POTENCIA ESTADÍSTICA Y CÁLCULO DEL TAMAÑO DEL EFECTO EN G*POWER: COMPLEMENTOS A LAS PRUEBAS DE SIGNIFICACIÓN ESTADÍSTICA Y SU APLICACIÓN EN PSICOLOGÍA

STATISTICAL POWER AND EFFECT SIZE CALCULATING IN G*POWER: COMPLEMENTARY ANALYSIS OF STATISTICAL SIGNIFICANCE TESTING AND ITS APPLICATION IN PSYCHOLOGY

DOI: 10.22199/S07187475.2014.0002.00006

Recibido: 16 de Junio del 2014 | Aceptado: 08 de Agosto del 2014

MANUEL CÁRDENAS CASTRO[1]; HÉCTOR ARANCIBIA MARTINI[2]
(UNIVERSIDAD DE VALPARAÍSO, Valparaíso, Chile)

RESUMEN

El uso de pruebas de significación estadística es una estrategia que se encuentra muy arraigada en la investigación psicológica. Sin embargo, se han sobrevalorado las bondades de dichas pruebas al considerarlas como un indicador suficiente de la veracidad de una hipótesis, omitiendo la cuantificación de las diferencias encontradas. Así, las conclusiones resultan erróneas al homologar diferencia significativa con diferencia "grande", "importante" o "relevante". Dada la creciente importancia de los indicadores del tamaño del efecto y la potencia estadística, en este artículo desarrollamos un breve marco conceptual del análisis estadístico de la potencia y el tamaño del efecto, así como ejemplos prácticos de su cálculo utilizando el programa G*Power 3.1.6, para estimular y facilitar su inclusión en futuras publicaciones.

PALABRAS CLAVE: Inferencia estadística, potencia estadística, tamaño del efecto.

ABSTRACT

The use of statistical significance test is deeply rooted in psychological research. However, it has been on estimating the benefits of such tests considering itself a sufficient indicator of the accuracy of a hypothesis. It has tended to ignore the quantification of the differences found and has tended to draw the wrong conclusions to approve significant difference with "large", "important" or "relevant" difference. Given the increasing importance of indicators of effect size and statistical power, in this article we provide a brief conceptual framework of the statistical analysis of the power and effect size and practical examples of its calculation using the program G * Power 3.1.6 to help alleviate the lack of many researchers of how to perform such analyzes

KEY WORDS: Statistical inference, statistical power analysis, effect size.

1. Afiliado a la Universidad de Valparaíso, Chile E-mail: manuel.cardenas@uv.cl
2. Afiliado a la Universidad Autónoma de Madrid, España. E-mail: ps.arancibia@gmail.com

Entre las recomendaciones que ya hace bastante tiempo ha venido realizando la American Psychological Association (APA, 1994, 2008, 2011) se encuentra a) la utilización como práctica habitual de los intervalos de confianza (IC; límites probables entre los que se encuentra la verdadera diferencia entre dos medias); b) la exposición de los valores de las medias y desviaciones típicas (*DT*) de cada grupo; c) la entrega de los valores exactos de probabilidad (y no los tradicionales $p<.05$ ó $p<.01$); d) informar la potencia estadística de la prueba o diseño utilizado; y d) realizar el cálculo complementario del tamaño del efecto que cuantifica la magnitud de la diferencia entre dos medias (Wilkinson, 1999). Lo cierto es que hasta la fecha buena parte de estas recomendaciones no son consideradas.

Actualmente parte importante de las revistas científicas en psicología publican artículos que no informan del tamaño del efecto y omiten sistemáticamente los cálculos del tamaño de la muestra y la potencia estadística del diseño (Bezeau & Graves, 2001; Crosby et al., 2008; Fidler, 2002; García, Ortega & De la Fuente, 2008; Kirk, 1996; Vacha-Haaze & Ness, 1999; Vacha-Haaze & Thompson, 1998). Estas omisiones ponen en cuestión la credibilidad de los hallazgos que puedan derivarse de dichos estudios y representan, en el decir de algunos de los más reputados especialistas del área, una de las mayores muestras de ignorancia colectiva (Cohen, 1988).

La ausencia de cuantificadores del tamaño del efecto lleva a tomar decisiones fundadas en el desconocimiento de una parte importante de la información, aquella que cuantifica la magnitud de los efectos encontrados (e.g. ¿estaríamos dispuestos a aceptar como dato suficiente la significación de las correlaciones sin su respectivo índice de magnitud?). Del mismo modo, la ausencia de información sobre la potencia de los diseños elude hacerse cargo de los denominados errores de tipo II ("falsos negativos) que constituyen la prueba más relevante de la validez de nuestro diseño de estudio.

Aunque diversas razones explican estas omisiones dos de ellas son fundamentales: a) la ausencia en los paquetes estadísticos de mayor uso de módulos para su valoración y, b) la falta de exigencias editoriales para su inclusión como información imprescindible (Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000).

En el diseño de la investigación debiese considerarse el tamaño de la muestra y la potencia estadística que lograríamos con ella. No obstante, en aquellos estudios donde dicho paso ha sido omitido resulta importante al menos exigir el cálculo y especificación del tamaño del efecto como estrategia de análisis post-hoc. Aunque esta práctica ha sido cuestionada (la de realizar cálculos post hoc), dado que no permite planear apropiadamente el estudio ni corregir los errores de diseño, acompañar las pruebas de significación con la cuantificación de la magnitud del efecto alcanzado permite, al menos, comprender adecuadamente los resultados de dichos análisis.

Así, el objetivo de este artículo es doble. Primero demostrar la relevancia de la inclusión de estas estrategias, y, complementariamente, explicar paso a paso la manera de calcularlos para cada una de las principales pruebas de significación estadística empleadas en el campo de la psicología, utilizando el software de distribución gratuita G*Power (disponible desde enero del 2007 en la v. 3.0.0.).

Pruebas de hipótesis, sensibilidad y potencia de la significación estadística

Se entiende por sensibilidad de una prueba a su capacidad para detectar diferencias o efectos allí donde los haya y potencia estadística al grado de probabilidad de rechazar una hipótesis nula cuando esta es

realmente falsa, es decir, a la capacidad de una prueba para detectar diferencias entre grupos cuando estas están presentes. Por su parte, las pruebas de contraste de hipótesis determinan si la hipótesis nula, que se plantea en términos de no-diferencias o no-relación, puede ser rechazada con cierto nivel de confianza o si, por el contrario, debe ser mantenida. Si se rechaza se asume que la diferencia detectada por un "tratamiento" no es atribuible al azar o no ha ocurrido por mera casualidad, aceptándose que ha producido un efecto real. Para dicho rechazo se recurre a una convención, que aunque no exenta de importantes cuestionamientos (Morrison & Henkel, 2006), fija el nivel de confianza de la estimación en el campo de la psicología en α=0.05 ($p<0.05$; que corresponde a un 5% de error). La hipótesis nula se rechaza si la probabilidad es igual o menor al nivel alfa que hemos fijado a priori.

Este proceso de decisión puede conducir, sin embargo, a dos potenciales errores (Figura 1). El error Tipo I (falsos positivos) consiste en la probabilidad (α) de rechazar una hipótesis nula que es en realidad verdadera asumiendo erróneamente que el tratamiento ha producido un efecto, y el error Tipo II (falsos negativos) que consiste en la probabilidad (definida como β) de mantener una hipótesis nula que en realidad es falsa asumiendo que no existen efectos de tratamiento cuando en realidad sí los hay. De este modo, la potencia estadística de una prueba no es sino el complemento de la probabilidad de error de tipo II (1-β). En ambos casos hay una falta de sensibilidad de la prueba de significación.

FIGURA 1.
Posibilidades de error en pruebas de significación estadística (Lipsey, 1990).

	Situación en la población	
Conclusión de la prueba estadística	Tratamiento y control difieren	Tratamiento y control no difieren
Diferencias significativas (rechazar H_0)	Conclusión correcta Probabilidad = 1-β (Potencia)	Error de tipo I Probabilidad = α
Diferencias no significativas (aceptar H_0)	Error de tipo II Probabilidad = β	Conclusión correcta Probabilidad = 1-α

En la Figura 1 podemos ver graficadas las opciones de acierto y error en las pruebas de significación estadística cuando se comparan los estadísticos obtenidos en la muestra con los parámetros poblacionales. La potencia estadística esperada convencionalmente para un análisis es del 80% (1-β=.80). Es decir, existe un 20% de probabilidad de aceptar la hipótesis nula cuando esta es en realidad falsa (β=.20). Se estima que un valor inferior implicaría un riego demasiado grande de incurrir en un error Tipo II. Un valor superior, como se verá más adelante, implicaría ampliar excesivamente la muestra. Así, la potencia estadística constituye un índice de la validez de nuestros resultados estadísticos (Cohen, 1992; Bono & Arnau Gras, 1995).

Relación entre potencia estadística y tamaño del efecto

La potencia estadística (PE) se calcula sobre la base de tres cifras: tamaño de la muestra (*n*), nivel de error (α) y tamaño del efecto (TE). En términos generales podemos afirmar que cuanto mayor sea la muestra, mayor será la potencia estadística (manteniendo constante el TE y α), dado que el error aleatorio de medida es menor (Lipsey, 1990; Cohen, 1988). El tamaño del efecto representa el grado en que la hipótesis nula es falsa. Cuando el TE es grande la PE aumenta (Cohen, 1988, 1992). Al incrementar el error de Tipo I la potencia también aumenta y cuanto más pequeño es el valor de α, más baja será la potencia. Es por ello que debe equilibrarse la probabilidad de cometer errores de Tipo I y II (Sedlmeier & Gigerenzer, 1989).

Estimar el tamaño del efecto, que responde a la magnitud de las diferencias encontradas en el estudio, y la potencia estadística, que responde al grado de validez que tienen los hallazgos de la investigación, es importante y constituye cada vez más una exigencia debido a razones éticas y técnicas (Cohen, 1998; Grissom & Kim, 2012; Murphy, Myors, & Wolach 2009; Nickerson, 2000). Éticas ya que no resulta correcto realizar estudios que no sean lo suficientemente estrictos para determinar el efecto real de un "tratamiento" debido a su falta de potencia. Técnicamente tampoco sería apropiado dado el derroche de recursos que implica reclutar más participantes de los necesarios para lograr verificar los objetivos del estudio.

Críticas a las pruebas de significación y potencia estadística

Al realizar el contraste de hipótesis se deberían responder tres preguntas básicas: ¿Podemos afirmar que hay diferencia? ¿Es grande la diferencia? ¿Es importante la diferencia? Las pruebas de significación nos permiten responder tan sólo a la primera. El tamaño del efecto permite dar cuenta de la segunda y la tercera sólo es respondida mediante un criterio de relevancia clínica.

Entre las críticas que más se han hecho sentir están las que indican que por sí mismas, las pruebas de significación estadísticas constituyen una pobre estrategia científica (Meehl, 1978; Cohen, 1994; Thompson & Snyder, 1997), ya que su énfasis es poco informativo, dado que con un número suficiente de casos y con medidas medianamente fiables la hipótesis nula siempre será falseable, al margen de la verdad o falsedad de la teoría sustantiva. Un valor elevado en una prueba de contraste de hipótesis sólo indica que la probabilidad de que las diferencias detectadas debidas al azar sea muy alta. Nada nos indica del tamaño de dichas diferencias, por lo que valores muy altos o muy bajos de α no deberían interpretarse jamás en sentido de diferencias/ no diferencias importantes. Por otra parte, adoptar un nivel de confianza fijo para el rechazo de hipótesis transforma en una decisión dicotómica lo que en realidad es un continuo de incertidumbre (Kirk, 1996). Así, confiar en la significación estadística como si fuera un índice de certeza es incorrecto, ya que el nivel de significación no informa sobre la magnitud de las diferencias ni sobre su importancia práctica (Cohen & Hyman, 1979).

El cálculo del TE es un análisis complementario de las pruebas de significación que contribuye a subsanar las limitaciones anteriormente expuestas. El "efecto" refiere al resultado de un tratamiento experimental. El tamaño del efecto es la diferencia más pequeña que el investigador está dispuesto a aceptar como clínicamente relevante (Prajapati, Dunne, & Armstrong, 2010) y nos indica cuánto de la variable dependiente se puede explicar, predecir o controlar por la variable independiente (Snyder & Lawson, 1993). De otro modo, informa el grado en que la hipótesis nula es falsa y lo hace mediante un índice en una métrica común que indica

la magnitud de una relación o efecto (Cohen, 1988). Una diferencia significativa no es una diferencia necesariamente grande o importante, para ello se debe cuantificar la magnitud de dicha diferencia significativa (Grissom & Kim, 2012).

Familias, índices y fórmulas para el cálculo del tamaño del efecto

Se distinguen en la literatura tres familias de índices del tamaño del efecto: diferencias estandarizadas de medias, coeficientes de correlación e índices de riesgo para tablas de contingencia. Las diferencias estandarizadas de medias (Cohen, 1988) señalan el grado de diferencia entre dos medias en un lenguaje que permite compararlas con los resultados obtenidos en otros estudios, con independencia del tamaño de las muestras utilizadas en ellos. Los coeficientes de correlación (Furr, 2004) incluyen los índices que expresan el grado de asociación existente entre dos variables, así como las que expresan la proporción de varianza explicada. Los índices de riesgo para tablas de contingencia (Kline, 2004) cuantifican la asociación entre variables nominales dicotómicas, proporcionando una estimación de la proporción de sujetos que experimentan un determinado resultado. En la Figura 2 se presentan los índices del tamaño del efecto más habituales dentro de cada familia, además de su respectiva fórmula.

FIGURA 2.
Índices y fórmulas para el cálculo del tamaño del efecto.

Familia	Tipo de prueba	Fórmula para cálculo del tamaño del efecto	Cálculo de la desviación típica común
Diferencia estandarizadas de medias	Índice d para Pruebas t (muestras independientes)	$d = \frac{\bar{Y}1 - \bar{Y}2}{S_{común}}$	$S_{común} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n1+n2-2}}$
	Índice d para Pruebas t (muestras relacionadas)	$d = \frac{\bar{Y}pre - \bar{Y}post}{S_{pre}}\left(1 - \frac{3}{4n-5}\right)$	
	Índice d para diseño de pre y post test con grupo de control	$d = \frac{(\bar{Y}expre - \bar{Y}expost) - (\bar{Y}conpre - \bar{Y}conpost)}{S_{pre}}$	$S_{pre} = \sqrt{\frac{(n_1-1)(S_{expre})^2 + (n_2-1)(S_{conpre})^2}{n1+n2-2}}$
Índices de correlación	Índice w para pruebas de asociación Chi-cuadrado	$w = \sqrt{\sum_{i=1}^{rc} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	
	Índice f para análisis de varianza de un factor	$f = \frac{\sigma_m}{\sigma}$	$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{Sigma^2}{k}}$
	Índices Eta cuadrado parcial y f para ANOVA factorial	$\eta_p^2 = \frac{suma\ de\ cuadrados_{Tratamiento}}{Suma\ de\ cuadrados_{Tratamiento} + Suma\ de\ cuadrados_{Error}}$	$f = \sqrt{\frac{eta^2}{1 - eta^2}}$
	Índice f^2 para análisis de regresión múltiple	$f^2 = \frac{R^2}{1 - R^2}$	$f^2 = \frac{Suma\ de\ Cuadrados_{Efecto}}{Suma\ de\ Cuadrados_{Error}}$

El cálculo de los tamaños del efecto permite interpretar las diferencias encontradas y compararlas de un estudio a otro independientemente de las variaciones de diseño o de las diferencias del tamaño muestral. De allí la relevancia que estos índices tienen en los estudios de meta-análisis cuyo fin es sistematizar la información disponible en un determinado campo.

Interpretación de los índices del tamaño del efecto

La interpretación de los índices del tamaño del efecto se presenta en la Figura 3. En ella se señalan los valores referenciales para las principales pruebas estadísticas incluidas en las familias de índices antes referidas.

FIGURA 3.
Valores referenciales para el tamaño del efecto de las diferentes pruebas de significación estadística.

Prueba	símbolo	Pequeño	Mediano	Grande
Pruebas t	d	.20	.50	.80
ANOVA unifactorial	f	.10	.25	.40
ANOVA factorial	η_p^2 / f	.01	.06	.14
Chi cuadrado	w / φ	.10	.30	.50
Regresión múltiple	f^2	.02	.15	.35

Si bien se trata de valores consensuados, se espera que la magnitud de la diferencia sea interpretada por el investigador de acuerdo a los resultados obtenidos y a la evidencia existente. En cualquier caso, la valoración de los tamaños del efecto puede hacerse de diversas formas: interpretación del valor absoluto, del valor relativo y valoración de coste-beneficio. El valor absoluto remite a la tabla presentada y que es construida sobre la base de los percentiles a los que cada punto de corte remite. Un valor del tamaño del efecto *d*=.30 indicaría que al compararlo con las tablas, el 62% de las personas quedaría por debajo de dicho resultado. El valor relativo se relaciona con la relevancia práctica al dar cuenta de la comparación con los valores encontrados en otros estudios similares. Finalmente, la valoración de coste-beneficio se refiere a la importancia que pequeños tamaños de efecto encontrados pudiesen tener cuando sus costes de implementación no son elevados.

Ahora bien, el cálculo del TE también resulta importante en algunos casos en que las pruebas de significación no muestran resultados significativos. Por ejemplo la posible falta de significación podría deberse al tamaño muestral en investigaciones que consideren pocos participantes. Esta es sin duda una razón fundamental para estimar el tamaño muestral necesario para detectar las posibles diferencias.

Cálculo del tamaño del efecto utilizando G*Power v. 3.1.6

G*Power es un programa estadístico, de descarga gratuita, diseñado para realizar estimaciones de la potencia estadística y del tamaño del efecto (Erdfelder, Faul, & Buchner, 1996; Faul, Erdfelder, Lang, & Buchner, 2007). El programa requiere de un procesador con una velocidad mínima de 2.46 Mb y puede ser descargado gratuitamente en el sitio web de los autores: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/. Permite realizar los tradicionales análisis a priori (calcula el *n* muestral apropiado para alcanzar una determinada potencia con un determinado TE y α) y post hoc (dónde las estimaciones de la PE, el Error Standard (*ES*) y TE realizan en diseños ya terminados) de estimación de la potencia de una prueba, pero también entrega análisis de compromiso (calcula α y PE cuando los otros dos términos son conocidos), sensitividad (estima cuál es el efecto mínimo que la prueba es sensible para detectar) y criterio (calcula el α necesario para lograr una determinada PE cuando *n* y TE son conocidos). En síntesis, el programa permite realizar diversos cálculos tales como el del tamaño del efecto, de la potencia esperada de un test, de la muestra necesaria para lograr una determinada potencia, y permite verificar la significación respecto de las posibilidades reales del estudio.

En los apartados siguientes se ejemplificará el cálculo de TE y PE mediante pruebas *t* para muestras independientes (familia de diferencia de medias estandarizadas). Posteriormente revisaremos ejemplos de las familias de correlaciones (ANOVAS de un factor y factorial, así como regresión lineal múltiple) y de índices de riesgo para tablas de contingencia (Chi-cuadrado). Todos los análisis previos fueron realizados utilizando el paquete estadístico SPSS v. 20.0.

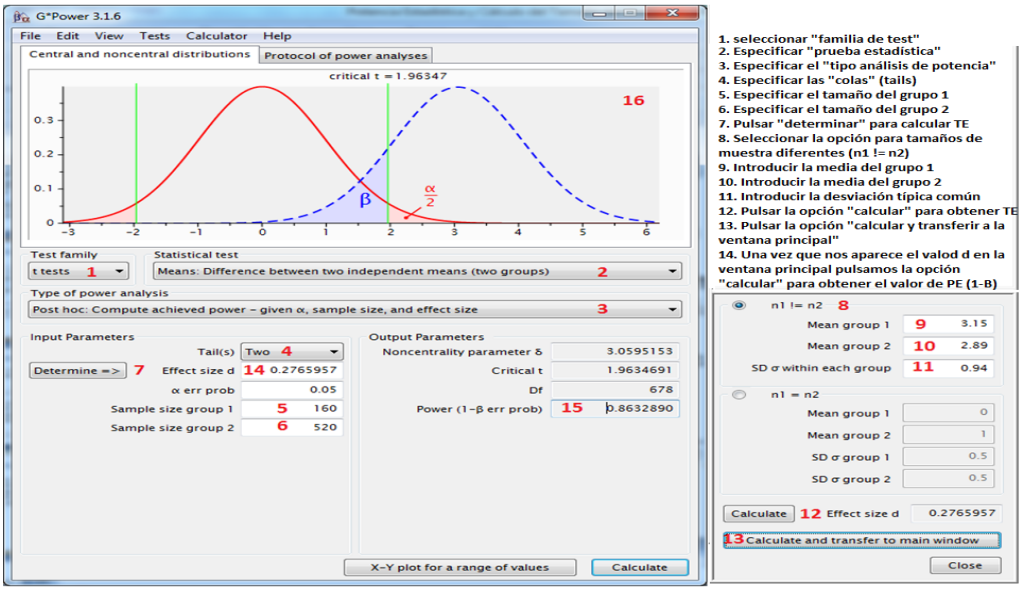
El índice d para el caso de dos medias independientes

Para ejemplificar los procedimientos utilizados se utilizaron las puntuaciones de un inventario sobre crecimiento post traumático en dos grupos de personas afectadas por eventos estresantes en los últimos seis meses. A estas personas se les pedía que evaluaran la percepción de cambios o mejoras en su vida fruto del esfuerzo cognitivo por adaptarse a los acontecimientos estresantes. El grupo de personas que accedieron a ayuda psicológica (n=160; M=3.15 y DT=.94) puntuó significativamente más alto que el grupo de personas que no recibió dicha atención (n=520; M=2.89; y DT=.94). Estas diferencias fueron significativas estadísticamente ($t_{(678)}$=3.06; p=.002; IC95% [.094, .430]). En la Figura 4 se presentan cada uno de los pasos necesarios para obtener el valor de d en G*Power. El índice d representa el grado de separación entre la hipótesis nula y la hipótesis alternativa (o grado en que las dos hipótesis no se superponen).

Se recomienda utilizar el contraste bilateral para estos análisis (ver paso 4 de la gráfica 1) dado que éstos requieren una mayor diferencia para detectar una misma potencia. Los contrastes unilaterales, en principio, sólo son permisibles cuando conocemos el sentido de las diferencias que se quieren detectar.

En el ejemplo que analizamos se aprecia que el tamaño del efecto (d=.27) puede ser considerado mediano ya que se encuentra en torno a .30 que es el valor fijado convencionalmente. El único cálculo en que debemos incurrir para el caso en que las muestras sean de tamaño diferente es la desviación típica común (ver Tabla 2). La potencia estadística (1-β=.86) supera los niveles mínimos exigidos (80%), constatándose en la gráfica que la probabilidad de cometer un error de tipo II es del 14%. La gráfica se interpreta siguiendo las siguientes coordenadas: la sombra más clara (roja en el visor del programa) representa la posibilidad de error Tipo I (α); la sombra oscura (azul en el visor del programa) la probabilidad de error Tipo II (β); la curva de línea continua representa la distribución poblacional (roja en el visor del programa); la línea discontinua la distribución muestral (azul en el visor del programa); y la línea vertical (verde en el visor del programa) corresponde a los puntos críticos de t.

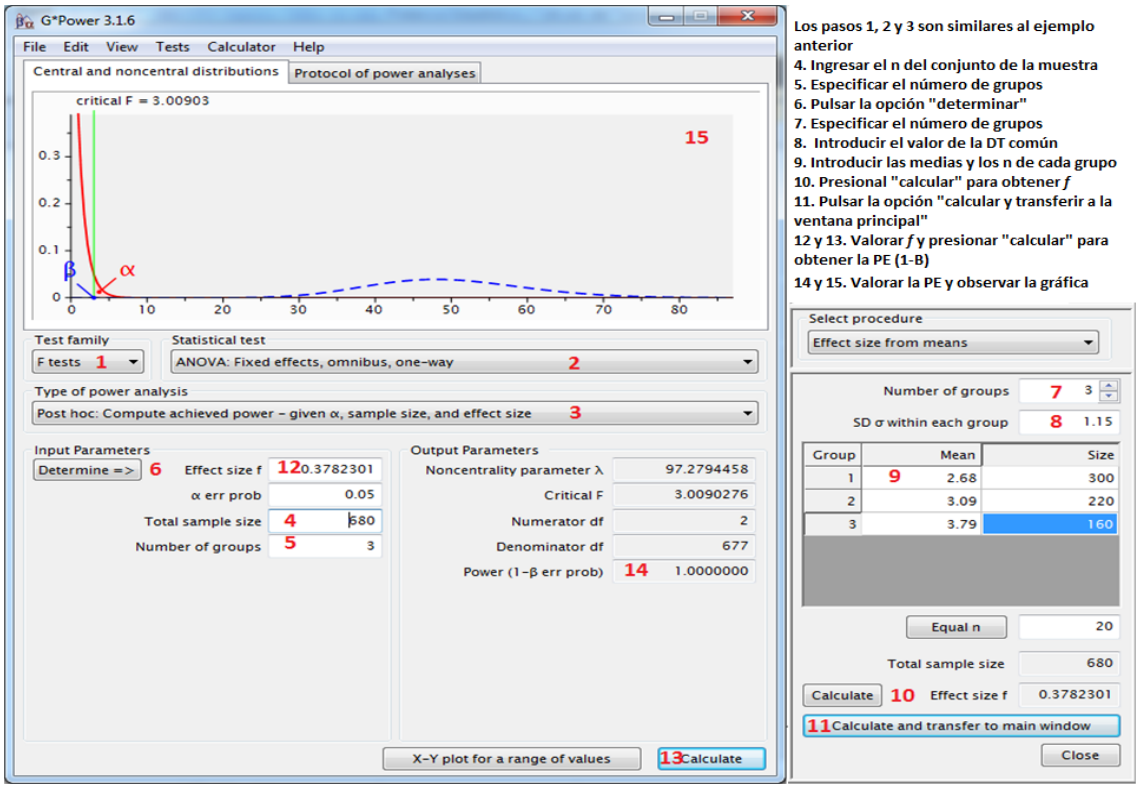
FIGURA 4.
Pruebas t para muestras independientes: Cálculo del tamaño del efecto y la potencia estadística en G*Power (análisis post-hoc).



El estadístico f para el caso de comparación de medias en más de dos grupos

El análisis característico utilizado para analizar la comparación de medias en más de dos grupos corresponde al análisis de varianza (ANOVA) de un factor. A modo de ejemplo comparamos las puntuaciones de tres grupos. El primero estaba formado por personas que accedieron a servicios psicológicos (*n*=160); el segundo por personas que, habiendo accedido, abandonaron tempranamente el tratamiento (*n*=220); y tercero a personas que no solicitaron dicha ayuda (*n*=300). El análisis estadístico indica que el grupo que reporta mayores puntuaciones es el de aquellos que accedieron y terminaron su tratamiento (*M*=3.79; *DT*=1.02), seguido de aquellos que lo discontinuaron (*M*=3.09; *DT*=1.16). Las puntuaciones más bajas las obtuvieron las personas que no tuvieron tratamiento (*M*=2.68; *DT*=1.27). De acuerdo al ANOVA realizado, diferencias entre grupos fueron significativas estadísticamente ($F_{(2, 677)}$=45.50; *p*=.000; IC95% [2.88, 2.08]). Los análisis post hoc (Tukey) también indicaron que las medias de los tres grupos diferían significativamente entre sí.

FIGURA 5.
ANOVA de un factor: Cálculo del tamaño del efecto y la potencia estadística en G*Power (análisis post-hoc).

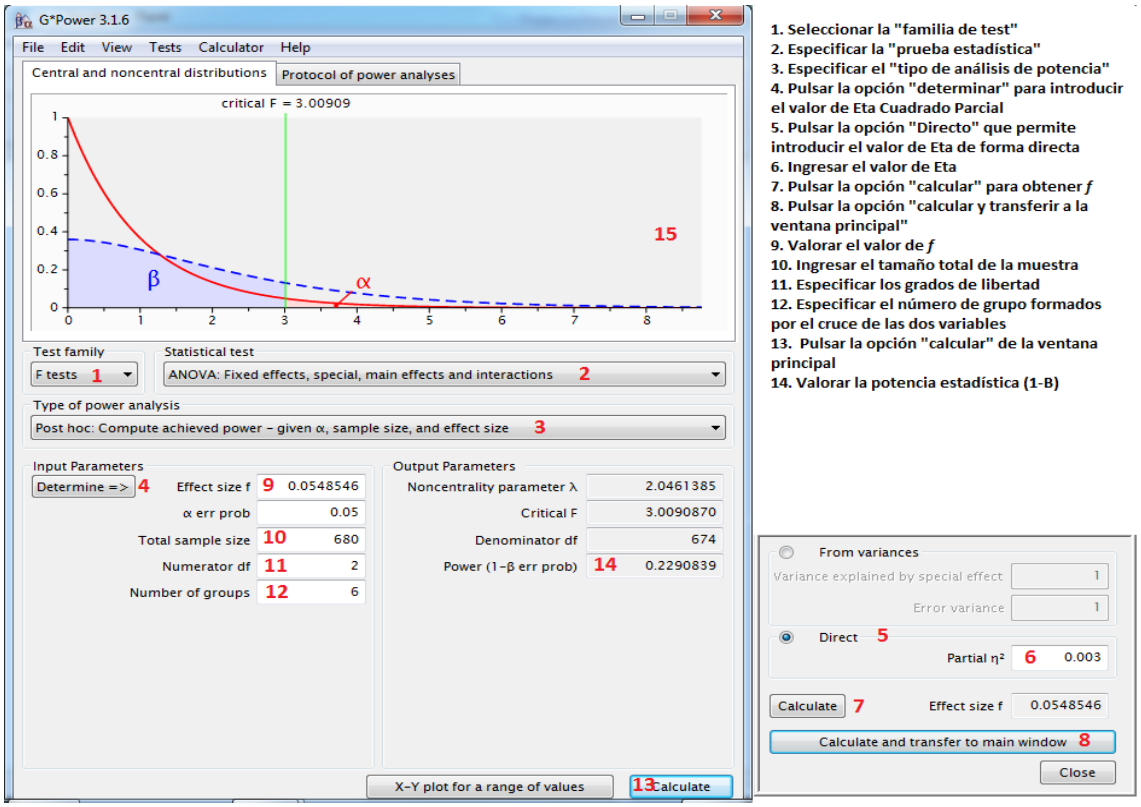


En la Figura 5 se presentan los pasos para obtener el valor de *f* en G*Power. Los resultados indican que para estos análisis la probabilidad de cometer un error de Tipo II es casi inexistente y que la magnitud de las diferencias entre las medias de los grupos es alta.

El estadístico f para el análisis de varianza factorial

Los resultados del paquete estadístico SPSS para este tipo de análisis permiten obtener un índice del tamaño del efecto denominado eta cuadrado parcial (n_p^2). Este se interpreta como proporción de varianza de la variable dependiente que es explicada por las variables predictoras o independientes. Siguiendo con el ejemplo antes expuesto, para calcular el tamaño del efecto se segmentó la muestra por las variables sexo (hombre/mujer) y grado de acceso a servicios psicológicos (atención continuada/ atención discontinuada/ sin atención). En realidad podríamos calcular los tamaños del efecto para cada una de las variables, así como para la interacción (del mismo modo que se realiza un contraste de hipótesis para cada una de ellas). Los pasos para el cálculo de TE y PE en G*Power se pueden realizar directamente sobre el valor de eta-cuadrado parcial (Figura 3). En el ejemplo, dicha interacción no es significativa y la proporción de varianza explicada por Eta es extremadamente baja ($F_{(2,\ 674)}$=.98; p=.37; η^2=.003) por lo que no es esperable encontrar un valor de *f* importante. La única precisión que se debe hacer es sobre la introducción de los grados de libertad en la ventana principal de G*Power y que corresponde al valor que se muestra frente a cada variable e interacción en la salida del programa SPSS (generalmente corresponde al número de dimensiones de cada variable menos uno, salvo para el caso de las interacciones).

FIGURA 6.
ANOVA factorial: Cálculo del tamaño del efecto y la potencia estadística en G*Power (análisis post-hoc).

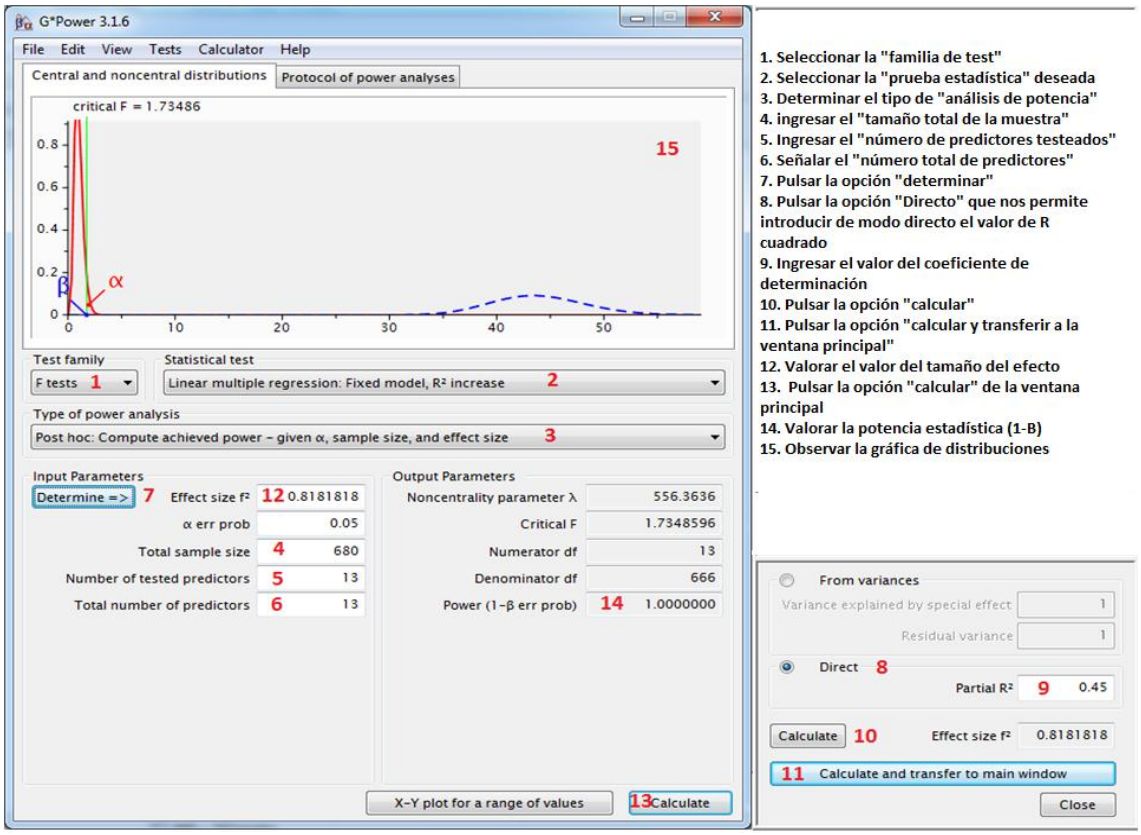


Los análisis que hemos realizado ratifican un tamaño del efecto bajo y agregan información relevante sobre la potencia estadística de la prueba. Hemos afirmado que la convención indica que la potencia debe ser superior al 80%, o de lo contrario la validez del diseño puede ser puesta en duda. En este caso el ANOVA indica que no existe efecto de interacción significativo entre variables y el valor del tamaño del efecto (*f*) viene a confirmar este resultado. En el hipotético caso de que la prueba de significación hubiese entregado valores $p<.05$ habría que ser muy cautos al momento de extraer conclusiones si el tamaño del efecto fuera de la magnitud encontrada.

El estadístico f^2 para el caso de múltiples variables predictoras

El estadístico f^2 es utilizado en caso de procedimientos de regresión lineal múltiple y se estima a partir del coeficiente de regresión al cuadrado (R^2). Es, como en el caso anterior, una transformación desde un índice que cuantifica la proporción de varianza de la variable dependiente que es explicado por el conjunto de las variables predictoras. A partir de su cálculo se puede definir, sobre la base de su comparación con unos valores referenciales consensuados, si el tamaño del efecto puede ser considerado alto, medio o bajo. El procedimiento de cálculo en G*Power (Figura 7) es extremadamente simple si se trabaja de forma directa con el valor de R^2 parcial (se puede obtener con el análisis de regresión múltiple realizado en SPSS).

FIGURA 7.
Regresión multiple: Cálculo del tamaño del efecto y la potencia estadística en G*Power (análisis post-hoc).



El ejemplo anterior informa que, en conjunto, las trece hipotéticas variables predictoras incorporadas en el modelo explican el 45% de la varianza total de la variable dependiente (niveles de crecimiento post traumáticos reportados). El valor del tamaño del efecto f^2= .81. Se trata de un valor alto que indica que el efecto de las variables incorporadas en el modelo es sustantivo. También informa de la potencia estadística y de la muy baja posibilidad de cometer un error de Tipo II (la gráfica muestra como las distribuciones se encuentran totalmente separadas y que la posibilidad de un error β está muy alejada de nuestra curva de la distribución muestral).

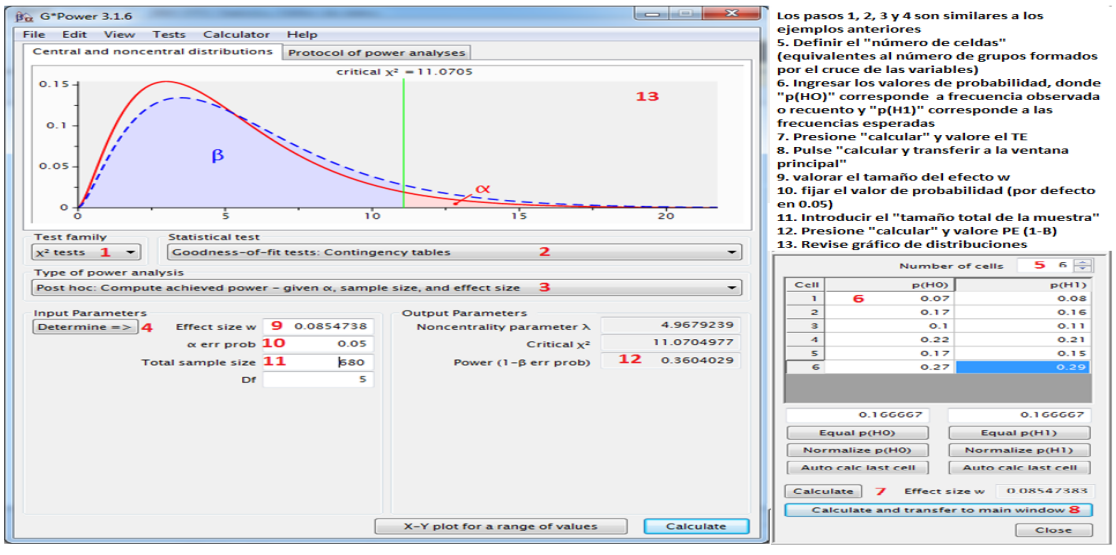
Con los ejemplos anteriores hemos presentado los análisis más típicos en las familias de diferencias estandarizadas e índices de correlación. Si bien, se suele incorporar en esta familia el cálculo del coeficiente Chi-cuadrado, lo presentaremos como un ejemplo de índices para tablas de contingencia (aunque tradicionalmente aquí deberían expresarse también los índices de riesgo relativo y odds ratios).

Índice w para coeficiente de asociación en tablas de contingencia

Se trata de un índice de tamaño del efecto para pruebas de asociación, típicamente Chi-cuadrado (X^2). Para ejemplificar mediante G*Power hemos contrastado la hipótesis de que no existen diferencias entre hombres y mujeres, con el reporte de niveles altos, moderados o bajos de crecimiento post traumático. Para el caso de matrices cuadradas (variables dicotómicas) SPSS entrega el valor de los coeficientes Phi, que adopta valores entre 0 y 1, y su interpretación es similar al coeficiente de correlación de Pearson. En el caso de que una de las variables tenga más de dos niveles (como en nuestro ejemplo), phi puede tomar valores superiores a 1 (pues el valor X^2 puede ser mayor que el tamaño muestral). Aunque este valor de phi debería ser suficiente para cuantificar el efecto encontrado, en la Figura 8 se muestra cómo realizar el cálculo del índice *w*, para lo cual será necesario fijarse en los valores observados de las frecuencias esperadas y observadas de la tabla de contingencia, los cuales deben ser transformados en proporciones (simplemente dividiéndolos por el tamaño total de la muestra).

Para el caso de nuestro ejemplo los resultados obtenidos nos indican que no existe asociación entre las variables ($X^2_{(2)}$=4.69; p=.09; Phi=.08). Es decir, el reporte de niveles altos, medios o bajos de crecimiento post traumático no se relaciona con el sexo de quien responde. Como podemos apreciar el valor del tamaño del efecto obtenido es *w*=.08, similar el coeficiente phi que entrega SPSS.

FIGURA 8.
Chi-cuadrado: Cálculo del tamaño del efecto y la potencia estadística en G*Power (análisis post-hoc).



Vemos también que con una probabilidad de α=.05, *n*=680 y *w*=.08, la potencia estadística apenas es de 1-β=.36, lo que indicaría que existe una alta probabilidad (64%) de cometer errores de tipo II si se rechaza la hipótesis nula.

DISCUSIÓN

La estrategia de significación estadística y rechazo de hipótesis nula es probablemente una de las más arraigadas en investigación en psicología. Resulta sumamente llamativo que los investigadores y publicaciones hayan transformado este procedimiento en la estrategia científica privilegiada en la investigación en psicología, dada la acotada información que es capaz de ofrecer.

La ritualizada práctica de entregar la significación estadística de los contrastes, sin especificar el tamaño del efecto o la lateralidad del contraste, conduce la mayor parte de las veces a predicciones triviales. Así, se construye todo un andamiaje teórico que termina por sobre valorar hallazgos y por anidar resultados contradictorios, que podrían haber sido resueltos con facilidad, si el nivel de las exigencias se elevara mínimamente siguiendo las recomendaciones que desde hace mucho viene haciendo la APA sobre los resultados referidos a pruebas de significación. Aun hoy esta exigencia sigue siendo relevante pues observamos cómo cada día más sofisticados análisis estadísticos de datos se utilizan como criterio de verdad o relevancia. En este sentido debemos seguir buscando la significación práctica y no únicamente una de carácter estadístico ya que casi todas las hipótesis nulas pueden eventualmente ser rechazadas con muestras suficientemente amplias, no pudiéndose afirmar entonces que dichos hallazgos resulten “importantes” ni dar cuenta sobre la magnitud de dichas diferencias. Esto más bien llevaría a confundir sistemáticamente una diferencia estadísticamente significativa con una diferencia relevante.

Las pruebas de significación están lejos de ser un índice de certeza y constituyen un criterio pobre para aceptar o rechazar resultados de investigación. De hecho, la falta de significación no significa que la hipótesis nula sea verdadera ni que los efectos de los dos grupos sean equivalentes. La ausencia de evidencia nunca es evidencia de ausencia de efectos (Altman & Bland, 1995).

Cualquier prueba de significación estadística que no vaya acompañada de un cálculo del tamaño del efecto carece de los parámetros necesarios para juzgar la importancia del hallazgo. De otro modo, lo que hacemos al rechazar una hipótesis nula, particularmente en el caso de la comparación de medias, es afirmar que existe una diferencia. Mientras más baja sea la probabilidad asociada y mayor el valor del estadístico de contraste, más probable será que la diferencia de medias sea distinta de cero. Eso sí, nada hemos dicho de la magnitud ni de la importancia de dicha diferencia. Es decir, hemos afirmado con bastante confianza que las medias son diferentes (la diferencia sería mayor de lo puramente aleatorio), pero ¿Cuán grande e importante es dicha diferencia? Esta es una pregunta que no es posible contestar sin recurrir a un análisis del tamaño del efecto.

En este artículo pretendemos contribuir a subsanar algunas de las deficiencias de la investigación psicológica, particularmente aquellas referidas a vacíos de formación teórica, a problemas de acceso a programas estadísticos apropiados y de ejecución práctica de los análisis. La facilidad proverbial con la que puede ser subsanada la omisión del cálculo del tamaño del efecto y la potencia estadística es evidente y hoy en día no existen excusas para no informarlas.

Aunque nosotros hemos mostrado las posibilidades del cálculo post-hoc de la potencia de un contraste, esta debería preferentemente plantearse a priori ya que

de la otra forma nada puede hacerse contra los problemas de diseño que ya se hayan cometido. Abordar la potencia desde cálculos a priori ayuda a orientar el diseño y a definir el tamaño muestral de cada grupo en referencia a los valores medios del tamaño del efecto obtenidos en otros estudios (los cuales nos obliga a conocer de antemano). En cualquier caso, debemos tener en consideración que las violaciones del supuesto de aleatoriedad de la muestra son recurrentes y deberían marcar un claro límite a la generalización de nuestros hallazgos dada su escasa representatividad. La preocupación de la potencia está íntimamente vinculada al error de medida y nos obliga a procurar que la fiabilidad de los instrumentos utilizados quede debidamente verificada. Aún es posible ver como se usan instrumentos sin acompañarlos del reporte de la fiabilidad o sin entregar indicaciones de su validez para la muestra en la que se utilizan.

La significación estadística no está relacionada con el impacto práctico de un estudio. Un efecto relevante no es algo discernible sólo con información estadística, es ante todo necesario comprender y explicar subjetivamente la realidad que impregna el fenómeno. Debemos comprender que el producto primario de una investigación cuantitativa no es un valor de probabilidad (*p*) sino una o más medidas del tamaño del efecto (Cohen, 1962, 1992).

Por sobre todo estos datos y análisis deben ser útiles para explicar los fenómenos y realizar predicciones sobre la realidad. En este sentido, el valor de la estimación del tamaño del efecto debe ser interpretado en el contexto de un estudio y área concreta de investigación ya que un pequeño tamaño del efecto puede ser de gran importancia en determinados ámbitos (Frías-Navarro, Llobet, & García 2000).

El presente artículo presenta una serie de limitaciones que esperamos subsanar en futuros estudios. Lo primero es la apretada exposición y discusión teórica que hemos podido hacer del tema, asunto que hemos intentado subsanar por la vía de remitir a la bibliografía primaria sobre esta discusión. Lo segundo es la perspectiva y ejemplos limitados al análisis post hoc del tamaño del efecto. Una presentación de los cálculos a priori sigue siendo necesaria debido a que es en el momento del diseño donde se puede asegurar una adecuada potencia (los casos que presentamos sólo sirven para verificarla una vez concluido el estudio, y no para asegurarla).

REFERENCIAS

Altmand, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal* 311, 485.

American Psychological Association (1994). P*ublication Manual of the American Psychological Association* (4th ed). Washington, DC: Author.

American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association (2008). Reporting Standards for Research in Psychology. Why Do We Need Them? What Might They Be? *American Psychologist 63*(9), 839-851.

Bezeau, S. & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23*(3), 399-406.

Bono, R. y Arnau Gras, J. (1995). Consideraciones generales en torno a los estudios de potencia. *Anales de Psicología 11*(2), 193-202.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*(3), 145-153.

Cohen, J, (1988). Statistical Power Analysis for the Behavioral Sciences. (2nd ed.), New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1992). Cosas que he aprendido (hasta ahora). *Anales de Psicología, 8*(1-2), 3-18.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cohen, S. A., & Hyman, J. S. (1979). Learning for Mastery: Ten Conclusions after 15 Years and 3,000 Schools. *Educational Leadership*, *37*(2), 104-109.

Crosby, R.D., Wonderlich, S.A., Mitchell, J.E., de Zwaan, M., Engel, S.G., Connolly, K., Flessner, C., Redlin, J., Markland, M., Simonich, H., Wright, T.L., Swanson, J.M., & Taheri, M. (2008). An empirical analysis of eating disorders and anxiety disorders publications (1980-2000)– part II: Statistical hypothesis testing. *International Journal of Eating Disorders, 39*(1), 49-54.

Erdfelder, E., Faul, F., & Buchner, A. (1996). G*POWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1-11.

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Fidler, F. (2002). The Fifth edition of the APA Publication Manual: Why its Statistics Recommendations are so Controversial. *Educational and Psychological Measurement, 62*(5), 749-770.

Frías-Navarro, D., Llobet, L. P. y García, J.F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicothema 12*(2), 236-240.

Furr, R. M. (2004). Interpreting effect sizes in contrast analysis. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences, 3*, 1-25.

García, J., Ortega, E., & De la Fuente, L. (2008). Tamaño del Efecto en las revistas de Psicología Indizadas en Redalyc. *Informes Psicológicos, 10*(11), 173-188.

Grissom, R.J., & Kim, J.J. (2012). *Effect sizes for research: Univariate and Multivariate Applications* New York: Routledge.

Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Kline, R.B. (2004). Beyond significance testing: Reforming data analysis methods En *Behavioral Research*, (pp. 3-17). Washington, DC, US: American Psychological Association, xii, 325 pp.

Lipsey, M.W. (1990). *Design sensivity: Statistical power for experimental research.* Newbury Park, CA. Sage.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of consulting and clinical Psychology*, *46*(4), 806-834.

Morrison, D.E. & Henkel, R.E. (Eds.). (2006). *The significance test controversy: A reader*. Transaction Publishers.

Murphy, K.R., Myors, B., & Wolach, A.H. (2009). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Routledge.

Nickerson, R.S. (2000). Null hypothesis significance testing: a review of an old and

continuing controversy. *Psychological methods*, *5*(2), 241.

Prajapati, B., Dunne, M., & Armstrong, R. (2010). Sample size estimation and statistical power analyses. *Optometry Today*, *16*(7).

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies?. *Psychological Bulletin*, *105*(2), 309-316.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education*, *61*(4), 334-349.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the Journal of Experimental Education. *The Journal of Experimental Education*, *66*(1), 75-83.

Vacha-Haase, T. & Ness, C.M. (1999). Statistical significance testing as it relates to practice: Use within Professional Psychology: Research and Practice. *Professional Psychology: Research and Practice, 30*(1), 104-105.

Vacha-Haase, T., & Thompson, B. (1998). Further Comments on Statistical Significance Tests. *Measurement and Evaluation in Counseling and Development*, *31*(1), 63-67.

Vacha-Haase, T., Nilsson, J. E., Reetz, D.R., Lance. T.S., & Thompson, B. (2000). Reporting Practices and APA Editorial Policies Regarding Statistical Significance and Effect Size. *Theory and Psychology, 10*(3), 413-425.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604.