada y son un caso particular de las matrices Tipo H. A partir de los trabajos de Huynh (1978) y Mendoza (1980), el cumplimiento conjunto de los supuestos de homogeneidad de las matrices de covarianza y esfericidad de las mismas ha sido referido como esfericidad multimuestral. El valor de la esfericidad (ε) está comprendido entre $1/(q\text{-}1)$ y 1. Cuanto más se desvíe de 1, más desviación de la esfericidad existe. Cuando existe esfericidad el valor de ε es 1 y la matriz se dice que es esférica. El modelo Multivariado (AMVAR) no requiere este supuesto, pero sí es imperativo para que se pueda calcular que el número de sujetos dentro de cada grupo sea mayor que el número de medidas repetidas. A pesar de las ventajas reseñadas anteriormente (ver también Davis, 2002; Finney, 1990) estos diseños no están exentos de dificultades. En primer lugar, su estructura favorece la aparición de efectos residuales (distinto grado de correlación entre las medidas repetidas) y, por ende, la violación del supuesto de esfericidad. En segundo lugar, en ciertos ambientes aplicados el tamaño de los grupos es muy dispar, lo que propicia que las matrices de dispersión sean heterogéneas y que los datos no se distribuyan normalmente (Micceri, 1989; Sawilowsky y Blair, 1992; Wilcox, 2001). Por último, el investigador frecuentemente no ejerce un control estricto sobre las circunstancias en las que se registran los datos, lo que le ocasiona pérdida de sujetos y de ocasiones de medida. Así las cosas, la validez de la conclusión estadística (esto es, la posibilidad de cometer un mayor error de Tipo I que el supuesto a priori -α- o de Tipo II -escasa potencia de prueba-) resulta comprometida en el AVAR y en el AMVAR en el primer y segundo caso respectivamente y para ambos en los restantes.

Para solventar estos problemas se han desarrollado técnicas alternativas de análisis, clásicas unas, de reciente desarrollo otras, pero todas ellas con un común denominador, garantizar la robustez y la potencia en la prueba de las hipótesis de los efectos del diseño. Como ninguna técnica de análisis es uniformemente mejor que el resto en todos los casos, conviene saber cual de ellas es más acertado utilizar en función de las características que una matriz de datos concreta tiene. Así pues, los objetivos de este trabajo teórico (Montero y León, 2005) son:

– Exponer diferentes técnicas de análisis, qué problema tratan de acometer y con qué eficacia. Ser utilizadas con frecuencia y estar disponibles en los paquetes

estadísticos estándar (SAS, SPSS) son los dos avales que justifican la decisión de los estadísticos que exponemos. Para ello, además de apoyarnos en las investigaciones realizadas sobre estas técnicas, y con la intención de ser pedagógicos en la exposición, hemos efectuado experimentos de simulación Monte Carlo *ad hoc* (5000 replicaciones) con datos normalmente distribuidos ejemplificando en cada uno de ellos el comportamiento con respecto al error de Tipo I de las diferentes técnicas en varias situaciones que se irán relatando.

– De otra parte, destilar conclusiones con ánimo de proporcionar a los investigadores de la psicología aplicada unas líneas básicas acerca de qué técnica de análisis utilizar en una situación determinada, además de exhortar al esmero tanto en la planificación de la investigación como en la exposición de los resultados.

Características de las simulaciones efectuadas

Todas las simulaciones se llevan a cabo para el modelo no aditivo (condición de Ho verdadera para el tratamiento intra-sujeto y la interacción del tratamiento intra-sujeto con el tratamiento entre-sujetos), de un diseño de medidas repetidas factorial mixto -el mismo que subyacía en los dos ejemplos descritos anteriormente- con una variable entre-sujetos de tres niveles y una variable intra-sujeto de cuatro niveles para un nivel de significación $\alpha = 0{,}05$. El tamaño total de la muestra (N) fue 30 salvo en alguna situación en que fue 60. La decisión de utilizar $N = 30$ fue porque para tres grupos y cuatro medidas repetidas resulta un tamaño por grupo de 10 sujetos (teniendo el mismo número de sujetos por grupo) que es idóneo para que tanto el AVAR como el AMVAR funcionen bien cuando la distribución de los datos satisface las asunciones que sendos estadísticos requieren.

Las medidas empíricas de la probabilidad de cometer errores Tipo I se obtuvieron tabulando el número de veces que cada estadístico excede el valor crítico (α) y dividiendo por el número de replicaciones efectuadas. Se tomó como criterio de robustez $\alpha \pm 3$ errores estándar (ES). En nuestro caso $\alpha \pm 3$ ES constituye el intervalo (0,04-0,06). Así las cosas, el estadístico cuyo error de Tipo I sea inferior a 0,04 será considerado conservador, si excede de 0,06 será liberal y si está dentro del intervalo tendrá un comportamiento robusto.

Técnicas de análisis: pruebas univariadas

Vamos a suponer que tenemos que analizar los datos que han sido recogidos de las investigaciones que en la introducción se pusieron de ejemplo. La ecuación lineal univariada que subyace al modelo mixto del AVAR con dos factores de efectos fijos (la variable A intra-sujeto y la variable B entre-sujetos) y uno aleatorio (la variable π sujetos) es la que sigue:

$$y_{ijk} = \mu + \alpha_j + \beta_k + \pi_{i(j)} + (\alpha\beta)_{jk} + (\beta\pi)_{ki(j)} + \varepsilon_{ijk} \qquad (1)$$

donde, y_{ijk} representa la medida en *k* (nivel de la variable intra-sujeto) dentro del grupo *j* (nivel de la variable entre-sujetos) del sujeto *i* (uno de los sujetos del conjunto *N*); μ es la media general; α_j es el efecto del *j*th nivel de la variable entre-sujetos; β_k es el efecto del *k*th nivel de la variable intra-sujeto; $\pi_{i(j)}$ es el efecto del *j*th sujeto medido en el *j*th nivel de la variable entre-sujetos; $(\alpha\beta)_{jk}$ es el efecto de la interacción entre la variable intra y entre-sujetos; $(\beta\pi)_{ki(j)}$ es el efecto de la interacción de la *k*th medida con el *i*th sujeto dentro del *j*th nivel de la variable entre-sujetos y ε_{ijk} es el componente de error aleatorio.

Tres son las hipótesis nulas que podemos poner a prueba, las referidas a cada una de las fuentes de variación o efectos, esto es, la del tratamiento intra-sujeto (variable A), la del tratamiento ente-sujetos (variable B) y la de la interacción entre ambas (AxB).

Los estadísticos asociados con los efectos se distribuyen de acuerdo a la *F* ordinaria (Scheffé, 1956) con los *gl* siguientes. Formalmente sería:

$$F_B = CM_B / CM_{B\times S/A} \sim \mathrm{F}[\alpha;(q-1),(N-p)(q-1)], \quad (2)$$
$$F_{A\times B} = CM_{A\times B} / CM_{B\times S/A} \sim \mathrm{F}[\alpha;(p-1)(q-1),(N-p)(q-1)].$$
$$F_A = CM_A / CM_{S/A} \sim \mathrm{F}[\alpha;(p-1),(N-p)],$$

donde ~ denota que la *F* empírica "se distribuye como" la *F* teórica.

Si se cumplen los supuestos de normalidad conjunta multivariada, independencia entre los vectores de observaciones de las diferentes unidades experimentales, homogeneidad de las matrices de dispersión y esfericidad, el modelo univariado de la varianza (AVAR) es robusto para poner a prueba las respectivas hipótesis nulas sea o no el diseño balanceado, es decir, los valores *p* (probabilidad asociada a los datos asumiendo la Ho verdadera) constituyen un reflejo exacto de la probabilidad de rechazar las hipótesis nulas cuando de hecho son verdaderas, incluso es robusto frente a la ausencia de normalidad (Keselman, Lix y Keselman, 1996). También es uniformemente más potente que cualquier otro enfoque para detectar los efectos de los tratamientos.

Ahora bien, como anteriormente apuntamos, la falta de equilibrio de los grupos además de ser frecuente generalmente va acompañada de una ausencia de'homogeneidad entre las matrices de dispersión (Kazdin, 2001; Keselman *et al*., 1996; Kowalchuck, Lix y Keselman, 1996; Vallejo *et al*., 2002). En esta situación el AVAR abandona la robustez y las hipótesis nulas pueden ser falsamente rechazadas (comportamiento liberal) o falsamente retenidas (comportamiento conservador) en función del tipo de relación existente entre el tamaño de los grupos y el tamaño de las matrices de dispersión. Cuando la relación es positiva (a mayor tamaño de grupo mayores varianzas) se vuelve muy conservador. Cuando la relación es negativa (a mayor tamaño de grupo menores son las varianzas) se vuelve sustancialmente liberal. En la Tabla 1 se muestra el comportamiento del AVAR cuando se satisface el supuesto de esfericidad tanto bajo homogeneidad de las matrices de dispersión como de heterogeneidad, y ambas para un diseño balanceado y no balanceado. Las matrices esféricas utilizadas en nuestra simulación son de simetría combinada (en este caso, pues, ε = 1). Nos resta añadir además

de lo anteriormente comentado sobre ellas, que en estas matrices la correlación entre las diferentes medidas intra-sujeto es constante, esto es, que las variables aleatorias del modelo están igualmente correlacionadas para todos los pares de observaciones de cada sujeto.

TABLA 1. Error de Tipo I para los estadísticos: AVAR y AVAR con los grados de libertad corregidos mediante Greenhouse y Geisser (1959) y Huynt y Feldt (1976) con la corrección de Lecoutre (1991). $\alpha = 0,05$; $N = 30$.

Σ = Tipo S (ε =1)		$\Sigma_1 = \Sigma_2 = \Sigma_1$			$\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$	
V.Entre	AVAR	0,0508[1]	0,0532[2]	**0,0670[1]**	**0,0242[3]**	**0,1488[4]**
V.Intra	AVAR	0,0502[1]	0,0511[2]	0,0567[1]	0,0481[3]	**0,0688[4]**
Interacción	AVAR	0,0503[1]	0,0513[2]	**0,0700[1]**	**0,0154[3]**	**0,2271[4]**
Σ = Tipo NE (ε =0,50)		$\Sigma_1 = \Sigma_2 = \Sigma_1$			$\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$	
V.Entre	AVAR	0,0486[1]	0,0497[2]	**0,0667[1]**	**0,0242[3]**	**0,1529[4]**
V.Intra	AVAR	**0,0976[1]**	**0,1000[2]**	**0,1099[1]**	**0,0986[3]**	**0,0866[4]**
Interacción	AVAR	**0,1167[1]**	**0,1180[2]**	**0,1283[1]**	**0,0602[3]**	**0,2469[4]**
V.Entre	G-G	0,0499[1]	0,0518[2]	**0,0651[1]**	**0,0261[3]**	**0,1455[4]**
V.Intra	G-G	0,0489[1]	0,0532[2]	0,0519[1]	0,0500[3]	**0,0631[4]**
Interacción	G-G	0,0418[1]	0,0553[2]	**0,0667[1]**	**0,0259[3]**	**0,1554[4]**
V.Entre	H-F.L	0,0497[1]	0,0515[2]	**0,0652[1]**	**0,0271[3]**	**0,1462[4]**
V.Intra	H-F.L	0,0541[1]	0,0523[2]	0,0548[1]	0,0572[3]	**0,0642[4]**
Interacción	H-F.L	0,0508[1]	0,0576[2]	**0,0695[1]**	**0,0262[3]**	**0,1559[4]**

Nota. AVAR=Análisis Univariado de la Varianza estándar; G-G= AVAR con los *gl* corregidos mediante Greenhouse y Geisser (1959); H-F.L= AVAR con los *gl* corregidos mediante Huynt y Feldt (1976) con la corrección de Lecoutre (1991). Tipo S= Matriz de Simetría Combinada; Tipo NE= Matriz No Estructurada; $\Sigma_1 = \Sigma_2 = \Sigma_1$=Matrices de dispersión homogéneas; $\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$=Matrices de dispersión heterogéneas; 1= apareamiento nulo en un diseño balanceado (10-10-10) –sujetos en cada uno de los tres grupos-; 2= apareamiento nulo en un diseño no balanceado (6-10-14); 3= diseño no balanceado apareamiento positivo (6-10-14); 4= diseño no balanceado apareamiento negativo (14-10-6); bandas de robustez ±3 ES (0,04-0,06), en negrita se muestra el comportamiento no robusto sea conservador o liberal.

En la Tabla 1 podemos observar que cuando las matrices S son homogéneas el AVAR es robusto para las tres fuentes de variación sea el diseño o no balanceado. Sin embargo, advertimos que cuando las varianzas son heterogéneas las variables intra y entre-sujetos tienen un comportamiento liberal cuando el tamaño de los grupos es igual. Cuando el tamaño de los grupos es diferente y el apareamiento es negativo las tres fuentes de variación tienen un comportamiento liberal, siendo para la variable entre-sujetos y para la interacción extremadamente liberal. Si el apareamiento es positivo la variable entre-sujetos y la interacción se manifiestan muy conservadoras.

Si se incumple el supuesto de esfericidad, con independencia de que el diseño esté o no equilibrado y las matrices de dispersión sean homogéneas o no, lo usual es que la hipótesis nula sea falsamente rechazada más veces de lo debido, incrementándose la liberalidad conforme las matrices de covarianza se desvían del patrón de esfericidad requerido (Collier, Baker, Mandeville y Hayes, 1967) y también conforme incrementa el grado de heterogeneidad, en mayor medida cuando el apareamiento es negativo. Las matrices que no satisfacen el criterio de esfericidad pueden obedecer a diferentes es-

tructuras. La utilizada en nuestra simulación es No Estructurada (NE). Una matriz NE es una matriz simétrica donde tanto las varianzas como las covarianzas varían sin ninguna estructura definida y la correlación entre las diferentes medidas intra-sujeto es arbitraria, careciendo por ello de esfericidad. En nuestra simulación el valor de ε =0,50.

En la Tabla 1 también se expone el comportamiento del AVAR cuando las matrices de dispersión no satisfacen el supuesto de esfericidad, tanto bajo homogeneidad como de heterogeneidad, y ambas para un diseño balanceado y no balanceado. En ella podemos observar que, con excepción del error de Tipo I para la variable entre-sujetos cuando las matrices son homogéneas, el resto se encuentran alteradas, su comportamiento es muy liberal tanto para la variable intra-sujeto como para la interacción.

Terminamos de ver cómo la validez de la conclusión estadística del AVAR ha resultado mermada cuando las matrices Σ son heterogéneas, y cómo todavía resulta más afectada cuando además hay ausencia de esfericidad. ¿Hay algún modo de corregir esto? La respuesta es sí. El investigador dispone de varias alternativas univariadas que dividimos en tres bloques: Estadísticos orientados a corregir los valores críticos de la *F* univariada, el enfoque del Modelo Lineal Mixto (MLM) y el modelo bootstrap-*F*.

Muchos han sido los ajustes que se han propuesto para corregir la liberalidad de la tradicional prueba *F* (Fernández, 1995; Quintana y Maxwell, 1994). Tres son los que pasamos a exponer: prueba de Greenhouse y Geisser (1959), prueba de Huynh y Feldt (1976) y el enfoque de la Aproximación General Mejorada (AGM) Huynh (1978). Todos estos procedimientos implican modificar los *gl* para la variable intra-sujeto y para la interacción multiplicándolos por un valor ε (valor que, como ya señalamos, indica la desviación de la esfericidad de la matriz de dispersión Σ).

Pruebas univariadas que asumen la igualdad de las matrices de dispersión

–Prueba de Greenhouse y Geisser (1959).

Basándose en la premisa de que el ajuste mediante el límite inferior de ε que es 1/(*q*-1) para un diseño con sólo una variable intra-sujeto propuesto inicialmente por Geisser y Greenhouse (1958) resulta en exceso conservador, estos autores sugieren modificar los *gl* del numerador y denominador de las razones *F* para el tratamiento intra-sujeto y la interacción utilizando el estimador muestral $\hat{\varepsilon}$ del parámetro de esfericidad ε desarrollado por Box (1954) del siguiente modo:

$$F_B = CM_B / CM_{B\times S/A} \approx \mathrm{F}[\alpha;(q-1)\hat{\varepsilon},(N-p)(q-1)\hat{\varepsilon}], \qquad (3)$$
$$F_{A\times B} = CM_{A\times B} / CM_{B\times S/A} \approx \mathrm{F}[\alpha;(p-1)(q-1)\hat{\varepsilon},(N-p)(q-1)\hat{\varepsilon}].$$

donde $\approx$ denota que la *F* empírica "se distribuye aproximadamente" como la *F* teórica y

$$\hat{\varepsilon} = \frac{p^2(\overline{\sigma}_{jj} - \overline{\sigma}_{..})^2}{(p-1)\left(\sum_{i=1}^{n}\sum_{j=1}^{p}\sigma_{jj'}^2 - 2p\sum_{j=1}^{p}\overline{\sigma}_{j.}^2 + p^2\overline{\sigma}^2{}_{..}\right)} \qquad (4)$$

donde $\bar{\sigma}_{jj}$ es la media de los elementos de la diagonal principal de la matriz de dispersión Σ; $\sigma_{..}$ es la media de todos los elementos de la matriz Σ; $\bar{\sigma}_{j.}$ es la media de la fila *j* de Σ y $\sum_{i=1}^{n} \sum_{j=1}^{p} \sigma_{jj}^2$ es la suma de cada uno de los elementos de la matriz elevada al cuadrado.

– Prueba de Huynh y Feldt (1976).

Cuando la violación de la esfericidad es ligera ($\varepsilon \geq 0{,}75$) diversos autores han observado que $\hat{\varepsilon}$ puede subestimar el verdadero valor de ε. Para solventar este comportamiento conservador Huynh y Feldt (1976) propusieron una nueva estimación del parámetro de esfericidad a partir del valor de $\hat{\varepsilon}$ calculando los valores *F* como sigue

$$F_B = CM_B / CM_{B\times S/A} \approx \mathrm{F}[\alpha;(q-1)\tilde{\varepsilon}, (n-p)(q-1)\tilde{\varepsilon}], \qquad (5)$$
$$F_{A\times B} = CM_{A\times B} / CM_{B\times S/A} \approx \mathrm{F}[\alpha;(p-1)(q-1)\tilde{\varepsilon}, (N-p)(q-1)\tilde{\varepsilon}],$$

donde el estimador del parámetro de esfericidad viene dado por

$$\tilde{\varepsilon} = \frac{N(q-1)\hat{\varepsilon} - 2}{(q-1)[N-p-(q-1)\hat{\varepsilon}]}, \qquad (6)$$

donde *p* denota los niveles del factor entre grupos, *q* denota los niveles de la variable intrasujeto y $\hat{\varepsilon}$ el estimador de Greenhouse y Geisser (1959).

Sin embargo, cuando el tamaño de los grupos es pequeño y el número de grupos grande $\tilde{\varepsilon}$ puede sobreestimar el verdadero valor de ε incluso puede llegar a exceder la unidad. Así las cosas, Lecoutre (1991) propuso modificar el numerador de la Ecuación (6)

$$\tilde{\varepsilon}_L = \frac{(N-p+1)(q-1)\hat{\varepsilon} - 2}{(q-1)[N-p-(q-1)\hat{\varepsilon}]}, \qquad (7)$$

La investigación empírica ha puesto de relieve que cuando la matriz de covarianza promediada no es esférica las prueba de Greenhouse y Geisser y Huynh y Feldt modificada por Lecoutre son estimadores robustos a la violación de la esfericidad multimuestral siempre que las matrices sean homogéneas y el diseño balanceado. Sin embargo, ningún procedimiento basado en reducir los *gl* asociados con el estadístico *F* del tradicional AVAR resulta robusto cuando el diseño no disfruta de homogeneidad de las matrices de covarianza, estén los grupos equilibrados o no (Keselman *et al*., 2001).

También en la Tabla 1 se muestra el comportamiento del AVAR con los *gl* corregidos mediante los procedimientos de Geisser y Greenhouse y Huynh y Feldt modificado por Lecoutre cuando las matrices Σ son homogéneas y heterogéneas y carecen de esfericidad. En ella observamos que ambos estadísticos se comportan de modo correcto

cuando las matrices Σ son homogéneas sea el diseño balanceado o no. Cuando las matrices Σ son heterogéneas y el tamaño de los grupos es el mismo, tanto para la variable entre-sujetos como para la interacción ambos estadísticos se alejan levemente de la robustez. Cuando el tamaño de los grupos es diferente la estimación es conservadora si el apareamiento es positivo y muy liberal si el apareamiento es negativo. En condiciones de heterogeneidad la variable intra-sujeto sólo se aleja de la robustez cuando el apareamiento es negativo, mostrándose levemente liberal.

Pruebas univariadas que no asumen la igualdad de las matrices de dispersión

– El enfoque de la Aproximación General Mejorada (AGM).

Cuando además de ausencia de esfericidad existe heterogeneidad Huynh (1978) recomendó modificar los parámetros de los valores críticos F_B y F_{AXB} de la Ecuación (2) en términos de las matrices de covarianza y del tamaño de los grupos como sigue:

$$F_B \approx b F[\alpha; h', h] \qquad (8)$$
$$F_{AxB} \approx c F[\alpha; h'', h]$$

Si se cumple el supuesto de esfericidad las ecuaciones (2) y (8) son equivalentes.

Finalmente, para conducirse con la heterogeneidad de covarianza a través de los grupos, la razón F_A aproximada viene dada por $cF[\alpha; h'^{\bullet}, h^{\bullet}]$, donde los estimadores del sesgo por la falta de homogeneidad y de los *gl* coinciden con los valores críticos de Box (1954). Los estimadores de *c, b, h, h', h'', h*˙ y *h*˙ y la corrección debida a Lecoutre los encontramos en Algina (1994), Algina y Oshima (1995) y Keselman y Algina (1996).

Kesekman, Algina, Wilcox y Kowalchuck (2000) han desarrollado el procedimiento AGM con estimadores robustos (con medias recortadas). Algina y Oshima (1995), Kesekman, Keselman y Lix (1995), Keselman, Algina, Kowalchuk y Wolfinger (1998), Kowalchuck, Keselman, Algina y Wolfinger (2004) y Livacic (2005) señalan que el estadístico AGM es robusto y potente tanto para el tratamiento como para la interacción frente a la violación de la esfericidad cuando las matrices Σ son heterogéneas incluso cuando los datos se asientan en distribuciones no normales, sólo es levemente liberal cuando el apareamiento entre Σ y n_j es negativo. En la Tabla 2 se muestra el comportamiento del enfoque AGM en ausencia de esfericidad multimuestral y heteroscedasticidad de las matrices de dispersión. En ella podemos apreciar cómo el estadístico AGM es robusto para las tres fuentes de variación cuando las varianzas son heterogéneas y tenemos el mismo tamaño de muestra en cada uno de los tres grupos de la variable entre-sujetos. También es robusta cuando el diseño es no balanceado pero el apareamiento es positivo. Sólo es levemente liberal tanto para la variable intra-sujeto como para la interacción cuando el apareamiento es negativo.

TABLA 2. Error de Tipo I para los estadísticos: AGM, MLM y Bootstrap-F. α = 0,05; N = 30.

Σ = Tipo NE (ε=0,50)			$\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$	
V.Entre	AGM	0,0498[1]	0,0502[3]	0,0491[4]
V.Intra	AGM	0,0525[1]	0,0496[3]	**0,0692[4]**
Interacción	AGM	0,0498[1]	0,0478[3]	**0,0701[4]**
V.Entre	Proc Mixed AIC	0,0550[1]	**0,0395[3]**	**0,0725[4]**
V.Intra	Proc Mixed AIC	0,0510[1]	**0,0370[3]**	**0,0803[4]**
Interacción	Proc Mixed AIC	0,0560[1]	**0,0330[3]**	**0,1084[4]**
V.Entre	Proc Mixed BIC	**0,0600[1]**	**0,0285[3]**	**0,1140[4]**
V.Intra	Proc Mixed BIC	0,0575[1]	**0,0170[3]**	**0,1454[4]**
Interacción	Proc Mixed BIC	**0,0690[1]**	**0,0260[3]**	**0,1624[4]**
V.Entre	Proc Mixed MCI	0,0500[1]	0,0455[3]	0,0550[4]
V.Intra	Proc Mixed MCI	0,0475[1]	0,0535[3]	0,0530[4]
Interacción	Proc Mixed MCI	0,0445[1]	**0,0355[3]**	**0,0682[4]**
V.Entre	Bootstrap-F	**0,0364[1]**	0,0476[3]	0,0590[4]
V.Intra	Bootstrap-F	0,0424[1]	0,0514[3]	**0,0644[4]**
Interacción	Bootstrap-F	0,0417[1]	0,0488[3]	0,0596[4]

Nota. AGM= Aproximación General Mejorada; Proc Mixed AIC= MLM mediante el procedimiento PROC MIXED del SAS utilizando el criterio de información de Akaike (1974) de selección de la matriz de dispersión; Proc Mixed BIC= MLM mediante el procedimiento PROC MIXED del SAS utilizando el criterio de información de Schwarz (1978) de selección de la matriz de dispersión; Proc Mixed MCI= MLM mediante el procedimiento PROC MIXED del SAS utilizando la matriz Correctamente Identificada; Bootstrap-F= Procedimiento Bootstrap-F; Resto: ver Tabla 1.

– El Modelo Lineal Mixto (MLM). Análisis orientado a corregir los valores críticos de la F univariada y a modelar las matrices de dispersión.

El Modelo Lineal Mixto general para analizar medidas repetidas extiende el modelo clásico a situaciones donde los supuestos de independencia entre las puntuaciones de diferentes sujetos y homogeneidad de las matrices de dispersión no son requeridos. El MLM tiene muchas ventajas que le convierten en una estrategia sólida para realizar inferencias más exactas tanto de los parámetros del modelo -efectos de los tratamientos e interacción- (Kowalchuck *et al*., 2004; Littel, 2002; Wolfinger, 1996) como de sus errores estándar (Núñez-Antón y Zimmerman, 2001), a saber: a) Con independencia de que el número de niveles de la variable intra-sujeto exceda o no el número de sujetos (esto es, sin importar el tamaño de la muestra), permite analizar los datos cuando se tienen observaciones perdidas (datos incompletos), es decir, cuando para algunos sujetos no se dispone de todos los registros -quizás porque no acudieron a la cita, quizás por algún error del experimentador-. También permite que para cada sujeto se realicen diferentes número de registros (tal vez porque se ha logrado determinada meta, por ejemplo, que hayan alcanzado una determinada destreza). También permite que los intervalos de tiempo entre cada aplicación de tratamiento sean específicos para cada sujeto. b) Permite encontrar cual es la estructura de la matriz Σ que mejor se ajuste a los datos antes de proceder al análisis, posibilitando modelar las variaciones intra y

entre-sujetos tanto para datos completos como incompletos (Vallejo, Fernández y Secades, 2004). El procedimiento PROC MIXED del SAS (SAS Institute, 2002) ofrece varias estructuras de covarianza para ajustar, entre ellas, de Simetría Combinada (S), de Huynh-Feldt (H), No Estructurada (NE) y Autorregresiva de primer orden (AR1) utilizando los criterios de información de Akaike (1974) (AIC) y de Schwarz (1978) (BIC) (ver Keselman *et al*., 1998; Littell, Milliken, Stroup y Wolfinger, 1998, pp. 101-102). La elección del procedimiento de estimación para el análisis del MLM y los estadísticos de prueba están descritos en Littell *et al*. (1998, pp. 498-502). Ya hemos explicado qué es una matriz de Tipo S, de Tipo H y NE. ¿Qué es una matriz autorregresiva? En estos diseños que descansan en el registro sucesivo de conductas de cada uno de los sujetos es probable que la condición S ó H no se cumpla. Que la matriz sea NE es esperable cuando la variable independiente intra-sujeto son diferentes tratamientos que se administran de modo aleatorio para cada uno de los sujetos y se distancia su aplicación de un modo prudencial para evitar efectos residuales. Ahora bien, cuando la variable intra-sujeto es la edad o el tiempo es posible que haya efectos residuales y/o autocorrelación entre las puntuaciones (son muchos los investigadores que advierten que una muy pequeña autocorrelación puede viciar completamente las pruebas de significación, incrementando el error de Tipo I o de Tipo II si es positiva o negativa, respectivamente). Es imperativo añadir que cuando la investigación implica además algún proceso de aprendizaje la maduración puede hacer mella haciéndose notar con la presencia añadida de heterogeneidad de las varianzas de los tratamientos. Lo que sucede es que siempre que existe autocorrelación hay ausencia de esfericidad y ninguno de los estadísticos anteriormente expuestos asume que la correlación entre las observaciones en distintos puntos del tiempo sea una función de la distancia temporal entre ellos. c) Permite que las medidas repetidas sean de naturaleza categórica (Davis, 2002), incluso permite el manejo de covariadas cuantitativas y categóricas cambiantes a lo largo del tiempo. Esto último es de capital importancia, supongamos que en la investigación realizada por Méndez *et al*. (2004) la evolución y desaparición de la fobia dependa en parte de si los padres trabajan o no a turnos y de lo que acerca de eso acontecía en el momento en que se realizasen los registros (dado que éstos fueron pieza fundamental en la aplicación de la terapia).

Sin embargo, también tiene puntos débiles, dos en concreto: De una parte, las dificultades que en ocasiones tiene para identificar correctamente la estructura de la matriz Σ subyacente dado que los criterios AIC y BIC no la seleccionan siempre correctamente (Keselman, Algina, Kowalchuck y Wolfinger, 1999a y 1999b; Livacic, 2005; Vallejo, Fernández y Ato, 2003). De otra, que los estimadores de precisión e inferencia se basan en su distribución asintótica -con muestras muy grandes ajusta bien- (Wolfinger, 1996), por lo que puede ocasionar serios problemas cuando se trabaja con muestras reducidas (Vallejo *et al*., 2002). Cuando disponemos de pequeños tamaños de muestra, en lugar de utilizar la prueba *F* por defecto del MLM las soluciones desarrolladas por Fai y Cornelius (S) (1996) y Kenward y Roger (KR) (1997) son recomendadas por muchos autores.

La investigación metodológica ha puesto de manifiesto que cuando se selecciona correctamente la matriz Σ el procedimiento MLM controla la tasa de error cuando

existe heterogeneidad en las matrices sean los tamaños de los grupos igual o diferentes y sea como sea la relación entre las matrices y el tamaño (Livacic, 2005, Vallejo, Fernández y Velarde, 2001). Sin embargo, cuando la muestra es reducida muchos autores advierten que es deficiente (Keselman *et al*., 1999b; Keselman, Kowalchuck y Boik, 2000; Wright y Wolfinger, 1996). Anteriormente expusimos que en estos casos los estimadores KR y S eran más apropiados. Pues bien, de la investigación se desprende que el estimador MLM (KR) resulta potente y robusto en condiciones normales y no normales (Kowalchuck *et al*., 2004; Vallejo, Fernández, Herrero y Conejo, 2004) aunque si las condiciones son severamente sesgadas y el apareamiento es negativo, la estimación del efecto de la interacción se vuelve conservadora. El estimador MLM (S) tiene un buen comportamiento para la variable intra-sujeto pero no para la interacción (Keselman *et al*., 1999a; Livacic, 2005).

En la Tabla 2 también se muestra el comportamiento PROC MIXED en ausencia de esfericidad multimuestral cuando la matriz Σ se ajusta mediante los criterios de búsqueda AIC y BIC, y cuando la matriz se identifica correctamente (MCI). Se observa cómo el método de búsqueda AIC es menos liberal y menos conservador que el método de búsqueda BIC cuando el apareamiento es negativo y positivo, respectivamente. También su excelente comportamiento cuando la matriz es correctamente identificada.

– El enfoque bootstrap–F. Enfoque del Modelo Mixto de Scheffé con los valores críticos determinados empíricamente.

Se basa en derivar la distribución muestral empírica del estadístico de interés remuestreando repetidamente con reposición desde la muestra disponible. Vallejo, Cuesta, Fernández y Herrero (2006) describen cómo se aplica la metodología bootstrap para contrastar las hipótesis de un diseño multigrupo de medidas repetidas p x q con cuatro operaciones sumamente sencillas. A diferencia de lo que sucedía con los enfoques anteriores, bajo este método los valores críticos derivados desde la teoría normal son innecesarios. Berkuvits, Hancok y Nevitt (2000), Keselman *et al*. (2000) y Vallejo *et al*. (2002) señalan que, como esta técnica no requiere para su aplicación hipótesis respecto de la forma de la distribución de la población, permite obtener estimadores más precisos cuando los datos provienen de distribuciones no normales y la matrices Σ son heterogéneas, sobre todo cuando tienen distinto tamaño. Keselman, Wilcox y Lix (en prensa) y Livacic (2005) señalan que tiene un excelente comportamiento con respecto al error de Tipo I para el tratamiento en condiciones normales y no normales, sin embargo resultaba conservador para la interacción en condiciones no normales. La potencia en cualquier situación no era muy elevada. En la Tabla 2 podemos ver su excelente comportamiento cuando el apareamiento entre el tamaño y la forma de la matriz es positivo. Sólo es levemente liberal su estimación para la variable intra-sujeto y cuando el apareamiento es negativo.

Técnicas de análisis: pruebas multivariadas

Como la ausencia de esfericidad es más la norma que la excepción cuando los datos están recogidos secuencialmente, el investigador puede optar por alternativas de

análisis multivariadas que no requieren averiguar cómo es la estructura de la matriz de dispersión subyacente. Todas ellas requieren el cumplimiento de los supuestos de normalidad conjunta multivariada, independencia entre los vectores de observaciones y que el tamaño de la muestra, para que se pueda calcular, sea suficientemente mayor que el número de medidas repetidas (N-$p \geq q$-1).

Enfoque multivariado clásico:asume la igualdad de las matrices de dispersión

Como expusimos en la introducción, el análisis multivariado de la varianza (AMVAR) junto al análisis univariado de la varianza (AVAR) son las técnicas de análisis generalmente utilizadas por los investigadores para analizar los datos recogidos de un diseño de medidas repetidas. A priori sendos pueden elegirse con mucha lógica. Si el investigador manipula la variable intra-sujeto para asignar sus niveles de modo aleatorio para cada uno de los sujetos y lo hace de modo espaciado uno de otro para evitar los efectos de orden y residuales, puede esperar que la matriz Σ que subyace a sus datos sea una de Tipo S o de Tipo H y aplicar el AVAR. Si así es, hará una elección correcta (siempre que se cumplan el resto de los supuestos que la técnica requiere). Sin embargo, a menudo la variable intra-sujeto son observaciones registradas a lo largo del tiempo y la variable entre-sujetos consiste en la presencia o ausencia de tratamiento (grupo experimental y grupo de control, respectivamente) como en la investigación realizada por Méndez *et al.* (2004) que nos sirvió de ejemplo. Es estas situaciones es esperable que las medidas de la variable dependiente estén intercorrelacionadas (y por tanto no haya esfericidad). El investigador entonces necesita una técnica de análisis que sea más flexible con respecto a la forma de la matriz de varianzas-covarianzas. En esta situación elegir el análisis multivariado para analizar sus datos es acertado porque permite a la matriz de dispersión tener cualquier estructura, siempre que se cumpla además de los supuestos anteriormente descritos, que las matrices de dispersión sean homogéneas.

Para una exposición sencilla de la formulación implicada en el análisis multivariado remitimos al lector interesado a Arnau (1990), Vallejo (1991) y Timm (2002). Cuando las varianzas son homogéneas (sean o no equilibrados los grupos), no existen observaciones perdidas y el tamaño de muestra no es excesivamente reducido, la literatura empírica ha puesto de relieve que el enfoque multivariado mantiene la tasa de error controlada al nivel nominal estipulado indistintamente de la forma de la matriz Σ. Esto lo podemos observar en la Tabla 3. Sin embargo, también en la misma Tabla observamos que su comportamiento no es robusto cuando las varianzas son heterogéneas sea el tamaño de los grupos igual o diferente, del siguiente modo: si el diseño es equilibrado, con independencia de la forma de la distribución, se vuelve sensiblemente liberal. Cuando el diseño está desequilibrado se comporta de una manera excesivamente conservadora cuando la naturaleza de la relación entre el tamaño de los grupos y el tamaño de las matrices de dispersión es positiva y excesivamente liberal cuando la relación es negativa con independencia de la forma de la matriz Σ.

TABLA 3. Error de Tipo I para los estadísticos: AMVAR, WJ, WJR y BF. α = 0,05.

Σ = Tipo NE (ε=0,50)			$\Sigma_1=\Sigma_2=\Sigma_3$	
V.Entre	AMVAR(*N*=30)	0,0486[1]	0,0497[3]	0,0492[4]
V.Intra	AMVAR(*N*=30)	0,0500[1]	0,0528[3]	0,0505[4]
Interacción	AMVAR(*N*=30)	0,0415[1]	0,0430[3]	0,0485[4]
Σ= Tipo NE (Σ=0,50)			$\Sigma_1\neq\Sigma_2\neq\Sigma_3$	
V.Entre	AMVAR(*N*=30)	**0,0667[1]**	**0,0242[3]**	**0,1529[4]**
V.Intra	AMVAR(*N*=30)	**0,0745[1]**	**0,0128[3]**	**0,2305[4]**
Interacción	AMVAR(*N*=30)	**0,0855[1]**	**0,0190[3]**	**0,2550[4]**
V.Entre	WJ(*N*=30)	0,0547[1]	0,0479[3]	0,0529[4]
V.Intra	WJ(*N*=30)	0,0525[1]	0,0481[3]	**0,0688[4]**
Interacción	WJ(*N*=30)	**0,0667[1]**	0,0581[3]	**0,1186[4]**
V.Entre	WJ(*N*=60)	0,0527[1]	0,0465[3]	0,0481[4]
V.Intra	WJ(*N*=60)	0,0505[1]	0,0495[3]	0,0539[4]
Interacción	WJ(*N*=60)	0,0478[1]	0,0453[3]	**0,0623[4]**
V.Entre	WJR(*N*=30)	0,0505[1]	0,0478[3]	0,0488[4]
V.Intra	WJR(*N*=30)	**0,0368[1]**	**0,0387[3]**	0,0427[4]
Interacción	WJR(*N*=30)	**0,0360[1]**	**0,0342[3]**	0,0463[4]
V.Entre	WF(*N*=30)	0,0530[1]	0,0457[3]	0,0519[4]
V.Intra	WF(*N*=30)	0,0502[1]	0,0475[3]	0,0410[4]
Interacción	WF(*N*=30)	0,0483[1]	0,0462[3]	0,0400[4]
V.Entre	WF(*N*=60)	0,0530[1]	0,0457[3]	0,0519[4]
V.Intra	WF(*N*=60)	0,0539[1]	0,0481[3]	0,0570[4]
Interacción	WF(*N*=60)	0,0517[1]	0,0480[3]	0,0547[4]

Nota. AMVAR= Análisis Multivariado de la Varianza; WJ= procedimiento Welch-James; WJR= procedimiento Welch-James con estimadores robustos; WF= procedimiento de Brown-Forsythe; N= 30 y *N*= 60, tamaños totales de la muestra; Cuando *N*= 60 el tamaño de los grupos es 1= apareamiento nulo en un diseño balanceado (20-20-20) –sujetos en cada uno de los tres grupos-; 3= diseño no balanceado apareamiento positivo (12-20-28); 4= diseño no balanceado apareamiento negativo (28-20-12); Resto: ver Tabla 1.

En lo que se refiere a la falta de normalidad, a pesar de que este método de análisis exhibe un comportamiento ligeramente liberal cuando se analizan modelos aditivos (no hay interacción entre las variables intra y entre-sujetos) y ligeramente conservador cuando se analizan modelos no-aditivos (sí hay interacción significativa), podemos decir que es relativamente robusto al incumplimiento de dicho supuesto, en especial, si se dispone de un tamaño de muestra razonable ($n_{j(min)}/(k-1)>3$ ó 4). La falta de robustez del AMVAR cuando no hay homogeneidad de las matrices de dispersión ha impulsado el desarrollo de estadísticos robustos a la violación de esta asunción. De ellos se cuenta a continuación.

Estadísticos robustos multivariados para la ausencia de homogeneidad de las matrices de dispersión

– Prueba multivariada de Welch-James (WJ) y prueba multivariada de Welch-James con estimadores robustos (WJR).

La solución multivariada Welch-James (WJ) fue propuesta por Keselman, Carriere y Lix (1993) a partir del trabajo Johansen (1980). Más tarde, en el año 2000, estos mismos autores sugieren que la robustez del procedimiento WJ se puede incrementar sustancialmente reemplazando las usuales medias y matrices de covarianza con medias recortadas y matrices de covarianza winsorizadas cuando el tamaño de muestra sea reducido y/o los datos hayan sido extraídos desde distribuciones sesgadas.

La literatura empírica (Algina y Keselman, 1997; Algina y Oshima, 1995; Keselman *et al*., 1999a, 1999b; Keselman, *et al*., 1996; Kowalchuck *et al*., 2004) indica que la prueba WJ es robusta y potente tanto para el tratamiento como para la interacción y, por ende, provee valores válidos *p* frente a la violación de la esfericidad cuando las matrices Σ son heterogéneas, incluso cuando los datos se asientan en distribuciones no normales, sólo es levemente liberal cuando el apareamiento entre Σ y n_j es negativo. Livacic (2005) apunta esto último para la interacción bajo distribuciones leve o severamente sesgadas. Vallejo *et al*. (2002) matiza que su mejor comportamiento lo alcanza cuando $n_{j(min)}/(q\text{-}1) \geq 6$, es decir, necesita un tamaño de muestra elevado. Algina y Keselman (1997) y Keselman *et al*. (1993) también aconsejan cómo debe ser el tamaño de los grupos.

En la Tabla 3 se muestra el comportamiento de los dos enfoques, el enfoque WJ y el enfoque WJR cuando se incumple el supuesto de esfericidad multimuestral (matriz NE, ε=0,50 y matrices de dispersión heterogéneas $\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$) y cuando sólo se incumple el supuesto de esfericidad (matriz NE, ε=0,50 y matrices de dispersión homogéneas $\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$). Hemos considerado dos tamaños muestrales, moderado (N = 30) y grande (N = 60) para el procedimiento WJ y sólo el primero para el procedimiento WJR. Esto nos da una idea de la tendencia en su comportamiento en función del tamaño de la muestra. Podemos observar cómo cuando el tamaño de la muestra es moderado el procedimiento WJ es ligeramente liberal para la interacción cuando sólo el supuesto de esfericidad es violado, y muy liberal cuando a la ausencia de esfericidad se le suma la heterogeneidad de las matrices de dispersión y el apareamiento es negativo. Para las variables intra y entre-sujetos es robusto. Podemos advertir cómo un incremento en el tamaño de la muestra mejora sensiblemente el resultado anterior. El procedimiento WJR con un moderado tamaño de la muestra subsana el exceso de error de Tipo I para la interacción cuando el apareamiento es negativo, sin embargo, se muestra conservador para el tratamiento intra-sujeto y para la interacción en las otras dos situaciones.

– Prueba multivariada de Brown-Forsythe (BF).

La versión multivariada del enfoque de Brown-Forsythe fue propuesta por Vallejo y Ato (en prensa) a partir de lo trabajo de Vallejo, Fidalgo y Fernández (2001) y Vallejo y Livacic-Rojas (2005). La investigación empírica concluye que el procedimiento BF controla la tasa de error cuando existe heterogeneidad en las matrices Σ sea el tamaño

de los grupos igual o diferente y sea la relación que sea entre las matrices y el tamaño, además de que no exige un gran tamaño muestral (Livacic, 2005; Vallejo *et al*., 2001; Vallejo y Livacic-Rojas, 2005).

En la Tabla 3 también se muestra el comportamiento del procedimiento BF en las mismas condiciones que los estadísticos WJ y WJR anteriormente expuestos. Podemos observar su buen comportamiento cuando el tamaño de la muestra es moderado excepto para la interacción cuando el apareamiento es negativo, que se vuelve conservador. Un incremento del tamaño de la muestra la convierte en una prueba estadística excelente para poner a prueba las hipótesis de todas las fuentes de variación.

Discusión y conclusiones

Mediante la experimentación con simulación las técnicas estadísticas dirimen sus diferencias y muestran sus competencias. Este testigo anteriormente expuesto es el que ahora recogemos para abordar el segundo objetivo que nos planteamos. Partiendo de que los fines sin los medios adecuados son frustrantes, y los medios sin los fines son tan improductivos como inútiles, hemos de tener conciencia de que una investigación siempre se debe diseñar para que pueda hacer frente a las circunstancias más desfavorables y de que la ausencia de precauciones metodológicas y estadísticas ha conducido a soluciones equivocadas a los problemas reales (Keselman, Othoman, Wilcox y Fradette, 2004; Ramos-Álvarez, Valdés-Conroy y Catena, 2006; Wilcox, 2003). Así pues, extendemos las siguientes recomendaciones:

- Dado que un error en la premisa aparecerá siempre en la conclusión, en primer lugar se debe realizar una correcta planificación de la investigación atendiendo a la validez en todos sus aspectos. Debemos procurar que la validez interna y de constructo sean exquisitas, lo mismo que la fiabilidad en las variables de registro. Algunos puntos al amparo de la validez de la conclusión estadística son los que a continuación se citan.
- Se ha de cuidar el tamaño de la muestra: ha sido ampliamente demostrada la baja potencia de todos los estadísticos con grupos menores de 30 sujetos. Tras un análisis exhaustivo de los resultados de investigación Keselman *et al*. (1993) aconsejan para garantizar la potencia que $n_{j(min)}$ debe ser 2 ó 3 y 3 ó 4 veces mayor que (q-1) para poner a prueba el tratamiento y la interacción respectivamente bajo distribución normal. Estos dígitos deben ser mayores (3 ó 4 y 5 ó 6) si la distribución es no normal. Blanca Mena (2004), Keselman *et al*. (1993), Keselman *et al*. (1996), Keselman *et al*. (2001) y Vallejo *et al.* (2002) entre otros, también han exhortado esta advertencia.
- Se debe mimar la arquitectura de la muestra: cuidar que no exista pérdida de unidades experimentales (sujetos), de ocasiones de medida o de registro de las mismas durante el desarrollo de la investigación, además de hacerlo en intervalos regulares, todo ello con ánimo de que el diseño esté convenientemente equilibrado (ver Vallece y Green, 2002).
- Es necesario y en este orden, tener en cuenta el número de grupos y las carac-

terísticas de los datos en ellos recogidos. La elección de la prueba estadística se debe basar en ese examen. Aconsejamos proceder del siguiente modo. Si sólo tenemos un grupo, se debe atender a los puntos 1 y 2. Si tenemos más de uno primero tomaremos el pulso a la heterogeneidad de matrices de covarianza (debemos sospechar que exista si el diseño es no balanceado). Si los grupos son homogéneos atenderemos a los puntos 1 y 2, si no lo son, sólo contemplaremos el punto 3. Los puntos 1, 2 y 3 son los que siguen:

(1) Si se cumplen todos los supuestos (independencia, normalidad, homogeneidad y esfericidad), sea el diseño o no balanceado, y el tamaño de muestra (N) es suficientemente mayor que el número de medidas repetidas (q): AVAR o AMVAR. Si N es reducido con respecto a q: AVAR.

(2) Si se cumplen todos los supuestos del modelo excepto el de esfericidad, sea el diseño o no balanceado: si N es suficientemente mayor que q: estadístico de Huynh-Feldt con la corrección de Lecoutre si $\varepsilon<0{,}75$ y estadístico de Greenhouse y Geisser si $\varepsilon\geq0{,}75$. AMVAR en cualquier caso. Si N es reducida con respecto a q, uno de los dos primeros en función del valor de ε.

(3) Si no se cumple el supuesto de homocedasticidad de las matrices de covarianza, exista o no esfericidad y sea o no el diseño balanceado: si el tamaño de la muestra es suficientemente grande: BF, AGM o WJ. Si el tamaño de la muestra es reducido, BF o AGM, mejor el primero porque no impone ninguna restricción sobre la forma de la matriz de dispersión.

– Implementación de los estadísticos: AVAR, AMVAR, estadístico de Huynh-Feldt con la corrección de Lecoutre y estadístico de Greenhouse y Geisser (SAS, SPSS); MLG (PROC MIXED, SAS); AGM (Programa en SAS/IML de Algina, 1997); WJ (Programa en SAS/IML de Lix y Keselman, 1995; Keselman *et al*., 2001); BF (Programa en SAS/IML de Vallejo, Moris y Conejo, 2006).
– Siempre que se aplique un estadístico que no sea de naturaleza multivariada es imperativo examinar y seleccionar adecuadamente la estructura de la matriz de covarianza que subyace a los datos. Sólo esto otorgará una base racional para realizar inferencias exactas y eficientes de los parámetros del modelo y con ello obtener pruebas más potentes de los efectos (Vallejo *et al*., 2002). No aconsejamos el MLM aplicando los usuales criterios de búsqueda semiautomática de la matriz Σ cuando no existen datos perdidos. Sin embargo, enfatizamos su utilidad cuando hay pérdida de observaciones, covariadas dependientes del tiempo, etc.
– Al exponer los resultados es necesario informar qué prueba estadística se ha utilizado y los *gl* a ella asociados (Keselman *et al*., 2001), además de sustituir los habituales asteriscos que se adjuntan al resultado empírico por el valor de p.

Blanca Mena (2004), Keselman *et al*. (1993), Keselman *et al*. (2001), Keselman *et al*. (2000) y Kowalchuck, Keselman y Algina (2003) han hecho otras recomendaciones.

Los puntos anteriores van orientados a garantizar tanto la potencia como la robustez para el tratamiento y para la interacción (no hacemos referencia a los contrastes de

medias) de los diseños de medidas repetidas, pero sólo en aquellas situaciones donde la escala de medida es continua, se efectúa un solo registro en cada ocasión, no hay datos discordantes, ni pérdida de ocasiones de medida para algún sujeto, ni covariadas cambiantes con el tiempo, y el diseño está convenientemente equilibrado. Todo eso que aquí no se ha contemplado nos hace conscientes de que la realidad admite una enorme complejidad, no obstante, nos debe tranquilizar que su estudio esté en plena ebullición.

Es verdad que el texto que se escribe nunca es el fiel precipitado de la intención inicial, sino el punto de encuentro entre el propósito y el resultado. Que es difícil aunar profundidad con claridad, rigor con estilo y análisis con síntesis. Sin embargo, esperamos con éste corregir la miopía que surte de exclusivamente utilizar el AVAR o el AMVAR en el análisis de los datos de los diseños de medidas repetidas.

Referencias

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716-723.

Algina, J. (1994). Some alternative approximate tests for a split-plot design. *Multivariate Behavioral Research, 29*, 365-384.

Algina, J. (1997). Generalization of improved general approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. *British Journal of Mathematical and Statistical Psychology, 50,* 243-252.

Algina, J. y Keselman, H.J. (1997). Testing repeated measures hypotheses when covariances are heterogeneous: Revisiting the robustness of the Welch-James test. *Multivariate Behavioral Research, 32*, 255-274.

Algina, J. y Oshima, T.C. (1995). An improved general approximation test for the main effect in a split-plot design. *British Journal of Mathematical and Statistical Psychology, 48*, 149-160.

Arnau, J. (1990). *Diseños experimentales multivariables*. Madrid: Alianza.

Berkovits, I., Hancock, G. y Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robusrness to sphericity and normality violations. *Educational and Psychological Measurement, 60,* 877-892.

Blanca Mena, M. (2004). Alternativas de análisis estadístico en los diseños de medidas repetidas. *Psicothema, 16,* 509-518.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effects of inequality of variance in the one–way classification. *Annals of Mathematical Statistics, 25*, 290-403.

Collier, R.O., Baker, F.B., Mandeville, G.K. y Hayes, T.F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures designs. *Psychometrika, 32,* 339-353.

Davis, Ch. (2002). *Statistical methods for the analysis of repeated measurements*. Nueva York: Springer-Verlag.

Fai, A. y Cornelius, P. (1996). Approximate F-Tests of multiple degree of freedom hypotheses in generalized least squares analyzes of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation, 54,* 363-378.

Fernández, P. (1995). *Comparación mediante simulación de la potencia y robustez del enfoque multivariado de medidas repetidas frente al correspondiente univariado con la estructura del error modelada a través de procesos autorregresivos.* Tesis Doctoral no publicada. Universidad de Oviedo.

Finney, D. (1990). Repeated measurement: What id measurement and what effects. *Statistics in Medicine, 9*, 639-644.

Geisser, S. y Greenhouse, S.W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics, 29,* 885-891.

Greenhouse, S.W. y Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika, 43*, 161-165.

Huynh, H. y Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements design have exact F-distributions. *Journal of the American Statistical Association, 65*, 1582-1585.

Huynh, H. y Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom from sample data randomized block and split-plot designs. *Journal of Educational Statistics, 1,* 69-82.

Johansen, S. (1980). The Welch-James approximation of the distribution of the residual sum of squares in weighted linear regression. *Biometrika, 67*, 85-92.

Kazdin, A. (2001). *Métodos de investigación en Psicología Clínica*. México, D.F.: Prentice Hall.

Kenward, M.G. y Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983-997.

Keselmam, H.J. y Algina, J. (1996). The analysis of higher-order repeated measures designs. En B. Thompson (Ed.), *Advances in social science methodology* (vol. 4) (pp. 45-70). Greenwich, CT: JAI Press.

Keselman, H.J., Algina, J. y Kowalchuck, R.K. (2001). The analysis of the repeated measures design: A review. *British Journal of Mathematical and Statistical Psychology, 54,* 1-20.

Keselman, H.J., Algina, J., Kowalchuck, R.K. y Wolfinger, R.D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics. Simulation and Computation, 27*, 591-604.

Keselman, H.J., Algina, J., Kowalchuck, R.K. y Wolfinger, R.D. (1999a). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology, 52*, 63-78.

Keselman, H.J., Algina, J., Kowalchuck, R.K. y Wolfinger, R.D. (1999b). The analysis of repeated measurements: A comparison of mixed-model Satterthwaite *F* tests and a nonpooled adjusted degrees of freedom multivariate test. *Communications in Statistics-Theory and Methods, 28*, 2967-2999.

Keselman, H.J., Algina, J., Wilcox, R.K. y Kowalchuk, R.K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test again. *Educational and Psychological Measurement, 60,* 925-938.

Keselman, H.J., Carriere, M.C. y Lix, L.M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. *Journal of Educational Statistics, 18*, 305-319.

Keselman, H.J., Keselman, J.C. y Lix, L.M. (1995). The analysis of repeated measurements: Univariate tests, multivariate tests, or both? *British Journal of Mathematical and Statistical Psychology, 48,* 319-338.

Keselman, H.J., Kowalchuck, R. y Boik, R. (2000). An examination of robustness of the empirical bayes and other approaches for testing main interaction effects in repeated measures designs. *British Journal of Mathematical and Statistical Psychology, 53*, 51-67.

Keselman, J., Lix, L y Keselman H.J. (1996). The analysis of repeated measurements designs: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology, 49,* 275-298.

Keselman, H.J., Othoman, A., Wilcox, R. y Fradette, K. (2004). The new and improvement two-sample *t* tests. *Psychologycal Science, 15,* 47-51.

Keselman, H.J., Wilcox, R. y Lix, L. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. Recuperado de www:/umanitoba.ca/faculties/arts/psychology.

Kowalchuck, R.K., Keselman, H.J. y Algina, J.Y. (2003). Repeated measures interaction test with aligned ranks. *Multivariate Behavioral Research, 38*, 963-974.

Kowalchuck, R.K., Keselman, H.J., Algina, J.Y. y Wolfinger, R.D. (2004). The analysis of repeated measures with mexed-model adjusted *F* test. *Educational and Psychological Measurements, 64*, 224-242.

Kowalchuck, R.K., Lix, L.M. y Keselman, H.J. (1996, junio). *The analysis of repeated measures designs*. Comunicación presentada en el 61 st. Annual Meeting of the Psychometric Society, Banff, Alberta, Canada.

Lecoutre, B. (1991). A correction for the approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics, 16*, 371-372.

Littell, R.C. (2002). Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics, 7*, 472-490.

Littell, R.C., Milliken, G.A., Stroup, W.W. y Wolfinger, R.D. (1998). *SAS system for mixed models.* Cary, NC: SAS Institute.

Livacic, P. (2005). *Una evaluación empírica de procedimientos alternativos de análisis y diseños de medidas repetidas.* Tesis Doctoral no publicada. Universidad de Oviedo.

Lix, L. y Keselman, H. (1995). Approximate degrees of freedom test: A unified perspective on testing for mean equality. *Psychological Bulletin, 117*, 547-560.

Méndez, X., Orgiles, M. y Espada, J.P. (2004). Escenificaciones emotivas para la fobia a la oscuridad: un ensayo controlado. *International Journal of Clinical and Health Psychology, 4*, 505-520.

Mendoza, J.L. (1980). A significance test for multisample sphericity. *Psychometrika, 45,* 495-498.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 92*, 778-785.

Montero, I. y León, O.G. (2005). Sistema de clasificación del método en los informes de investigación en Psicología. *International Journal of Clinical and Health Psychology*, *5,* 115-127.

Núñez-Antón, V. y Zimmerman, D.L. (2001). Modelización de datos longitudinales con estructuras de covarianza no estacionarias: modelos de coeficientes aleatorios frente a modelos alternativos. *Qüestiió, 25,* 225-262.

Palmero, F., Breva, A., Diago, J.L., Díez. J.L. y García, I. (2002). Funcionamiento psicofisiológico y susceptibilidad a la sintomatología premestrual en mujeres Tipo A y Tipo B. *International Journal of Clinical and Health Psychology, 2*, 111-136.

Quintana, S. y Maxwell, S.E. (1994). A Monte Carlo comparison of seven e-adjustment procedures in repeated measures designs with sample sizes. *Journal of Educational Statistics, 19*, 57-71.

Ramos-Álvarez, M.M., Valdés-Conroy, B. y Catena, A. (2006). Criteria of the peer-review process for publication of experimental and cuasi-experimental research in Psychology. *International Journal of Clinical and Health Psychology*, *6*, 773-787.

SAS Institute (2002). *SAS/STAT software: Version 9.0 (TS M0)*. Cary, NC: SAS Institute.

Sawilowski, S.S. y Blair, R.C. (1992). A more realistic look at the robustness and Type II error probabilities of the *t* test to departures from population normality. *Psychological Bulletin, 111*, 352-360.

Scheffé, H. (1956). A mixed model for the analysis of variance. *Annals of Mathematical Statistics, 27,* 23-36.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Timm, N.H. (2002). *Applied multivariate analysis*. Nueva York: Springer-Verlag.

Vallace, D. y Green, B.S. (2002). Analysis of repeated measures designs with linear mixed models. En D.S. Moskowitz y S.L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 135-170). Mahwah, NJ: Lawrence Erlbaum Associates.

Vallejo, G. (1991). *Análisis univariado y multivariado de los diseños de medidas repetidas de una sola muestra y de muestras divididas.* Barcelona: PPU.

Vallejo, G., Arnau, J., Bono, R., Cuesta, M., Fernández, P. y Herrero, J. (2002). Análisis de diseños de series temporales cortas. *Metodología de las Ciencias del Comportamiento, 4,* 301-323.

Vallejo, G. y Ato, M. (en prensa). Modified Brown-Forsythe procedure for testing interaction effects in Split-Plot designs. *Multivariate Behavioral Research, 41.*

Vallejo, G., Cuesta, M., Fernández, P. y Herrero, J. (2006). A comparison of the bootstrap-F, improved general approximation and Brown-Forsythe multivariate approaches in a mixed repeated measures design. *Educational and Psychological Measurement, 66*, 35-62.

Vallejo, G., Fernández, P. y Ato, M. (2003). Tasas de potencia de dos enfoques robustos para analizar datos longitudinales. *Psicológica, 24,* 109-122.

Vallejo, G., Fernández, P., Herrero, J., y Conejo, N. (2004). Alternative procedures for testing fixed effects in repeated measures designs when assumptions are violated. *Psicothema*, *16*, 498-508.

Vallejo, G., Fernández, J.R., y Secades, R. (2004). Application of a mixed model approach for assessment of interventions and evaluation of programs. *Psychological Reports, 95,* 1095-1118.

Vallejo, G., Fernández, P. y Velarde, H. (2001). Un estudio comparativo de pruebas robustas para el análisis de datos longitudinales. *Metodología de Ciencias del Comportamiento, 3*, 35-52.

Vallejo, G., Fidalgo, A.M., y Fernández, P. (2001). Effects of covariance heterogeneity on three procedures for analysing multivariate repeated measures designs. *Multivariate Behavioral Research*, *36*, 1-27.

Vallejo, G. y Livacic-Rojas, P. (2005). A comparison of two procedures for analyzing small sets of repeated measures data. *Multivariate Behavioral Research*, *40,* 179-205.

Vallejo, G., Moris, J. y Conejo, N (2006). A SAS/IML program for implementing the modified Brown-Forsythe procedure in repeated measures designs. *Computer Methods and Programs in Biomedicine, 83,*169-177.

Wilcox, R.R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Nueva York: Springer.

Wilcox, R.R. (2003). *Applying contemporary statistic techniques.* San Diego: Academic Press.

Wolfinger, R.D. (1996). Heterogeneous variance-covariance structures for repeated measurements. *Journal of Agricultural, Biological, and Enviromental Statistics, 1*, 205-230.

Wright, S.P. y Wolfinger, R.D. (1996, octubre). *Repeated measures analysis using mixed models: Some simulations results*. Comunicación presentada en la Conference on Modelling Longitudinal and Spatially Correlated data: Methods, Applications, and Future directions, Nantucket, MA, Estados Unidos.