VNIVERSITAT DE VALÈNCIA

Facultad de Psicología

Programa de Doctorado en Investigación en Psicología

Tesis Doctoral con Mención Internacional

Tamaño del efecto y su intervalo de confianza y meta-análisis en Psicología

Presentada por: Laura Badenes Ribera

Dirigida por:

Dra. María Dolores Frías Navarro

Dra. Amparo Bonilla Campos

Valencia, Octubre 2016

La Dra. María Dolores Frías Navarro y la Dra. Amparo Bonilla Campos,

de la Universitat de València,

DECLARAN:

Que el trabajo titulado *Tamaño del efecto y su intervalo de confianza y meta-análisis en Psicología*, que presenta Laura Badenes Ribera para la obtención del título de doctor/a, se ha realizado bajo nuestra dirección y cumple los requisitos para poder optar a Mención Internacional.

Y para que así conste y tenga los efectos oportunos, firmamos el presente documento.

María Dolores Frías Navarro Amparo Bonilla Campos

Valencia, 10 de octubre de 2016

Agradecimientos

Para la realización de la Tesis Doctoral con mención internacional presentada, Laura Badenes Ribera ha disfrutado de:

- La ayuda para la formación de personal investigador en formación de carácter predoctoral, en el marco del programa "VALi+d" (ACIF/2013/167), de la Conselleria d´Educació, Cultura i Esport, de la Generalitat Valenciana (España).

La presente Tesis Doctoral se engloba dentro del proyecto de investigación:

- *Impacto de la reforma estadística en Educación y Psicología: de la significación estadística a la estimación de efectos* (Ministerio de Ciencia e Innovación español, Dirección General de Investigación y Gestión del PlanNacional de I+D+i, código EDU2011-22862) Departamento de Metodología de las Ciencias del Comportamiento de la Universitat de València (España).

Para la obtención de la Tesis Doctoral presentada con mención de "Tesis Internacional", Laura Badenes Ribera ha realizado una estancia internacional en:

- Dipartimento di Psicologia, Università degli studi di Torino (Turín, Italia), bajo la dirección del Dr. Claudio Longobardi (desde 01/03/2015 hasta 01/06/2015).

A Manolo y Alfreda, mis padres

A Noelia, mi hermana

A Diego, mi hermano y compañero,

siempre estás en mi corazón y en lo mejor de mis pensamientos.

A mi familia, por lo que soy y quien soy.

ÍNDICE DE CONTENIDOS

RESUMEN

Introducción

La Práctica Basada en la Evidencia[1] (PBE) se define como "*la integración de la mejor evidencia disponible con la experiencia clínica en el contexto de las características, cultura y preferencias del paciente"* (*American Psychological Association* (APA), Presidencial Grupo de Trabajo sobre la Práctica Basada en la Evidencia, 2006, p. 273). Por definición, la PBE se basa en la utilización de la investigación científica en la toma de decisiones en un esfuerzo por producir los mejores servicios posibles en la práctica clínica (Babione, 2010; Sánchez-Meca y Botella, 2010). En consecuencia, la PBE requiere de los profesionales nuevas habilidades como la capacidad para evaluar y jerarquizar la calidad de la evidencia o las investigaciones psicológicas, para proporcionar el mejor servicio posible a los pacientes mediante la incorporación de la mejor evidencia en la experiencia o el juicio profesional, junto a las opiniones de los pacientes (Sackett, Straus, Richardson, Rosenberg, y Haynes, 2000).

Dentro de este proceso de evaluación crítica de la evidencia es crucial conocer y comprender el proceso de la prueba de significación de la hipótesis nula (*Null Hypothesis Significance Testing*, NHST) como herramienta para el análisis de datos, dado que este procedimiento goza de una considerable difusión en la investigación en Psicología, siendo utilizado en la mayor parte de los artículos publicados en revistas del área (Cumming y cols., 2007). En consecuencia, saber cómo interpretar los valores *p* de probabilidad es una competencia básica del profesional en Psicología y en cualquier disciplina en que se aplique la inferencia estadística.

El valor de *p* relacionado con los resultados de una prueba estadística es la probabilidad de obtener los datos observados o un valor más extremo si la hipótesis nula es verdadera (Kline, 2013). La definición es clara y precisa, sin embargo, los conceptos erróneos de los valores *p* siguen siendo numerosos y repetitivos (Badenes-Ribera, Frías-

[1] De acuerdo con Frías-Navarro y Pascual-Llobell (2003) y Sánchez-Meca y Botella (2010) la traducción más correcta del nombre en inglés, *Evidence-Based Practice* sería Práctica Basada en Pruebas, más que Práctica Basada en la Evidencia. Sin embargo, dado que PBE es el término que más se utiliza, es el que se mantiene en esta Tesis Doctoral.

Navarro, y Pascual-Soler, 2015; Falk y Greenbaum, 1995; Haller y Krauss, 2002; Kühberger, Fritz, Lermer, y Scherndl, 2015; Oakes, 1986).

Los errores de interpretación más comunes del valor *p* son la "falacia de la probabilidad inversa", la "falacia de la replicación", la "falacia del tamaño del efecto" y la "falacia de la significación clínica o práctica" (Carver, 1978; Cohen, 1994; Harrison, Thompson, y Vannest, 2009; Kline, 2013; Nickerson, 2000; Wasserstein y Lazar, 2016).

La "falacia de la probabilidad inversa" es la falsa creencia de que el valor de *p* indica la probabilidad de que la hipótesis nula (H_0) es cierta, dado ciertos datos (Pr (H_0|Datos)). Esto significa confundir la probabilidad del resultado, asumiendo que la hipótesis nula es verdadera, con la probabilidad de que la hipótesis nula sea verdadera, dados ciertos datos (Kline, 2013; Wasserstein y Lazar, 2016).

La "falacia de la replicación" vincula el valor de *p* con el grado de replicabilidad del resultado de un estudio. Supone creer erróneamente que el valor de *p* indica el grado de replicabilidad del resultado y su complemento, 1-*p*, a menudo se interpreta como indicación de la probabilidad exacta de replicación (Carver, 1978; Nickerson, 2000).

La "falacia del tamaño del efecto" relaciona la significación estadística con la magnitud del efecto detectado. En concreto, supone creer erróneamente que el valor de *p* proporciona información directa sobre el tamaño del efecto (Carver, 1978). Es decir, que cuanto más pequeño es el valor de *p* más grandes son los tamaños del efecto. Sin embargo, el valor de *p* no informa sobre la magnitud de un efecto. Éste sólo puede ser determinado mediante la estimación directa de su valor con los estadísticos apropiados y su intervalo de confianza (Cumming, 2012; Cumming, Fidler, Kalinowski, y Lai, 2012; Kline, 2013; Wasserstein y Lazar, 2016).

La "falacia de la significación clínica o práctica" es la falsa creencia de que el valor de *p* indica la importancia de los hallazgos (Nickerson, 2000; Wasserstein y Lazar, 2016). De esta manera, un efecto estadísticamente significativo es interpretado como un efecto importante. Sin embargo, un resultado estadísticamente significativo no indica que el resultado sea importante, de la misma manera que un resultado no estadísticamente significativo todavía podría ser importante.

Dados los errores de interpretación del valor de *p* y otras críticas sobre el uso y abuso de del procedimiento de la NHST (e.g., Monterde-i-Bort, Frías-Navarro, y Pascual-Llobell, 2010; Wasserstein y Lazar, 2016), la APA (2001, 2010a) recomendó

reportar los estadísticos de tamaño del efecto y sus intervalos de confianza, que, en conjunto, transmiten más claramente la magnitud de los hallazgos de investigación (Ferguson, 2009).

Existen docenas de estadísticos del tamaño del efecto disponibles (Henson, 2006; Kline, 2013), los cuales se pueden clasificar en dos grandes grupos: las medidas de diferencias de medias y las medidas de la fuerza de las relaciones entre variables (Frías-Navarro, 2011b; Kline, 2013; Rosnow y Rosenthal, 2009). El primero se basa en la diferencia de medias estandarizadas (e.g., *d* de Cohen, *g* de Glass, *g* de Hedges, *f* de Cohen, etc.) y el segundo se basa en la proporción de varianza explicada o la correlación entre dos variables (e.g., R^2/r^2, η^2, w^2).

Los estadísticos del tamaño del efecto reportados con mayor frecuencia son la R^2, *d* de Cohen, y η^2 (e.g., Peng y Chen, 2014). Estos estadísticos han sido criticados por su sesgo (es decir, que tienden a estar positivamente sesgados), su falta de robustez a los valores atípicos, y su inestabilidad bajo las violaciones de los supuestos estadísticos (Grissom y Kim, 2012; Kline, 2013; Wang y Thompson, 2007).

Por último, dentro de este contexto de cambio y avances metodológicos, las revisiones sistemáticas y meta-analíticas han ganado una considerable relevancia y prevalencia en las revistas de mayor prestigio (APA, 2010a; Borenstein, Hedges, Higgins, y Rothstein, 2009). Los estudios meta-analíticos ofrecen varias ventajas sobre las revisiones narrativas: el meta-análisis implica un proceso de investigación con base científica que depende del rigor y la transparencia de cada una de las decisiones tomadas durante su elaboración, y permite dar una respuesta definitiva acerca de la naturaleza de un efecto cuando hay resultados contradictorios (Borenstein y cols., 2009). Los meta-análisis facilitan estimaciones del tamaño del efecto más precisas, permiten evaluar la estabilidad de los efectos, y ayudar a los investigadores a contextualizar los valores de los tamaños del efecto obtenidos en su estudio (Cumming y cols., 2012). Sin embargo, los estudios meta-analíticos no están libres de sesgos, por ejemplo, el sesgo de publicación, que es una de las mayores amenazas para la validez de este tipo de estudios, cuya consecuencia es una sobreestimación del tamaño del efecto (Borenstein y cols., 2009; Sánchez-Meca y Marín-Martínez, 2010). Así, Ferguson y Branninck (2011) analizaron 91 estudios de meta-análisis publicados en la *American Psychological Association* y en la *Association for Psychological Science Journal* y encontraron que de 91 estudios analizados, 26 (41%) reportaron evidencia del sesgo de

publicación. Por lo tanto, los investigadores, los profesionales de la Psicología y, en general, los lectores de los estudios meta-analíticos deben conocer métodos para detectar este tipo de sesgo. En este sentido, el *funnel plot* es una gráfica que se utiliza con frecuencia como método de detección de sesgo de publicación en las Ciencias de la Salud (Sterne, Gavaghan, y Egger, 2005).

En definitiva, es necesario llevar a cabo investigaciones sobre el grado de conocimiento metodológico que los psicólogos académicos y profesionales tienen sobre la calidad metodológica de las evidencias y de la investigación psicológica para la correcta aplicación del enfoque de la PBE y la adquisición de un conocimiento científico válido. Este tipo de investigación puede aportar luz sobre estos problemas y dar lugar a programas de formación para tratar de corregirlos o minimizarlos.

Objetivos

El primer objetivo de este trabajo fue detectar los errores de razonamiento estadístico que los psicólogos académicos y profesionales españoles cometen cuando se les presentan los resultados de una prueba de inferencia estadística. Con este fin, se analizaron dos cuestiones: la primera fue la extensión de los errores más comunes de interpretación con respecto al valor de *p* y la segunda fue el grado en que se interpretan correctamente los valores de *p* por parte de ambos colectivos.

El segundo objetivo fue analizar lo que los psicólogos académicos y profesionales españoles conocen sobre los tamaños del efecto, sus intervalos de confianza y los estudios de meta-análisis, teniendo en cuenta que esta es una de las principales recomendaciones propuestas por la APA (2010a) para mejorar la práctica estadística en la investigación psicológica y favorecer la acumulación de conocimiento y la replicación de los hallazgos.

Por último, se trató de comprobar si los resultados de la investigación sobre los errores de interpretación del valor de *p* y el nivel de conocimiento sobre los tamaños del efecto, sus intervalos de confianza y los meta-análisis, realizados en los psicólogos académicos españoles, son constantes, para lo cual llevamos a cabo sendos estudios de replicación con una muestra de psicólogos académicos chilenos e italianos.

Método

Procedimiento

Se realizaron una serie de estudios transversales mediante encuesta *on-line*. Para ello, se registraron las direcciones de correo electrónico de los psicólogos académicos españoles, chilenos e italianos a través de la consulta de las webs de las universidades en estos países. Los potenciales participantes fueron invitados a completar una encuesta a través del uso de un sistema CAWI (*Computer Assisted Web Interviewin*g). Se envió un mensaje de seguimiento dos semanas después a los potenciales participantes que no habían contestado a la encuesta. La recogida de datos se llevó a cabo durante el año académico 2013-2014 para la muestra española y desde marzo a mayo de 2015 para la muestra chilena e italiana.

En cuanto a la muestra española de psicólogos profesionales, se envió un e-mail a los Colegios Oficiales de Psicólogos invitándoles a participar en la encuesta *on-line* sobre práctica profesional en Psicología. Los potenciales participantes fueron invitados a completar una encuesta a través del uso de un sistema CAWI. Tres semanas después se envió un mensaje de seguimiento. La recogida de datos se llevó a cabo durante los meses de mayo a septiembre de 2015.

Participantes

La muestra de psicólogos académicos españoles estuvo formada por 472 participantes. La media de años de los profesores en la Universidad fue de 13.56 años (*DT* = 9.27). Los hombres representaron 45.8% (*n* = 216) y las mujeres 54.2% (*n* = 256).

La muestra de psicólogos académicos chilenos e italianos estaba compuesta por 194 participantes. De estos 194 participantes, 159 eran italianos y 35 chilenos. De los 159 participantes italianos, 45.91% eran hombres y 54.09% mujeres, con una edad media de 47.65 años (*DT* = 10.47). El número medio de años que los profesores habían pasado en el ámbito académico fue de 12.90 años (*DT* = 10.21). De los 35 psicólogos académicos chilenos, los hombres representaron el 45.71% de la muestra y las mujeres el 54.29. Además, la edad media de los participantes fue de 43.60 años (*DT* = 9.17). El número medio de años que los profesores habían pasado en el ámbito académico fue de 15 años (*DT* = 8.61).

Por último, la muestra de psicólogos profesionales españoles estuvo formada por 77 participantes (68.8% mujeres, 31.2% hombres, edad media de 41.44 años, *DT* = 9.42).

Instrumento

El instrumento aplicado consistió en una encuesta dividida en dos secciones. La primera sección incluía ítems relacionados con información sobre el sexo, la edad y los años de experiencia como psicólogo académico, el área de conocimiento a la que está adscrita, y el tipo de Universidad (pública/privada). Además, para los psicólogos profesionales españoles, la primera sección también incluyó ítems relacionados con los años de experiencia como psicólogo profesional, el entorno clínico (pública/privada), y el grado de familiaridad con el movimiento de la PBE.

La segunda sección incluyó ítems relacionados con el conocimiento sobre aspectos metodológicos relacionados con la PBE, como por ejemplo, la interpretación del valor *p*, el nivel de conocimiento de los estadísticos del tamaño del efecto, intervalos de confianza, estudios de meta-análisis, y las listas de comprobación de la calidad metodológica de los estudios.

Análisis de datos

Todos los estudios incluyeron estadísticos descriptivos de las variables objeto de evaluación, tales como frecuencias y porcentajes. Además, los análisis incluyeron la estimación del intervalo de confianza para los porcentajes. Para el cálculo del intervalo de confianza se utilizaron los métodos de puntuación basados en la obra de Newcombe (2012).

Todos los análisis se realizaron con el programa estadístico SPSS v. 20 de IBM para Windows.

Resultados y conclusiones

Los resultados indican que la comprensión de muchos conceptos estadísticos sigue siendo problemática entre los psicólogos académicos y profesionales españoles, y también entre los psicólogos académicos chilenos e italianos. Los errores metodológicos de interpretación y los pobres conocimientos de determinados estadísticos y procedimientos han sido y continúan siendo una fuente de amenaza directa para una adecuada implementación de la PBE en la práctica profesional y para la adquisición de un conocimiento científico válido.

En cuanto a los errores de interpretación del valor de *p*, la "falacia de la probabilidad inversa" fue la interpretación errónea más prevalente entre los psicólogos académicos españoles, italianos y chilenos. Esto significa que algunos psicólogos académicos confunden la probabilidad de obtener un resultado dado o un resultado más extremo si la hipótesis nula es verdadera (Pr (Datos|H_0) con la probabilidad de que la hipótesis nula sea cierta dados algunos datos (Pr (H_0|Datos).

Además, los psicólogos académicos españoles, italianos y chilenos adscritos al área de Metodología no fueron inmunes a las interpretaciones erróneas del valor de *p*, lo que puede dificultar la formación estadística de los estudiantes y facilitar la transmisión de estas falsas creencias, así como su perpetuación (Haller y Krauss, 2002; Kirk, 2001; Kline, 2013; Krishnan y Idris, 2014). Estos resultados son consistentes con estudios previos (Haller y Krauss, 2002; Lecoutre, Poitevineau, y Lecoutre, 2003; Monterde-i-Bort y cols., 2010).

Por otra parte, la "falacia de la significación clínica o práctica" fue la interpretación errónea más frecuente entre los psicólogos profesionales españoles. Sin embargo, un resultado estadísticamente significativo no indica que el resultado es importante, de la misma manera que un resultado no estadísticamente significativo aún podría ser importante (Nickerson, 2000; Wasserstein y Lazar, 2016). La importancia clínica se refiere a la utilidad práctica o aplicada o a la importancia del efecto de una intervención. Es decir, si produce alguna diferencia real (auténtica, palpable, práctica, notable) para los clientes o para otros con los que interactúan en la vida cotidiana (Kazdin, 1999, 2008).

Las pruebas de significación estadística tienen un propósito y responden a unos problemas y no a otros. Una prueba de significación estadística no indica la importancia de un resultado, la replicabilidad del mismo, o incluso la probabilidad de que un resultado sea debido al azar (Carver, 1978). El valor de *p* nos informa de si existe un efecto, pero no revela el tamaño del efecto, ni su significación clínica/práctica (Ferguson, 2009; Sullivan y Feinn, 2012). El tamaño del efecto sólo puede ser determinado mediante la estimación directa de su valor con los estadísticos apropiados y su intervalo de confianza (Cohen, 1994; Cumming, 2012; Kline, 2013; Wasserstein y Lazar, 2016).

Sin embargo, interpretar un resultado estadísticamente significativo como importante o útil, confundir el nivel de significación de alfa con la probabilidad de que la hipótesis nula sea cierta, relacionar el valor de *p* con la magnitud del efecto, y creer que la probabilidad de replicación de un resultado es 1-*p* son interpretaciones erróneas o falsas creencias que siguen existiendo entre los psicólogos académicos y psicólogos profesionales, como muestran los resultados.

Estos conceptos erróneos son problemas de interpretación y no son un problema del procedimiento de la NHST en sí mismo (Leek, 2014). Detrás de estas interpretaciones erróneas existen algunas creencias y atribuciones acerca de la significación estadística de los resultados. Por lo tanto, es necesario mejorar la enseñanza de la estadística, la formación de los psicólogos y el contenido de los manuales de estadística con el fin de garantizar una formación de alta calidad a los futuros profesionales (Babione, 2010; Cumming, 2012; Kline, 2013; Haller y Krauss, 2002).

Los problemas en la comprensión del valor *p* influencian las conclusiones que los profesionales extraen de sus datos (Hoekstra, Morey, Rouder, y Wagenmakers, 2014), poniendo en peligro la calidad de los resultados de la investigación psicológica (Frías-Navarro, 2011a). El valor de la evidencia científica depende de la calidad de los análisis estadísticos y de su interpretación (Faulkner, Fidler, y Cumming, 2008).

Por otro lado, la mayoría de los participantes en los estudios realizados afirmaron utilizar estudios meta-analíticos en su práctica profesional y tener un conocimiento adecuado sobre los mismos, así como de los estadísticos del tamaño del efecto. Sin embargo, reconocieron que tienen un pobre conocimiento de los gráficos que se utilizan en los meta-análisis, como por ejemplo, el *forest plot* y el *funnel plot*, lo cual puede llevar a una mala interpretación de los resultados y, por lo tanto, dar lugar a una mala práctica, teniendo en cuenta que la mayoría de los participantes declaró que usaba estudios meta-analíticos en su práctica profesional. Como varios autores señalan, la presentación gráfica de los resultados es una parte importante de un meta-análisis y se ha convertido en la principal herramienta para la presentación de los resultados de múltiples estudios sobre la misma pregunta de investigación (Anzures-Cabrera y Higgins, 2010; Borenstein, y cols., 2009). De este modo, el *forest plot* y el *funnel plot* son gráficos utilizados en los estudios de meta-análisis para presentar las estimaciones del tamaño del efecto medio y el sesgo de publicación, respectivamente.

El sesgo de publicación es una importante amenaza para la validez de los estudios meta-analíticos, ya que las estimaciones meta-analíticas derivadas podrían ser imprecisas, típicamente, sobreestimando el efecto. A ese respecto, el *funnel plot* se utiliza como método de detección del sesgo de publicación en las Ciencias de la Salud (Sterne y cols., 2005). Por lo tanto, investigadores, académicos y profesionales deben tener un conocimiento adecuado de este tipo de gráfica, que es una herramienta básica de los estudios de meta-análisis para detectar el sesgo de publicación y la heterogeneidad de los tamaños de efecto.

Con respecto al tipo de estadístico del tamaño del efecto que conocen los participantes, estos mencionaron en mayor medida los estadísticos de la familia de las diferencias de medias estandarizadas y η^2 (estadísticos del tamaño del efecto paramétricos). Sin embargo, estos estadísticos del tamaño del efecto han sido criticados por su falta de robustez frente a los valores atípicos o desviación de la normalidad, y la inestabilidad bajo las violaciones de los supuestos estadísticos (Algina, Keselman, y Penfield, 2005; Grissom y Kim, 2012; Kline, 2013; Peng y Chen, 2014; Wang y Thompson, 2007). Hay razones teóricas y evidencia empírica de que los valores atípicos y las violaciones de los supuestos estadísticos son comunes en la práctica (Erceg-Hurn y Mirosevich, 2008; Grissom y Kim, 2001).

Los resultados sugieren que la mayoría de los psicólogos académicos y profesionales españoles y los psicólogos académicos italianos y chilenos no conocen las alternativas para los estadísticos del tamaño del efecto paramétricos, tales como los estadísticos no paramétricos (e.g., correlación de Spearman), los estadísticos robustos de la diferencia de medias estandarizada (basados en las medias recortadas y varianzas winsorizada), la probabilidad de superioridad (PS), el número necesario a tratar (NNT) o el área bajo la curva ROC (AUC) (Erceg-Hurn y Mirosevich, 2008; Ferguson, 2009; Grissom y Kim, 2012; Keselman, Algina, Lix, Wilcox, y Deerin, 2008; Kraemer y Kupfer, 2006; Peng y Chen, 2014; Wilcox, 2010; Wilcox y Keselman, 2003). Como Erceg-Hurn y Mirosevich (2008) señalaron esto podría ser debido a la falta de exposición a estos métodos. De esta manera, "*el plan de estudios de estadística en Psicología, los artículos de las revistas, los manuales populares, y el software están dominados por la estadística desarrollada antes de la década de 1960*" (*op. cit*., p. 593).

En cuanto a las listas de control de la calidad metodológica de los estudios, de nuevo la mayor parte de los participantes dijeron no tener conocimiento sobre ellas. Sin embargo, éste es un campo en expansión y actualmente existen listas de comprobación para estudios primarios (por ejemplo, CONSORT), para estudios de meta-análisis clásicos (por ejemplo, AMSTAR) y para estudios de meta-análisis en red (por ejemplo, PRISMA-NMA).

Por otro lado, el análisis del comportamiento de los investigadores asociado con sus prácticas metodológicas señala que, en las tres muestras de psicólogos académicos, los participantes que podían nombrar algún estadístico del tamaño del efecto presentaron un perfil más cerca de las buenas prácticas estadísticas y de diseño de investigación. Sin embargo, hay tres temas de alerta en relación al conocimiento que los psicólogos académicos españoles, chilenos e italianos tienen acerca del tamaño del efecto y la validez de la conclusión estadística: asocian erróneamente el tamaño del efecto con la importancia de un hallazgo ("falacia de la significación clínica o práctica"), siguen utilizando en una alta proporción expresiones del valor *p* que giran en torno al oráculo del valor alfa, y no conocen el propósito de planificar *a priori* la potencia estadística en un estudio.

Por último, dos acontecimientos que han permitido el debate en la ciencia sobre procedimientos estadísticos, el progreso hacia una reforma estadística y una mayor transparencia y calidad de los estudios, como son el debate abierto sobre los usos y abusos de las pruebas de significación estadística (que comenzó casi desde el inicio de su uso) y el desarrollo de herramientas de verificación como los listados de comprobación (CONSORT, STROBE, PRISMA...), siguen siendo desconocidos para una alta proporción de psicólogos académicos españoles, italianos y chilenos y entre los psicólogos profesionales españoles.

Por lo tanto, el presente trabajo proporciona evidencia de la necesidad de formación estadística de los psicólogos académicos y profesionales españoles, y de los psicólogos académicos chilenos e italianos, teniendo en cuenta los problemas relacionados con la interpretación adecuada de los resultados obtenidos con el procedimiento NHST y el pobre conocimiento de términos estadísticos del tamaño del efecto, estudios meta-analíticos y listas de control de la calidad metodológica.

La PBE requiere tener un conocimiento adecuado sobre los fundamentos de la metodología de investigación con el fin de ser capaces de evaluar críticamente los tests y las evidencias que los estudios incluyen en sus informes. Los problemas de comprensión del valor *p* de probabilidad, de los estadísticos del tamaño del efecto y de los estudios meta-analíticos, influyen en las conclusiones que los profesionales extraen de los datos, lo que pone en peligro la calidad de los resultados de la investigación psicológica y una adecuada implementación de una PBE en la práctica profesional. Como Faulkner y cols. (2008) señalan, el valor de la evidencia científica depende de la calidad de los análisis estadísticos realizados y de su interpretación. Por lo tanto, la interpretación de los resultados es un filtro de calidad que no puede ser sometido a las creencias erróneas o pobres interpretaciones del procedimiento estadístico.

No obstante, varias limitaciones en la serie de estudios realizados en este trabajo deben ser reconocidas. Por ejemplo, la baja tasa de respuesta podría afectar a la representatividad de las muestras y, por lo tanto, a la generalización de los resultados entre los psicólogos académicos y profesionales. Sin embargo, es posible que los participantes que respondieron a la encuesta se sintieran más seguros de su conocimiento estadístico que aquellos que no respondieron. Si este fuera el caso, los resultados podrían subestimar las barreras a la PBE. Además, los resultados de nuestra investigación sobre concepciones erróneas del valor *p* están de acuerdo con los resultados de estudios anteriores sobre este tema en muestras de psicólogos académicos y estudiantes de Psicología (Badenes-Ribera, Frías-Navarro y Pascual Soler, 2015; Falk y Greenbaum, 1995; Haller y Krauss, 2002; Kühberger y cols., 2015; Monterde-i-Bort, y cols., 2010; Oakes, 1986).

Por otra parte, los resultados de la investigación sobre el nivel de conocimiento de la magnitud del efecto y los estudios de meta-análisis en las muestras de psicólogos españoles (ambos grupos, psicólogos profesionales y académicos) fueron consistentes con los resultados del estudio sobre estos temas en la muestra de psicólogos académicos italianos y chilenos.

Todo esto lleva a concluir en la necesidad de formar adecuadamente a los psicólogos para mejorar la práctica profesional. La PBE requiere de profesionales que evalúen críticamente los resultados de la investigación psicológica. Para ello, se requiere una formación adecuada en conceptos estadísticos, metodología y diseños de

investigación, así como en los resultados de las pruebas de inferencia estadística y en los estudios de meta-análisis.

Por ejemplo, los manuales de estadística deberían incluir una sección sobre el actual debate y las críticas del procedimiento NHST, en términos de si las pruebas de significación estadística son la mejor manera de avanzar en la adquisición de un conocimiento científico válido. Además, deberían añadir información sobre cómo calcular e informar el tamaño del efecto y sus intervalos de confianza, tanto en los resultados estadísticamente significativos y como en los resultados no estadísticamente significativos. Y, por último, los autores de los manuales deberían dar ejemplos de cómo decidir si un resultado estadísticamente significativo tiene importancia práctica o clínica (Gliner, Leech, y Morgan, 2002). Por otra parte, los programas de software estadístico deberían actualizarse para incluir en sus menús otras técnicas como la estimación de los intervalos de confianza de los estadísticos del tamaño del efecto paramétricos, y la estimación de estadísticos del tamaño el efecto más resistentes a los valores extremos (*outliers*) y a las violaciones de los supuestos de las pruebas paramétricas (normalidad de la variable y homogeneidad de la varianza), tales como los estadísticos robustos modernos y sus intervalos de confianza. En ese sentido, hay varios sitios web que ofrecen programas para el cálculo de los estimadores del tamaño del efecto y sus intervalos de confianza (ver Frías-Navarro, 2011b; Fritz, Morris, y Richler, 2012; Grissom y Kim, 2012; Kline, 2013; Peng, Chen, Chiang y Chiang, 2013).

En definitiva, el objetivo de esta serie de estudios ha sido especialmente hacer hincapié en la necesidad de una re-educación estadística de los psicólogos profesionales y académicos, que incluye la difusión del uso de las listas de control, como una herramienta para evaluar la calidad metodológica de los estudios, y motivar el desarrollo de manuales que describan conceptualmente las pruebas estadísticas y señalen las consecuencias de una mala práctica estadística en la acumulación de conocimientos científicos válidos. Además, el propósito ha sido tener en cuenta la necesidad de incorporar los modernos estadísticos robustos del tamaño del efecto a los programas estadísticos como el SPSS.

En la actualidad existe un debate científico y social abierto que podría cambiar el curso de las prácticas estadísticas entre los investigadores de la Psicología y las Ciencias de la salud. Por ejemplo, durante los últimos tres años las críticas contra el procedimiento de inferencia estadística clásica basada en el valor de probabilidad *p* y la

decisión dicotómica para mantener o rechazar la hipótesis nula se han endurecido (Allison, Brown, George, y Kaiser, 2016; Nuzzo, 2014; Wasserstein y Lazar, 2016). Además, la baja proporción de estudios de replicación, el sesgo de publicación que conducen a una sobreestimación de la magnitud de los efectos, las prácticas estadísticas cuestionables (*Questionable Research Practices*, QRPs) dirigidas a alcanzar resultados estadísticamente significativos como no informar de los resultados de todas las variables dependientes medidas en el estudio, informar solamente de los resultados estadísticamente significativos, eliminar los valores extremos o 'outliers' y aumentar la muestra hasta lograr la significación estadística (*p-hacking*) y el fraude también son temas actuales de discusión (Earp y Trafimow, 2015; Ioannidis, 2005a, 2005b; Kepes, Banks, y Oh, 2014). Debates a los que ha tratado de contribuir la realización del presente trabajo, aportando evidencias del actual estado de la cuestión, en lo que se refiere al conocimiento y las prácticas de los psicólogos académicos y profesionales en relación a la metodología y los diseños de investigación.

Los hallazgos del presente trabajo son una prueba empírica de todas las conductas inapropiadas que rodean al proceso de inferencia estadística y que durante décadas han sido objeto de estudio por los investigadores, como son las interpretaciones inadecuadas y el mal uso que se realiza de las técnicas de inferencia debido a las falacias estadísticas y de tamaño del efecto que la rodean. Profesores, científicos y profesionales de la Psicología no son inmunes a tales creencias. El problema no se ha resuelto a pesar de las recomendaciones y alertas que de manera permanente se han detallado en las publicaciones científicas. La re-educación estadística que corrija los errores de interpretación de las diferentes falacias y la incorporación de una práctica estadística basada en la evidencia, orientada al uso consciente y explícito de todos los elementos que rodean al proceso de inferencia estadística, es esencial para interpretar de forma crítica sus resultados.

La literatura que se ha desarrollado sobre el razonamiento estadístico y su educación tiene toda una línea de investigación abierta sobre los errores de interpretación de los valores de *p* (Beyth-Maron, Fidler y Cumming, 2008; Garfield, Ben-Zvi, Chance, Medina, Roseth, y Zieffler, 2008; Garfield, y Franklin, 2011; Garfield, Zieffler, Kaplan, Cobb, Chance, y Holcomb, 2011), a la cual se pretende sumar la presente investigación, poniendo en evidencia su importancia, su vigencia y sus implicaciones en el desarrollo y la transmisión del conocimiento científico válido.

ABSTRACT

Introduction

Evidence-Based Practice (EBP) is defined as "*the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences*" (APA, Presidential Task Force on Evidence Based Practice, 2006, p. 273). By definition, EBPP relies on the utilization of scientific research in decision making in an effort to produce the best possible services in clinical practice (Babione, 2010; Sánchez-Meca & Botella, 2010). Consequently, EBP requires to professionals new skills as the ability to critically evaluate and rank the quality of evidence or psychological research to provide the best possible service to patients by incorporating the best evidence into experience or professional judgment and opinions of patients (Sackett et al., 2000).

Within this process of critical evaluation of evidence it is crucial knowing and understanding the process of the Null Hypothesis Significance Testing (NHST) as tool to data analysis, given that this procedure enjoys considerable diffusion in Psychology (Cumming et al., 2007). For example, these authors found that 97% of the articles published in Psychology journals use the NHST. Consequently, knowing how to interpret *p* values of probability is a core competence of the professionals in Psychology and any discipline where statistical inference is applied.

The *p*-value linked to the results of a statistical test is the probability of witnessing the observed result or a more extreme value if the null hypothesis was true (Kline, 2013). The definition is clear and precise, however, the misconceptions of the *p*-value continue to be numerous and repetitive (Badenes-Ribera, Frías-Navarro, & Pascual-Soler, 2015; Falk & Greenbaum, 1995; Haller & Krauss, 2002; Kühberger et al., 2015; Oakes, 1986; Wasserstein & Lazar, 2016).

The most common misconceptions of the *p*-value are the "inverse probability fallacy", the "replication fallacy", the "effect size fallacy" and the "clinical or practical significance fallacy" (Carver, 1978; Cohen, 1994; Harrison et al., 2009; Kline, 2013, Nickerson, 2000; Wasserstein & Lazar, 2016).

The "inverse probability fallacy" is the false belief that the p-value indicates the probability that the null hypothesis (H_0) is true, given certain data ($Pr(H_0|$ Data)). It means confusing the probability of the result, assuming that the null hypothesis is true, with the probability of the null hypothesis, given certain data (Kline, 2013; Wasserstein & Lazar, 2016).

The "replication fallacy" links p-value to the degree of replicability of the result. Consequently, it is the false belief that the p-value indicate the degree of replicability of the result and its complement, 1-*p*, is often interpreted as an indication of the exact probability of replication (Carver, 1978; Nickerson, 2000).

The "effect size fallacy" relates statistical significance to the size of the detected effect. Specifically, it involves the false belief that the *p*-value provides direct information about the effect size (Carver, 1978). That is, supposing that the smaller is the *p* value are larger the effect sizes. However, the *p*-value does not report on the magnitude of an effect. The effect size can only be determined by directly estimating its value with the appropriate statistic and its confidence interval (Cumming, 2012; Cumming et al., 2012; Kline, 2013; Wasserstein & Lazar, 2016).

The "clinical or practical significance fallacy" is the false belief that the *p*-value indicates the importance of the findings (Nickerson, 2000; Wasserstein & Lazar, 2016). In this way, a statistically significant effect is interpreted as an important effect. Nevertheless, a statistically significant result does not indicate that the result is important, in the same way that a non-statistically significant result might still be important.

Given the misconceptions of the *p*-value and other critics about the use and abuse of NHST (e.g., Monterde-i-Bort et al., 2010; Wasserstein & Lazar, 2016), the American Psychological Association (APA, 2001, 2010a) strongly recommended the reporting of effect sizes (ES) and their confidence intervals (CIs), which, taken together, clearly convey the importance of the research findings (Ferguson, 2009).

There are dozens of effect size measures available (Henson, 2006; Kline, 2013). Nevertheless, they can be classified into two broad groups: measures of mean differences and measures of strength of relations (Frías-Navarro, 2011b; Kline, 2013; Rosnow & Rosenthal, 2009). The former is based on the standardized group mean difference (e. g. Cohen's *d*, Glass's *g*, Hedges' *g*, Cohen's *f*); the latter is based on the

proportion of variance accounted for or correlation between two variables (e. g., R^2/r^2, η^2, w^2).

The most frequently reported ES measures are the unadjusted R^2, Cohen's *d*, and η^2 (e.g., Peng & Chen, 2014). These statistics have been criticized for bias (i.e., they tend to be positively biased), lack of robustness to outliers, and instability under violations of statistical assumptions (Grissom & Kim, 2012; Kline, 2013; Wang & Thompson, 2007).

Finally, within this context of change and methodological advances, systematic, meta-analytic reviews of studies have gained considerable relevance and prevalence in the most prestigious journals (APA, 2010a; Borenstein et al., 2009). Meta-analytic studies offer several advantages over narrative reviews: meta-analysis involves a scientifically-based research process that depends on the rigor and transparence of each of the decisions made during its elaboration, and it can provide a definitive answer about the nature of an effect when there are contradictory results (Borenstein et al., 2009). Meta-analyses facilitate more precise ES estimations, they make it possible to rate the stability of the effects, and they help researchers to contextualize the ES values obtained in their study (Cumming et al., 2012). Nevertheless, meta-analytic studies are not free of bias, such as the publication bias, which is one the greatest threats to the validity of meta-analytic reviews. For example, Ferguson & Branninck (2011) analyzed 91 meta-analytic studies published in American Psychological Association and Association for Psychological Science journal and they found that of the 91 studies analyzed, 26 (41%) reported evidence of publication bias. The consequence of publication bias is an overestimation of effect size (Borenstein et al., 2009; Sánchez-Meca & Marín-Martínez, 2010). Therefore, researchers and readers of meta-analytic studies (such as, practitioner psychologists) should know methods for detecting this bias. In this way, funnel plot is a graphical that is used frequently as publication bias detection method in the health sciences (Sterne et al., 2005).

Therefore, it is needed to carry out research on the degree of methodological knowledge that academic psychologists and practitioner psychologists have about methodological quality of evidence (or psychological research) for proper implementation of EBP approach. This kind of research may bring light about these issues and lead to develop training programs.

Objectives

The first purpose of these works was to detect the statistical reasoning errors that Spanish academic psychologists and Spanish practitioner psychologists make when presented with the results of a statistical inference test. To this end, two questions have been analyzed. The first was the extension of the most common misconceptions of the *p* value and the second was the extent to which *p* values are correctly interpreted.

The second purpose was to analyze what Spanish academic psychologists and Spanish practitioner psychologists know about ES, their CIs, and meta-analyses, given that this is one of the main recommendations proposed by the APA (2010) to improve statistical practice and favor the accumulation of knowledge and the replication of findings.

Finally, to check whether the results of the research on misconception of the *p*-value and the level of knowledge of effect sizes, confidence intervals and meta-analysis conducted in Spanish academic psychologists are reliable, it has been carried out a replication study with a sample of Chilean and Italian academic psychologists.

Method

Procedure

Several cross-sectional studies were carried out through on-line survey. For this purpose, the e-mail addresses of Spanish, Chilean and Italian academic psychologists were found by consulting the webs of the universities at these countries. Potential participants were invited to complete a survey through the use of a CAWI (Computer Assisted Web Interviewing) system. A follow-up message was sent two weeks later to non-respondents. The data collection was performed during the 2013-2014 academic year for Spanish sample and from March to May 2015 for Chilean and Italian sample.

Regarding Spanish practitioner psychologists sample, it was send an e-mail to Spanish Psychological Associations inviting them to participate in the on-line survey on professional practice in Psychology. Potential participants were invited to complete a survey through the use of a CAWI system. A follow-up message it was sent three weeks later. The data collection was performed from May to September 2015.

Participants

The sample of Spanish academic psychologists consisted of 472 participants. The mean number of years of the professors at the University was 13.56 years (*SD* = 9.27). Men represented 45.8% (*n* = 216) and women 54.2% (*n* =256).

The sample of Chilean and Italian academic psychologists was comprised of 194 participants. Of these 194 participants, 159 were Italian and 35 were Chilean. Of the 159 Italians participants, 45.91% were men and 54.09% were women, with a mean age of 47.65 years (*SD* = 10.47). The mean number of years that the professors had spent in academia was 12.90 years (*SD* = 10.21). Of the 35 Chilean academic psychologists, men represented 45.71% of the sample and women54.29%. In addition, the mean age of the participants was 43.60 years (*SD* = 9.17). The mean number of years that the professors had spent in academia was 15 years (*SD* = 8.61).

Finally, the sample of Spanish practitioner psychologists consisted of 77 participants (68.8% women and 31.2% men, average age of 41.44 years, *SD* = 9.42).

Instrument

The instrument applied consisted of a survey divided in two sections. The first one included items related to information about sex, age and years of experience as academic psychologist, Psychology knowledge area, kind of university (public/private). In addition, for Spanish practitioner psychologists the first section included items related to years of experience as practitioner psychologist, clinical setting (public or private), and degree of familiarity with EBP movement.

The second section included items related to the knowledge on methodological issues associated with EBP, such as misconceptions of the *p*-value, level of knowledge about effect size statistics, confidence intervals, meta-analysis studies, and checklists of methodological quality of the studies.

Data analysis

All of the studies included descriptive statistics for the variables under evaluation such as frequencies and percentage. In addition, they included confidence interval for percentages (CIs). To calculate the CIs for percentages we used score methods based on the works of Newcombe (2012).

All analyses were performed with the statistical program IBM SPSS v. 20 for Windows.

Results and conclusions

The findings indicate that the comprehension of many statistical concepts continues to be problematic among Spanish academic and practitioner psychologists, and among Chilean and Italian academic psychologists. The methodological errors and the poor methodological knowledge have been and continue to be a source of direct threat to properly implement the EBP in professional practice and getting valid scientific knowledge.

Regarding misconceptions of the *p*-value, the "inverse probability fallacy" was the most frequently observed misinterpretation among Spanish, Italian and Chilean academic psychologists. This means that some academic psychologists confuse the probability of obtaining a result or a more extreme result if the null hypothesis was true ($Pr(Data|H_0)$) with the probability that the null hypothesis is true given some data ($Pr(H_0|Data)$).

In addition, Spanish, Italian and Chilean academic psychologists from the area of Methodology were not immune to erroneous interpretations of the *p*-value, and this can hinder the statistical training of students and facilitate the transmission of these false beliefs, as well as their perpetuation (Haller & Krauss, 2002; Kirk, 2001; Kline, 2013; Krishnan & Idris, 2014). These findings are consistent with previous studies (Haller & Krauss, 2002; Lecoutre et al., 2003; Monterde-i-Bort et al., 2010).

On the other hand, "clinical or practical significance fallacy" was the most frequently observed misinterpretation among Spanish practitioner psychologists. Nevertheless, a statistically significant result does not indicate that the result is important, in the same way that a non-statistically significant result might still be important (Nickerson, 2000; Wasserstein & Lazar, 2016). Clinical significance refers to the practical or applied value or importance of the effect of an intervention. That is, whether it makes any real (e.g., genuine, palpable, practical, noticeable) difference to the clients or to others with whom they interact in everyday life (Kazdin, 1999, 2008).

Statistical significance tests have a purpose and respond to some problems and not to others. A statistical significance test does not speak about result importance, replicability, or even the probability that a result was due to chance (Carver, 1978). *P*-

value informs us whether an effect exists, but the *p*-value does not reveal the size of the effect, and neither the clinical/practical significance of the effect (Ferguson, 2009; Sullivan & Feinn, 2012). The effect size can only be determined by directly estimating its value with the appropriate statistic and its confidence interval (Cohen, 1994; Cumming, 2012; Kline, 2013; Wasserstein & Lazar, 2016).

Nevertheless, interpreting a statistically significant result as important or useful, confusing the alpha's significance level with the probability that the null hypothesis is true, relating *p*-value to magnitude effect, and believing that the probability of replicating a result is 1-*p* are erroneous interpretations or false beliefs that continue to exist among academic psychologists and practitioner psychologists, like the results of the studies conducted show.

These misconceptions are interpretation problems and they are not a problem of NHST itself (Leek, 2014). Behind these erroneous interpretations are some beliefs and attributions about the significance of the results. Therefore, it is necessary to improve the statistical education and training of psychologists and the content of statistics textbooks in order to guarantee high quality training of future professionals (Babione, 2010; Cumming, 2012; Kline, 2013; Haller & Krauss, 2002).

Problems in understanding the *p* value influence the conclusions that professionals draw from their data (Hoekstra et al., 2014), jeopardizing the quality of the results of psychological research (Frías-Navarro, 2011a). The value of the evidence depends on the quality of the statistical analyses and their interpretation (Faulkner et al., 2008).

On the other hand, most of the participants reported using meta-analytic studies in their professional practice and having adequate knowledge about them, including effect size statistics. Nevertheless, they acknowledged having a poor knowledge of graphical displays for meta-analyses, such as, forest plot and funnel plot, which may become in a misinterpretation of results and, therefore, lead to bad practice, taking into account that most of the participants said that they used meta-analytic studies in their professional practice. As several authors point out, the graphical presentation of results is an important part of a meta-analysis and it has become the primary tool for presenting the results of multiple studies on the same research question (Anzures-Cabrera & Higgins, 2010; Borenstein, et al., 2009, Botella & Sánchez-Meca, 2015). In this way,

forest plot and funnel plot are graphics used in meta-analytic studies to present pooled effect size estimates and publication bias, respectively.

Publication bias is an important threat to the validity of meta-analytic studies, since meta-analytically derived estimates could be inaccurate, typically overestimated. The funnel plot is used as a publication bias detection method in the health sciences (Sterne et al., 2005). Therefore, researchers, academics, and practitioners must adequately know funnel plots, which is a basic tool of meta-analytic studies to detect bias publication and heterogeneity of effect sizes.

With regard to type of effect size statistic they know, the participants mentioned to a greater degree the effect size statistics from the family of standardized differences in means and η^2 (parametric effect size statistics). Nevertheless, these effect size statistics have been criticized for lack of robustness against outliers or departure from normality, and instability under violations of statistical assumptions (Algina et al., 2005; Grissom & Kim, 2012; Kline, 2013, Peng & Chen, 2014; Wang & Thompson, 2007). There are theoretical reasons and empirical evidence that outliers and violations of statistical assumptions are common in practice (Erceg-Hurn & Mirosevich, 2008; Grissom & Kim, 2001). The findings suggest that the most of the Spanish academic psychologists, Spanish practitioner psychologists and Italian and Chilean academic psychologists do not know the alternatives for parametric effect size statistics such as, non-parametric statistics (e.g., Spearman correlation), the robust standardized mean difference (trimmed means and winsorized variances), the probability of superiority (PS), the number needed to treat (NNT), or the area under the ROC Curve (AUC) (Erceg-Hurn & Mirosevich, 2008; Ferguson, 2009; Grissom & Kim, 2012; Keselman et al., 2008; Kraemer & Kupfer, 2006; Peng & Chen, 2014; Wilcox, 2010; Wilcox & Keselman, 2003). As Erceg-Hurn and Mirosevich (2008) pointed out this might be due to lack of exposure to these methods. In this way, "the psychology statistics curriculum, journal articles, popular textbooks, and software are dominated by statistics developed before the 1960s" (*op. cit*., p.593).

Concerning the methodological quality checklists, again most of the participants said not having knowledge about them. Nevertheless, this is an expanding field and currently there are checklists for primary studies (e.g., CONSORT), for meta-analytic studies (e.g., AMSTAR) and for network meta-analytic studies (e.g., PRISMA-NMA).

On the other hand, the analysis of the researcher's behavior associated with its methodological practices point out that Spanish, Chilean and Italian academic psychologists who could give a name of effect size statistics presented a profile more close to good statistical practices and design research. Nevertheless, three issues alert on the knowledge that both groups of academics have about effect size and validity of statistical conclusion in general: they associate wrongly effect size with the importance of a finding (clinical or practical significance fallacy), they continue to use in a high proportion *p*-value expressions that revolve around the oracle of the value of alpha, and they don't know the purpose of planning a priori statistical power in a study.

Finally, two events that have allowed the science debate on statistical procedures, progress towards a statistical reform and greater transparency and quality of studies, such as the open debate on the uses and abuses of statistical significance tests (which started almost since the beginning of its use) and the development of check tools such checklists (CONSORT, STROBE, PRISMA…), continue to be unknown in a high proportion by Spanish academic psychologists, Spanish practitioner psychologists and Italian and Chilean academic psychologists.

Therefore, the present work provides evidence of the need for statistical training, given the problems related to adequately interpreting the results obtained with the NHST procedure and the poor knowledge of effect size statistical terms, meta-analytic studies and methodological quality checklists that Spanish academic psychologists, Spanish practitioner psychologists and Italian and Chilean academic psychologist have.

The EBP requires having adequate knowledge about the fundamentals of research methodology in order to be able to critically evaluate the tests or evidence that studies include in their reports. The problems in understanding the *p*-value of probability, effect size statistics and meta-analytic studies influence the conclusions that professionals draw from the data, which jeopardizes the quality of the results of psychological research and a proper implementation of EBP in professional practice. As Faulkner et al. (2008) point out, the value of the evidence depends on the quality of the statistical analyses and their interpretation. Therefore, the interpretation of the findings is a quality filter that cannot be subjected to erroneous beliefs or poor interpretations of the statistical procedure.

Nevertheless, it must be acknowledged several limitations in this series of studies. For instance, the low response rate might affect the representativity of the sample and, therefore, the generalizability of the findings among academic and practitioner psychologists. Nevertheless, it is possible that the participants who responded to the survey felt more confident about their statistical knowledge than those who did not respond. Should this be the case, the results might underestimate the barriers to EBP.

In addition, the findings of the research on misconceptions of the *p*-value agree with the results of previous studies about this topic in samples of academic psychologists and undergraduates of Psychology (Badenes-Ribera, Frías-Navarro & Pascual-Soler, 2015; Falk & Greenbaum, 1995; Haller & Krauss, 2002, Kühberger et al., 2015; Monterde-i-Bort et al., 2010; Oakes, 1986).

Furthermore, the findings of the research on knowledge level of effect size and meta-analytic studies in the samples of Spanish psychologists (both groups, practitioner and academic psychologists) were consistent with the results of the study on these topics in Italian and Chilean sample.

All of this leads us to conclude with the need to adequately training psychologists to improve the professional practice. EBP requires professionals to critically evaluate the findings of psychological research and studies. In order to do so, training is necessary in statistical concepts, research design methodology, and the results of statistical inference tests and meta-analytic studies.

For example, textbooks of statistics should include a section on the current debate and criticisms of the NHST procedure, in terms of whether statistical significance tests are the best way to advance the body of valid scientific knowledge. Moreover, they should add information about how to calculate and report the effect size and its confidence intervals, both in statistically significant results and in the non-significant ones. And finally, the authors should give examples in order to decide whether the result has practical or clinical importance (Gliner et al., 2002). On the other hand, statistical software programs should be updated to include in their menus other techniques such as the estimation of confidence intervals for parametric effect size statistics, and the estimation of effect size statistics more resistant to extreme values (outliers) and violations of the assumptions of the parametric tests (normal distribution

and homogeneity of variance), such as modern robust effect size statistics and their confidence intervals. There are several websites that offer routines/programs for computing general or specific effect size estimators and their confidence intervals (see Frías-Navarro, 2011b; Fritz et al., 2012; Grissom y Kim, 2012; Kline, 2013; Peng et al., 2013).

To conclude, the purpose of these studies has been especially to emphasize the need for statistical re-education among practitioner and academic psychologists, to disseminate the use of checklists, as a tool for assessing methodological quality of studies, and to motivate the development of manuals that conceptually describe statistical tests and point out the consequences of bad statistical practice on the accumulation of scientific knowledge. Also, the purpose has been to note the need for incorporating the modern robust effect size statistics in statistical programs, such as SPSS.

Currently there is an open scientific and social debate that could change the course of statistical practices among researchers. For example, during the last three years criticism against the classical statistical inference procedure based on the probability value *p* and the dichotomous decision to keep or reject the null hypothesis has been hardened (Allison et al., 2016; Nuzzo, 2014; Wasserstein & Lazar, 2016). In addition, the low proportion of replication studies, publication bias that lead to an overestimation of the magnitude of effects, questionable statistical practices (Questionable Research Practices, QRPs) leading to find statistically significant results (called *p*-hacking), such as recording many response variables and deciding which to report after the analysis, reporting only statistically significant results, remove outliers and increase sample size to get statistical significance, and fraud also are current issues of discussion (Earp & Trafimow, 2015; Ioannidis, 2005a; Kepes et al., 2014).

The realization of this study has tried to contribute to this debate, providing evidence of the current state of affairs, in what refers to the knowledge and practices of academic and professional psychologists in relation to the methodology and research designs.

The findings of the present work are an empirical evidence of all inappropriate behaviors surrounding the process of statistical inference and that for decades have been studied by researchers, such as misinterpretations and misuse of statistical inference

techniques due to statistic and effect size fallacies that surround it. Academics, scientists and professionals are not immune to such beliefs. The problem has not been resolved despite the recommendations and alerts that have been permanently detailed in scientific publications. Statistical-reeducation to correct the errors of interpretation of the various fallacies and incorporating an Evidence Based Statistical Practice oriented to the conscious and explicit use of all elements surrounding the process of statistical inference is essential to interpret critically the results of statistical inference.

The literature that has been developed on statistical thinking and its education has a whole line of research open on this issue (Beyth-Maron et al., 2008; Garfield et al., 2008; Garfield, & Franklin, 2011; Garfield et al., 2011), to which might be added this investigation, highlighting its importance, its validity and its implications for the development and transmission of scientific knowledge.

INTRODUCCIÓN

El Código ético de la Asociación Americana de Psicología (APA, 2010b) y el Código Deontológico del Psicólogo español (Consejo General de Colegios Oficiales de Psicólogos, 2010) destacan la responsabilidad de los profesionales de la Psicología de ofrecer tratamientos con apoyo empírico que garanticen la validez de sus efectos. Por lo tanto, de acuerdo con estas directrices, la práctica profesional en el ámbito de la Psicología debe estar basada en las mejores pruebas o evidencias científicas. Y el logro de las mejores evidencias (entendidas como pruebas) requiere la planificación y ejecución de diseños de investigación que aporten resultados válidos representativos de la realidad estudiada.

La investigación basada en la evidencia implica la aplicación de procedimientos objetivos, rigurosos y sistemáticos que aporten conocimiento válido y fiable. La calidad del conocimiento científico requiere que los investigadores planifiquen adecuadamente su investigación, la ejecuten eficientemente, analicen los datos correctamente, interpreten bien los resultados y presenten de forma clara las conclusiones (Frías-Navarro, 2011a, Wasserstein y Lazar, 2016). Y además requiere que los avances científicos se difundan adecuadamente y faciliten la actividad de los profesionales.

Históricamente, los científicos sociales de Ciencias de la Educación, Ciencias Biológicas y especialmente los psicólogos, han confiado en el procedimiento de la "Prueba de Significación de la Hipótesis Nula" (*Null Hypothesis Significance Testing*, NHST) como la técnica por excelencia en el análisis de datos (Gigerenzer, 2004; Huberty, 1993; Lovell, 2013; Ludbrook, 2013; Perezgonzalez, 2015).

El procedimiento de contraste de la significación de la hipótesis nula (NHST) parte de la suposición de que la hipótesis nula es cierta (generalmente '*nil hypothesis*' o hipótesis de efecto cero) y bajo dicho supuesto se calcula la distribución del estadístico en todas las posibles muestras de la población. A partir de ahí se calcula la probabilidad del valor concreto (o un dato más extremo) de un estadístico obtenido en la muestra. Posteriormente, se observa la probabilidad del estadístico dentro de su distribución y se decide si se mantiene la hipótesis nula ($p >$ alfa) o se puede rechazar ($p \leq$ alfa), asumiéndose un error de Tipo II y Tipo I respectivamente. Por lo tanto, el valor *p* es la

probabilidad de obtener el valor del estadístico (o más extremo) en caso de ser cierta la hipótesis nula.

De acuerdo con esta lógica, se entiende que un nivel de significación del 5% indica que, en promedio, 5 de cada 100 veces que la hipótesis nula sea cierta se rechazará por azar, cuantificando de este modo el error aleatorio (pero no dice que 5 de cada 100 veces que rechacemos la hipótesis nos estaremos equivocando, pues se asume que H_0 es cierta y la prueba no demuestra ni su falsedad ni su certeza). Finalmente, si el valor *p* del estadístico se juzga como pequeño (incompatible con el modelo de la hipótesis nula bajo cuya certeza se estima el valor del estadístico de los datos), entonces se rechazará la hipótesis nula y se aceptará la hipótesis alternativa (siempre con probabilidad de equivocación), concluyendo que hay un efecto diferente a cero que se considera estadísticamente significativo, no interpretándose en este caso en la toma de decisiones (comparando con el nivel de alfa prefijado) como error aleatorio o azar, sino como un efecto sistemático.

Sobre estos principios de inferencia estadística se ha construido gran parte de la Ciencia que se desarrolla en las investigaciones empíricas como la Psicología o las Ciencias de la Salud. Sin embargo, la comprensión de dichos principios no es una tarea sencilla, tal y como se ha demostrado repetidamente con debates sobre el uso y abuso de las pruebas de significación estadística y los problemas de comprensión que conducen a los investigadores a interpretaciones incorrectas de la información que ofrece un valor *p* vinculado a un estadístico.

El tema de las falacias estadísticas ha sido debatido y analizado durante décadas, prácticamente desde sus orígenes y en la segunda década del siglo XXI sigue llenando libros y artículos que alertan de la necesidad de la re-educación estadística de científicos, profesionales y del material docente que reciben los estudiantes, junto con la necesidad de actualizar los programas estadísticos para que incorporen en sus menús otras técnicas, como el intervalo de confianza o la estimación de la potencia estadística *a priori*, de tal manera que la presencia de la estimación del tamaño del efecto cobre una mayor relevancia y facilite el desarrollo de un pensamiento meta-analítico que reflexione sobre las magnitudes de los efectos, su variabilidad y su contexto de investigación.

En las últimas décadas han crecido exponencialmente las publicaciones que critican la aplicación inadecuada de esta estrategia analítica, lo insatisfactoria que resulta para alcanzar la acumulación de conocimiento científico válido, y la utilización, casi exclusiva, de la significación de los resultados como único criterio de interpretación (Cumming, 2012; Falk, 1998; Frías, Pascual, y García, 2000; Kline, 2013; Krueger, 2001; Monterde-i-Bort y cols., 2010; Nickerson, 2000; Pascual, Frías y García, 2000; Perezgonzalez, 2015). Por ejemplo, recientemente, Wasserstein y Lazar (2016) detallan las recomendaciones de la *American Statistical Association* (ASA) sobre los valores *p* y la significación estadística, destacando las interpretaciones incorrectas y el mal uso que se hace de las pruebas de contraste estadístico, que incluso ha llevado a que algunas revistas las desaconsejen y/o prohíban (e.g., *Basic and Applied Social Psychology)* y a que algunos científicos recomienden su abandono (e.g., Cumming, 2012). Entre sus recomendaciones destacan la idea de no fundamentar los juicios científicos en una decisión dicotómica basada en el valor *p,* sino que es necesario valorar factores contextuales que también forman parte del proceso de inferencia estadística, como es el diseño del estudio, la calidad de la medición, la evidencia externa para el fenómeno objeto de estudio y la validez de las asunciones que subyacen al análisis de datos. Destacan las recomendaciones que a veces se requiere una decisión binaria "sí / no" por cuestiones prácticas, pero eso no significa que considerar solamente los valores *p* asegura que una decisión es correcta o incorrecta o que se ha logrado un hallazgo científico, pues este tipo de creencias conduce a una distorsión del proceso de investigación científica. Además, la significación estadística no es equivalente a la significación científica, del mismo modo que valores de *p* pequeños tampoco implican necesariamente la presencia de efectos grandes o efectos importantes, ni valores de *p* grandes implican que el hallazgo sea poco importante o que no haya efecto. En definitiva, continúan las recomendaciones de la ASA afirmando que en sí mismo el valor de *p* no es una buena medida de la evidencia de un modelo o de una hipótesis. Concluyen las recomendaciones con la referencia a las buenas prácticas estadísticas como un componente esencial de la buena práctica científica, enfatizando el uso de un buen diseño del estudio y la interpretación de los resultados en un contexto determinado.

En parte debido a estas críticas, la APA creó el 28 de febrero de 1996 el grupo de trabajo sobre inferencia estadística (*Task Force on Statistical Inference*, TFSI) con el objetivo de examinar las practicas estadísticas en Psicología. Entre sus recomendaciones destacan tres directrices que marcarán el contenido de la quinta edición del Manual de Publicación de la APA (2001): (1) acompañar la presentación, análisis e interpretación de los datos con otros estadísticos como la estimación del tamaño del efecto; (2) informar de los intervalos de confianza de los tamaños del efecto, y (3) utilizar procedimientos gráficos que mejoren la interpretación y comunicación de los resultados. Estas directrices, han sido incorporadas al Manual de Publicación de la APA en sucesivas ediciones (2001, 2010a).

La reforma estadística propuesta implica un cambio importante en la conducta de los investigadores ya que supone cambiar el punto de mira desde "cómo es de probable o improbable el resultado muestral" (aplicación de las pruebas tradicionales de significación estadística y las decisiones dicotómicas basadas en ellas) hacia nuevas estrategias de análisis que estudien el tamaño del efecto y sus intervalos de confianza, la significación práctica o clínica de los hallazgos, y que favorezcan su replicabilidad (Balluerka, Vergara y Arnau, 2009; Cumming, 2014; Cumming y Finch, 2005; Frías-Navarro, 2011; Frías-Navarro, Pascual-Soler, Badenes-Ribera, y Monterde-i-Bort, 2014; Hoekstra, Johnson, y Kiers, 2012; Kline, 2013; Rosnow, y Rosenthal, 2009; Vacha-Haase, 2001; Wilkinson y TFSI, 1999). Es decir, es necesario 'evaluar' el valor del tamaño del efecto estimado junto con sus intervalos de confianza y su utilidad (su grado de importancia práctica), y para ello hay que considerar el contexto de la investigación y comparar los resultados de forma explícita y directa con los obtenidos en el área de estudio donde se enmarca el trabajo, promoviendo de este modo el desarrollo del pensamiento meta-analítico (Cumming, 2012; Henson, 2006; Kline 2013). Dentro de este contexto de cambio y avance metodológico los estudios de revisión sistemática tipo meta-análisis han cobrado una alta relevancia y presencia en las revistas más prestigiosas (Borenstein y cols., 2009; Cumming, 2012; Sánchez-Meca y Botella, 2010).

El valor de las investigaciones psicológicas depende en gran medida de la calidad de los análisis estadísticos y la interpretación de los resultados va más allá de las pruebas de significación estadística, tal y como señala la reforma estadística (APA, 2010a; Cumming, 2014; Cumming y cols., 2012, Kelley y Preacher, 2012; Kline, 2004).

Si además tenemos en cuenta, tal y como se ha constatado por muchas investigaciones, que la aplicación de muchos conceptos estadísticos continúa siendo incorrecta y que la aplicación de muchas técnicas estadísticas es imprecisa (Balluerka, Gómez y Hidalgo, 2005; Bakker, 2014; Bakker y Wicherts, 2011; Campitelli, 2015; Caperos y Pardo, 2013; Coulson, Healey, Fidler y Cumming, 2010; Pascual y cols., 2000; Palmer y Sesé, 2013; Peng y Chen, 2014; Peng y cols., 2013; Lecoutre y cols., 2003; Sun, Pan y Wang, 2013; Wasserstein y Lazar 2016) y, además, no se atiende a los requisitos o supuestos de aplicación de dichas pruebas (Faulkner y cols., 2008; Hoekstra, Kiers y Johnson, 2012) entonces el estudio de la calidad metodológica de las pruebas obtenidas con la investigación psicológica se convierte en un tema urgente, actual y necesario.

Por ello, es preciso llevar a cabo investigaciones sobre el grado de conocimiento que psicólogos académicos y profesionales tienen sobre la calidad metodológica de las evidencias en la investigación psicológica para la correcta aplicación del enfoque de la Práctica Basada en la Evidencia y la adquisición de un conocimiento científico válido. Este tipo de investigación puede aportar luz sobre estos problemas y contribuir a promover programas de formación para tratar de corregirlos o minimizarlos.

La Tesis Doctoral que aquí se presenta tiene como principal objetivo abordar las interpretaciones inadecuadas y el mal uso que se realiza de las técnicas de inferencia estadística, debido a las falacias estadísticas y de tamaño del efecto que la rodean, a las que no son inmunes profesores, científicos y profesionales.

De acuerdo con ello, este trabajo tiene tres objetivos específicos:

El primer objetivo es detectar los errores de razonamiento estadístico que psicólogos académicos y profesionales españoles cometen cuando se les presentan los resultados de una prueba de inferencia estadística. Pues su visión e interpretación de los hallazgos es un filtro de calidad que no debiera estar sometido a creencias o interpretaciones erróneas del procedimiento estadístico que representa la herramienta fundamental para obtener conocimiento científico. Con este fin, se analizan dos cuestiones: la primera es la extensión de los errores más comunes de interpretación con respecto al valor *p* y la segunda es el grado en que se interpretan correctamente los valores *p* por parte de ambos colectivos.

El segundo objetivo es analizar el conocimiento que psicólogos académicos y profesionales españoles tienen sobre los tamaños del efecto, sus intervalos de confianza y los estudios de meta-análisis, teniendo en cuenta que ésta es una de las principales recomendaciones propuestas por la APA (2010a) para mejorar la práctica estadística en la investigación psicológica y favorecer la acumulación de conocimiento y la replicación de los hallazgos.

Por último, se trata de comprobar si los resultados de la investigación sobre los errores de interpretación del valor *p* y el nivel de conocimiento sobre los tamaños del efecto, sus intervalos de confianza y los meta-análisis, en psicólogos académicos españoles, son fiables, es decir, si se pueden replicar con una muestra de psicólogos académicos chilenos e italianos.

Para enmarcar los estudios realizados (que se recogen en la segunda parte del trabajo, en los Capítulos 4 y 5), en la primera parte de la Tesis Doctoral se presenta una exposición extensa y detallada de la lógica subyacente al procedimiento de la NHST, así como su origen y las principales críticas dirigidas a este procedimiento que están relacionadas con los problemas de interpretación del valor *p* y su uso inadecuado (Capítulo 1); le sigue un repaso a las recomendaciones metodológicas propuestas por la APA (2001, 2010a) frente a tales críticas, como son el uso de los índices del tamaño del efecto y sus intervalos de confianza como método complementario a las pruebas de inferencia estadística en el análisis de datos, junto a los estudios de replicación (que se tratan en el Capítulo 2), así como la importancia de los estudios de meta-análisis (que se abordan en el Capítulo 3) como métodos para favorecer la acumulación de un conocimiento científico válido.

1. LA PRUEBA DE SIGNIFICACIÓN ESTADÍSTICA DE LA HIPÓTESIS NULA (NHST): EL VALOR *P*

El objetivo de toda investigación científica es la búsqueda de explicación de los fenómenos y con ello poder derivar predicciones sobre la realidad, elaborando teorías sobre el comportamiento de dichos fenómenos. Ya sea para comprobar teorías o para estimar efectos de un tratamiento, los investigadores tienen que realizar un proceso de comprobación o contraste de hipótesis traduciendo la hipótesis científica a hipótesis estadística.

Se pueden distinguir dos estrategias a la hora de realizar un contraste de hipótesis (1) *contraste de hipótesis mediante la ejecución de las pruebas de significación estadística*; (2) *contraste de hipótesis mediante los intervalos de confianza.*

El presente capítulo se enmarca dentro del contraste de hipótesis mediante las pruebas de significación estadística, dado que, históricamente, los psicólogos han confiado en el 'procedimiento de la Prueba de Significación de la Hipótesis Nula (*Null Hypothesis Significance Testing*, NHST), entendida como tamaño de efecto cero ("*nil hypothesis*"), como la técnica por excelencia para llevar a cabo un contraste de hipótesis (Frías-Navarro y Gómez-Frías, 2014; Gigerenzer, 2004; Huberty, 1993; Lovell, 2013; Ludbrook, 2013; Nickerson, 2000; Perezgonzalez, 2015). Por ejemplo, el 97% de los artículos publicados en 10 revistas internacionales de Psicología utilizaron la prueba de la NHST (Cumming y cols., 2007). También Gigerenzer y Marewski (2015) señalan que, en la revista *Nature*, el 89% de los artículos publicados durante 2011 que analizan cuestiones conductuales, neuropsicológicas y médicas informan del valor *p* sin proporcionar información sobre el tamaño del efecto, la potencia estadística o el modelo de estimación.

Por lo tanto, el objetivo de este capítulo es presentar la lógica subyacente al procedimiento de la NHST, su origen y las principales críticas a este procedimiento que están relacionadas con sus problemas de interpretación y su uso inadecuado. Pues, como se expondrá en los siguientes capítulos, estas críticas han dado lugar a recomendaciones metodológicas para cambiar las prácticas estadísticas de los psicólogos hacia nuevas estrategias de análisis de datos.

1.1. La lógica del actual procedimiento de la NHST

Históricamente, los psicólogos han confiado en el procedimiento de la NHST como técnica para el contraste de hipótesis a través de las pruebas específicas de inferencia estadística (e.g., t de Student, ANOVA, Chi cuadrado, correlación de Pearson, etc.).

Las pruebas de inferencia estadística se utilizan para estimar la probabilidad o valor *p* que tiene el resultado obtenido en una investigación dentro de una distribución de datos que atribuye la diferencia o la relación observada al azar (error aleatorio), conocida como la "distribución o modelo de la hipótesis nula" (Frías-Navarro, 2011a).

Durante el proceso de inferencia estadística, la distribución de la hipótesis nula se asume que es cierta desde el principio y con el procedimiento NHST se estima el valor *p* del resultado de la investigación empírica (o un resultados más extremo), esto es, la probabilidad de los datos (o datos más extremos) bajo dicho supuesto. Y se rechaza la hipótesis nula cuando la probabilidad asociada al resultado es tan pequeña que podemos concluir que, con una alta probabilidad, ese dato no puede ser atribuible al error aleatorio (si lo fuera, se cometería un error de Tipo I), concluyendo que el azar o el error aleatorio no es el responsable del efecto detectado, dado que es un dato extraño y altamente improbable dentro de dicho modelo.

De acuerdo con Frías-Navarro (2011a), los pasos a seguir en el procedimiento actual de la NHST para evaluar empíricamente las hipótesis son:

Pasos a priori (antes de la recogida de datos):

1. Establecer *a priori* el criterio de nivel de significación estadística o alfa (α) vinculado a la decisión estadística de rechazar o no rechazar la hipótesis nula. Error de Tipo I (probabilidad de rechazar la hipótesis nula siendo realmente cierta).
2. Formular la hipótesis nula (H_0) y asumir de partida que es cierta.

Pasos a posteriori (después de la recogida de datos):

3. Seleccionar la prueba estadística óptima (la más potente) que permita comprobar la hipótesis formulada en la hipótesis nula.
4. Calcular el valor *p* de probabilidad del resultado vinculado a la prueba estadística seleccionada, es decir, calcular la probabilidad del resultado observado (o más extremo) asumiendo que la hipótesis nula es cierta.

5. Tomar una decisión estadística: Rechazar o mantener la Hipótesis nula (H_0) en función del valor *p* vinculado a la prueba estadística y del valor del nivel de alfa asumido a priori (α). Es decir, se compara el valor de *p* (probabilidad del resultado bajo el modelo de la hipótesis nula) con el nivel de alfa elegido *a priori* (probabilidad de rechazar la hipótesis nula siendo realmente cierta) y se toma una de las dos siguientes decisiones:

*Rechazar la hipótesis nula (H_0) si $p \leq \alpha$: si el valor de *p* vinculado a la prueba estadística es menor o igual al valor de significación estadística o alfa asumido *a priori*, se rechaza H_0 puesto que el resultado observado (el dato obtenido) es improbable que ocurra dentro de la distribución de la hipótesis nula. Todo ello conduce a rechazar el azar o el error aleatorio como causa probable de las diferencias o relaciones encontradas entre las variables, asumiendo que la hipótesis nula es cierta y asumiendo un nivel de riesgo de error de tipo I o alfa determinado. En otras palabras, se trata de un resultado extraño o incompatible con la distribución de la H_0.

Por ejemplo, un valor de $p = .001$ indica que hay una entre mil posibilidades de que ocurra ese dato bajo el supuesto de verdad de la hipótesis nula (que plantea efecto cero y cualquier efecto lo atribuye al error aleatorio). Es decir, es un dato altamente improbable de tal manera que por azar no ocurrirá (normalmente) y si de hecho ocurre, el investigador hipotetiza que se debe a la presencia de un efecto sistemático (asumiendo un cierto nivel de error llamado de Tipo I, generalmente el 5% de error de rechazar la hipótesis nula siendo realmente cierta).

*Mantener la hipótesis nula (H_0) si $p > \alpha$: si el valor de *p* vinculado a la prueba estadística es mayor al valor de la significación estadística o alfa asumido *a priori*, se mantiene H_0 puesto que el resultado es compatible con la distribución de la hipótesis nula. En este caso, el resultado no es concluyente, siendo incorrecto realizar inferencias de igualdad entre los grupos o ausencia de relaciones entre las variables o afirmaciones de aceptar la hipótesis nula.

Por lo tanto, el procedimiento NHST sólo permite que el investigador realice una decisión dicotómica: rechazar la hipótesis nula o no rechazarla y no cuantifica la evidencia a favor de dicha hipótesis y, por ello, es erróneo concluir que "se acepta" la hipótesis nula cuando el valor de *p* es mayor al de alfa (Wilkinson, y TFSI, 1999). La

distribución de la hipótesis nula asume que es cierta desde el principio y solamente se puede mantener o rechazar con el proceso de inferencia estadística.

Conviene tener siempre presente dos cuestiones: 1) que el valor *p* es la probabilidad condicionada de un resultado (de los datos) y nunca es la probabilidad de la hipótesis nula ni de la hipótesis alternativa. El valor *p* del resultado se estima en función de un modelo de distribución de la hipótesis nula conocido que plantea efecto cero (potencia estadística igual a cero). 2) Nunca se produce un rechazo absoluto de la hipótesis nula ni se acepta de forma absoluta la hipótesis alternativa pues siempre existen las probabilidades del error de Tipo I y del error de Tipo II; en este contexto mantener o rechazar la hipótesis nula se entiende siempre en términos probabilísticos.

1.2. Origen y expansión del procedimiento actual de la NHST

El procedimiento actual de la NHST, descrito en el apartado anterior, es el resultado de la combinación de las perspectivas del "test de significación" (*Significance Test*) de Fisher y del "test de la hipótesis estadística" (*Statistical Hypothesis Test*) de Neyman y Pearson. Es decir, es un híbrido entre dos escuelas estadísticas enfrentadas: la escuela de Fisher y la escuela de Neyman-Pearson (Anderson, Burnham, y Thompson, 2000; Borges, 1997; Borges, San Luis, Sánchez, y Cañadas, 2001; Gigerenzer, 2004; Gill, 1999; Hager, 2013; Perezgonzalez, 2014, 2015; Spielman, 1978; Valera-Espín, Sánchez-Meca y Marín-Martínez, 2000; Wagenmarkers, 2007).

Los enfoques de Fisher y Neyman-Pearson se hibridaron en los libros de Estadística publicados en el ámbito de la Psicología durante los años 1940-1960 (Gigerenzer y Murray, 1987). A este respecto, Halpin y Stam (2006) señalan el manual de Lindquist publicado en 1940 como la fuente original de la hibridación de los enfoques de Fisher y Neyman-Pearson. El libro de Lindquist describe un único procedimiento de prueba estadística en el que se presenta de forma indiscriminada aspectos del enfoque de Fisher y de Neyman-Pearson, sin presentar ningún enfoque en su totalidad.

Durante este periodo de 1940-1960 conocido como "revolución de la inferencia", los manuales de Estadística de la época presentaron el modelo híbrido de la prueba NHST a los investigadores, y la revolución de la inferencia consistió en un incremento exponencial de la aplicación del procedimiento NHST por parte de los investigadores, donde la inferencia de la muestra a la población llegó a ser considerado

el punto crucial de la investigación, institucionalizándose el ritual de la hipótesis nula (Gigerenzer, 1987, 2004; Gigerenzer & Murray, 1987). Como Cohen (1990) afirma, el hecho de que el procedimiento de la NHST rápidamente se convirtiera en la base para la inferencia estadística en las Ciencias del Comportamiento no es sorprendente. El procedimiento de la NHST es muy atractivo, ofrece un esquema determinista, mecánico y objetivo, independiente del contenido, y que lleva a tomar decisiones dicotómicas (sí/no) a partir de un claro punto de corte (el santificado nivel mágico de .05). Por su parte, Nester (1996) ofreció varias razones de su rápida expansión: (1) parece ser objetivo y exacto (como ya señaló Cohen, 1990); (2) las pruebas específicas de inferencia estadística (e.g., ANOVA, *t* de Student, etc.) están disponibles en los paquetes estadísticos comerciales; (3) todos parecen utilizarlo; (4) se enseña en las universidades; y (5) algunos editores de revistas y directores de tesis exigen su uso. Por ejemplo, en el clásico editorial de Melton (1962), editor de la revista *Journal of Experimental Psychology*, se destacaba que los trabajos con resultados estadísticamente no significativos serían difíciles de publicar mientras que aquellos que obtengan los valores *p* más bajos serían dignos de formar parte de su revista.

De este modo, los psicólogos científicos adoptaron un modelo híbrido de la prueba estadística que perdura hasta la actualidad. Como Thompson (1999) señala:

> Vemos el uso de combinaciones híbridas de las lógicas reflejadas en los informes publicados en los que los investigadores declaran un alfa fijado pero, no obstante, tienen en cuenta que sus resultados "se acercaron" a la significación estadística y en los informes donde los autores presentan los valores de *p* calculados específicos (a veces los índices del tamaño del efecto) independiente de las decisiones de rechazo de hipótesis. No tengo objeciones intrínsecas a la utilización de las lógicas híbridas.
>
> Pero sí me parece que el campo nunca resolvió con éxito las tensiones entre estas dos escuelas de pensamiento. Y ahora los autores en un artículo dado parecen moverse de forma espontánea e inconscientemente a través de diferentes filosofías analíticas. Deseo vivamente que resolvamos con más éxito las actuales controversias a través del continuo debate y la reflexión profunda (p. 159).

A continuación se expondrá sucintamente en qué consisten cada uno de estos enfoques de inferencia estadística.

1.2.1. El test de Significación de Ronald Fisher

De acuerdo con Perezgonzalez (2015), Fisher propuso el test de significación (*Significance test*) como una herramienta para identificar los resultados de investigación de interés, definidos como aquellos con una baja probabilidad de ocurrencia (valores *p* pequeños) bajo el modelo de la hipótesis nula.

Para ello, el test de significación localiza los resultados de investigación dentro de la distribución y evalúa su probabilidad teórica (Perezgonzalez, 2014). Los resultados con valores pequeños de *p* son tomados como evidencia en contra de la hipótesis nula, por lo que cuanto menor sea el valor de *p* más fuerte es la evidencia que proporciona.

Desde esta lógica, es plausible "*graduar los niveles de significación (como significativo y altamente significativo), ya que reflejan la fuerza relativa de las evidencias contra la hipótesis nula"* (Perezgonzalez, 2014, p. 855), de tal manera que a mayor nivel de significación estadística mayor evidencia en contra de la H_0 (Perezgonzalez, 2015).

Siguiendo a Perezgonzalez (2015), el test de la significación se puede resumir en 5 pasos:

1. Plantear una hipótesis nula (H_0). La hipótesis hace referencia, típicamente, al valor de algún parámetro en la población de referencia. La hipótesis nula no necesita ser una hipótesis de efecto cero (Gigerenzer, 2004).
2. Determinar la prueba estadística apropiada para el objetivo de la investigación y/o las variables evaluadas y su distribución bajo el supuesto de que la hipótesis nula es verdadera.
3. Calcular la probabilidad teórica de los resultados bajo la H_0 (valor de *p*).

Es decir, aplicar la prueba estadística a los datos obtenidos en la muestra para determinar el nivel de significación estadística (valor *p*) de los resultados (o más extremos) que corresponde a la distribución de la prueba estadística utilizada bajo el supuesto de que la hipótesis nula es verdadera.

4. Evaluar la significación estadística de los resultados:

Desde la perspectiva de Fisher, un resultado de investigación con una baja probabilidad de ocurrencia (valores pequeños de *p*) puede ser tomado como evidencia

en contra de la hipótesis nula (es decir, como prueba de que la hipótesis nula no puede explicar estos resultados de manera satisfactoria).

La cuestión es determinar qué valor de *p* es suficientemente pequeño para que sea considerado un resultado de interés o estadísticamente significativo (Gill, 1999). En este sentido, el valor de *p* que se debe considerar estadísticamente significativo depende en gran medida de la pregunta del investigador y del contexto del problema, y puede variar de una investigación a otra (Gigerenzer, 2004; Perezgonzalez, 2015). La valoración de los valores *p* también se puede dejar al albedrío del lector, por lo que informar de los valores de *p* exactos es muy informativo desde esta aproximación (Gigerenzer, 2004; Perezgonzalez, 2015; MacDonald, 1997).

Desde esta perspectiva la diferencia entre un nivel de significación de .05 y de .06 no es demasiado crítica (Perezgonzalez, 2014), o los valores de *p* de .049 y .051 tienen aproximadamente la misma significación estadística, en torno a un nivel de significación del 5% (Perezgonzalez, 2015).

En general, la evaluación de los resultados de la investigación se realiza comparando el valor de *p* obtenido con un nivel de significación estadística (o valor de *p* teórico, que no necesariamente se tiene que establecer *a priori*, ni tener un valor determinado y fijo para todas las ocasiones, pudiendo utilizarse los valores convencionales de .05 ó .01). No obstante, la obra de Fisher está llena de frases que hacen referencia a los de niveles de significación de .01 ó .05 (Gill, 1999).

De tal manera que:

• Si el valor *p* es aproximadamente igual o menor que el nivel de significación, el resultado se considera estadísticamente significativo.

• Si el valor *p* es mayor que el nivel de significación, el resultado se considera estadísticamente no significativo.

Un resultado estadísticamente significativo se puede interpretar de dos maneras: O bien es un resultado excepcionalmente raro, es decir, que ocurre con baja probabilidad dentro del modelo de la hipótesis nula, o bien la hipótesis nula no puede explicar el resultado obtenido de manera satisfactoria (Carver, 1978; Macdonald, 1997; Perezgonzalez, 2014, 2015). El test no nos dice cuál de estas dos posibles explicaciones es la correcta. Mientras que los resultados estadísticamente no significativos pueden ser

ignorados, todavía pueden proporcionar información útil, como por ejemplo si los resultados fueron en la dirección esperada y sobre su magnitud (Perezgonzalez, 2015).

5. Tomar la decisión estadística: Rechazar la H_0 o no llegar a ninguna conclusión.

*Rechazar H_0: si los resultados son estadísticamente significativos (valores de *p* pequeños). El valor *p* observado se toma como evidencia en contra de la hipótesis nula, por lo que cuanto menor sea el valor de *p* más fuerte es la evidencia que proporciona.

*No se llega a ninguna conclusión: si los resultados no son estadísticamente significativos (valores de *p* altos).

Por tanto, las únicas decisiones posibles son rechazar la hipótesis nula o declarar que la evidencia no es suficiente. De hecho, aunque siempre es posible rechazar la H_0, la H_0 nunca podría ser apoyada o establecida, es decir, aceptada.

1.2.2. La Prueba de Hipótesis Estadística de Neyman-Pearson

Jerzy Neyman y Egon Sharpe Pearson intentaron mejorar el procedimiento de contraste de hipótesis propuesto por Fisher, pero, finalmente, terminaron desarrollando un nuevo procedimiento llamado la "prueba de hipótesis estadística" (*Statistical Hypothesis Test*).

Neyman y Pearson introdujeron la hipótesis alternativa (H_1), los conceptos de Error de Tipo I y II y sus probabilidades asociadas (Alfa y Beta), y la potencia estadística de la prueba (Borges, 1997; Perezgonzalez, 2014, 2015).

El ***Error de Tipo I (o Alfa)*** se define como la probabilidad de rechazar la hipótesis nula siendo realmente verdadera, es decir, cuando la diferencia observada en las poblaciones está realmente provocada por el azar o el error aleatorio. Por tanto, este error se comete cuando se rechaza la hipótesis nula siendo verdadera.

El ***Error de Tipo II (o Beta)*** hace referencia a la probabilidad de mantener la hipótesis nula cuando realmente es falsa, es decir, sí existe una diferencia real entre las poblaciones. En consecuencia, este error se comete cuando se mantiene la hipótesis nula siendo falsa.

La ***potencia estadística (o 1-Beta***) se define como decisión correcta, donde se rechaza la hipótesis nula siendo realmente falsa, es decir, existe un efecto y la prueba realmente lo detecta. La potencia estadística es la probabilidad de rechazar correctamente la hipótesis nula en favor de la hipótesis alternativa (es decir, de aceptar

correctamente H_1). Matemáticamente, es lo opuesto al Error de Tipo II o Beta (por lo tanto, 1 - β) (Frías-Navarro, 2011a; Hubbard, 2004; Macdonald, 1997).

De acuerdo a Perezgonzalez (2015), la potencia estadística depende de tres factores: (1) la prueba estadística seleccionada (e.g., las pruebas paramétricas tienen mayor potencia que las pruebas no paramétricas, y las pruebas de una cola tienen mayor potencia estadística que las pruebas de dos colas), (2) el tamaño del efecto esperado (a mayor tamaño del efecto mayor potencia estadística), (3) Alfa (aumenta la potencia estadística cuanto más grande es α) y Beta (cuanto más pequeño es beta mayor es la potencia).

El procedimiento de Neyman-Pearson es un test orientado a la toma de decisiones (Borges, 1997; Gill, 1999) puesto que conduce a una decisión entre la hipótesis nula y la hipótesis alternativa (Perezgonzalez, 2015), y, además, se puede considerar como un test de aceptación, porque el interés es decidir entre la aceptación de la hipótesis nula o la aceptación de la hipótesis alternativa (Borges, 1997; Perezgonzalez, 2014, 2015; Spielman, 1978), siempre con los márgenes de error alfa y beta.

De acuerdo con Perezgonzalez (2015), el procedimiento de la hipótesis estadística se puede resumir en 8 pasos, 6 de los cuales se llevan a cabo antes de la recogida de datos (*a priori*) y 2 después de la recogida de datos (*a posteriori*).

Pasos a priori (antes de la recogida de datos)

1. Establecer el tamaño del efecto esperado en la población.
2. Formular dos hipótesis complementarias: una hipótesis de interés o hipótesis nula (H_0), y una hipótesis complementaria o hipótesis alternativa (H_1).
3. Determinar la prueba estadística óptima (la más potente) y su distribución bajo el supuesto de que la H_0 es cierta.
4. Especificar un nivel de significación (*alfa* o valor crítico) de la prueba estadística bajo el supuesto de que la H_0 es cierta.

El nivel de significación ayuda a establecer la región crítica (o región de rechazo) en la distribución de probabilidad de la hipótesis nula. Neyman y Pearson a menudo trabajaron con los niveles convencionales de alfa tales como 5% (α = .05) y 1% (α = .01), aunque se pueden establecer otros niveles de significación.

El nivel de alfa ayuda a identificar el valor crítico del test, el límite para decidir entre la H_0 y la H_1.

5. Calcular el tamaño de la muestra requerido para tener una buena potencia estadística (1 - beta).

6. Calcular el valor crítico del test.

El valor crítico será usado como punto de corte para la decisión entre las hipótesis.

Pasos a posteriori (después de la recogida de datos)

7. Calcular el valor del test para la investigación.

Desde la perspectiva de Neyman-Pearson también se puede utilizar el valor de *p* en el contraste de hipótesis, lo cual implica calcular la probabilidad teórica de los datos bajo la distribución de la hipótesis nula.

8. Tomar una decisión estadística: (1) Rechazar H_0 y aceptar H_1; (2) Aceptar H_0, o (3) No concluir ninguna cosa.

*Rechazar H_0 y aceptar la H_1: si el valor del test cae dentro de la región crítica (o el valor de *p* es inferior o igual al nivel de alfa), se rechaza la H_0 y se acepta la H_1 dado que dicho valor es poco probable bajo la hipótesis nula.

*Aceptar la H_0: si el valor del test cae fuera de la región crítica y el test tiene buena potencia estadística (o el valor de *p* es mayor al nivel de alfa), se acepta la H_0 dado que dicho valor del test es probable bajo la hipótesis nula.

*No concluir ninguna cosa: si el resultado cae fuera de la región crítica y el test no tiene buena potencia estadística (o el valor de *p* es mayor al nivel de alfa y el test no tiene buena potencia estadística), no se llega a ninguna conclusión.

1.2.3. Algunas diferencias entre los enfoques de Fisher y Neyman-Pearson

Hay diferencias importantes entre la prueba de Fisher y la prueba de Neyman y Pearson (una revisión más profunda de las diferencias entre ambos enfoques se encuentra en Hager, 2013).

Por ejemplo, en la prueba de significación de Fisher, no se identifica ninguna hipótesis complementaria explícita a la hipótesis nula H_0, es decir, no existe la hipótesis alternativa (H_1) y el valor *p* que resulta de los datos se evalúa como la fuerza de la

evidencia en contra de la hipótesis nula. Además, no hay una noción de la potencia estadística de la prueba, ni de aceptar hipótesis alternativas en la interpretación final.

Por el contrario, en la prueba de hipótesis de Neyman-Pearson se identifican dos hipótesis complementarias: la Hipótesis nula (H_0) y la Hipótesis alternativa (H_1), y el rechazo de una implica la aceptación de la otra. Y este rechazo se basa en un nivel alfa predeterminado (fijado *a priori*). Por otro lado, contempla los Errores Tipo I y Tipo II y sus probabilidades asociadas Alfa y Beta.

Además, existen diferencias en la naturaleza de la toma de decisiones. Desde la perspectiva de Fisher, la decisión es asimétrica, es decir, si un resultado es estadísticamente significativo, se rechaza la hipótesis nula, de lo contario no se puede extraer ninguna conclusión. Mientras que en el enfoque de Neyman y Pearson, la decisión es simétrica entre dos hipótesis complementarias, esto es, los resultados de la investigación pueden apoyar la hipótesis alternativa (rechazar la H_0 y aceptar H_1), o, si la potencia estadística es adecuada, apoyar a la hipótesis nula (aceptar H_0). En el caso de que la potencia no sea adecuada, no se pude concluir ninguna cosa (Gigerenzer, 2004; Perezgonzalez, 2014, 2015).

En definitiva, ni Fisher ni Neyman y Pearson estarían satisfechos con la hibridación de estos dos enfoques (Gigerenzer, 2004; Gill, 1999; Hager, 2013). Pues Fisher se opuso a la preselección del nivel de significación, así como al proceso de toma de decisiones de dos resultados obligatorios. Por su parte, Neyman y Pearson no estaban de acuerdo con la interpretación de los valores de *p* y su graduación.

1.2.4. NHST a través de las ideas de Fisher y Neyman-Pearson

En general, en el procedimiento actual de la NHST se establecen dos hipótesis estadísticas complementarias: la hipótesis nula (e.g., μ_A-μ_B = 0) (*Fisher y Neyman-Pearson*) y la hipótesis alternativa (e.g., μ_A-$\mu_B \neq 0$) (*Neyman y Pearson*) que son afirmaciones sobre parámetros de la población.

Las hipótesis estadísticas son evaluadas con un valor crítico definido *a priori* (alfa) o criterio de nivel de significación estadística (*Neyman-Pearson*), y posteriormente se selecciona la prueba estadística óptima, es decir, con la mayor potencia estadística posible (*Neyman-Pearson*) para calcular el valor *p* de probabilidad vinculado al resultado de la prueba estadística asumiendo que la hipótesis nula es cierta (*Fisher*). Y, finalmente, el investigador toma la decisión de rechazar o no rechazar la

hipótesis nula en función del valor *p* de probabilidad asociado a la prueba estadística (*Fisher*) (en función de si es igual o menor al nivel de significación estadística o alfa establecido *a priori*, antes de recoger los datos) (*Neyman y Pearson*), apareciendo el Error de Tipo I, el Error de Tipo II y la potencia estadística (decisión correcta, donde se rechaza la hipótesis nula siendo realmente falsa, es decir, existe un efecto y la prueba realmente lo detecta) (*Neyman y Pearson*). El valor de *p* no cuantifica la evidencia a favor de dicha hipótesis (*Neyman-Pearson*), por ello, es erróneo concluir que se acepta la hipótesis nula cuando el valor de *p* es mayor al de alfa (*Fisher*). Otros autores consideran que valores de *p* mayores a alfa dan lugar a la aceptación de la hipótesis nula (*Neyman y Pearson*).

Como se puede observar, el procedimiento descrito es una amalgama de teorías incompatibles (Borges, 1997; Gigerenzer, 2004; Halpin y Stam, 2006; Hager, 2013; Perezgonzalez, 2014, 2015; Wagenmarkers, 2007) que no está claramente definida. De hecho, en función del autor que describe el procedimiento de la NHST o el investigador que utilice dicho procedimiento, éste se asemeja más al procedimiento de Fisher o al procedimiento de Neyman y Pearson (Perezgonzalez, 2014, 2015). En este sentido, Perezgonzalez (2015) apunta que la *APA* (2010a) o Wilkinson y TFSI (1999), entre otros, están más cerca de la perspectiva de Fisher, mientras que autores como Cohen (1988), Cortina y Dunlap (1997), Frick (1996), Kline (2004), Nickerson (2000), Rosnow y Rosenthal (1989) Schmidt (1996), entre otros, están más cerca de la perspectiva de Neyman y Pearson.

Por ello, algunos autores desaconsejan el uso del procedimiento de la NHST para realizar un contraste de hipótesis y, en su lugar, aconsejan utilizar ya sea la prueba de significación de Fisher o la prueba de hipótesis estadística de Neyman y Pearson (e.g., Gigerenzer, 2004; Perezgonzalez, 2015, 2014).

1.3. Críticas al procedimiento de la NHST

Si bien la incorporación del procedimiento de la NHST al ámbito de la Psicología se produjo en la década de los años 1950 (Gigerenzer, 2004; Hubbard y Ryan, 2000; Perezgonzalez, 2015), las críticas a este procedimiento aparecieron tempranamente desde sus inicios en el año 1900 (e.g., Berkson, 1938; Boring, 1919) y desde 1950 el debate y la discusión sobre el procedimiento de la NHST ha crecido de forma

exponencial en los campos de Educación, Psicología, Ecología y Medicina (Anderson y cols., 2000; Pascual y cols., 2000; Wasserstein y Lazar, 2016).

Por ejemplo, Bakan (1966) concluyó que "*el test de significación estadística en la investigación psicológica puede tomarse como un ejemplo de un tipo de inconsciencia esencial en la realización de investigaciones*" (p. 436). Meehl (1978) señaló que el procedimiento de la NHST "*es una de las peores cosas que ha ocurrido en la historia de la Psicología*" (p. 817). Carver (1978) recomendó eliminar las pruebas de la NHST puesto que no sólo son inútiles sino también perjudiciales debido a los problemas de interpretación que plantean. Schmidt y Hunter (1997) afirmaron que las pruebas de la NHST "*retrasan el crecimiento del conocimiento científico; nunca hacen una contribución positiva*" (p. 37) y Thompson (1992) agregó que han creado "*un daño considerable en cuanto a la acumulación de conocimiento*" (p. 436). Finalmente, Tryon (1998) afirmó que el hecho de que "*los expertos en estadística e investigadores que publican en las mejores revistas no pueden interpretar sistemáticamente los resultados de estos análisis es muy preocupante*". Y continuó diciendo que "*Es difícil estimar la desventaja que el uso generalizado, incorrecto e intratable de un método analítico de datos primarios tiene en una disciplina científica, pero los efectos nocivos son, sin duda, considerables*" (p. 796).

Tales críticas han estimulado un debate profundo, que continúa en la actualidad, entre los detractores del procedimiento NHST que arguyen que éste debería ser abandonado y/o prohibido (e.g., Caver, 1978; Cumming, 2014; Meehl, 1978; Trafimow y Marks, 2015) y sus defensores (e.g., Abelson, 1997; Cortina y Dunlap, 1997; Frick, 1996; Sakaluk, 2016), que ha quedado reflejado en los números especiales de revistas prestigiosas (e.g., *Journal of Experimental Education* en 1993; *Psychological Science* en 1997; *Research in the Schools* en 1998; *Journal of Psychology* en 2009, *Perspective on Psychological Science*, 2012; *Journal of Experimental Social Psychology*, 2016); en simposios de las reuniones anuales de la *American Psychological Association*, *American Psychological Society* y *American Educational Research Association*; y finalmente, en la edición de libros exclusivamente dedicados al procedimiento de la NHST como por ejemplo "*The significance test controversy*" (Morrsion y Henkel, 1970) o "*What if there were no significance tests*? (Harlow, Mulaik, y Steiger, 1997).

Como Wagenmarkers (2007) señala, los temas de discusión en el ámbito de la Psicología se han focalizado en los problemas de interpretación del procedimiento de la NHST. En sus propias palabras:

> La discusión sobre la NHST en la literatura psicológica se ha centrado en los problemas de interpretación, mientras que en la literatura estadística se ha focalizado sobre los problemas de la construcción formal del procedimiento de la NHST. Por tanto, la perspectiva estadística sobre los problemas asociados con el procedimiento de la NHST es fundamentalmente diferente a la perspectiva psicológica (p. 779).

De acuerdo a Harrison y cols., (2009) y Wagenmarkers (2007) los temas que han predominado en la discusión sobre la NHST en la literatura psicológica son:

(1) *La prueba de la NHST no nos dice lo que queremos saber* (Aguinis y cols., 2010; Carver, 1993, Cohen, 1994; Kline, 2013).

Lo que los investigadores, profesionales de la Psicología y de las Ciencias de la Salud quieren saber, y por tanto, desean que la NHST indique, es la probabilidad de que, dados nuestros datos, la hipótesis nula sea verdadera en la población ($P(H_0/Datos)$). Sin embargo, lo que la prueba de la NHST nos dice es la probabilidad de obtener unos datos (o datos más extremos) si la hipótesis nula es verdadera ($P(Datos/H_0)$).

(2) *La H_0 siempre puede ser rechazada con tamaños muestrales grandes* (Falk y Greenbaum, 1995; Kalinowski y Fidler, 2010; Thompson, 1998).

La mayoría de los investigadores coincide en señalar que, dada una muestra lo suficientemente grande, por lo menos para una variable de resultado que está al menos en escala de intervalo, la hipótesis nula casi siempre será rechazada (Bakan, 1966; Falk y Greenbaum, 1995; Thompson, 2006). En otras palabras, la probabilidad de rechazar la hipótesis nula se incrementa a medida que aumenta el tamaño de la muestra. Por lo tanto, la NHST da más información acerca del tamaño de la muestra que de la hipótesis nula (Tressoldi, Giofre, Sella, y Cumming, 2013).

(3) *En el mundo real la hipótesis nula nunca es exactamente verdadera*

Algunos investigadores sostienen que la hipótesis nula siempre es falsa en la población (Cohen, 1994; Schmidt, 1996; Tressoldi y cols., 2013), por lo que siempre será rechazada cuando se incremente el número de observaciones o el tamaño de la muestra.

(4) *Un resultado estadísticamente significativo no implica un resultado importante* (Aguinis y cols., 2010; Thompson, 1996).

Como señalan Harrison y cols. (2009), la prueba NHST no indica la importancia de los resultados, porque los eventos o sucesos improbables no son necesariamente importantes. Muchos investigadores y profesionales del ámbito de la Psicología y de la Salud y lectores de la investigación malinterpretan el término "significativo" en el sentido de "resultados importantes". El término "estadísticamente significativo" no es sinónimo de importancia desde el punto de vista de su significación clínica o práctica (Fethney, 2010; Thompson, 1996).

(5) *Las pruebas de la NHST no pueden determinar la replicabilidad de los resultados* (Carver, 1978; Cumming, 2008; Tressoldi y cols., 2013).

Muchos investigadores confunden resultados improbables en el sentido de resultados reproducibles (Harrison y cols., 2009). Sin embargo, si se repite la investigación, es probable que el valor de *p* sea diferente, lo que significa que los valores de *p* son una pobre medida de la replicabilidad de los resultados (Tressoldi y cols., 2013).

En definitiva, desde la perspectiva psicológica como Frías, Pascual y García (2002) señalan, "*las críticas al procedimiento incluyen desde las falsas concepciones sobre la información que facilita el procedimiento, provocando interpretaciones erróneas de sus resultados, hasta su uso inadecuado como medio de obtención de datos de significación práctica, sustantiva o clínica*" (p. 181).

En otras palabras, hay dos focos principales del debate sobre la prueba de la NHST en el ámbito de la Psicología: (i) su lógica inherente y lo que puede y no puede hacer, y (ii) la forma en que se ha utilizado e interpretado por los psicólogos (Finch, Thomason y Cumming, 2002). La mayoría de estas críticas son errores de interpretación (falacias sobre el valor *p*), mientras otros son defectos específicos en la lógica de la síntesis del propio procedimiento. El debate sobre ambas cuestiones continúa en la actualidad, así como sobre cuáles son las consecuencias para la práctica (Cumming, 2014; Kehle, Bray, Chafouleas, y Kawano, 2007; Perezgonzalez, 2014; Téllez, García, y Corral-Verdugo, 2015; Tressoldi y cols., 2013).

1.4. Significación estadística: valor *p*

El valor *p* se define como la probabilidad del resultado observado (o resultados más extremos) en los datos de una muestra vinculado a una prueba estadística (ANOVA, *t* de Student, Chi Cuadrado, correlación, análisis de regresión…) dentro de la distribución de la hipótesis nula que es conocida. El valor *p* es la probabilidad del resultado observado (o resultados más extremos) dado que la hipótesis nula sea cierta (Gill, 1999; Hubbard y Lindsay, 2008; Frías-Navarro, 2011a; Johnson, 1999; Kline, 2004, 2013). Es decir, el valor *p* es la probabilidad de los datos (o más extremos) asumiendo que la hipótesis nula es cierta ($P(Datos/H_0)$), que no es lo mismo que la probabilidad de los datos (P(Datos). Por lo tanto, el valor de *p* es una probabilidad acumulada en lugar de una probabilidad exacta: cubre el área de probabilidad que se extiende desde los resultados observados hacia la cola de la distribución (Carver, 1978; Hubbard, 2004). La distribución de la hipótesis nula representa todos los valores posibles que se podrían esperar si realmente no hubiese efecto (efecto cero = hipótesis nula cierta). El valor *p* es la probabilidad de los datos, no de la hipótesis nula.

Cuando el resultado observado está unido a un valor de $p \leq .05$ está indicando que, si realmente no hay efecto en la población (la hipótesis nula es cierta), entonces únicamente aparecerá ese resultado o más extremo en menos del 5% de las ocasiones. Es decir, la probabilidad de obtener ese resultado (o un resultado mayor) será menor a .05 y no ofrece ningún tipo de información sobre la probabilidad de la hipótesis nula dados ciertos datos. Cuanto más pequeño es el valor *p* más improbable es el resultado observado si la hipótesis nula es cierta.

Por lo tanto, cuando el valor de *p* es menor o igual a .05 ($p \leq .05$) sólo se pueden concluir dos cosas: que el efecto observado es improbable o que ese efecto no corresponde a la distribución de la hipótesis nula. El valor *p* no señala cuál de las dos respuestas es la correcta. Se produce el rechazo de H_0 cuando, asumiendo que la hipótesis nula es cierta, también se asume un riesgo de equivocarse (valor de alfa) al concluir que el dato del estudio probablemente no corresponde a la distribución de la hipótesis nula, que plantea efecto cero y atribuye los efectos improbables al error aleatorio.

Así pues, cuando el investigador aplica el procedimiento NHST y obtiene que $p < .05$ concluye que el efecto observado tiene una probabilidad muy baja de ser el resultado del azar o del error aleatorio y por eso rechaza la hipótesis nula (asumiendo un riesgo de Error de Tipo I), pero, de rechazar la hipótesis nula ($p <$ alfa) no se sigue de forma directa que dicha hipótesis es falsa, ya que solamente podemos concluir que los datos serían improbables, si la hipótesis fuera verdadera. Pero nunca se puede demostrar ni su falsedad ni su certeza.

En el procedimiento actual de la NHST el valor *p* no cuantifica la evidencia a favor de la hipótesis nula y, por ello, es erróneo concluir que se acepta la hipótesis nula cuando el valor de *p* es mayor al de alfa (Wilkinson y TFSI, 1999). Además, la decisión de rechazo no es una implicación directa del valor *p* pues faltaría el valor de alfa y quizás aquí comienza la interpretación incorrecta que el investigador realiza del valor *p*.

1.5. Errores de interpretación del valor *p*

Uno de los mayores problemas con la NHST se refiere a la interpretación de la significación estadística de los resultados (Cumming, 2012; Frías-Navarro, 2011a; Gliner, Vaske y Morgan, 2001; Kline, 2013; Lambdin, 2012; Morgan, 2003; Verdam, Oort, y Sprangers, 2014; Wasserstein y Lazar, 2016), es decir, a la interpretación de los valores *p* de probabilidad vinculados a las pruebas significación estadística.

Como ya se ha comentado, el valor *p* se define como la probabilidad del resultado observado o un valor más extremo si la hipótesis nula es cierta (Fidler, 2005; Hubbard, 2004; Hubbard y Lindsay, 2008; Johnson, 1999; Kline, 2013; Lambdin, 2012)

La definición es clara y precisa. Sin embargo, las interpretaciones inadecuadas y, por tanto, las conclusiones improcedentes, han proliferado y siguen proliferando (e.g., Verdam y cols., 2014), lo que resulta lógico, como Borges y cols. (2001, p. 173) señalan, puesto que el contrate de hipótesis *"que se ha consolidado y que se maneja en la actualidad supone un maridaje extraño entre dos posturas esencialmente incompatibles que nunca pretendieron aunarse: la de Fisher, de una parte, y la de Neyman y Pearson de otra"* (en el mismo sentido Gigerenzer y Murray, 1987; Gigerenzer, 2004; Gill, 1999; Hubbard, 2004; Lambdin, 2012; MacDonald, 1997).

En la literatura se han descrito diversas interpretaciones erróneas del valor *p* (e.g., Badenes-Ribera y Frías-Navarro, 2014; Bakan, 1966; Carver, 1978; Cohen, 1994; Gigerenzer, 2004; Gill, 1999; Goodman, 1999, 2008; Greenland y Poole, 2011; Johnson, 1999; Kline, 2004, 2013; Nickerson, 2000; Vallecillos, 2002; Vallecillos, 2002; Vallecillos y Batanero, 1997; Verdam y cols., 2014; Wasserstein y Lazar, 2016).

Por ejemplo, Fidler (2005), Frías-Navarro (2011a) y Lambdin (2012) enumeran algunos de estos errores de interpretación:

1. Un valor de *p* es la probabilidad de que los resultados se replicarán si el estudio se realiza de nuevo.
2. Debemos tener más confianza en los valores *p* obtenidos con muestras más grandes que con muestras pequeñas.
3. Un valor de *p* es una medida del grado de confianza en el resultado obtenido.
4. Un valor de *p* automatiza el proceso de hacer una inferencia inductiva.
5. Las pruebas de significación dan objetividad al proceso inferencial.
6. Un valor de *p* es una inferencia desde los parámetros de la población a nuestra hipótesis de investigación.
7. Un valor de *p* es una medida de la confianza que debemos tener en la veracidad de nuestra hipótesis de investigación.
8. Un valor de *p* nos dice algo sobre los miembros de la muestra.
9. Un valor de *p* es una medida de la validez de las inducciones hechas en base a los resultados.
10. Un valor de *p* es la probabilidad de que la hipótesis nula sea verdadera (o falsa) dado los datos.
11. Un valor de *p* es la probabilidad de que la hipótesis alternativa sea verdadera (o falsa) dado los datos.
12. El valor *p* es un indicador del tamaño del efecto.
13. Un resultado estadísticamente no significativo es evidencia de efecto nulo.
14. El valor *p* es el valor de alfa.
15. El valor *p* indica la importancia del efecto.

Las cuatro interpretaciones más comunes son (Harrison y cols., 2009; Kalinowski y Fidler, 2010; Kehle y cols., 2007; Wasserstein y Lazar, 2016):

- "Falacia de la probabilidad inversa" (*inverse probability fallacy*)
- "Falacia de la replicación" (*replication fallacy*)
- "Falacia del tamaño del efecto" (*magnitude fallacy*)
- "Falacia de la significación clínica o práctica" (*clinical or practical significance fallac*y)
- A continuación se expondrán cada una de estas cuatro falsas creencias sobre el valor *p* o significación estadística de los resultados.

1.5.1. Falacia de la probabilidad inversa

La "falacia de la probabilidad inversa" (*inverse probability fallacy*) hace referencia a la creencia errónea de que el valor *p* indica la probabilidad de que la hipótesis nula (H_0) sea verdadera dado ciertos datos ($Pr(H_0|Datos)$). Supone creer que las pruebas de significación estadística proveen información sobre la veracidad de la H_0 dados los datos obtenidos en una muestra (Balluerka y cols., 2005; Frías y cols., 2002; Greenland y Poole, 2011). De acuerdo a Gill (1999), esta falsa creencia derivaría directamente de la perspectiva de Fisher. Recordemos que desde la perspectiva de Fisher, los valores pequeños de *p* son tomados como evidencia en contra de la H_0, por lo que cuanto menor sea el valor de *p* más fuerte es la evidencia que proporciona. Por ello, muchos investigadores tienen la creencia de que cuanto menor es el valor de *p* asociado a un test de inferencia estadística (Chi cuadrado, ANOVA, *t* de Student, etc.), mayor es la probabilidad de que la hipótesis nula sea falsa (Carver 1978; Cohen, 1994; Gill, 1999; Lambdin, 2012; Meehl, 1990). En consecuencia creen que valores pequeños de *p* suponen mayor evidencia de la falsedad de la H_0 ($P(H_0/Datos)$). Por ejemplo, Finch, Cumming, y Thomason (2001) encontraron que el 38% de los artículos publicados en la revista *Journal of Applied Psychology* en los últimos 60 años que informaron un resultado estadísticamente no significativo lo interpretaron como una demostración de que la hipótesis nula era verdadera.

Esta creencia errónea supone que las pruebas de inferencia estadística (NHST, test de significación, test de hipótesis estadística) calculan la probabilidad de que la H_0 sea verdadera dado ciertos datos ($P(H_0/Datos)$). Sin embargo, las pruebas de inferencia estadística solamente dan información sobre la probabilidad de obtener unos resultados iguales o más extremos que los obtenidos bajo el supuesto de que la H_0 es verdadera ($P(Datos/H_0)$). Es decir, parten de que la H_0 es verdadera, y entonces se preguntan cuál es la probabilidad de observar esos datos o más extremos, dada esa condición ($P(Datos/H_0)$). Por tanto, no ofrece información sobre la probabilidad condicional de la hipótesis nula basada en los datos obtenidos en un estudio (Kirk, 1996; Greenland y Poole, 2011, Shaver, 1993), que es lo que los investigadores quieren saber, es decir, si la hipótesis nula es verdadera dados los datos obtenidos en una muestra $P(H_0/Datos)$ (Aguinis y cols., 2010; Cohen, 1994; Kirk, 1996; Kline, 2013; Lambdin, 2012).

Las técnicas de inferencia bayesiana permiten dar respuesta a la probabilidad de la hipótesis nula dados ciertos datos así como cuantificar en qué grado los datos apoyan dicha hipótesis (Cumming, 2012, 2014; Gill, 1999; Nickerson, 2000; Wagenmakers, 2007).

Dentro de la "falacia de la probabilidad inversa", también encontramos la creencia errónea de que la prueba de la NHST permite hacer afirmaciones sobre **la probabilidad de la hipótesis alternativa o de investigación**. A este respecto, las pruebas de inferencia estadística no permiten hacer afirmaciones sobre la probabilidad de que la hipótesis alternativa sea correcta (Kline, 2004; Lambdin, 2012; Morgan, 2003), ni, por tanto, que la teoría subyacente sea confirmada (Wasserstein y Lazar, 2016). Según Cohen (1994), la decisión dicotómica de rechazar, o no rechazar, la hipótesis nula no permite comprobar una teoría psicológica, pues la hipótesis alternativa ni es especificada ni es comprobada por el procedimiento de la NHST. En otras palabras,

> el procedimiento de la NHST no especifica realizar predicciones sobre la hipótesis de investigación o científica sino que las predicciones se realizan sobre la hipótesis estadística de nulidad y será su rechazo la que dará crédito a la hipótesis de investigación, que ni es especificada ni es comprobada con dicho procedimiento estadístico (Frías y cols., 2002, p. 182)

Sin embargo, la prueba de la NHST proporciona la ilusión de que la hipótesis alternativa ha sido confirmada, ya que el rechazo de la hipótesis nula se ve a menudo (erróneamente) como evidencia directa, en lugar de indirecta, de la validez de la hipótesis de investigación (Gill, 1999).

Esta falsa creencia derivaría de la perspectiva de Neyman y Pearson donde el rechazo de la H_0 supone la aceptación de la H_1 (Perezgonzalez, 2014, 2015).

Como Kline (2004) señala, esta falacia refleja dos errores conceptuales: (1) rechazar la hipótesis nula en un único estudio no implica que se haya probado la hipótesis alternativa y, (2) que la hipótesis estadística de la hipótesis alternativa sea aceptada, no significa que la hipótesis sustantiva de la hipótesis alternativa también lo sea. Por tanto, en este punto, es crucial distinguir entre las hipótesis estadísticas (H_0 y H_1) y las hipótesis sustantivas o científicas, las cuales difieren en sus niveles de abstracción y en las implicaciones que siguen al rechazo de la hipótesis nula.

Como se ha comentado, la prueba de la NHST sólo informa de la probabilidad de los datos, dado que la hipótesis nula sea verdadera. La respuesta a esta pregunta no permite a los investigadores confirmar la hipótesis teórica de investigación (Morgan, 2003). Además, conviene tener presente que ante el rechazo de la hipótesis nula se acepta la hipótesis alternativa y como consecuencia el investigador vincula ésta a su hipótesis teórica (entre muchas posibles explicaciones alternativas) dada la calidad y validez de su diseño de investigación.

Como Lambdin (2012) afirma,

> La ecuación, $P(D|H_0) \neq P(H_1|D)$, implica que una baja probabilidad de un resultado (o datos) dada la verdad de la hipótesis nula no indica la probabilidad de la hipótesis alternativa dados los datos. Tal declaración equivocada se presta muy fácilmente a la creencia errónea de que el rechazo de la hipótesis nula indica que el tratamiento funciona. Que el tratamiento no funcione no es su hipótesis nula y que funcione el tratamiento no es la hipótesis alternativa. Su hipótesis nula es probablemente que $\mu_A-\mu_B = 0$ y su hipótesis alternativa es probablemente $\mu_A-\mu_B \neq 0$. Rechazar la H_0 sólo implica que $\mu_A-\mu_B \neq 0$, no que el tratamiento funcione. Hay casi un número infinito de razones por las que $\mu_A-\mu_B \neq 0$ (p. 75-76).
>
> Con una [hipótesis] nula de no diferencia y un α de .05, lo que indica un resultado [estadísticamente] significativo es que se encontraría la diferencia obtenida menos del 5% de las veces si en realidad $\mu_A-\mu_B = 0$. La [hipótesis] nula sólo se refiere a la hipótesis

> estadística. Que funcione el tratamiento es la hipótesis de investigación (p. 75-76, entre corchetes propio).

En definitiva, el apoyo a una hipótesis de investigación contra todas las hipótesis rivales que compiten a la hora de explicar un efecto observado en una investigación no viene dado por las pruebas de la NHST sino por la calidad y validez del diseño de la investigación y los estudios de replicación que corroboran la evidencia que demuestra el efecto en una variedad de situaciones (Carver, 1978; Frías-Navarro, 2011a; Lambdin, 2012).

1.5.2. Falacia de la replicación

A menudo, la significación estadística se toma como evidencia de la probabilidad de replicar (o de la fiabilidad de) los resultados obtenidos en una investigación (Borges, 1997; Carver, 1978; Kalinowski y Fidler, 2010; Nickerson, 2000). En este sentido, los valores pequeños de *p* son interpretados como indicadores de una fuerte probabilidad de obtener los mismos resultados estadísticamente significativos en otra investigación. Y, en algunos casos, el complemento de *p* (1 - *p*) es interpretado como un indicador de la probabilidad exacta de replicación.

Carver (1978) se refirió a esta creencia como la "fantasía de replicabilidad o de la fiabilidad". Por su parte, Falk y Greenbaum (1995) se refirieron a esta falsa creencia como la "falacia de replicación". Estos autores observaron que muchos investigadores creían que un valor pequeño de *p* tal como $p = .05$ implicaba que 95 de cada 100 repeticiones serían estadísticamente significativas.

Sin embargo, como Thompson (2003) afirma:

> Si el *p* calculado [valor de *p*] informara al investigador acerca de la verdad de la hipótesis nula en la población, a continuación, esta información podría probar directamente la posibilidad de reproducir los resultados. ... Desafortunadamente, esto no es lo que las pruebas de significación estadística evalúan, y no es lo que evalúan los *p* calculados (p. 96, entre corchetes propio).

En consecuencia, es falso afirmar 1 - *p* es la probabilidad de que los resultados sean replicables o fiables (Carver, 1978). El error es evidente cuando se recuerda que $p = P(Datos/H_0)$ y es por tanto una función de un único conjunto de datos que produce una prueba estadística. Lo que realmente nos interesa en términos de replicación es la distribución de esta prueba estadística en repetidas investigaciones.

Sin embargo, las pruebas de significación estadística no evalúan de un modo concluyente la replicabilidad de los datos. Para replicar los resultados se requieren otras estrategias como por ejemplo la realización de nuevos experimentos (Monterde-i-Bort y cols., 2010; Monterde-i-Bort, Pascual-Llobell y Frías-Navarro, 2006; Pascual y cols., 2000). El valor *p* no es un indicador de la confianza en la prueba estadística. Es un valor vinculado a un resultado concreto de una muestra y no ofrece información sobre la distribución del estadístico y los posibles resultados con otros conjuntos de datos (Gill, 1999).

No obstante, en determinadas condiciones los valores de *p* están relacionados con la replicación, pero la probabilidad de replicación generalmente no es 1- *p* (Kline, 2013). Como señalan Greenwald, Gonzalez y Guthiere (1996, citados en Kline, 2013, p. 98) existe una relación curvilínea entre los valores de *p* y la potencia estadística media en repeticiones aleatorias hipotéticas basadas en el mismo número de casos. Esta relación monotónica creciente entre replicabilidad y *p* no se mantiene si suponemos que la hipótesis de nula es verdadera. Por tanto, no existe método más objetivo de saber si un fenómeno es fiable que la replicación empírica del mismo, pues los efectos fiables serán repetibles en posteriores afirmaciones mientras que los efectos aleatorios no lo serán (Pascual y cols., 2000).

En definitiva, como Lambdin (2012) afirma, es una ilusión pensar que se puede aprender algo acerca de la replicabilidad de los resultados desde un valor *p*. Recordemos que el valor *p* no es la probabilidad de los datos, P(Datos); el valor *p* es la probabilidad de los datos (o datos más extremos) asumiendo que la hipótesis nula es cierta, $P(Datos/H_0)$), es por tanto una probabilidad condicionada.

1.5.3. Falacia del tamaño del efecto

Un error común en la interpretación de los resultados es el de equiparar la significación estadística (es decir, la observación de un resultado con un valor de *p* menor que el nivel de significación preestablecida, típicamente .05) con el tamaño o magnitud del efecto (Cohen, 1994; Frías y cols., 2002; Gliner y cols., 2001; Kalinowski y Fidler, 2010; Kline, 2004, 2013; Motulsky, 2015; Nickerson, 2000; Wasserstein y Lazar, 2016). Es decir, es un error suponer que si el valor de *p* es pequeño (por ejemplo $p < .05$), entonces el efecto debe ser grande (Kline, 2004, 2013; Wasserstein y Lazar, 2016).

Kline (2013) llama a esta creencia errónea la "falacia de la magnitud" (*magnitude fallacy*). Esta falacia del tamaño representa una de las críticas más fuertes contra las pruebas de significación de la hipótesis nula, y podría subyacer en la deficiencia de los informes científicos publicados en revistas de impacto a la hora de informar de estadísticos del tamaño del efecto. Los investigadores y los revisores de las revistas se podrían plantear la siguiente cuestión: ¿Por qué y para qué molestarse en informar de un tamaño del efecto cuando se cree que el valor *p* es un indicador del mismo? (Fidler, 2005; Kirk, 2001).

Una razón por la cual los valores de *p* no se pueden utilizar como índices del tamaño del efecto es porque en los valores *p* se confunde el valor del efecto con el tamaño de la muestra, puesto que tanto el tamaño del efecto como el tamaño de la muestra tienen un impacto en el valor de *p* (Thompson, 2006). Es decir, el valor *p* es producto del tamaño del efecto y del tamaño de la muestra (Botella y Sánchez-Meca, 2015; Frías y cols., 2002; Grissom y Kim, 2012). En consecuencia, es posible obtener valores de *p* pequeños (y resultados estadísticamente significativos), con muestras muy grandes y tamaños del efecto pequeños o triviales, y también se pueden obtener valores de *p* pequeños con muestras pequeñas y tamaños del efecto grandes (Kalinowski y Fidler, 2010). Por lo que el valor *p* nada dice acerca del tamaño del efecto.

Dicho de otro modo, con una muestra grande o con una pequeña variabilidad, incluso un efecto muy pequeño puede llegar a ser estadísticamente significativo, y con una muestra pequeña o con gran variabilidad incluso un efecto grande puede dejar de alcanzar el nivel convencional de significación estadística ($p < .05$) (Nickerson, 2000). En el mismo sentido, Fidler (2005) advierte que se pueden encontrar tamaños del efecto pequeños pero estadísticamente significativos en muestras de alta potencia estadística (por ejemplo, muestras suficientemente grandes), y tamaños del efecto grandes sin significación estadística en investigaciones con pobre diseño y baja potencia estadística.

Por ejemplo, Kalinowski y Fidler (2010) citan un estudio con alta potencia, efecto trivial (irrisorio) pero estadísticamente significativo. Este estudio analiza el efecto que tienen las dosis bajas de aspirina en la reducción de los ataques al corazón en una muestra de 22,071 hombres. En él, a 11,034 hombres se les dio una aspirina que tenían que tomar cada 2 días (grupo experimental), mientras que a los otros 11,037 hombres les dieron un placebo (grupo control). Los resultados indicaron una diferencia estadísticamente significativa entre los grupos ($p < .000001$) en la reducción de los

ataques al corazón a favor del grupo experimental. Sin embargo, el tamaño del efecto fue pequeño (r^2 = .0011). Este pequeño tamaño del efecto podría tener relevancia práctica o clínica si la aspirina ayudara a prevenir los infartos y salvar vidas con una inversión económica baja. Por lo tanto, la significación estadística, el tamaño del efecto y la importancia clínica o práctica son conceptos diferentes.

En definitiva, el valor de *p* no es un indicador directo del tamaño del efecto, puesto que en el valor de *p* se confunde el valor de la magnitud del efecto y el tamaño de la muestra (Botella y Sánchez-Meca, 2015; Nickerson, 2000). De hecho, la prueba de inferencia estadística basada en muestras de gran tamaño casi siempre produce resultados estadísticamente significativos (Borges, 1997; Gill, 1999; Harrison y cols., 2009, Macdonald, 1997; Oakes, 1986). En consecuencia, se pueden obtener valores de *p* pequeños (resultados estadísticamente significativos), aumentando el tamaño de la muestra (Cohen, 1994). Esta conducta de aumentar el tamaño de la muestra para obtener resultados estadísticamente significativos entra dentro de lo que actualmente se conoce como "*p-hacking*" (o "pirateo de los valores *p*") (Head, Holman, Lanfear, Kahn, y Jennions, 2015; Motulsky, 2015; Nuzzo, 2014), del que se hablará sucintamente en el último apartado de este capítulo.

Por tanto, como señalan Aguinis y cols. (2010), la significación estadística de los resultados no dice nada acerca de la fuerza o magnitud del efecto (en el caso de los diseños experimentales en los que la causalidad puede ser inferida) o de la relación (en diseños no experimentales en los que solamente se puede hacer inferencia sobre la covariación de las variables). Por ejemplo, si en un estudio sobre colaboración entre los miembros de un equipo y el rendimiento del equipo los resultados indican que existe una relación estadísticamente significativa entre ambas variables, esta significación estadística no indica en qué grado la colaboración del equipo se relaciona con el rendimiento del equipo. Si el resultado es estadísticamente significativo, se llega a la conclusión de que estas variables están relacionadas entre sí, pero no se sabe en qué magnitud. Para ello, tenemos que calcular las estimaciones de la magnitud de su relación (*op. cit*.).

Así pues, el tamaño del efecto sólo puede ser conocido si es estimado (Cumming, 2012; Kalinowski y Fidler, 2010; Kline, 2013)**.** El tamaño del efecto es una expresión estadística de la magnitud de la relación entre dos variables, o la magnitud de la diferencia entre los grupos con respecto a algunos atributos de interés. Por lo tanto,

los tamaños del efecto contienen información acerca de la magnitud o la dirección de los resultados de los estudios de investigación cuantitativa, o ambos (Lipsey y Wilson, 2001). Además de proporcionar información importante acerca del impacto de un tratamiento sobre el resultado de interés o la asociación entre variables, los índices del tamaño del efecto también proporcionan una métrica común para comparar la dirección y la fuerza de la relación entre las variables entre los estudios, lo que es clave para la realización de estudios de meta-análisis (Frías-Navarro, 2011b; Kelley y Praecher, 2012; Valera-Espín y Sánchez-Meca, 1997).

1.5.4. Falacia de la significación clínica o práctica

Otro error común en la interpretación de los resultados es la "falacia de la significación clínica o práctica", la cual vincula la significación estadística de los resultados de una investigación (valor *p*) con la significación clínica o practica de los mismos (Berben, Sereika, y Engberg, 2012; Fethney, 2010; Fidler, 2005; Frías y cols., 2002; Gelman y Stern, 2006; Head y cols., 2015; Kalinowski y Fidler, 2010; Wasserstein y Lazar, 2016). Es decir, se equipara la significación estadística y la significación (importancia) clínica (Aguinis y cols., 2010; Verdam y cols., 2014; Wasserstein y Lazar, 2016). De este modo, los resultados estadísticamente significativos son considerados importantes, con relevancia clínica o práctica. Y los resultados no estadísticamente significativos son considerados como efecto nulo o sin importancia. Sin embargo, un resultado estadísticamente significativo no indica necesariamente que sea un resultado importante desde el punto de vista clínico, práctico o sustantivo y viceversa (Gliner y cols., 2002; Kirk, 1996; Morgan, 2003; Nickerson, 2000; Palmer y Sesé, 2013). Un ejemplo citado en Gelman y Stern (2006), supongamos que el efecto estimado de un medicamento para disminuir la presión arterial es de 0.10 con un error estándar de 0.03, esto sería estadísticamente significativo, pero probablemente no tendría importancia práctica. Sin embargo, a la inversa, un efecto estimado de 10 con un error estándar de 10 no sería estadísticamente significativo, pero tendría la posibilidad de ser importante en la práctica.

Como Kazdin (2008) afirma

> La significación estadística es una función del tamaño de la muestra y la variabilidad dentro y entre los sujetos. La diferencia requerida para la significación [estadística] en el resultado (por ejemplo, en las medidas de la ansiedad, la discordia marital) puede no reflejar una diferencia detectable o real en la vida cotidiana de cualquier cliente

> individual o incluso del grupo. En resumen, las conclusiones sobre el tratamiento que se basan en estudios que muestran diferencias estadísticas son difíciles de traducir en efectos sobre la vida de los participantes en el estudio, y mucho menos generalizar a los pacientes atendidos en la práctica (p. 148, entre corchetes propio).

Por lo tanto, en la valoración de los resultados de la investigación hay que distinguir la significación estadística de los resultados (valor *p*) y la significación clínica o práctica de los mismos (McGouh y Faraone, 2009). Para establecer esta distinción ayudaría emplear la expresión "estadísticamente significativo" en lugar de solamente "significativo". Pues como Thompson (1999) afirma:

> […] los autores deben emplear la frase completa, "estadísticamente significativo", en lugar de simplemente "significativo" porque siento que muchos investigadores confunden el significado coloquial de "significativo" con el significado técnico de "estadísticamente significativo (p. 163-164)

Dado que como se ha expuesto la significación estadística no es sinónima de significación clínica, veamos pues qué se entiende por significación clínica.

1.6. Significación clínica

En 1984 Jacobsen, Follette y Revenstorf propusieron el término "significación clínica" como una manera de determinar el valor práctico de los tratamientos en contraposición a la significación estadística de los resultados. De acuerdo con Kazdin (1999), la "significación clínica" se define como:

> El valor o importancia práctica o aplicada del efecto de la intervención –es decir, si la intervención tiene un diferencia real (por ejemplo, genuina, palpable, practica, notable) en la vida diaria de los clientes o de los otros con quienes los clientes interactúan (p. 332).

Así pues, la importancia o significación clínica o práctica de los resultados se refiere al grado en que el efecto de la intervención se traduce en un cambio significativo para las personas, la práctica clínica o la sociedad (Durlak, 2009; Greenfield, Kuhn, y Wojtys, 1996; Kazdin, 1999, 2001, 2008).

Pero, ¿cómo valorar la significación clínica o práctica de los resultados? No existen directrices bien establecidas para valorar la importancia clínica o práctica de los hallazgos (Fethney, 2010). Algunos autores afirman que los resultados no pueden ser clínicamente significativos si no son estadísticamente significativos; por lo que la

significación estadística se erige como una condición previa necesaria para la determinación de la significación o importancia clínica (e.g., Greenstein, 2003). En cambio, otros autores consideran que la significación estadística no necesariamente equivale a importancia clínica de los resultados (e.g. Kazdin, 1999).

Como se ha comentado, mientras que las pruebas de significación estadística proporcionan información importante acerca de los resultados del estudio, no reflejan necesariamente la importancia clínica de los resultados (Berben y cols., 2012; Frías y cols., 2002; Jacobsen y cols., 1984; Page, 2014).

Por ello, los profesionales de la salud (psicólogos clínicos, médicos, etc.) deberían estar más interesados en conocer la significación clínica o práctica de los resultados, que en conocer la significación estadística de los mismos (Fethney, 2010; Page, 2014). Pues una diferencia puede ser estadísticamente significativa y, al mismo tiempo, tener poca o ninguna importancia para la salud, el bienestar o la calidad de vida de los pacientes afectados por una determinada enfermedad. Y, por el contrario, diferencias clínicamente relevantes podrían considerarse estadísticamente insignificantes (Ferrill, Brown, y Kyle, 2010; Frías y cols., 2002; Frías-Navarro, 2011a; Gliner y cols., 2001; Morgan, 2003; Nickerson, 2000; Thompson, 1999, 2002a). Por tanto, es posible que efectos con significación o importancia práctica no alcancen la significación estadística y por tanto sean rechazados. Y, al contrario, puede haber efectos con poca significación o importancia práctica pero que hayan alcanzado la significación estadística y por ello se tomen, a menudo, como significativos o importantes (Armijo-Olivo, Warren, Fuentes, y Magee, 2011; Kirk, 1996).

Por ejemplo, una intervención o tratamiento para la depresión puede no tener ningún impacto apreciable estadísticamente en lo que respecta a diferenciar a los participantes del grupo que recibieron el tratamiento para la depresión y los participantes del grupo control, pero todavía puede hacer mucho para ayudar a las personas a lidiar con sus síntomas o para mejorar la calidad de vida (Thompson, 2002a).

Otro ejemplo citado por Frías-Navarro (2011a), en un estudio sobre distintos tratamientos para perder peso en personas obesas, se encontró que los participantes obesos que recibían terapia de grupo perdían más peso que los participantes que recibían terapia individual, siendo la diferencia entre ambos grupos estadísticamente significativa. Sin embargo, la pérdida de 1.9kg no fue clínicamente significativa. El

Instituto Nacional de Salud de los Estados Unidos señala que sólo una reducción de peso del 10% o más es clínicamente significativa. Por lo tanto una reducción de peso de 1.9kg es una pérdida pequeña en términos de utilidad clínica, práctica o sustantiva.

En definitiva, los valores de *p* no evalúan la importancia del resultado, y, en consecuencia, no deberían ser utilizados como un vehículo eficaz para soslayar la responsabilidad de hacer un juicio subjetivo sobre el valor real, práctico, de los resultados (Thompson, 1999). Pues como Thompson (1993) señaló,

> Los estadísticos se pueden utilizar para evaluar la probabilidad de un suceso. Pero la importancia es una cuestión de valores humanos, y las matemáticas no pueden ser utilizadas como una vía de escape atávica (a la *Miedo a la libertad* de Fromm) de la responsabilidad humana existencial de hacer juicios de valor. Si el paquete del ordenador no le preguntó por sus valores antes del análisis, él (el ordenador) podría no haber considerado su sistema de valores en el cálculo de *p*, y por tanto, *p* no puede ser alegremente utilizado para inferir el valor de los resultados de la investigación (p. 365).

Por otra parte, la magnitud del tamaño del efecto se ha interpretado como un índice de relevancia o importancia clínica (Kirk, 1996). En este sentido, cuanto más grande es el tamaño del efecto, más grande es la diferencia entre los grupos y por tanto más grande se ha considerado que es la importancia clínica de los resultados (Musselman, 2007). Cohen (1988) sugirió unas reglas generales para la interpretación de los valores de los estadísticos del tamaño de las diferencias de medias estandarizadas (pequeñas = 0.20, medias = 0.50 y grandes = 0.80) y de las correlaciones (pequeñas = .10, medias = .30 y grandes = .50). Desde este punto de vista, por ejemplo, los investigadores podrían considerar una magnitud del efecto de r = .40 como clínicamente relevante, ya que este valor del tamaño del efecto podrían representar un efecto moderado que podría ser de interés para la práctica clínica dentro de un contexto determinado.

Sin embargo, los estadísticos del tamaño del efecto únicamente proporcionan información acerca de la magnitud, pero el efecto no siempre tiene importancia o relevancia clínica (Grissom y Kim, 2012; Kalinowski y Fidler, 2010; Kazdin, 2001; Ogles, Kirk, Lunnen, y Bonesteel, 2001). Como Nickerson (2000) señala,

> un gran efecto no es una garantía de importancia clínica o práctica, no más que un pequeño valor de *p*; aunque, como regla general, un gran efecto parece más probable que sea importante que uno pequeño, al menos desde un punto de vista práctico (p. 257).

Así pues, los índices del tamaño del efecto pueden ayudar a evaluar la significación clínica o práctica de los resultados, puesto que los juicios sobre la importancia clínica o práctica de los resultados son en gran medida subjetivos (Aguinis y cols., 2010; Frías-Navarro, 2011b; Kirk, 2001; Thompson, 1993). Por esta razón, el tamaño del efecto debe interpretarse en el contexto clínico del problema que se está estudiando y requiere por lo tanto del juicio del experto sobre la temática (Page, 2014; Schulz y cols., 2002).

Finalmente, existe un debate acerca de cuáles son los indicadores que se pueden utilizar para evaluar la significación clínica o practica de los resultados de una investigación (Fethney, 2010; Kazdin, 1999, 2001; Kendall, 1999). Schulz y cols. (2002) postulan un enfoque inclusivo que abarque múltiples indicadores. Así, estos autores sugieren cuatro criterios para valorar la significación clínica de los hallazgos:

(1) la sintomatología, es decir, grado en que las personas vuelven a su funcionamiento normal o experimentan un cambio en los síntomas;
(2) la calidad de vida, es decir, en qué medida las intervenciones o tratamientos en términos generales mejoran la calidad de vida de un individuo;
(3) la importancia social, es decir, grado en que los resultados son importantes para la sociedad; y
(4) la validez social, es decir, grado en que los objetivos del tratamiento, los procedimientos y los resultados son aceptables según la evaluación de los clientes/pacientes o los expertos y su impacto en la vida de los participantes.

Cabe señalar que los investigadores y los profesionales de la salud pueden tener diferentes puntos de vista sobre la importancia relativa de un tipo de indicador sobre otro, aunque es evidente que existe un cierto solapamiento entre estas categorías.

Pero, ¿cómo cuantificar la significación clínica o práctica de unos resultados?

1.6.1. Métodos para valorar la significación clínica o práctica

Una forma en la que se puede valorar la importancia clínica o práctica de los resultados es a través de lo que se conoce en la literatura como la "diferencia mínima clínicamente importante" (*minimum clinically important difference*, MCID).

La MCID es un valor umbral para determinar la existencia de un cambio importante. En este sentido, cualquier cantidad de cambio mayor que el umbral MCID

se considera que es significativa o importante (Copay, Subach, Glassman, Polly, y Schuler, 2007).

De acuerdo con Fethney (2010) existen tres métodos para valorar la MCID: (1) enfoque basado en anclajes, (2) enfoque basado en la distribución y (3) enfoque basado en el panel de expertos.

Los tres enfoques miden un cambio cuantificable en los resultados, pero la elección específica del enfoque decidirá el tipo de cambio medido (Copay y cols., 2007). A continuación se detalla sucintamente en qué consiste cada uno de estos enfoques.

(1) ***Enfoque basado en anclajes***: utiliza un criterio externo para interpretar si una determinada magnitud de cambio es significativa o no. Para ello, compara el cambio en una variable de interés con alguna otra variable medida considerada como el anclaje o criterio externo (Copay y cols., 2007; Fethney, 2010; Turner y cols., 2010). Por ejemplo, la MCID se puede determinar por los pacientes en las medidas de calidad de vida, o por los médicos en los índices de la enfermedad-actividad (Turner y cols., 2010). De acuerdo a Fethney (2010), entre ambas medidas, el resultado de la variable de interés y el resultado de la variable que actúa como anclaje, debe existir una relación o asociación.

Como Copay y cols. (2007) señalan, por lo general, se comparan las respuestas de los pacientes con otra evaluación subjetiva, normalmente una clasificación de la evaluación global en la que los pacientes se autocalifican como "mejor", "sin cambios" o "peor".

(2) ***Enfoque basado en la distribución de las puntuaciones observadas en una muestra relevante:*** consiste en comparar la magnitud de cambio observado (evaluado a través de las respuestas de los pacientes) con alguna medida de variabilidad, por ejemplo, la desviación típica (1/2 desviación típica), el error estándar de la medida, el tamaño del efecto o la diferencia mínima detectable (Jacobsen y cols., 1984; Copay y cols., 2007; Musselman, 2007). Dentro de este enfoque, también estarían las comparaciones normativas, esto es, comparar los datos de los individuos tratados con los datos de individuos normativos (e.g., Kendall, Marrs-Garcia, Nath, y Sheldrick, 1999) para determinar si el cambio observado es clínicamente significativo. Como se

observa, se trata de una aproximación puramente estadística (Baicus y Cariol, 2009; Turner y cols., 2010).

(3) ***El enfoque de panel de expertos****:* invita a los expertos en el campo a leer la literatura relevante y tratar de llegar a un consenso en cuanto a la valoración de la MCID.

Sin embargo como Kazdin (2001) afirma, la MCID evaluada a través de estos métodos no tiene necesariamente que ser equivalente a un cambio real, palpable en la vida cotidiana de la persona, porque, por ejemplo:

> Un cliente que cambia de modo que él o ella está en el rango normativo sobre una medida (evaluada) o cambia mucho (por ejemplo, 2 desviaciones típicas) puede o no estar funcionando mejor o estar funcionando bien en la vida cotidiana. Por el contrario, un cliente que no ha cambiado lo suficiente como para entrar en un rango normativo o no ha cambiado mucho en la medida (evaluada) bien puede haber cambiado de manera que afecte en gran medida a su vida cotidiana. […] Lo que hace que una diferencia sea clínicamente significativa para los investigadores puede ser discrepante con lo que hace que una diferencia sea clínicamente significativa para los clientes y aquellos con los que interactúan (p. 456-457).

Concluyendo, la significación estadística no es sinónimo de relevancia clínica o práctica de un efecto o relación entre variables. La significación clínica debe ser determinada por el investigador o profesional dentro del contexto de la investigación (o campo de especialización) atendiendo al cambio producido en la sintomatología, en la calidad de vida, etc. Esta valoración de la significación clínica debe efectuarse con una evaluación multimétodo de la diferencia mínima clínicamente significativa, puesto que la literatura muestra que los diferentes métodos producen diferentes resultados y presentan limitaciones (Baicus y Cariol, 2009; Terwee, Roorda, Knol, De Boer, y De Vel, 2009; Turner y cols., 2010). Por ejemplo, si se utiliza el tamaño del efecto como índice de significación clínica (método basado en la distribución) junto con un criterio externo (método basado en anclaje), el tamaño del efecto debería ser pequeño en pacientes que no informaron cambio alguno en el método basado en anclaje y, por el contrario, el tamaño del efecto debería ser grande en pacientes que informaron una gran mejora (Copay y cols., 2007). Finalmente, la valoración de la significación clínica también debe tener en cuenta que la MCID realmente haya producido un cambio en la vida cotidiana de la persona (Kazdin, 2001).

1.7. Estudios previos sobre errores de interpretación del valor *p*

Los errores de interpretación del valor *p* o significación estadística de los resultados han estado y siguen estando presentes desde que se comenzó a utilizar el contraste de hipótesis mediante la ejecución de pruebas de significación estadística (e.g., *t* de Student, ANOVA…), a pesar de los constantes debates sobre el significado del valor *p*. Así lo atestiguan los estudios sobre las interpretaciones erróneas del valor *p* que se han realizado desde los años 60 en el ámbito de la Psicología y que se exponen a continuación.

En 1963, Rosenthal y Gaito llevaron a cabo uno de los primeros estudios sobre la interpretación de los valores *p* en una muestra formada por 9 investigadores y 10 estudiantes graduados de Psicología de los Estados Unidos de América (EEUU). Los participantes tenían que evaluar su grado de confianza, evaluado en una escala tipo-*Likert* con 6 puntos de anclaje (desde nada de confianza "0" hasta extrema confianza o creencia "5"), en 12 valores *p* diferentes (desde $p = .001$ hasta $p = .90$) asociados a dos diferentes tamaños de muestra ($n = 10$ y $n = 100$). Estos autores encontraron que el grado de confianza tanto de los investigadores como de los estudiantes era una función decreciente del nivel de *p*, y que para algunos niveles de *p* el tamaño de la muestra más grande siempre dio lugar a una mayor confianza. Estos autores concluyeron que existía un efecto del énfasis en el nivel *p* (valor *p*) sobre el grado de confianza (llamado "*Cliff effect*"), es decir, una brusca caída de la confianza en valores de *p* por encima de .05. Más tarde, Nelson, Rosenthal y Rosnow (1986) en una muestra de 85 investigadores del campo de la Psicología de EEUU encontraron este mismo efecto de caída brusca de los niveles de confianza ("*Cliff effect"*) en los valores *p* por encima de .05.

En 2001, Poitevineau y Lecoutre replicaron el estudio de Rosenthal y Gaito (1963) en una muestra de 18 investigadores de Psicología de Francia. Los participantes tenían que evaluar su grado de confianza en 12 valores *p* diferentes (.001, .01, .03, .05, .07, .10, .15, .20, .30, .50, .70, .90) asociados con dos diferentes tamaños de muestra ($n = 10$ y $n = 100$). Estos autores también detallaron la existencia del efecto Cliff entre los participantes, si bien, en una proporción menor que en el estudio original de Rosenthal y Gaito (1963).

En 1986, Oakes realizó un estudio con 70 profesores universitarios de Psicología en el Reino Unido. El estudio consistió en presentar a los participantes una situación de investigación donde los resultados del contraste de hipótesis ejecutado a través de la prueba *t* de Student arrojaban un valor de p =.01. La tarea de los participantes consistía en señalar como verdadera o falsa un conjunto de siete afirmaciones.

El cuestionario de Oakes decía:

> Suponga que tiene un tratamiento que usted sospecha que puede alterar el rendimiento en una tarea determinada. Compare las medias del grupo control y experimental (digamos 20 sujetos en cada muestra). Además, suponga que utiliza t-test para muestras independientes y el resultado es (t = 2,7, df = 18, p = .01). Por favor, marque cada una de las declaraciones siguientes como "verdadero" o "falso". "Falso" significa que la declaración no se sigue lógicamente de las premisas anteriores. También tenga en cuenta que varias o ninguna de las afirmaciones puede ser correcta.
>
> 1) Ha refutado absolutamente la hipótesis nula (es decir, no hay diferencia entre las medias poblacionales). [] Verdadero/falso []
>
> 2) Ha encontrado la probabilidad de que la hipótesis nula sea verdadera. [] Verdadero/ falso []
>
> 3) Se ha demostrado absolutamente su hipótesis experimental (que hay una diferencia entre las medias poblacionales). [] Verdadero/falso []
>
> 4) Puede deducir la probabilidad de que la hipótesis experimental sea cierta. [] Verdadero/falso []
>
> 5) Ya sabe, si decide rechazar la hipótesis nula, la probabilidad de tomar una decisión equivocada. [] Verdadero/falso []
>
> 6) Tiene un hallazgo experimental fiable en el sentido de que si, hipotéticamente, el experimento se repite un gran número de veces, obtendría un resultado significativo en el 99% de las ocasiones. [] Verdadero/falso []
>
> 7) Se conoce la probabilidad de los datos dada la hipótesis nula [] Verdadero/falso []

De las siete afirmaciones solamente una, el ítem 7, era (y es) verdadera. El resto de afirmaciones (desde el ítem 1 hasta el ítem 6) eran (y son) concepciones erróneas, falacias, ilusiones o fantasías, en definitiva, creencias erróneas sobre el significado del valor *p*.

Los resultados del estudio de Oakes señalaron que el 97% de los profesores percibieron como verdadera al menos una de las seis opciones falsas del significado del valor *p*. Las afirmaciones que mayor respaldo recibieron fueron los ítems 4, 5 y 6. Es decir, un valor $p = .01$ significa que: "Puede deducir la probabilidad de que la hipótesis experimental sea cierta"; "Ya sabe, si decide rechazar la hipótesis nula, la probabilidad de tomar una decisión equivocada" y "Tiene un hallazgo experimental fiable en el sentido de que si, hipotéticamente, el experimento se repite un gran número de veces, obtendría un resultado significativo en el 99% de las ocasiones".

Finalmente, sólo el 11.3% de los profesores ($n = 8$) percibió como correcta la única afirmación que realmente lo era. Esto es, el valor *p* es la probabilidad de los datos dado que la hipótesis nula sea verdadera.

Por su parte, Falk y Greenbaum (1995) en una muestra de estudiantes universitarios de Israel, y usando una prueba similar a la de Oakes (1986), encontraron resultados comparables. Estos autores añadieron un ítem que hacía referencia explícita a que ninguna de las 6 alternativas era correcta ("*Ninguna de las afirmaciones es correcta*"). Como medida adicional, estos autores hicieron leer a sus alumnos el clásico artículo de Bakan (1966), que advierte explícitamente en contra de las conclusiones erróneas sobre el valor *p*. Sin embargo, sólo el 13% de sus participantes optó por la alternativa correcta. Falk y Greenbaum llegaron a la conclusión de que "*si no se toman medidas fuertes en la enseñanza de la estadística, las posibilidades de superar este error son bajas en la actualidad*" (p. 93).

En el año 2002, Haller y Krauss llevaron a cabo una replicación directa del estudio de Oakes (1986) en una muestra de profesores universitarios y estudiantes de Psicología de Alemania. En concreto, la muestra estaba formada por 30 profesores del área de Metodología que impartían docencia sobre la prueba de la NHST, 39 profesores de Psicología que no impartían docencia sobre Metodología y 44 estudiantes que habían cursado con éxito distintas asignaturas de Estadística en las que se había impartido docencia sobre el procedimiento de la NHST. Haller y Krauss observaron que el 80% de los profesores del área de Metodología, el 89.7% del resto de profesores y el 100% de los estudiantes de Psicología cometieron algún tipo de error en la interpretación del valor *p*.

La Figura 1 muestra los hallazgos generales de los estudios de Oakes (1986) y Haller y Krauss (2002) en el porcentaje de participantes que cometieron al menos un error de interpretación del valor *p*.

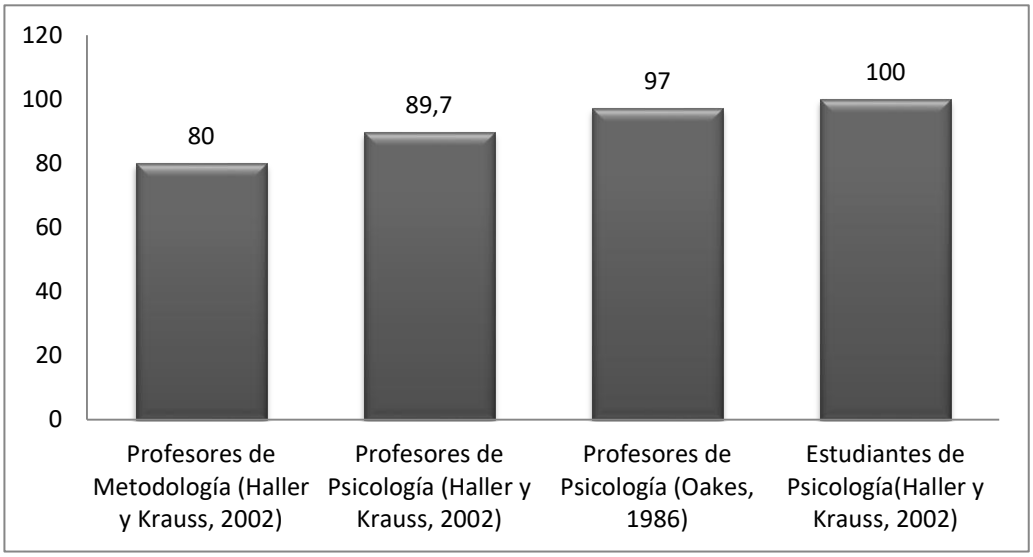


Figura 1. Porcentaje de participantes que cometieron al menos un error de interpretación del valor *p* en los estudios de Haller y Krauss (2002) y Oakes (1986) (Elaboración propia a partir de los datos de Haller y Kraus, 2002).

Se puede observar que, a pesar de haber transcurrido quince años entre los estudios de Oakes y el de Haller y Krauss, quince años de debates y críticas al procedimiento de la NHST, el porcentaje de interpretaciones erróneas del valor *p* apenas había variado entre el profesorado de Psicología.

La Tabla 1 muestra el porcentaje de acuerdos que recibieron cada una de las afirmaciones en los estudios de Haller y Kraus (2002) y Oakes (1986). Se puede observar que tanto en el estudio de Oakes (1986) como en el estudio de Haller y Krauss (2002), las afirmaciones 1 y 3 se identificaron con mayor frecuencia como falsas. Esto es, los participantes consideraron correctamente que un valor $p < .01$ no significa que se "Ha refutado absolutamente la hipótesis nula" ni que "Se ha demostrado absolutamente la hipótesis experimental". Aun así, hasta un tercio de los estudiantes y entre el 10% y 15% del grupo de profesores estuvieron de acuerdo con estas afirmaciones.

Tabla 1. Porcentaje de acuerdo en cada una de las afirmaciones en los estudios de Haller y Krauss (2002) y Oakes (1986).

Afirmaciones abreviadas	Metodólogos	No metodólogos	Estudiantes	Oakes (1986)
1. H_0 es absolutamente rechazada	10	15	34	1
2. Se demostró la probabilidad de la H_0	17	26	32	36
3. H_1 es absolutamente confirmada	10	13	20	6
4. Se demostró la probabilidad de la H_1	33	33	59	66
5. Probabilidad del Error Tipo I	73	67	68	86
6. Probabilidad de replicación	37	49	41	60

Además, cuando las falsas afirmaciones se refieren a la probabilidad de la hipótesis alternativa (ítem 4) era más aceptado por los estudiantes y profesores como una conclusión válida que cuando se refiere a la probabilidad de la hipótesis nula (ítem 2).

Gigerenzer, Kraus y Vitouch (2004) sugieren que el hecho de que los participantes sean más propensos a creer que el nivel de significación estadística determina la probabilidad de H_1 en lugar de la de H_0 se puede explicar en el hecho de que "*el enfoque de los investigadores está en la hipótesis experimental (H_1) y que el deseo de hallar la probabilidad de H_1 impulsa este fenómeno*" (p. 6).

Por otra parte, las afirmaciones que recibieron mayor respaldo por parte de los participantes fueron los ítems 4, 5 y 6. Esto es, ítem 4 "Puede deducir la probabilidad de que la hipótesis experimental sea cierta", ítem 5 "Ya sabe, si decide rechazar la hipótesis nula, la probabilidad de tomar una decisión equivocada" e ítem 6 "Tiene un hallazgo experimental fiable en el sentido de que si, hipotéticamente, el experimento se repite un gran número de veces, obtendría un resultado significativo en el 99% de las ocasiones". Estos errores de interpretación son prácticamente iguales en los tres grupos (profesores de Metodología, profesores de Psicología y estudiantes), lo que sugiere ser, como Gigerenzer y cols. (2004) señalan, "*una fantasía colectiva que parece viajar por la transmisión cultural de maestro a alumn*o" (p. 4) tanto en la muestra del estudio de Haller y Krauss (2002) como en los participantes del estudio de Oakes (1986).

En definitiva y en general, se observa una leve mejoría en la comprensión del valor *p* en el estudio de Haller y Krauss respecto del estudio de Oakes, si bien un alto porcentaje de profesores universitarios siguieron cometiendo errores de interpretación (afirmaciones 4, 5 y 6). Por lo que, como Gigerenzer (2004) señala, "*el número de ilusiones que tenían sigue siendo impresionante*" (p. 596).

Diversos autores analizaron los artículos publicados en revistas psicológicas de prestigio. Por ejemplo, Vacha-Haase y Ness (1999) revisaron 256 artículos publicados en la revista *Professional Psychology: Research and Practice* entre 1990 y 1997 y encontraron que en el 77% de los mismos se utilizaron pruebas de significación estadística para realizar el contraste de hipótesis, pero en menos del 20% de los artículos se usó correctamente el término significación estadística. Por su parte, Finch y cols., revisaron los artículos publicados en el *Journal of Applied Psychology* durante los

últimos sesenta años y encontraron que en el 38% de los estudios los resultados estadísticamente no significativos eran interpretados como que la hipótesis nula se consideraba verdadera. Finalmente, Hoekstra, Finch, Kiers y Johnson (2006) revisaron 266 artículos publicados en la revista *Psychonomic Bulletin and Review* entre 2002 y 2004 encontrando que más de la mitad de los investigadores interpretaban erróneamente la ausencia de significación estadística de los resultados. En concreto, los autores de estos estudios vinculaban la ausencia de significación estadística como prueba de evidencia de ausencia de efecto y, además, el 20% de los investigadores vinculó la presencia de significación estadística como prueba del efecto.

En el año 2010, Monterde-i-Bort y cols. realizaron una encuesta al personal docente e investigador de las Universidades Españolas del área de Psicología replicando el estudio que años antes habían llevado a cabo Gordon (2001) y Mittag y Thompson (2000) en los Estados Unidos de América con miembros de la *American Educational Research Association* (AERA). El objetivo de su estudio era, entre otras cuestiones estadísticas, analizar las interpretaciones que el personal docente e investigador (PDI) hacía de los valores *p* de probabilidad asociada a las pruebas de significación estadística. En concreto se analizaron: la "falacia de replicación", la "falacia del tamaño del efecto" y la "falacia de la significación clínica o práctica". Los ítems que componían cada una de las referidas falacias sobre el valor *p* fueron: (Los ítems que están entre comillas son considerados falsos)

- "*Falacia del tamaño del efecto*": percepción de las probabilidades estadísticas como una medida del tamaño del efecto

1.- «Valores de *p* pequeños ofrecen evidencia directa de que los efectos del tratamiento han sido grandes».

2.- «Si una docena de investigadores estudiaron el mismo fenómeno usando la misma hipótesis nula, y ninguno de sus estudios arrojó resultados estadísticamente significativos, significa que los efectos investigados no son destacables ni importantes».

3.- Los valores de *p* obtenidos en diferentes pruebas estadísticas no pueden ser directamente comparados, porque estos valores dependen de los tamaños de muestra utilizados en cada prueba.

- "*Falacia de la significación clínica o práctica*": percepción de los valores *p* como medidas directas de la importancia del resultado

1.- «Un resultado con una $p < .05$ indica que ese resultado es importante».

2.- Los estudios con resultados no-significativos pueden ser aún muy importantes.

3.- «Los resultados improbables son generalmente los más importantes y destacables».

- "*Falacia de la replicación*": percepción de los valores *p* como medidas de la probabilidad de replicación de los resultados.

1.- «Cuanto más pequeños son los valores de *p* más frecuentemente resultarán replicados dichos hallazgos en el futuro».

Los resultados señalaron que el PDI cometía la falacia del tamaño del efecto, esto es, el creer que el valor *p* indica la magnitud de un efecto, si bien, su presencia no fue relevante. Sin embargo, no ocurrió lo mismo con la falacia de la significación clínica o práctica, esto es, la creencia falsa de que el valor *p* está vinculado con la importancia del hallazgo obtenido. Bajo esta falsa creencia, los resultados no estadísticamente significativos fueron considerados como no importantes; lo que en algunos casos puede ser verdad, pero no necesariamente. Como Monterde-i-Bort y cols. (2006) señalan,

La improbabilidad de un dato no es señal inequívoca de su importancia, entre otras razones porque un resultado no significativo en una nueva investigación con mayor tamaño de muestra puede alcanzar significación estadística. Se sabe, además, que los resultados nulos pueden ser en algunos casos interesantes por sí mismos o expresión fehaciente de la falta de potencia estadística (p. 853).

Finalmente, respecto de la falacia de replicación, los hallazgos del estudio de Monterde-i-Bort y cols. (2010) pusieron de manifiesto la necesidad de clarificar el concepto de replicación y su relación con el valor *p* de probabilidad.

Recientemente, Kühberger y cols. (2015) en una muestra de estudiantes universitarios de Psicología de Austria analizaron la "falacia de la magnitud" y la "falacia de la significación clínica o práctica de los resultados". Los resultados mostraron que los estudiantes cometieron la "falacia de la magnitud", es decir, interpretaron los valores pequeños de *p* como que tienen un tamaño del efecto más alto que los valores de *p* más altos. Y, también, la "falacia de la significación clínica o

práctica", en el sentido de que los resultados no estadísticamente significativos fueron interpretados como evidencia de que no hay efecto o que el efecto es insignificante. Por tanto, los estudiantes confundieron la significación estadística con la significación de los resultados (tanto de su magnitud como de su importancia clínica).

En definitiva, los estudios analizados sugieren que los investigadores, profesores y estudiantes de Psicología no entienden muy bien lo que hacen o no hacen las pruebas de significación estadística. Las interpretaciones incorrectas del valor *p* siguen siendo abundantes y repetitivas, a pesar del constante debate sobre el citado valor. Sin embargo, las falacias sobre el valor *p* son problemas de interpretación y no del procedimiento NHST.

1.8. Otros problemas con el valor *p*: el *p-hacking*

Existe una creciente preocupación por que muchos de los resultados publicados sean falsos positivos o resultados estadísticamente significativos erróneos (Bakker, 2014; Ioannidis, 2005a; Wasserstein y Lazar, 2016), y por la escasa replicación en el ámbito de la Psicología (Henson, 2006; Koole y Lakens, 2012; Pashler y Wagenmakers, 2012; Stangor y Lemay, 2016).

Los falsos positivos pueden ser el resultado de una conducta, consciente o inconsciente, conocida como *p-hacking* (pirateo de los valores *p*).

Los científicos, las instituciones de investigación, los países, las organizaciones internacionales y las revistas científicas se evalúan cada vez más en función del número de artículos que publican y las citas que reciben que por los descubrimientos importantes que hacen en la ciencia (Fanelli, 2012). En este contexto, los artículos que reportan resultados estadísticamente significativos (resultados positivos) atraen más interés y se citan más a menudo, por lo que los editores de las revistas y los revisores externos tienden a favorecerlos. Mientras que los artículos que reportan resultados no estadísticamente significativos (resultados negativos) son menos propensos a ser publicados y citados (Fanelli, 2010). De hecho, hay evidencia de que las revistas con alto prestigio y factor de impacto publican de manera desproporcionada resultados estadísticamente significativos (e.g., Fanelli, 2012; Francis, 2012a; Ioannidis y Trikalinos, 2007). Por ejemplo, más del 90% de artículos publicados que utilizan el procedimiento de la NHST informaron de resultados estadísticamente significativos (Fanelli, 2010), lo que resulta extraño, al menos en el ámbito de la Psicología, dado que

los tamaños del efecto psicológicos no suelen ser grandes y además la mayoría de las publicaciones se suelen llevar a cabo con muestras pequeñas y con poca potencia estadística (Bakker, 2014).

La presión por publicar puede conducir a los científicos a realizar prácticas de investigación cuestionables (*Questionable Research Practices, QRPs*) dirigidas a alcanzar resultados estadísticamente significativos como el *p-hacking* (pirateo de los valores *p o falsos positivos*) o el *harking* (formular las hipótesis después de que los resultados sean conocidos, *Hypothesizing After the Results are Known*); sobre todo, si la carrera de los investigadores se evalúa mediante el recuento de los artículos publicados y el factor de impacto de las revistas donde han publicado dichos artículos (Fanelli, 2010; Hales, 2016; Stroebe, 2016).

Dentro de las prácticas más comunes de pirateo de los valores *p*, consciente o inconsciente, destacan: (1) recoger los datos, analizarlos y si se observa un resultado no estadísticamente significativo pero en la dirección esperada, se recogen más datos hasta que el resultado alcance la significación estadística; (2) registrar muchas variables e informar sólo de aquellas que han alcanzado resultados estadísticamente significativos; (3) decidir si se deben incluir o eliminar los valores extremos o atípicos (*outliers*) o aplicar transformaciones o pruebas no paramétricas, y (4) detener la exploración de datos si el análisis arroja un resultado con un valor de *p* estadísticamente significativo (resultado positivo) (Gadbury y Allison, 2014; Head y cols., 2015; Motulsky, 2015). Por ejemplo, John, Loewenstein y Prelec (2012) proporcionaron evidencia sobre estas prácticas cuestionables en una muestra de psicólogos. Estos autores encontraron que: el 15.6% de los mismos informó detener la recogida de los datos antes de lo previsto porque los resultados ya habían alcanzado la significación estadística, el 27.7% no reportó los resultados sobre todas las condiciones del estudio y, finalmente, el 38.2% informó decidir si excluía algunos datos después de ver el impacto que la exclusión de tales datos tenía sobre los resultados.

En consecuencia, la cuantificación del *p-hacking* es importante porque la publicación de falsos positivos obstaculiza el progreso científico dificultando la acumulación de un conocimiento científico válido sobre un determinado problema de investigación (Stroebe, 2016). En este sentido, los falsos positivos pueden conducir a explorar campos con efectos cero, promover inversiones económicas en programas de investigación infructuosos, y dificultar la realización de estudios de replicación, etc.

(Head y cols., 2015). Por ejemplo, Stangor y Lemay (2016) señalan que solamente el 18% de los estudios con un valor de *p* mayor de .04 fueron replicados mientras que el 63% de los estudios que tenían un valor de *p* menor a .001 sí fueron replicados. Sin embargo, la replicación, entendida como la confirmación de los resultados y conclusiones obtenidos en una investigación en otro estudio realizado de forma independiente, se considera la piedra angular de la ciencia acumulativa (Asendorpf y cols., 2013; Johnson, 1999).

La curva de los valores *p* es una herramienta útil para examinar la evidencia de *p-hacking* y evaluar la fiabilidad de las investigaciones publicadas (Head y col., 2015; Sakaluk, 2016). La curva de los valores *p* (p-curve) desarrollada por Simonsohn, Nelson y Simmons (2014a, 2014b) examina la distribución de los valores *p* reportados en un conjunto de estudios y detecta la manipulación inadecuada de los análisis estadísticos para producir valores *p* estadísticamente significativos cuando los análisis iniciales produjeron resultados que eran casi, pero no del todo, estadísticamente significativos (Gadbury y Allison, 2014).

1.9. Conclusión

El debate sobre la adecuación (utilidad y validez) de la prueba NHST para alcanzar un conocimiento científico válido sigue vigente a día de hoy, tanto por los problemas en la interpretación de los valores *p* como por su uso inadecuado como procedimiento hibridado y por las prácticas cuestionables de investigación como el *p-hacking*.

Los estudios empíricos sugieren que investigadores, profesores y estudiantes de Psicología no entienden muy bien lo que hacen o no hacen las pruebas de significación estadística. La definición del valor *p* es clara y precisa, sin embargo, las interpretaciones incorrectas de los valores *p* siguen siendo abundantes y repetitivas. No obstante, las falacias sobre el valor *p* son problemas de interpretación del investigador y no del procedimiento NHST. También es cierto que existen otro tipo de críticas centradas en el mecanismo metodológico del proceso de significación de la hipótesis nula, como la propia falsedad de la hipótesis nula, de tal manera que es cuestión de aumentar el tamaño de la muestra hasta lograr la significación estadística, lo que da lugar al *p-hacking*.

Respecto del *p-hacking*, los estudios aportan evidencia sobre la existencia de falsos positivos en la literatura científica. No obstante, como sucede con las falacias sobre el valor *p*, el *p-hacking* no es un problema inherente al procedimiento de la NHST sino de la ética del investigador y relacionado con el sesgo de publicación, quien llevado por la presión de publicar más artículos científicos en revistas prestigiosas y con factor de impacto, puede piratear, consciente o inconscientemente por falta de conocimiento, los valores de *p* para obtener resultados positivos, esto es, resultados estadísticamente significativos, que son más fácilmente publicables, lo que dificulta la acumulación de un conocimiento científico válido sobre un determinado problema de investigación. En este punto, la replicación es necesaria para validar los resultados positivos e invalidar los falsos positivos, puesto que los estudios de replicación permiten separar las conclusiones que son dignas de confianza de los resultados que son poco fiables. Así pues, una disciplina que invierte en estudios de replicación se inmuniza contra las prácticas erróneas.

Finalmente señalar que cuando el procedimiento de la NHST se utiliza con buen criterio puede ser una ayuda eficaz para interpretar los datos de las investigaciones. Por otro lado, el procedimiento de la NHST es uno más dentro de una amplia gama de técnicas de análisis estadístico de los datos que pueden y deben ser combinadas para proporcionar mayor comprensión del resultado empírico obtenido y facilitar la integración de resultados en futuros estudios de meta-análisis.

En definitiva, la calidad del conocimiento científico generado en una disciplina requiere que los investigadores planifiquen adecuadamente su investigación, la ejecuten eficientemente, analicen los datos correctamente, interpreten bien los resultados y presenten de forma clara las conclusiones. En consecuencia, es crucial tener un adecuado conocimiento sobre lo que el procedimiento NHST ofrece a los investigadores y profesionales de la salud y lo que no puede ofrecer. Por lo tanto, interpretar correctamente los valores de *p* es una competencia básica e indispensable para los investigadores, profesores, estudiantes y profesionales de la salud para poder implementar correctamente en la práctica profesional los hallazgos de las investigaciones.

2. MÁS ALLÁ DE LA PRUEBA NHST (I): TAMAÑO DEL EFECTO Y SU INTERVALO DE CONFIANZA Y REPLICACIÓN

Como ya se ha comentado en el capítulo anterior las críticas hacia las pruebas de significación estadística son casi tan antiguas como los mismos métodos (e.g., Berkson, 1938; Boring, 1919). Tales críticas sobre el uso y abuso de las pruebas de significación estadística, los problemas de interpretación de los valores *p* de probabilidad, la escasa potencia estadística de los estudios, y el *p-hacking* (entre otras cuestiones) han estimulado un debate profundo sobre la aplicación de las pruebas de significación estadística en Psicología (y otras ciencias) que perdura en la actualidad.

Dentro de este debate, se han sugerido métodos de análisis de datos alternativos a las pruebas de la NHST, o al menos complementarios. Así, se han defendido con firmeza el uso de (entre otros métodos como los basados en el remuestreo y las pruebas de permutación, la perspectiva bayesiana, el modelado estadístico, etc.):

(1) ***índices del tamaño del efecto*** (e.g., APA 1994, 2001, 2010a; Cohen, 1988, 1990, 1994; Cumming, 2012, 2014; Cumming, Williams, y Fidler, 2004; Ellis, 2010; Fidler, 2005; Frías-Navarro, 2011b; Frías-Navarro y cols., 2014; Kirk, 1996, 2001; Kline, 2013; Palmer y Sesé, 2013; Thompson, 1996, 1998, 1999, 2002a, 2002b, 2006, 2007; Valera-Espín y Sánchez-Meca, 1997; Wilkinson y TFSI, 1999).

(2) ***intervalos de confianza para los índices de tamaños del efecto*** (e.g**.,** APA, 2001, 2010; Brandstätter, 1999; Cohen, 1994; Cumming, 2012, 2014; Cumming y cols., 2004; Fidler, 2005; Frías-Navarro, 2011b; Gardner y Altman 2000; Palmer y Sesé, 2013; Thompson, 1999; 2002a, 2002b, 2006, 2007; Wilkinson y TFSI, 1999).

(3) ***estudios de replicación*** (e.g., Asendorpt y cols., 2013; Carver 1978; Cumming, 2008; Cumming y cols., 2004; Hubbard, 2004; Hubbard y Lindsay, 2008; Kline, 2013; Nickerson, 2000; Valera-Espín y Sánchez-Meca, 1997; Wilkinson y TFSI, 1999).

(4) ***estudios de meta-análisis*** (e.g., APA, 2010a; Borenstein y cols., 2009; Bustamante y Delgado, 1994; Cooper, 1989; Cumming, 2012; Gill, 1999; Hedges y Olkin, 1985; Kline, 2013; Rosenthal, 1984; Schmidt, 1996),

(5) y, sobre todo, un pensamiento crítico honesto (Perezgonzalez, 2015).

En este capítulo, se expondrán los estadísticos del tamaño del efecto y sus intervalos de confianza así como de los estudios de replicación, dejando para el siguiente capítulo el análisis de los estudios de meta-análisis.

Pero antes se hará un breve recorrido histórico sobre la respuesta que la Asociación Americana de Psicología (APA) ha dado a los problemas planteados por las pruebas de significación estadística en las distintas ediciones de sus Manuales de Publicación y en el informe del grupo de trabajo sobre inferencia estadística (Wilkinson y TFSI, 1999).

2.1. Recomendaciones de la Asociación Americana de Psicología (APA): Manual de publicación

En 1929 la Asociación Americana de Psicología (APA) publica en *Psychological Bulletin* una serie de instrucciones (denominadas "*Instrucciones relacionadas con la preparación de trabajos*") orientadas a apoyar a los investigadores en la preparación de los informes científicos. Esta publicación surge de la necesidad y el interés de editores y administradores de las revistas científicas por reglamentar o estandarizar un estilo en los informes de investigación (Santana-Cárdenas, 2006). Sin embargo, como Finch y cols. (2002) señalan, este documento no hizo referencia o mención a cómo se debían informar los resultados de las investigaciones.

Después de mucho revisar y modificar, en el año de 1952, las "Instrucciones relacionadas con la preparación de trabajos" se convirtieron en lo que actualmente se conoce como Manual de Publicación de la APA (primera edición).

En 1974 apareció una segunda edición de dicho Manual de publicación, modificada y ampliada con 136 páginas; en 1986 se publicó la tercera edición, de 208 páginas; en 1994 apareció la cuarta edición, con 368 páginas; en 2001 se presentó la quinta edición, y finalmente, en el año 2010 se publicó la sexta y última edición con 260 páginas.

A pesar de que las críticas aparecieron desde los inicios de la implementación de la prueba de la NHST no fue hasta la cuarta edición del Manual de publicación cuando la APA tomó medidas para abordar las cuestiones planteadas. Así, en 1994 en la cuarta edición del Manual de Publicación la APA "animó" el reporte de los estadísticos del tamaño del efecto (p. 18). A continuación, en 1996, la APA formó el Grupo de Trabajo

sobre inferencia estadística para investigar la conveniencia de prohibir las pruebas de significación estadística dado el intenso debate que existía sobre su uso y abuso (véase Thompson, 2007). Wilkinson y TFSI (1999) hicieron varias recomendaciones metodológicas, sin llegar a prohibir el uso del procedimiento de la NHST. En 2001, la quinta edición del Manual de Publicación de la APA incluyó la siguiente declaración: "*Casi siempre es necesario incluir algún índice del tamaño del efecto o la fuerza de la relación en la sección de resultados*" (pp. 25-26.). Y, finalmente, en la sexta edición del Manual de la APA se deja constancia de la necesidad de acompañar los valores *p* de probabilidad con un estadístico del tamaño del efecto y su intervalo de confianza.

A continuación se presentan brevemente cada uno de estos hitos históricos.

2.2.1. Primera Edición del Manual de Publicación de la APA (1952)

En la Primera edición del Manual de Publicación de la APA (1952) aparecieron las primeras pautas a la hora de informar de los resultados en los siguientes términos: "*La sección de resultados debe dar datos suficientes para justificar las conclusiones. Se debe prestar especial atención a las pruebas de significación estadística y a la lógica de la inferencia y generalización a partir de las observaciones empíricas* " (p. 397). Además, respecto a la presentación de las tablas, se afirmó que: "*No son necesarias amplias tablas de resultados no significativos. Por ejemplo, si sólo 2 de 20 correlaciones son significativamente diferentes de cero, las 2 correlaciones significativas se pueden mencionar en el texto, y el resto reportadas con unas pocas palabras*" (APA, 1952, p. 414).

Esta Primera edición del Manual de la APA no hizo referencia a las críticas hacia la prueba NHST y los métodos de análisis alternativos propuestos.

2.1.2. Segunda Edición del Manual de Publicación de la APA (1974)

La Segunda edición del Manual de Publicación de la APA (1974) hizo explícitos algunos detalles para reportar los resultados de las pruebas de inferencia estadística (NHST). En concreto el Manual señala que: "*Al informar sobre las pruebas de significación […] se debe incluir información relativa a la magnitud obtenida o el valor de la prueba, los grados de libertad, el nivel de probabilidad, y la dirección del efecto*" (p. 18).

Además, proporcionó ejemplos de cómo se debía informar sobre los resultados de la prueba NHST. Por ejemplo, "*Como se predijo, las niñas de primer grado reportaron significativamente un mayor gusto por la escuela que los chicos de primer grado, t (22) = 2.62, p < .01*" (p. 39).

Como señalan Finch y cols. (2002), los ejemplos del Manual indican que el término 'nivel de probabilidad' se refiere al valor *p* y el término 'magnitud' se refiere al valor de la prueba estadística (por ejemplo, $t = 2.62$), en lugar de a una medida del tamaño del efecto. Además, se observa que no hay ninguna referencia a las medias de tendencia central y desviaciones típicas de los grupos (chicos y chicas), o sobre la diferencia entre los grupos. Tampoco recoge información acerca de la cantidad de niñas y niños a los que les gustaba la escuela o lo mucho que diferían en su afición por la escuela. Por otro lado, en el apartado de resultados a los autores se les dijo simplemente "resumir los datos recogidos" (APA, 1974, p. 18), pero no había ningún requerimiento explícito de incluir medidas de tendencia central o de dispersión. Asimismo, el apartado de resultados se centró más en la elección entre tablas y gráficos que en el tipo de estadísticos descriptivos que se debía informar.

Sin embargo, de nuevo, el Manual de la APA no hizo mención al debate sobre el uso y abuso de la prueba NHST y los métodos de análisis alternativos. Por lo que los psicólogos que siguieron el Manual como una guía para la práctica estadística continuaron sus prácticas con graves deficiencias (Finch y cols., 2002).

2.1.3. Tercera Edición del Manual de Publicación de la APA (1983)

La Tercera edición del Manual de Publicación de la APA (1983) mantuvo su énfasis sobre cómo escribir con claridad, el estilo editorial y los detalles para la presentación de manuscritos; prestando poca atención a la presentación e interpretación de los datos y los resultados de la prueba NHST.

En este sentido, el Manual de la APA (1983) apenas hizo cambios respecto del Manual de 1974 en cuanto a la presentación de los resultados estadísticos. Así, los ejemplos de cómo informar sobre los resultados de las pruebas de inferencia estadística fueron extraídos del Manual de 1974 con la adición de los valores de las medias de la muestra (APA, 1983, p. 81). Además, se hizo explícita la necesidad de reportar los estadísticos descriptivos (medias y desviaciones típicas) de los grupos.

Sin embargo, el Manual no hizo ninguna referencia a los tamaños del efecto, a los intervalos de confianza, a la potencia estadística, a los trabajos de meta-análisis, o a las interpretaciones erróneas de la de NHST. A pesar de las continuas críticas sobre el uso y abuso de la prueba de la NHST en Psicología, y el debate sobre la necesidad de aportar medidas del tamaño del efecto expuesto en un amplio conjunto de revistas científicas de Psicología entre 1974 y 1983.

Finalmente, respecto de cómo informar sobre los resultados no estadísticamente significativos, el Manual indicaba que estos resultados debían ser aceptados como tal sin ser explicados y ser interpretados como debidos al azar (Finch y cols., 2002).

Una vez más el Manual no se hizo eco de los mensajes de reforma en las prácticas estadísticas de los investigadores, las reiteradas críticas a la prueba de la NHST y los mensajes sobre métodos alternativos de análisis de datos (Finch y cols., 2002).

2.1.4. Cuarta Edición del Manual de Publicación de la APA (1994)

En la Cuarta edición del Manual de Publicación de la APA (1994) se produjeron cambios sustanciales respecto de las ediciones anteriores. Esta edición del Manual fue la *primera en mencionar la necesidad de proporcionar datos sobre la potencia estadística de los análisis y los tamaños del efecto,* pero su introducción fue breve y se dieron pocos consejos prácticos (Fidler, 2010).

Respecto del tamaño del efecto el Manual señala:

> Ninguno de los dos tipos de valores de probabilidad (entiéndase alfa y el valor *p*) refleja la importancia (magnitud) de un efecto o la fuerza de una relación porque ambos valores de probabilidad dependen del tamaño de la muestra. Se puede estimar la magnitud del efecto [...] con una serie de medidas que no dependen del tamaño de la muestra. [...] Se le anima a proporcionar información sobre los tamaños del efecto aunque, en la mayoría de los casos, estas medidas sean fáciles de obtener cuando se proporcionen los estadísticos de prueba (por ejemplo, *t* y *F*) y los tamaños de la muestra (p. 18).

Sin embargo, como el estudio de Vacha-Haase, Nilsson, Reetz, Lance y Thompson (2000) muestra, este estímulo tuvo poco o ningún efecto sobre las prácticas estadísticas de los psicólogos en la elaboración de sus informes de investigación.

Además, el meta-análisis sólo se mencionó en la sección sobre la preparación de la lista de referencias. Por otra parte, no se hizo mención alguna a la estimación de parámetros mediante intervalos de confianza. Y, finalmente, el manual no hizo ninguna referencia a los problemas de interpretación por parte de los investigadores de la prueba NHST (Finch y cols., 2002). Por ejemplo, no se advirtió del mal uso del término "significativo" (Boring, 1919; Carver, 1978).

2.1.5. Grupo de Trabajo de Inferencia Estadística de la APA (1999)

Después de 10 años de críticas hacia la prueba NHST, de continuos debates sobre los usos y abusos de esta técnica de análisis y de las repetidas peticiones de un cambio en las prácticas estadísticas por parte de diversos autores (e.g., Carver, 1978; Cohen, 1994, Schmidt, 1996; Thompson, 1996), el 28 de febrero de 1996, la Junta de Asuntos Científicos (*Board of Scientific Affairs*, BSA) de la APA creó el denominado Grupo de Trabajo sobre Inferencia Estadística (*Task Force on Statistical Inference*, TFSI).

El propósito original del TFSI era investigar las críticas a la prueba de la NHST, incluyendo una propuesta para prohibir su uso en las revistas de la APA (Fidler 2002, 2010; Finch y cols., 2002; Thompson, 2006). Sin embargo, el TFSI estableció una agenda más amplia para sí mismo: "*considerar la modificación de las prácticas actuales en el tratamiento cuantitativo de los datos en la ciencia de la Psicología*" (Fidler, 2002, p. 751).

En palabras del propio TFSI, su objetivo fue "*clarificar algunas de las cuestiones polémicas que rodean las prácticas estadísticas, incluyendo las pruebas de significación estadística de la hipótesis nula (NHST) y sus alternativas; modelos alternativos subyacentes y transformación de datos, y nuevos métodos posibles gracias a los potentes ordenadores*" (Wilkinson y TFSI, 1999, p.594).

El grupo de trabajo estaba formado inicialmente por Robert Rosenthal, Robert Abelson y Jacob Cohen (directores del grupo), los cuales coincidieron en la conveniencia de contar con diferentes tipos de especialistas que incluyera a estadísticos, profesores de estadística, editores de revistas, autores de libros de estadísticas, expertos en informática, y personas sabias (*wisee elders*). En consecuencia, nueve personas más fueron invitadas a formar parte de este grupo de trabajo. Éstos fueron: Leona Aiken, Mark Appelbaum, Gwyneth Boodoo, David A. Kenny, Helena Kraemer, Donald Rubin,

Bruce Thompson, Howard Wainer, y Leland Wilkinson. Además, como supervisores fueron nombrados Lee Cronbach, Paul Meehl, Frederick Mosteller y John Tukey.

El TFSI se reunió dos veces en dos años. Después de la primera reunión, el grupo de trabajo hizo circular un informe preliminar que indicaba su intención de examinar las cuestiones más allá de la prueba de significación estadística de la hipótesis nula. En su segunda reunión, el TFSI recomendó revisar las secciones estadísticas del Manual de Publicación de la APA de 1994 y propuso reformar las prácticas de análisis de datos y presentación de informes.

Entre sus recomendaciones destacan tres directrices que marcarán las de la quinta edición del Manual de Publicación de la APA (2001): (1) acompañar la presentación, análisis e interpretación de los datos con otros estadísticos como la estimación del tamaño del efecto; (2) informar de los intervalos de confianza de los tamaños del efecto, y (3) utilizar procedimientos gráficos que mejoren la interpretación y comunicación de los resultados.

A pesar de cierta presión, el Grupo de Trabajo no recomendó prohibir el uso de la prueba NHST en las revistas de Psicología:

> Algunos habían esperado que este grupo de trabajo recomendaría la prohibición total de la utilización de las pruebas de significación en revistas de psicología. Aunque esto podría eliminar algunos abusos, el comité pensó que había suficientes contraejemplos (por ejemplo, Abelson, 1997) para justificar su tolerancia. Por otra parte, el comité cree que los problemas planteados en su cargo fueron más allá de la simple cuestión de la prohibición de las pruebas de significación (Wilkinson y TFSI, pp. 602-603).

Pero, el grupo de trabajo TFSI sí aceptó en gran medida las críticas de los llamados reformadores sobre la prueba de la NHST y recomendó reducir la excesiva confianza y dependencia del procedimiento de la NHST a favor, como se ha dicho, de la estimación de los tamaños del efecto, los intervalos de confianza y los procedimientos gráficos para la presentación y comunicación de los resultados de las investigaciones (Fidler, 2005, 2010; Thompson, 2006). Así, el TFSI advertía:

> Es difícil imaginar una situación en la que una decisión dicotómica aceptar-rechazar sea mejor que informar del valor de *p*, o, mejor aún, un intervalo de confianza.

> Nunca utilice la desafortunada expresión "aceptar la hipótesis nula.” Siempre proporcione alguna estimación del tamaño del efecto al informar de un valor de *p* (Wilkinson y TFSI, p. 599).

En concreto el informe del grupo TFSI estableció las siguientes recomendaciones:

Respecto de los tamaños del efecto:

> Los tamaños del efecto siempre tienen que estar presentes en los resultados primarios. Si las unidades de medida son significativas a un nivel práctico (por ejemplo, número de cigarrillos fumados por día), entonces por lo general se prefiere una medida no estandarizada (coeficiente de regresión o diferencia de medias) a una medida estandarizada (*r* o *d*). Esto ayuda a añadir breves comentarios que sitúan a estos tamaños del efecto en un contexto práctico y teórico.
>
> [...] Debemos acentuar otra vez que reportar e interpretar los tamaños del efecto en el contexto de previos tamaños del efecto informados es esencial para una buena investigación. Esto permite a los lectores evaluar la estabilidad de los resultados a través de muestras, diseños y análisis. Reportar tamaños del efecto también informa el análisis de la potencia y los meta-análisis que son necesarios en futuras investigaciones (p. 599).

Respecto de los intervalos de confianza:

> Se deben dar estimaciones de intervalos para cualquier tamaño del efecto que implican los resultados principales. Proporcionar intervalos para las correlaciones y otros coeficientes de asociación o variación siempre que sea posible.
>
> Los intervalos de confianza están generalmente disponibles en el software estadístico; de lo contrario, los intervalos de confianza para los estadísticos básicos se pueden calcular a partir de un *output* típico. Comparar los intervalos de confianza del estudio actual con los intervalos de estudios previos, relacionados, ayuda a centrar la atención en la estabilidad entre los estudios (Schmidt, 1996). Recoger los intervalos entre los estudios también ayuda en la construcción de las regiones plausibles para los parámetros de la población. Esta práctica debería ayudar a evitar el error común de asumir que un parámetro está contenido en un intervalo de confianza (p. 599).

Finalmente, el informe del TFSI explícitamente reconoce que el rol del Manual de Publicación de la APA va más allá de marcar directrices sobre el estilo de presentación de los informes de investigación y que debe establecer los principios para una buena práctica estadística.

En definitiva, el informe debe ser reconocido como un gran paso hacia adelante en el proceso de reforma de las prácticas estadísticas en los psicólogos, tanto por la amplia agenda oficial que se establece para su examen como por el impulso que da hacia medidas políticas concretas para lograr un cambio generalizado en las prácticas metodológicas (Finch y cols., 2002).

Por ello, aunque el TFSI no prohibió el uso de la prueba NHST, su informe es citado muy a menudo en los artículos que versan sobre la mejora de la práctica estadística en Psicología.

2.1.6. Quinta Edición del Manual de Publicación de la APA (2001)

En su quinta edición, el Manual de Publicación de la APA (2001) trató de dar respuesta a las peticiones de cambio en las prácticas de análisis de datos y en elaboración de los informes de investigación y adoptó algunas de las recomendaciones que propuso el informe de Wilkinson y TFSI (1999) y que los reformadores habían reclamado desde hacía muchos años. Así, se pueden destacar las siguientes recomendaciones en esta edición del Manual:

Respecto del tamaño del efecto:

> Para que el lector comprenda mejor la importancia de sus hallazgos, casi siempre es necesario incluir algún índice de la magnitud del efecto o la fuerza de la relación en la sección de Resultados. Usted puede calcular estos índices mediante un número de estimadores usuales del tamaño del efecto, que incluyen (pero no se limitan a) r^2, η^2, ω^2, R^2, ϕ^2, *V* de Cramer, *W* de Kendall, *d* y *K* de Cohen, λ y γ de Goodmnan y Kruskal, las medidas de significación clínica propuestas por Jacobson y Truax (1991) y por Kendall (1999), y la Θ multivariada de Roy, así como la *V* de Pillai-Bartlett.
>
> Como regla general, los indicadores del efecto con múltiples grados de libertad tienden a ser menos útiles que los indicadores del efecto que descomponen pruebas con múltiples grados de libertad en efectos significativos de un grado de libertad –en particular cuando éstos son los resultados que se informan en la discusión. Sin embargo, el principio general que debe seguirse consiste en proporcionar al lector no sólo información acerca de la significación estadística, sino también suficiente información para evaluar la magnitud del efecto o de la relación observados (pp. 25-26).

Además el Manual de la APA (2001) etiquetó el hecho de "*no informar del tamaño del efecto*" en los informes de investigación como un defecto "*en el diseño y en el informe de investigación*" (p. 5).

Sin embargo, esta recomendación de incluir algún índice del tamaño del efecto en los informes de investigación no fue acompañada de algún ejemplo de cómo reportar tales medidas de la magnitud del efecto. No obstante, las recomendaciones que se hicieron en la misma sección sobre el uso de NHST sí que se acompañaron de varios ejemplos de cómo reportar los valores de *p*.

Finalmente, el informe del TFSI incluyó más información respecto del tamaño del efecto que no apareció recogida en la quinta edición del Manual (Fidler, 2002; Finch y cols., 2002). Por ejemplo, el informe TFSI explicaba la importancia de reportar los tamaños del efecto para futuros estudios sobre análisis de potencia estadística y de meta-análisis. También enfatizó la importancia de interpretar el tamaño del efecto dentro de un contexto práctico y teórico (p. 599), aspectos no recogidos en la quinta edición del Manual de la APA (2001).

Respecto de los intervalos de confianza:

La Quinta edición del Manual fue la primera en recomendar el uso de intervalos de confianza. En concreto señala que:

> Informar sobre los intervalos de confianza (para la estimación de parámetros, para las funciones de parámetros tales como diferencias de medias, y para los tamaños del efecto) puede ser una manera bastante eficaz de reportar los resultados. Dado que los intervalos de confianza combinan información sobre la localización y precisión y pueden ser, a menudo, directamente utilizados para inferir el nivel de significación estadística, son, en general, la mejor estrategia para informar. El uso de los intervalos de confianza es por tanto fuertemente recomendado (p. 22).

Sin embargo, una vez más, no ofreció ejemplos de cómo se debían calcular e informar los ICs (Fidler, 2002). Y de nuevo, los comentarios del informe del grupo TFSI sobre las estimaciones del ICs no fueron incluidas en el Manual. Por ejemplo, como se ha comentado, el TFSI hizo hincapié en la importancia de comparar los ICs obtenidos en distintos estudios relacionados, puesto que ayuda a centrar la atención en la estabilidad entre los estudios, y no en realizar el contraste de hipótesis mediante los ICs, es decir, observar si el valor de la hipótesis nula (efecto 0) se encuentra dentro del

intervalo. Finalmente, el informe TFSI también advirtió del "error común de asumir que un parámetro está contenido en un intervalo de confianza" (Wilkinson y TFSI, 1999, p. 599). No obstante, el Manual no alertó a los investigadores de este error.

A pesar de todo, desde la publicación del Manual de la APA (2001) muchas revistas de Psicología empezaron a requerir o recomendar fuertemente que los investigadores informaran no sólo del resultado de la prueba de significación estadística, sino también de un índice del tamaño del efecto y los intervalos de confianza, como el *Journal of Consulting and Clinical Psychology* (La Greca, 2005). Además, 22 revistas científicas requirieron la presentación de estadísticos del tamaño del efecto (Thompson, 2006): *Contemporary Educational Psychology; Early Childhood Research Quarterly; Educational and Psychological Measurement; Educational Technology Research and Development; Exceptional Children (EC); Health Psychology; Journal of Agricultural Education; Journal of Community Psychology; Journal of Counseling & Development (JCD); Journal of Early Intervention; Journal of Educational Psychology; Journal of Educational and Psychological Consultation; Journal of Experimental Education; Journal of Experimental Psychology: Applied; Journal of Learning Disabilities; Journal of Personality Assessment; Language Learning; Measurement and Evaluation in Counseling and Development; The Professional Educator; Reading and Writing; y Research in the Schools.*

Finalmente señalar que, a pesar de los avances en la mejora de las prácticas estadísticas que supuso la quinta edición del Manual, algunos reformadores sintieron que estos avances fueron insuficientes y otros se sintieron decepcionados (Fidler, 2002). Así, como Finch y cols. (2002) señalan "*La quinta edición del Manual (APA, 2001) es en gran medida, desde un punto de vista de reforma, una oportunidad vital perdida*" (p. 17).

2.1.7. Sexta Edición del Manual de Publicación APA (2010)

La Sexta Edición del Manual de Publicación (APA, 2010a) incluye los cambios de mayor alcance en las guías sobre las prácticas estadísticas desde el asesoramiento estadístico que se introdujo en la primera edición por el Consejo de Editores, en 1952 (Cumming y cols., 2012). En este sentido, la APA reconoce la importancia del procedimiento de la NHST como punto de partida en los análisis de los datos y anima a

complementar la información proveniente de las pruebas de inferencia estadística con la información sobre la magnitud del efecto y su precisión.

En palabras del propio Manual:

> Históricamente, los investigadores en el campo de la Psicología han tomado la prueba de la significación estadística de la hipótesis nula (NHST) como un punto de partida para muchas (aunque no para todas) aproximaciones analíticas. La APA enfatiza que la NHST no es más que un punto de partida y que los elementos adicionales para presentar información, como los tamaños del efecto, los intervalos de confianza y una extensa descripción, son necesarios para transmitir el significado más completo de los resultados. El grado en que cualquier publicación periódica apoye (o rechace) la NHST es una decisión de cada editor. Sin embargo, uno de los requerimientos de todas las publicaciones periódicas de la APA es una presentación completa de todas las hipótesis examinadas y las estimaciones de tamaños de efecto e intervalos de confianza adecuados (p.33).

Además, el actual Manual de la APA menciona los estadísticos en numerosos lugares, sobre todo en el capítulo 2 (donde se describen las secciones o partes que debe contener un artículo científico), en el capítulo 4 (donde se especifican los formatos para reportar los estadísticos), y en el capítulo 5 (dedicado a las figuras y a las tablas).

Finalmente el Manual presenta dos nuevos apéndices, donde se incluyen las normas de presentación de los informes de investigación con datos originales (artículos científicos) (*Journal Article Reporting Standards*, JARS) y los estudios de Revisiones Sistemáticas y de Meta-análisis (*Meta-Analysis Reporting Standards*, MARS). El JARS y el MARS son listados de comprobación (*Checklist*) que detallan lo que debe incluirse en los respectivos informes de investigación.

En cuanto a las recomendaciones que incluye la sexta edición del Manual de la APA (2010a), destacan:

Respecto de los tamaños del efecto:

> Para las pruebas de inferencia estadística (por ejemplo, *t*, *F*, o pruebas χ^2), incluya la magnitud obtenida o el valor de la prueba estadística, los grados de libertad, la probabilidad de obtener un valor tan extremo o más extremo que el que se ha obtenido (el valor de *p* exacto), así como el tamaño y la dirección del efecto (p. 34).

Y sigue diciendo la APA,

> Para que el lector aprecie la magnitud o importancia de los hallazgos de un estudio, casi siempre es necesario incluir alguna medida del tamaño del efecto en la sección Resultados. Siempre que sea posible, proporcione un intervalo de confianza para cada tamaño del efecto reportado para indicar la precisión de la estimación del tamaño del efecto. Los tamaños del efecto se pueden expresar en las unidades originales (por ejemplo, la media del número de preguntas contestadas correctamente; kg/mes para una pendiente de regresión) y a menudo son más fáciles de entender cuando se reportan en las unidades originales. A menudo puede ser valioso reportar un tamaño del efecto no sólo en las unidades originales, sino también en alguna unidad estandarizada o unidades libres de medida (por ejemplo, como el valor *d* de Cohen) o un coeficiente de regresión estandarizado. Los indicadores del tamaño del efecto con un grado de libertad múltiple no suelen ser tan útiles como los indicadores del tamaño del efecto que descomponen las pruebas con grados de libertad múltiples en efectos significativos con un grado de libertad –particularmente cuando estos últimos son los resultados que se reservan para los comentarios. Sin embargo, el principio general que debe seguirse es proporcionar al lector información suficiente para evaluar la magnitud del efecto observado (p. 34).

Además, el Manual hace hincapié en la importancia de informar sobre los tamaños del efecto para los efectos estadísticamente no significativos: "*Mencionar todos los resultados relevantes [...] asegúrese de incluir los tamaños del efecto pequeños (o resultados estadísticamente no significativos) [...] No oculte los resultados incómodos por omisión*." (p. 32).

Como señala Fidler (2010), esto es importante debido a la práctica común de informar sobre los resultados estadísticamente no significativos simplemente con la expresión "*ns*". Este modo de informar sobre los resultados no estadísticamente significativos dificulta la interpretación de los resultados y la realización de estudios de meta-análisis.

Finalmente, el Manual anima no solo a informar sobre algún índice del tamaño del efecto sino también a interpretarlo. Pues como se ha visto con anterioridad, el Manual establece que "*Siempre que sea posible, la discusión y la interpretación de los resultados debe basarse sobre las estimaciones puntuales y los intervalos*" (p. 34). Por tanto, las conclusiones del estudio no deben basarse simplemente en la decisión dicotómica de mantener o rechazar la H_0 en base al valor *p* de las pruebas de inferencia estadística.

Respecto de los intervalos de confianza:

> Cuando se proporcionan estimaciones puntuales (por ejemplo, medias de las muestras o coeficientes de regresión), incluir siempre una medida complementaria de la variabilidad (precisión), con una indicación de la medida específica utilizada (por ejemplo, el error estándar).
>
> La inclusión de los intervalos de confianza (para las estimaciones de los parámetros, para las funciones de parámetros tales como las diferencias en medias, y para los tamaños del efecto) puede ser una manera muy eficaz para reportar los resultados. Porque los intervalos de confianza combinan información sobre la ubicación y precisión y a menudo pueden utilizarse directamente para inferir los niveles de significación, son, en general, la mejor estrategia de presentar la información. El uso de intervalos de confianza es, por lo tanto, fuertemente recomendado. Como regla general, es mejor utilizar un solo nivel de confianza, especificado a priori (por ejemplo, un intervalo de confianza de 95% o 99%), a lo largo del manuscrito.
>
> Siempre que sea posible, la discusión y la interpretación de los resultados debe basarse sobre las estimaciones puntuales y los intervalos (p. 34).

Respecto de los estudios de meta-análisis:

Como se ha comentado, la Sexta edición ofrece un apéndice con un listado de comprobación de los elementos que deben conformar un estudio de meta-análisis (MARS). El MARS proporciona una lista detallada de lo que debe ser reportado en un estudio de meta-análisis.

Además, tres páginas del manual están dedicadas a un artículo científico: muestra de los estudios (Capítulo 2, pp. 57-59), y dos sub-secciones dan pautas de meta-análisis (capítulo 2, pp. 36-37, y capítulo 6, p. 183).

Finalmente, el manual ofrece una fuerte motivación para el pensamiento meta-analítico: "*Su trabajo será más fácilmente una parte del conocimiento acumulado del campo si se incluye suficiente información estadística para permitir su inclusión en el meta-análisis futuro*" (p 34).

En definitiva, la última edición del Manual de la APA (2010a) mantiene y refuerza su énfasis en la denominada reforma estadística destacando el uso de los tamaños del efecto y sus intervalos de confianza y las técnicas bayesianas, tratando de minimizar la confianza excesiva que los investigadores tienen sobre las pruebas de

significación estadística y las decisiones dicotómicas apoyadas en los valores *p* de probabilidad. Además, fomenta los estudios de meta-análisis en muchos lugares y da normas para la comunicación de los trabajos de meta-análisis.

A continuación se expondrán estas alternativas de análisis, empezando por el estudio del tamaño del efecto y sus intervalos de confianza.

2.2. Tamaño del efecto y sus intervalos de confianza

Como se ha comentado, las deficiencias en la interpretación de la pruebas de significación estadística se han tratado de paliar con la estimación de algún índice del tamaño del efecto y su intervalo de confianza, junto con los valores *p* de probabilidad.

2.2.1 Tamaño del efecto: definición y ventajas

En la literatura metodológica existen inconsistencias en cómo se define el tamaño del efecto (Preacher y Kelley, 2011). Kelley y Preacher (2012) proporcionan una definición del tamaño del efecto inclusiva de todas las definiciones previas. De acuerdo a Kelly y Preacher, el tamaño del efecto es "*una representación cuantitativa de un fenómeno que se utiliza para responder a una pregunta de interés*" (*op. cit*., p.140). La pregunta de interés podría referirse a variabilidad, asociación, diferencia, proporcionalidad, superioridad, etc. (Preacher y Kelley, 2011). En este sentido, un tamaño del efecto puede ser un estadístico (en una muestra) o un parámetro (en una población) con un objetivo, cuantificar un fenómeno de interés. Y un índice o "*medida del tamaño del efecto es el nombre de una expresión que mapea datos, estadísticos, o parámetros en una cantidad que representa la magnitud del fenómeno del interés*" (Kline, 2013, p. 124). Por lo tanto, un índice del tamaño del efecto es una expresión cuantitativa de la magnitud de, por ejemplo, la relación entre dos variables o de la diferencia entre los grupos con respecto a algún atributo de interés (Frías-Navarro, 2011b; Lipsey y Wilson, 2001).

Un buen índice o medida del tamaño del efecto debe tener las siguientes características (Kelley y Preacher, 2012; Preacher y Kelley, 2011; Kline, 2013):

(1) Los valores del efecto deberían ser escalados adecuadamente, dada la medida y la pregunta de interés. Sin una escala interpretable, es difícil usar el tamaño del efecto para comunicar los resultados de una manera significativa y útil.
(2) El valor del efecto debería acompañarse de su intervalo de confianza.

(3) El punto estimado del tamaño del efecto de la población debería ser independiente del tamaño de la muestra.
(4) Un índice del tamaño del efecto debe tener buenas propiedades estadísticas, esto es, debe ser insesgado (su valor esperado debería ser igual al valor poblacional), consistente (el valor estimado debe converger con el valor poblacional a medida que aumenta el tamaño de la muestra), y eficiente (el estimador debe tener un mínimo error de varianza).

Las principales ventajas de los índices del tamaño del efecto son (Botella y Sánchez-Meca, 2015; Ferguson, 2009; Frías-Navarro, 2011b; Valera-Espín y Sánchez-Meca, 1997):

(1) En general, los índices del tamaño del efecto no dependen del tamaño de la muestra, a diferencia de los valores *p* de probabilidad (excepción hecha, e.g., del coeficiente de *phi*, *V* de Cramer).
(2) Permiten trabajar con una métrica común (ya sea vía diferencia de medias estandarizada, correlaciones, índices de riesgo, probabilidades) lo que posibilita comparar los tamaños del efecto de diferentes estudios primarios que han utilizado una métrica distinta en la medida original de las variables dependientes, a diferencia de los valores *p* de probabilidad.
(3) Favorecen la realización de estudios de meta-análisis en los que se integran de forma cuantitativa los resultados de estudios primarios sobre una misma temática, expresados en términos de tamaño del efecto, ofreciendo un tamaño del efecto medio.

Existe un amplio número de índices del tamaño del efecto (Henson, 2006; Kirk, 1996). En general, los índices del tamaño del efecto se pueden clasificar en tres grandes categorías generales (Botella y Sánchez-Meca, 2015; Ellis, 2010; Fergusson, 2009; Frías-Navarro, 2011b, Grissom y Kim, 2012; Kline, 2004; 2013; Rosnow y Rosenthal, 2009, Vacha-Haase y Thompson, 2004): (1) índices de la familia de medias, (2) índices de la familia de la relación o asociación, y (3) índices de riesgo. A efectos didácticos de una exposición más clarificadora, se ha optado por clasificar los índices del tamaño del efecto en seis categorías. A saber:

(1) ***Índices de la familia de "diferencias de medias".*** Como su nombre indica, estas estimaciones tienen en cuenta la magnitud de la diferencia entre dos o más grupos (e.g., *d* de Cohen).

(2) ***Índices de la familia de los coeficientes de correlación***. Estas estimaciones suelen examinar la fuerza de la relación entre variables. Aquí se expondrán algunos de los coeficientes de correlación para dos variables (e.g., *r* de Pearson).

(3) ***Índices de proporción de varianza explicada.*** Cuantifican la proporción de variabilidad de la variable respuesta o variable dependiente que es explicada por el efecto de la variable predictora o variable independiente (e.g. eta cuadrado).

(4) ***Índices de la familia de la asociación***. Miden la dependencia entre variables (e.g., V de Cramer).

(5) ***Índices de la familia de las estimaciones de riesgo***. Comparan el riesgo de un resultado en particular entre dos o más grupos (e.g., Riesgo Relativo).

(6) ***Índices de probabilidad de superioridad o dominancia***. Miden la probabilidad de que una puntuación seleccionada de un grupo sea superior a otra puntuación seleccionada al azar de otro grupo (e.g., índice del lenguaje común).

En el apartado 2.3 se procederá a la descripción de las diferentes familias de los tamaños del efecto y a la estimación de los mismos junto con sus intervalos de confianza.

2.2.2. Intervalos de confianza para el tamaño del efecto: definición y ventajas

De acuerdo a Smithson (2011), un intervalo de confianza es "*una estimación del intervalo alrededor de un parámetro poblacional θ que, en virtud de repetidas muestras aleatorias de tamaño N, se espera que incluya el verdadero valor en 100 (1-alpha) de las veces*" (p. 283). Por lo tanto, con un intervalo de confianza "*se comprueba la confianza de acuerdo con una distribución de probabilidad de que el verdadero valor poblacional se encuentre comprendido dentro de un rango de estimaciones*" (Valera-Espín y Sánchez-Meca, 1997, p. 86). Es decir, un intervalo de confianza indica la precisión con la que se estima un parámetro poblacional mediante un estadístico, dado *N* y α (Smithson, 2011).

En este contexto, el intervalo de confianza "*asociado a un tamaño del efecto nos indica el rango dentro del cual es probable que se encuentre el efecto real en la*

población" (Valera-Espín y Sánchez-Meca (*op. cit*, p. 87). Los intervalos de confianza son generalmente considerados como más informativos que las pruebas de significación, ya que proporcionan un rango de valores de los parámetros que reflejan el grado de incertidumbre o precisión en la estimación (Fidler, 2005; Frías-Navarro, 2011b; Kalinowski y Fidler, 2010; Valera-Espín y Sánchez-Meca, 1997).

Se han señalado diversas ventajas de los intervalos de confianza (ICs) en la literatura (e.g., Balluerka y cols., 2005; Brandstätter, 1999; Cumming, 2012, 2014; Cumming y Finch, 2005; Frías-Navarro, 2011b;Valera-Espín y Sánchez-Meca, 1997; Téllez y cols., 2015):

(1) Los ICs dan cuenta de la incertidumbre asociada a la estimación puntual del tamaño del efecto. Esto es así porque los ICs contemplan un rango de valores plausibles para el efecto de la población (por ejemplo, diferencia de medias), siendo los valores que caen fuera del IC poco plausibles.

(2) Al contener estimaciones puntuales del tamaño del efecto de la población, los ICs ofrecen información sobre la precisión de la estimación del parámetro poblacional. En este sentido, la amplitud del intervalo de confianza indica el grado de precisión del mismo. Así pues, un amplio intervalo indica una falta de precisión (o imprecisión) en la estimación del efecto; mientras que un intervalo más estrecho indica una mejor precisión en la estimación. A este respecto, como Fidler (2005) señala, los estudios con amplios ICs y, por tanto, con pobre precisión no se deberían tomar como evidencia de efectos nulos, uno de los principales problemas asociados con los valores de *p* cuando los resultados son estadísticamente no significativos. Por su parte, Kalinowski y Fidler (2010) señalan tres aspectos que afectan a la precisión de los ICs: (1) El tamaño de la muestra (cuanto mayor sea el tamaño de la muestra, más preciso será el IC, ya que será más estrecho), (2) el error estándar de la distribución muestral del estadístico, y (3) el nivel de confianza elegido para obtener el IC (un intervalo al 95% de confianza tiene una mayor amplitud que un IC al 90%, por lo tanto, el IC al 95% es menos preciso que un IC al 90%. Y a su vez, un IC al 99% de confianza tiene mayor amplitud que un IC al 95% y 90%, en consecuencia, el IC al 99% es menos preciso que un IC al 95% y un IC al 90%, esto es, a mayor grado de confianza, mayor amplitud tiene el intervalo y, por lo tanto, menos precisión).

(3) Los ICs facilitan la interpretación de los datos y permiten detectar fácilmente efectos triviales, permitiendo a los investigadores tomar decisiones de una manera clínicamente más relevante (Téllez y cols., 2015).

(4) Los ICs facilitan el pensamiento meta-analítico (Cumming, 2012, 2014; Cumming y Finch, 2005; Fidler, 2005; Fidler y Thompson, 2001). Es decir, los ICs pueden ayudar a pensar a través de los resultados de los estudios independientes, reconociendo información previa con un énfasis en el tamaño del efecto, en lugar de tomar decisiones dicotómicas ("rechazar" o "no rechazar" la hipótesis nula) basadas en los resultados de los estudios individuales.

(5) Los ICs también pueden utilizarse para hacer el contraste de hipótesis, es decir, rechazar o no la hipótesis nula, señalando sí o no el valor de la hipótesis nula cae dentro del intervalo. Por ejemplo, si el IC al 95% de un tamaño del efecto no incluye el valor de la H_0 (normalmente el valor cero) entonces se puede rechazar la H_0 con un nivel de significación del .05.

(6) Los ICs son más informativos que el valor *p* de probabilidad de las pruebas de significación estadística (Gill, 1999, Valera-Espín y Sánchez-Meca, 1997). Los ICs contienen toda la información proporcionada por una prueba de significación estadística, además de un rango de valores dentro del cual es probable que se encuentre la verdadera diferencia. Esta información facilita la comprensión de la "magnitud del efecto" a los investigadores y los profesionales de la salud y ofrece una fuente rica de información, además de la simple dicotomía sí/no de la prueba NHST (Armijo-Olivo y cols., 2011; Balluerka y cols., 2005; Kalinowski y Fidler, 2010). Por ejemplo, en el caso de intervalos de confianza para las diferencias entre los parámetros, los ICs no sólo permiten hacer el contraste de hipótesis de ausencia de diferencia (H_0), y rechazar la H_0 cuando el intervalo no incluye cero, sino que además, indican la dirección y la magnitud de la diferencia observada (Balluerka y cols. 2005). Las pruebas de significación estadística sólo informan de la dirección del efecto, pero no de su magnitud.

En resumen, como algunos autores señalan (e.g., Cumming, 2012, 2014; Balluerka y cols., 2005), los intervalos de confianza para los tamaños del efecto son más informativos que la prueba de la NHST porque contienen información sobre el tamaño del efecto y su precisión y, además, se pueden utilizar para llevar a cabo el contraste de hipótesis, si fuera necesario (simplemente determinando si el valor nulo está dentro o

fuera del intervalo). Por su parte, las pruebas de significación estadística requieren la información separada sobre los valores de *p*, el tamaño del efecto, y un cálculo de la potencia estadística para proporcionar información equivalente (Kalinowski y Fidler, 2010).

En el siguiente apartado se describen las familias de los tamaños del efecto y los índices que las componen, y en el apartado 2.4 se proporcionan algunas herramientas para la estimación de los tamaños del efecto y sus intervalos de confianza.

2.3. Estimación del tamaño del efecto y su intervalo de confianza

La mayoría de los programas estadísticos estándares como eel SPSS o SAS no incluyen la estimación de los tamaños del efecto y sus intervalos de confianza de forma directa (Frías-Navarro, 2011b).

En general, para calcular un intervalo de confianza se requiere la estimación del tamaño del efecto, la puntuación en la distribución de probabilidad que le corresponde al tamaño del efecto al nivel de confianza deseado y el error estándar o típico del estadístico (Valera-Espín y Sánchez-Meca, 1997).

Con estos datos, la obtención de los dos límites de un IC supone sumar y restar al estadístico del tamaño del efecto obtenido en una muestra ($\hat{\theta}$), un término de error que depende del error estándar de la distribución muestral del estadístico y del nivel de confianza asumido en la definición del intervalo (Fidler y Thompson, 2001).

Sin embargo, estimar el intervalo de confianza del tamaño del efecto (e.g., los ICs de los tamaños del efecto estandarizados como diferencia de media estandarizada, coeficientes de correlación, índices de proporción de varianza explicada, etc.) no es sencillo porque las distribuciones apropiadas para construirlos no son centrales, sino más bien sus homólogos no centrales (Fidler y Thompson, 2001; Frías-Navarro, 2011b).

A diferencia del cálculo de los ICs para distribuciones centrales de los tests, calcular los ICs para la distribución no central es poco práctico sin un programa de ordenador relativamente sofisticado (Fidler y Thompson, 2001; Grissom y Kim, 2012; Kline, 2013). Las distribuciones centrales son definidas por un solo parámetro, los grados de libertad, pero las distribuciones no centrales se definen por dos parámetros, los grados de libertad y el parámetro de no centralidad que indica el grado en que la hipótesis nula es falsa (Kline, 2013).

En el siguiente apartado se exponen los distintos tamaños del efecto agrupados en familias en función de sus similitudes, así como sus intervalos de confianza, junto con una aproximación a su estimación.

2.3.1.Índices de la familia de "diferencia de medias"

Los índices de diferencia de medias se utilizan para comparar dos grupos en una variable cuantitativa medida. Es decir, cuando se dispone de una variable categórica dicotómica (que sirve para clasificar a los sujetos en las dos categorías de la variable) y una variable cuantitativa medida (en la que se quiere comparar las categorías o grupos). Estos índices se pueden aplicar tanto en diseños experimentales (donde hay asignación aleatoria de los sujetos a las condiciones de la investigación), como en diseños cuasi-experimentales y no experimentales (donde no hay asignación aleatoria); la diferencia entre ellos no está en la manera de estimar el tamaño del efecto sino en la interpretación que se pueda hacer del mismo. Mientras que en los diseños experimentales se pueden hacer inferencias de causalidad, en el diseño cuasi-experimental las inferencias deben ser de relación entre variables (Botella y Sánchez-Meca, 2015; Frías-Navarro, 2011a).

Como Botella y Sánchez-Meca (2015) señalan, la comparación más sencilla es la diferencia de las medias de dos grupos o categorías (*diferencia de media directa*). En este caso, cuanto mayor es la diferencia entre las medias, mayor es la importancia de la variable categórica (o factor de clasificación de los grupos). Por el contrario, si las medias no difieren entre sí significa que el factor que distingue a los grupos es irrelevante.

Sin embargo, lo más habitual es utilizar una diferencia de medias estandarizada o tipificada, es decir, la diferencia de las dos medias en términos de unidades de desviación típica. Una diferencia de media estandarizada de 0.50 significa que la diferencia entre los dos grupos es equivalente a la mitad de una desviación típica, mientras que una diferencia de media estandarizada de 1 significa que la diferencia entre las medias es igual a una desviación típica. Por lo tanto, cuanto más grande sea la diferencia estandarizada, mayor será el efecto. Una ventaja de los tamaños del efecto estandarizados es que están libres de la escala de medida, lo que permite hacer comparaciones entre distintos estudios.

En este apartado se describirá cómo se estiman los índices de diferencia de medias directa y los índices de la diferencia de medias estandarizadas o tipificadas más utilizados y los métodos robustos.

Diferencia de medias directa

La diferencia de medias directa en la población se define como $\Delta=\mu_1-\mu_2$. La estimación de la diferencia de medias, D[2], depende del diseño de investigación, es decir, si se trata de un diseño de grupos independientes o de un diseño de grupos relacionados (o medidas repetidas). En un diseño de dos grupos independientes, la diferencia de medias directa se define como la diferencia entre las puntuaciones medias de un grupo en una variable de resultado y las puntuaciones medias del otro grupo. Mientras que en los diseños de medidas repetidas (o grupos relacionados, por ejemplo, diseños de investigación pretest y postest o cualquier forma de emparejamiento), la diferencia de medias se define como la diferencia entre, por ejemplo, las puntaciones medias de la fase de pretest y de la fase de postest en la variable de resultado (Borenstein y cols., 2009).

En ambos casos, para la estimación del intervalo de confianza de la diferencia de medias, se requiere estimar la varianza de la diferencia, a partir de la cual estimar el error estándar necesario para construir el intervalo de confianza al nivel deseado.

Diferencias de medias estandarizadas

Como en el caso de la diferencia de medias directa, la estimación de la diferencia de medias estandarizada depende del diseño de investigación, es decir, si se trata de un diseño de grupos independientes o de un diseño de grupos relacionados (o medidas repetidas).

Diferencias de medias estandarizadas para grupos independientes

El tamaño del efecto paramétrico que se desea estimar se basa en la diferencia de medias tipificada o estandarizada (δ). Puede ser estimado por la diferencia de medias estandarizada o tipificada a partir de los datos de una muestra (d). Los valores de la d pueden oscilar desde $-\infty$ hasta $+\infty$, donde 0 indica ausencia de diferencia entre las medias de los grupos. Sin embargo, lo habitual es que la d tome valores entre -3 y +3.

[2]Siguiendo a Borenstein y cols. (2009) se utiliza la letra D (en mayúsculas) para describir la diferencia de medias directa, por tanto, en valores directos, y se utiliza la *d* (en minúscula) para los valores estandarizados.

Los tres principales estimadores del tamaño del efecto paramétrico (δ) son: la *d* de Cohen (1988), la *g* de Hedges (1981) y la delta de Glass (1976), las cuales difieren en la desviación típica que utilizan en el denominador. Los tres estadísticos estiman el mismo parámetro y se diferencian en el modo de realizar la estandarización o tipificación de la diferencia de medias (Li, 2015; Peng y Chen, 2014; Ruscio, 2008a).

1. *d* de Cohen

La *d* de Cohen es el estimador del tamaño del efecto paramétrico más utilizado en Psicología cuando la investigación incluye una variable cuantitativa medida (variable dependiente) y una variable categórica dicotómica (Frías-Navarro, 2011b). Se define como la diferencia entre las medias de los dos grupos dividida por la desviación típica común.

La *d* de Cohen asume que la variable cuantitativa sigue una distribución normal en ambos grupos y que existe igualdad de la varianza poblacional, asunciones de las pruebas de inferencia estadística paramétrica. Por tanto, la *d* de Cohen es muy sensible a las violaciones de estos supuestos. Se sabe que las violaciones de estos supuestos son bastante comunes en la práctica en el ámbito de la Psicología (Erceg-Hurn y Mirosevich, 2008; Grissom y Kim, 2012).

Además, la *d* de Cohen es un estimador del tamaño del efecto que está positivamente sesgado en muestras pequeñas (Botella y Sánchez, 2015; Fritz y cols., 2012), es decir, tiende a sobreestimar el tamaño del efecto de la población en muestras pequeñas, especialmente cuando el tamaño de la muestra es menor de 20 observaciones ($n < 20$) (Kline, 2013) o menor de 10 observaciones por grupo (Nakagawa y Cuthill, 2007). En estos casos se debe aplicar el factor corrector propuesto por Hedges (1981).

2. Delta de Glass (o Δ)

Como se ha dicho, la *d* de Cohen asume la normalidad y la igualdad de la varianza poblacional (homoscedasticidad). Por lo tanto, cuando las varianzas poblacionales no son iguales (heterocedasticidad), es preferible aplicar la delta de Glass ya que solamente asume la normalidad de la variable.

La delta de Glass se define como la diferencia entre las medias de los dos grupos dividida por la desviación típica de la población control o de comparación.

Se asume que la varianza del grupo control es un estimador más adecuado de la varianza poblacional dado que la varianza experimental puede estar afectada por los efectos de la intervención. La varianza del grupo control no estará afectada y por ello representará mejor la varianza poblacional (Grissom y Kim, 2012; Lipsey y Wilson, 2001).

Finalmente, la delta de Glass, al igual que la *d* de Cohen, es un estimador del tamaño del efecto que está positivamente sesgado en muestras pequeñas (Grissom y Kim, 2012), es decir, tiende a sobreestimar el tamaño del efecto de la población en muestras pequeñas.

3. *g* de Hedges (o *d* insesgada, *d unbiased*, d^u)

El sesgo positivo de la *d* de Cohen y de la delta de Glass se puede corregir multiplicando el estimador (*d* o delta) por el factor corrector propuesto por Hedges (1981). El valor del factor de corrección se aproxima a 1 a medida que aumenta el tamaño de la muestra (Kline, 2013). Como Botella y Sánchez-Meca (2015) señalan, con un valor de factor de corrección próximo a 1 la corrección resulta inapreciable. Por ello, la corrección es muy pequeña cuando el tamaño de la muestra es grande (sólo el 3% para 25 grados de libertad), pero es más grande, cuando el tamaño de la muestra es pequeño (8% para 10 grados de libertad) (Fritz y cols., 2012).

Finalmente, para construir el intervalo de confianza se requiere estimar la varianza de la diferencia de medias estandarizada, a partir de la cual estimar el error estándar necesario para estimar los límites del intervalo de confianza al nivel deseado. El lector interesado en la estimación del intervalo de confianza puede consultar a Botella y Sánchez-Meca (2015) y Borenstein y cols. (2009) los cuales proporcionan las ecuaciones matemáticas necesarias para realizar dicho cómputo.

Diferencia de medias estandarizada para grupos dependientes o relacionados

La diferencia de medias estandarizada para grupos dependientes o relacionados, al igual que la diferencia de media directa de diseños de medidas repetidas, se aplica cuando se trabaja con diseños de investigación pretest y postest, o con cualquier forma de emparejamiento (Botella y Sánchez-Meca, 2015).

Se pueden distinguir dos situaciones para la estimación de la diferencia de medias estandarizada, cuando en el diseño de investigación no existe grupo de control

(cambio medio tipificado) y cuando en el diseño de investigación sí que hay grupo de control (diferencia de cambios medios tipificados).

1. Diseños de investigación sin grupo de control: cambio medio tipificado.

El cambio medio tipificado hace referencia a la diferencia entre las medias del pretest y postest de un grupo dividida por una desviación típica (Becker, 1988).

Se han propuesto diferentes estimadores del tamaño del efecto paramétrico que varían en la desviación típica que utilizan en el denominador para su cómputo.

1. $\boldsymbol{d_{c1}}$: La desviación típica de las puntuaciones de cambio (diferencia entre el pretest y el postest) (Gibbson, Hedeker y Davis, 1993).
2. $\boldsymbol{d_{c2}}$**:** La desviación típica del pretest (Becker, 1988; Morris, 2000; Morris y DeShon, 2002).
3. $\boldsymbol{d_{c3}}$**:** El promedio de las desviaciones típicas del pretest y del postest (Dunlap, Cortina, Vaslow, y Burke, 1996; Taylor y White, 1992).

Los dos primeros estimadores (d_{c1}, y d_{c2}) están sesgados por lo que se les aplica un factor corrector (véase Botella y Sánchez-Meca, 2015; Borenteisn y cols., 2009).

Para construir el intervalo de confianza se requiere estimar la varianza de la diferencia de medias estandarizada, a partir de la cual estimar el error estándar necesario para estimar los límites del intervalo de confianza al nivel deseado. De acuerdo con Botella y Sánchez-Meca (2015), no se dispone de una fórmula que permita calcular la varianza del índice d_{c3}. Por lo tanto, no es posible estimar su intervalo de confianza.

El lector interesado en la estimación del intervalo de confianza puede consultar a Botella y Sánchez-Meca (2015) y Borenstein y cols. (2009), los cuales proporcionan las ecuaciones matemáticas necesarias para realizar dicho cómputo.

2. Diseños de investigación con grupo de control: Diferencia de cambios medios tipificados

Cuando se trabaja con diseños pretest y postest con grupo de control o diseños de dos grupos con medidas pretest y postest se habla de diferencia de cambios medios tipificados. En estos supuestos se analiza un cambio diferencial, es decir, un cambio pre-post diferente en los dos grupos.

El tamaño del efecto paramétrico que se desea estimar se basa en la comparación de los cambios pre-post tipificados de ambos grupos (Borenstein y cols., 2009; Botella y Sánchez-Meca, 2015).

Como en los diseños de investigación sin grupo de control, se han propuesto diferentes estimadores del tamaño del efecto paramétrico que varían en la desviación típica que utilizan en el denominador para el cómputo del tamaño del efecto.

A. d_{g1}: La desviación típica del pretest (Becker, 1988; Morris, 2008; Morris y DeShon, 2002).

B. d_{g2}: El promedio de las desviaciones típicas del pretest de los grupos experimental y control (Carlson y Schmidt, 1999; Morris, 2008), si se puede asumir que las desviaciones típicas del pretest en ambos grupos son similares.

C. d_{g3}: La desviación típica de las puntuaciones de cambio.

Para construir el intervalo de confianza de cada uno de estos índices de diferencias de medias estandarizadas, se requiere estimar su varianza, a partir de la cual estimar su error estándar necesario para estimar los límites del intervalo de confianza al nivel deseado. El lector interesado en la estimación del intervalo de confianza puede consultar a Botella y Sánchez-Meca (2015) y Borenstein y cols. (2009), los cuales proporcionan las ecuaciones matemáticas necesarias para realizar dicho cómputo.

En definitiva, los índices de estimación de la diferencia de medias estandarizadas (*d*) difieren en el denominador, esto es, el tipo de desviación típica que utilizan en la estimación del tamaño del efecto, arrojando valores del tamaño del efecto distintos. Por ello, se debe especificar claramente cómo se calcularon las estimaciones del tamaño del efecto, independientemente de qué símbolo se utilice, de modo que el lector pueda interpretar correctamente los resultados, y estos puedan ser más fácilmente integrados en posteriores estudios meta-analíticos (Botella y Sánchez-Meca, 2015; Ellis, 2010; Frías-Navarro, 2011b; Fritz y cols., 2012; Kline, 2013). Además, es aconsejable utilizar múltiples índices de diferencia de medias estandarizadas como estimadores del tamaño del efecto (Grissom y Kim, 2012).

Diferencia de medias estandariza mediante métodos robustos modernos

Existen estadísticos de diferencia de medias estandarizada robustos (d_r) para calcular la magnitud de un efecto entre dos grupos independientes y dos grupos relacionados.

El término "estadísticos robustos" se refiere a procedimientos que son capaces de mantener la tasa de error de tipo I de una prueba en su nivel nominal y también mantener la potencia estadística de la prueba, incluso cuando se incumple el supuesto de normalidad y de homoscedasticidad (igualdad de varianza poblacional).

La diferencia de media estandarizada mediante métodos estadísticos robustos supone reemplazar las medidas de localización y dispersión sensibles como la media y la varianza, por medidas robustas como por ejemplo la media recortada o la mediana y la varianza winsorizada (e.g., Algina y cols., 2005; Keselman y cols., 2008). La estimación de la diferencia estandarizada robusta basada en la media recortada y la varianza winsorizada implica la eliminación de una parte de las puntuaciones altas y bajas de una muestra, lo que elimina el problema de los valores extremos (*outliers*) que generalmente conduce a variaciones extremas y distribuciones asimétricas.

Para la estimación de los intervalos de confianza de las diferencias de medias estandarizadas robustas expuestas (y otras) existe un programa estadístico de acceso libre desde la página web de Algina (http://plaza.ufl.edu/algina/index.programs.html) que estima el error estándar mediante los métodos de remuestreo (*bootstrap)*.

2.3.2. Índices de la familia del "coeficiente de correlación"

Los índices de la familia del coeficiente de correlación son probablemente una de la más conocidas formas de medir el tamaño del efecto, aunque muchos de los investigadores que utilizan estos estadísticos pueden no ser conscientes de que sean índices del tamaño del efecto (Ellis, 2010; Frías-Navarro, 2011b).

Los coeficientes de correlación (*r*) cuantifican la fuerza y la dirección de una relación entre dos variables (X e Y). Las variables pueden ser: dos variables cuantitativas al menos en escala de intervalo, dos variables al menos en escala ordinal, o bien, una variable cuantitativa al menos en escala de intervalo y otra variable categórica dicotómica o dicotomizada.

Los coeficientes de correlación son estadísticos estandarizados, es decir, libres de las unidades de medida, por lo que, en general, permiten su comparación entre diferentes estudios. Sus valores oscilan entre -1 (que indica una relación perfectamente lineal negativa o inversa) y 1 (lo que indica una relación lineal perfectamente positiva o directa), mientras que una correlación de 0 indica que no existe una relación lineal entre las variables (Ellis, 2010).

Dentro de la familia de los coeficientes de correlación se puede distinguir entre: (1) coeficientes paramétricos (los más conocidos y utilizados), (2) coeficientes no paramétricos (o robustos clásicos), y (3) coeficientes robustos modernos.

A continuación se describen cuáles componen cada uno de estos grupos o subfamilias de coeficientes de correlación.

2.2.2.1 Coeficientes de correlación paramétricos

Los coeficientes de correlación (*r*) paramétricos cuantifican la magnitud y dirección de la relación entre dos variables cuantitativas, o bien, una variable cuantitativa y otra variable categórica dicotómica natural o dicotomizada, asumiendo que existe una relación lineal entre las variables, que la(s) variable(s) cuantitativa(s) sigue(n) una distribución normal, y homocedasticidad (la varianza de los errores es constante).

Los coeficientes de correlación paramétricos más comunes son (para revisar otros índices de correlación ver Rosnow, Rosenthal y Rubien, 2000):

(1) Coeficiente de correlación producto-momento de Pearson (r_{xy}): cuantifica la fuerza y dirección de la relación entre dos variables cuantitativas (X e Y).

(2) **Coeficiente de correlación biserial-puntual (r_{pb}):** se utiliza para cuantificar la fuerza y dirección de la relación entre una variable dicotómica natural (X), es decir, una variable categórica que presenta dos categorías, y una variable cuantitativa en escala de intervalo o de razón (Y). Las dos categorías (o modalidades) de la variable categórica suelen codificarse con los valores 0 y 1 o bien 1 y 2.

A la hora de interpretar el valor del coeficiente r_{pb} se debe tener en cuenta a qué categoría o modalidad de la variable dicotómica natural X se le ha asignado el valor más alto. Si r_{pb} es positivo, significa que los sujetos pertenecientes a la categoría o modalidad de la variable X a la que se le ha asignado el número más alto en la codificación muestran unas puntuaciones más altas en la variable Y que los participantes de la otra categoría de la variable X. Por el contrario, si r_{pb} es negativo, significa que los sujetos pertenecientes a la categoría o modalidad de la variable X a la que se le ha asignado el número más alto muestran unas puntuaciones más bajas en la variable Y que los participantes de la otra categoría de la variable X (Grissom y Kim, 2012).

(3) Coeficiente de correlación biserial (r_b): se utiliza para cuantificar la fuerza y dirección de la relación lineal entre una variable dicotomizada (X) a partir de una variable cuantitativa que sigue una distribución normal y una variable cuantitativa (Y).

El coeficiente de correlación biserial es una estimación del valor del coeficiente de Pearson en el caso de que la variable no hubiera sido dicotomizada y la relación entre ellas fuera lineal (Grissom y Kim, 2012). Se interpreta del mismo que el coeficiente de correlación biserial-puntual.

Por otra parte, señalar que los coeficientes de correlación (r_{xy}, r_{pb} y r_b) están, de acuerdo con Grissom y Kim (2012), negativamente sesgados, es decir, tienden a infraestimar el tamaño del efecto en muestras pequeñas ($N \leq 20$), por lo que en estos casos, se debe aplicar el factor de corrección propuesto por Hedges y Olkin (1985).

Para la estimación del intervalo de confianza se utiliza la transformación de la Z de Fisher (Zr) (Borenstein y cols., 2009; Botella y Sánchez-Meca, 2015; Kline, 2013). Una vez obtenido el valor de la Zr y los límites inferior y superior de su intervalo de confianza, estos valores se deben devolver a la métrica original de la medida, aplicando la fórmula inversa (Borenstein y cols., 2009; Botella y Sánchez-Meca, 2015).

Existen páginas webs que directamente calculan los intervalos de confianza para los coeficientes de correlación aplicando la transformación Z de Fisher. Por ejemplo, la página http://vassarstats.net/rho.html estima los intervalos de confianza al 95% y 99% de los coeficientes de correlación dado un valor de *r* y un tamaño de la muestra.

2.2.2.2. Coeficientes de correlación no paramétricos (o robustos clásicos)

Como ya se ha dicho, los índices de la familia del "coeficiente de correlación" cuantifican la fuerza y la dirección de una relación entre dos variables (X e Y). Ejemplos de estos coeficientes, en el caso no paramétrico, son el coeficiente de correlación de Spearman (*rho*) y el coeficiente tau de Kendall (τ), los cuales "*reflejan la fuerza de una relación monótona más que lineal*" (Wilcox, 2010, p. 178). Estos estadísticos son robustos frente a los valores atípicos (*outliers*), pero no lo son cuando se utilizan para analizar datos heterocedásticos (Erceg-Hurn y Mirosevich, 2008; Wilcox, 2012).

(1) **Coeficiente de correlación de Spearman (*rho*)**: se utiliza cuando ambas variables X e Y está medidas al menos en escala ordinal. No asume ninguna distribución de los datos, es decir, no asume que los datos siguen una distribución normal, por lo que se puede utilizar cuando las variables continuas no siguen una distribución normal como alternativa al coeficiente de correlación de Pearson (Fritz y cols., 2012). Además, como se ha dicho, el coeficiente de Spearman es un estadístico resistente a los valores extremos o atípicos (*outliers*) (Wilcox, 2010). Es una adaptación del coeficiente de correlación de Pearson, que convierte las observaciones en rangos. Sus valores oscilan entre -1 (que indica una relación perfectamente lineal negativa o inversa) y +1 (lo que indica una relación lineal perfectamente positiva o directa), mientras que una correlación de 0 indica que no existe una relación lineal entre las variables.

Para la estimación del intervalo de confianza de *rho* se puede utilizar la transformación de la Z de Fisher, antes comentada. Sin embargo, Ruscio (2008b) señala que en ciertas condiciones (e.g. valores de *rho* > .30), este método infraestima el error estándar, por lo que el intervalo de confianza resultante es demasiado estrecho y proporciona una cobertura inferior a si hubiera sido estimado mediante métodos de *bootstrap.*

(2) **Coeficiente tau de Kendall (τ)**: se utiliza cuando ambas variables X e Y se miden en escala ordinal. De acuerdo con Long y Cliff (1997), el coeficinete tau puede tener propiedades superiores al coeficiente de correlación de Pearson *"cuando los datos son ordinales en naturaleza, y/o + cuando las relaciones de dos variables parecen no ser lineales, sino monótonas, y/o cuando las variables tienen distribuciones no normales*" (p. 31).

El coeficiente tau de Kendall (τ) proporciona una estimación del tamaño del efecto paramétrico basado en el número de pares concordantes en dos órdenes de rangos (Ferguson, 2009). Sus oscilan entre -1 (que indica una relación perfectamente lineal negativa o inversa) y 1 (lo que indica una relación lineal perfectamente positiva o directa), mientras que una correlación de 0 indica que existe independencia entre las variables (Wilcox, 2010).

Para la estimación de su intervalo de confianza también se puede utilizar la transformación de la Z de Fisher, si bien existen otros métodos para su estimación que, de acuerdo con Long y Cliff (1997), han mostrado buenas propiedades estadísticas.

Finalmente, resta mencionar la existencia de otros índices que también miden la fuerza de la relación entre variables en escala ordinal, como gamma y *d* de Somer (Agresti y Finley, 2009; Ferguson, 2009; Grissom y Kim, 2012). Gamma es una medida similar al tau de Kendall y mide la fuerza de la asociación entre los pares de rangos ordenados concordantes, mientras que la *d* de Somer es similar a gamma, aunque asimétrica. El tau de Kendall, gamma, *d* de Somer son índices con rangos de valores idénticos a r_{xy} (Ferguson, 2009).

2.2.2.3. Coeficientes de correlación robustos modernos

Como se ha comentado, los estadísticos robustos modernos ofrecen ventajas sobre los métodos paramétricos y no paramétricos clásicos en términos de control del Error de Tipo I y de potencia estadística (Erceg-Hurn y Mirosevich, 2008).

Entre los estadísticos de correlación robustos modernos el más conocido es la correlación winsorizada (r_{win}). La r_{win} se puede utilizar para analizar la fuerza de la relación entre dos variables cuantitativas medidas al menos en escala de intervalo (Wilcox, 2010, 2012).

Para la estimación del intervalo de confianza de la r_{win}, se puede utilizar la transformación de la Z de Fisher, dado que la estimación del estadístico finalmente se realiza mediante el coeficiente de correlación de Pearson (r_{xy}) aplicado a los datos una vez recortados y winsorizados.

2.3.3. Índices de la familia de la "proporción de varianza explicada"

Los índices de la familia de la proporción de varianza explicada cuantifican la proporción de variabilidad de la variable respuesta o variable dependiente (Y) que es explicada por el efecto de la variable predictora o variable independiente (X) (Frías-Navarro, 2011b; Grissom y Kim, 2012). Una de las ventajas de estos índices es su fácil interpretación, pues al multiplicarlos por 100 se puede hablar en términos de porcentaje de la variable dependiente explicada por el efecto de la variable predictora.

Dentro de esta familia los más comúnmente utilizados son (Ferguson, 2009, Frías-Navarro, 2011b; Fritz y cols., 2012): (1) coeficiente de determinación, (2) eta cuadrado, (3) eta cuadrado parcial, (4) eta cuadrado generalizada, (5) omega cuadrado, (6) omega cuadrado parcial, (7) épsilon, y (8) coeficiente múltiple de determinación.

La exposición de la aplicación de la mayoría de estos índices se focaliza en los diseños entre sujetos (o grupos independientes) y con un solo factor o variable independiente (diseños unifactoriales). No obstante, el lector debe saber que estos índices también son de aplicación en diseños factoriales y diseños de medidas repetidas (para una revisión ver, e.g., Grissom y Kim, 2012; Kline, 2013; Trigo-Sánchez y Martínez-Cervantes, 2016).

(1) Coeficiente de determinación (r^2): se describe como el cuadrado de algún índice de la familia de los coeficientes de correlación. De acuerdo a Grissom y Kim (2012) r^2 ha sido ampliamente utilizado como un estimador del coeficiente de determinación de la población.

Sus valores oscilan entre 0 y 1, siendo cero el efecto nulo, esto es, el caso en que las puntuaciones en la variable dependiente o predicha (Y) no son explicadas por la variabilidad de la variable independiente o predictora (X). Por el contrario, cuando r^2 = 1, toda la variabilidad observada en la variable dependiente (Y) es explicada por la variabilidad en la variable independiente o predictora (X).

Se debe tener en cuenta que r^2 es un estimador positivamente sesgado del tamaño del efecto en muestras pequeñas, por lo que sobreestima el tamaño del efecto en estos supuestos (Wang y Thompson, 2007). Se han propuesto diferentes métodos para corregir el sesgo del r^2, como el método de Ezequiel, el de Smith, el de Claudy, el Wherry, el Lord, etc. (para una revisión Wang y Thompson, 2007). De los diferentes métodos, Wang y Thompson (2007), encontraron en su estudio de simulación que el método de Ezequiel y el método de Smith eran los que obtuvieron los mejores resultados.

La corrección por el método de Ezequiel está disponible en el programa estadísticos SPSS a través del cálculo de r^2 ejecutando el procedimiento de regresión, pero la corrección Smith podría aplicarse fácilmente usando una hoja de cálculo.

La estimación de los intervalos de confianza se puede realizar usando los métodos descritos para la estimación de intervalos de confianza de los coeficientes de correlación (Grissom y Kim, 2012). Se remite al lector a dicho apartado para la construcción de los ICs para r^2.

(**2) Eta cuadrado (η^2)[3]:** describe la proporción de la variabilidad total de los datos que se explica por el efecto de que se trate. Se puede estimar desde la salida (*output*) del análisis de varianza (ANOVA). Se define como la relación de la suma de cuadrados para el efecto dividido por la suma total de cuadrados (Fritz y cols., 2012).

La η^2es un estimador del tamaño del efecto que está positivamente sesgado en muestras de tamaño pequeño, esto es, tiende a sobreestimar el tamaño del efecto (Grissom y Kim, 2012). En estos casos, se debe utilizar omega cuadrado. Sin embargo, la diferencia entre ambos estimadores es pequeña, y el sesgo disminuye a medida que aumenta el tamaño de la muestra (Grissom y Kim, 21012; Lakens, 2013).

Por otra parte, η^2 (al igual que la omega cuadrado) son útiles para comparar los tamaños del efecto dentro de un mismo estudio (Fritz y cols., 2012). Sin embargo, no sirven para comparar los tamaños del efecto entre estudios, debido a que la variabilidad total en un estudio depende del diseño del estudio, y aumenta cuando se manipulan variables adicionales (Fritz y cols., 2012; Lakens, 2013; Olejnik y Algina, 2003). Para comparar el efecto de un mismo factor entre estudios con diseños similares, se recomienda utilizar eta cuadrado parcial u omega cuadrado parcial, dado que eliminan la influencia de otros factores en el diseño (Olejnik y Algina, 2003), siempre que los términos del error sean comparables (Fritz y cols., 2012).

(3) Eta cuadrado parcial (η_p^2): se puede utilizar para comparar el tamaño del efecto de un mismo factor entre distintos estudios con similares diseños siempre que los términos del error sean comparables (Fritz y cols., 2012; Lakens, 2013; Olejnik y Algina, 2003) La η_p^2 expresa la suma de los cuadrados de los efectos en relación con la suma de los cuadrados de los efectos y la suma de los cuadrados del error asociado con el efecto (Fritz y cols., 2012). La estimación de η_p^2 se puede hacer desde los valores del estadístico *F* proporcionado por los investigadores en sus informes (véase a Bakeman, 2005; Fritz y cols., 2012; Lakens, 2013; Trigo-Sánchez y Martínez-Cervantes, 2016).

(4) Eta cuadrado generalizada (η_G^2): se utiliza para comprar los tamaños del efecto en diferentes estudios y con diferentes diseños. Como los índices anteriores, la η_G^2

[3] Como Fritz y cols. (2012) indican, algunos autores prefieren referirse a η^2como R^2 porque encaja con la convención estadística de reservar las letras griegas para los parámetros de población y debido a la similitud con R^2 de la regresión. Sin embargo, los programas estadísticos y los manuales de estadística utilizan con mayor frecuencia η^2 para el ANOVA y R^2 para los análisis de regresión. Por lo tanto, para hacer una exposición más clara en este capítulo se utilizará el η^2 relacionado con ANOVA y R^2 con los análisis de regresión.

proporciona una estimación de la proporción de variabilidad dentro de un estudio que está asociada con una variable, pero sin los efectos de distorsión de las variables introducidas en algunos estudios que no han sido consideradas en otros (Bakeman, 2005; Fritz y cols., 2012; Olejnik yAlgina, 2003; Trigo-Sánchez y Martínez-Cervantes, 2016).

Trigo-Sánchez y Martínez-Cervantes, (2016) han elaborado *scripts* para el programa estadístico SPSS que se pueden descargar desde la siguiente página web http://personal.us.es/trigo/suppmaterials.htm para calcular el índice eta cuadrado generalizado para diversos tipos de hipótesis, generales o específicas; tipos de diseños, univariables o factoriales; y con factores manipulados y/o medidos.

(5) Omega cuadrado (ω^2): se puede aplicar para estimar la proporción de varianza explicada en muestras de tamaño pequeño y para comparar el tamaño del efecto dentro de un mismo estudio.

(6) Omega cuadrado parcial (ω_p^2): es un índice alternativo a eta cuadrado parcial. Por lo tanto, se puede utilizar para comparar el tamaño del efecto de un mismo factor entre distintos estudios con muestras de tamaño pequeño siempre que sean los diseños similares y los términos de error comparables (Fritz y cols., 2012; Olejnik y Algina, 2003).

(7) **Épsilon cuadrado (ε^2):** es un índice alternativo a omega cuadrado (Ferguson, 2009), pero escasamente utilizado (Fritz y cols., 2012).

(8) Coeficiente múltiple de determinación (R^2): estima la asociación en la población entre Y y la óptima combinación lineal ponderada de las variables X (Grissom y Kim, 2012). De acuerdo a Fritz y cols. (2012), R^2 se define como el cuadrado de la correlación entre los valores observados y los valores predichos por la ecuación de regresión y se utiliza para informar de la proporción de la variabilidad de la variable dependiente que es predecible a partir de un conjunto de variables que se han introducido en la ecuación de regresión.

Como Fritz y cols. (2012) señalan, la R^2 describe el efecto de un conjunto de variables (una o varias) mientras que la η^2 describe el efecto de un solo factor o interacción.

La R^2 está positivamente sesgada, es decir, tiende a sobreestimar el tamaño del efecto. El sesgo incrementa a medida que aumenta el número de variables independientes que se incorporan al modelo de regresión y con muestras de tamaño muestral pequeño (Grissom y Kim, 2012). Para reducir este sesgo se ha propuesto un factor corrector que es aplicado por defecto en las salidas (*outputs*) de los programas estadísticos estándares como SPSS, SAS y SYSTAT. Por lo tanto, se recomienda el uso de la $R^2_{Ajustada}$, en lugar de la R^2.

Finalmente, la **estimación de los intervalos de confianza para los índices de proporción de varianza** explicada requiere el uso de **distribuciones no centrales.** Por lo tanto, el cálculo manual de los mismos no resulta práctico (Grissom y Kim, 2012). Smithson (2003) proporciona *scripts* para los programas estadísticos de SPSS, SAS, S-PLUS y R para construir los ICs para estos índices. En Fidler y Thompson (2001) el lector interesado puede encontrar una demostración de cómo se aplican los *scripts* mencionados en el programa estadístico SPSS.

2.3.4. Índices de la familia de los "índices de riesgo"

Los índices de riesgo son más comúnmente utilizados en la investigación biomédica (Ferguson, 2009; Frías-Navarro, 2011b; Rosnow y Rosenthal, 2003). Los índices de riesgo estiman la proporción de sujetos que experimentan un determinado resultado (Frías-Navarro, 2011b) o la diferencia en el riesgo de un resultado en particular (Ferguson, 2009).

Los índices de riesgo se aplican cuando las variables de resultados son dicotómicas naturales o han sido dicotomizadas (Sánchez-Meca, Marín-Martínez y Chacón-Mosoco, 2003). Por lo tanto, se utilizan en los datos que se organizan en tablas de contingencia 2x2.

Los tres estimadores de riesgo más más utilizados son: (1) Diferencias de Riesgo (o de proporciones), (2) Riesgo Relativo (o razón de proporciones), y (3) *Odds Ratio* (o razón de ventajas).

(**1) Diferencias de Riesgo (o de proporciones) (*DR*)**: se define como la diferencia entre las proporciones de la variable de resultado en los grupos formados por el factor ($p_1 - p_2$) y estima el parámetro $\pi_1.\pi_2$. Sus valores oscilan entre -1 y +1. El valor 0 refleja ausencia de efecto y se obtiene cuando las proporciones son iguales.

Para construir el intervalo de confianza se requiere estimar la varianza de la diferencia de riesgo, a partir de la cual estimar el error estándar necesario para estimar los límites del intervalo de confianza al nivel deseado. El lector interesado en la estimación del intervalo de confianza puede consultar a Borenstein y cols. (2009), Botella y Sánchez-Meca (2015), Newcombe (2012), los cuales proporcionan las ecuaciones matemáticas necesarias para realizar dicho cómputo.

(**2) Riesgo Relativo (o razón de proporciones) (*RR*):** se utiliza para valorar el rol de factores contextuales como covariables que representan un riesgo incrementado (factor de riesgo) o reducido (factor de protección) respecto de un potencial evento adverso (Botella y Sánchez-Meca, 2015). Es un índice no paramétrico (Frías-Navarro, 2011b).

Como Grissom y Kim (2012) señalan, el nombre de "riesgo relativo" se aplica a la investigación médica, en la que la categoría objetivo es la clasificación de las personas que tienen una enfermedad frente a la categoría de no presentar dicha enfermedad (la variable dependiente binomial). Una de las muestras tiene al menos un factor de riesgo (e.g., ser fumador) y la otra muestra no presenta este factor de riesgo (la variable independiente binomial). En la investigación psicológica, donde se representa el éxito de una terapia, en lugar de hablar de "riesgo relativo", es más apropiado hablar de "razones de tasas de éxito", o de "relación entre dos probabilidades independientes" o de "relación de dos proporciones independientes".

Los valores del *RR* oscilan entre 0 y +∞, siendo el valor 1 el efecto nulo. Por lo tanto, no es posible obtener valores de *RR* negativos, siendo todos sus valores positivos.

En cuanto a la interpretación de los valores del *RR*, los valores comprendidos entre 0 y 1 indican un menor riesgo en el grupo representado en el numerador, mientras que los valores > 1 (hasta el infinito) indican un mayor riesgo para este mismo grupo (Kline, 2013).

Como la *RR* oscila entre 0 y +∞, hace que la distribución sea asimétrica excepto con muestras de tamaño muy grande, pero la trasformación logarítmica de la *RR* sí se aproxima a la distribución normal. Por ello, para la estimación del intervalo de confianza para la *RR* se trabaja con su transformación logarítmica natural (Borenstein y cols., 2009; Botella y Sánchez-Meca, 2015; Kline, 2013). El lector interesado en la estimación del intervalo de confianza puede consultar a Botella y Sánchez-Meca (2015),

Borenstein y cols. (2009), Grissom y Kim (2012) y Kline (2013), los cuales proporcionan las ecuaciones matemáticas necesarias para realizar dicho cómputo.

(3) *Odds Ratio* (*OR*) (o razón de ventajas): Es un índice no paramétrico (Frías-Navarro, 2011b). El parámetro para la *OR* es $\omega = \Omega_1/\Omega_2$.

Como en el caso anterior, los valores de la *OR* oscilan entre 0 y +∞, siendo el valor 1 el efecto nulo. Por lo tanto, no es posible obtener valores de OR negativos, siendo todos sus valores positivos (Botella y Sánchez-Meca, 2015).

Se interpreta del mismo modo que el RR, por lo tanto, los valores comprendidos entre 0 y 1 indican un menor riesgo en el grupo representado en el numerador de la ecuación, mientras que los valores > 1 (hasta el infinito) indican un mayor riesgo para este mismo grupo (Kline, 2013).

Por otra parte, la *OR* tiende a los valores extremos cuando una de las celdas de la tabla de contingencia es cero. En estos casos, es decir, cuando la frecuencia en alguna celda de la tabla de contingencia sea 0, se recomienda ajustar la *OR* añadiendo una constante a cada una de las celdas para mejorar la *OR* como estimador de la *OR* paramétrica. El valor de la constante puede oscilar entre 10^{-8} y .5, si bien, se recomienda ajustar la *OR* por .5 (Grissom y Kim, 2012).

Al igual que antes se comentó para la *RR*, como la *OR* oscila entre 0 y +∞, hace que la distribución sea asimétrica excepto con muestras de tamaño muy grande, pero la trasformación logarítmica de la *OR* sí se aproxima a la distribución normal. Por ello, para la estimación del intervalo de confianza para la *OR* se trabaja con su transformación logarítmica natural. Una vez calculado el intervalo de confianza, los límites del intervalo de confianza se devuelven a su métrica original mediante la fórmula inversa (Botella y Sánchez-Meca, 2015; Borenstein y cols., 2009; Kline, 2013). El lector interesado en la estimación del intervalo de confianza puede consultar a Botella y Sánchez-Meca (2015), Borenstein y cols. (2009), Grissom y Kim (2012) y Kline (2013), los cuales proporcionan las ecuaciones matemáticas necesarias para realizar dicho cómputo.

(4) Numero Necesario a Tratar (NNT): El *NNT* mide el número de pacientes que, en promedio, tendrían que ser tratados con un tratamiento determinado para demostrar una ganancia adicional (o un fracaso menos) sobre un tratamiento estándar o placebo con el

que se compara (Furukawa y Leucht, 2011; Grissom y Kim, 2012; Manríquez, Villouta, y Williams, 2007).

El *NNT* es una medida absoluta y se calcula como la inversa de la reducción del riesgo (Ferguson, 2009; Grissom y Kim, 2012). Los valores teóricos entre los que puede oscilar el índice NNT son – ∞ y + ∞ (Grissom y Kim, 2012). El *NNT* ideal es 1, ya que implica que sólo se necesita tratar a un paciente para recibir un beneficio adicional. Cuanto mayor sea el *NNT*, menos eficaz es la intervención.

De acuerdo a Grissom y Kim (2012), una aproximación a la estimación del intervalo de confianza para el *NNT* es primero construir el intervalo de confianza para la diferencia de riesgo. Y después, estimar los recíprocos de los límites del intervalo de confianza de la diferencia de riesgo, lo que nos proporcionará los límites del intervalo de confianza para la *NNT*.

En cuanto a su interpretación, por ejemplo, si el 80% de las personas responden a un tratamiento experimental y el 30% responden a un tratamiento placebo, la diferencia de riesgo (diferencia entre estas dos tasas de respuestas a los tratamientos) sería de 50% o 0.5. El *NNT* es la inversa 1/0.5 = 2, lo que significa que dos personas necesitarían recibir tratamiento para que haya una respuesta más positiva en el tratamiento experimental que en el tratamiento placebo.

Otro ejemplo citado en Erceg-Hurn y Mirosevich (2008): imaginemos un ensayo controlado aleatorio en el que la terapia cognitiva-conductual se compara con una terapia psicoeducativa para el tratamiento de la depresión. El éxito en este supuesto se define como la remisión de la depresión en el postratamiento. En este caso, un NNT de 3 indicaría que es necesario tratar a tres pacientes con terapia cognitiva-conductual, para tener un paciente más en remisión que si el mismo número de pacientes fueran tratados con psicoeducación.

Otro ejemplo citado en Furukawa y Leucht (2011): si la tasa de respuesta en la fase aguda del tratamiento de un episodio depresivo mayor es del 60% en el grupo activo de drogas y del 30% en el grupo placebo, el NNT se calcula como 1/(0.6-0.3) = 33, lo que significa que de cada 33 pacientes tratados con el tratamiento experimental se evitaría 1 recurrencia de depresión mayor. Dicho de otro modo, se necesitan tratan a 33 personas para tener una ganancia sobre el grupo control.

El *NNT* es un índice razonablemente intuitivo, y fácilmente interpretable, para evaluar los resultados de los ensayos clínicos (Ferguson, 2009; McGough y Faraone, 2009). El *NNT* facilita la toma de decisiones clínicas en relación con la eficacia de un tratamiento, dado que permite, de una forma sencilla, comparar los beneficios para diferentes terapias. Por ello, el *NNT* ha sido ampliamente recomendado como una medida de la eficacia (Manríquez y cols., 2007).

El *NNT* es el índice del tamaño del efecto del tratamiento preferido en la medicina basada en la evidencia, mientras que el índice más común en la literatura médica cuando la variable de resultado es continua es la *d* de Cohen, el cual, como se ha dicho, expresa la magnitud de la diferencia entre dos grupos en unidades de desviación típica. Existen distintos métodos para transformar los valores de la *d* de Cohen en *NNT*, por ejemplo, el método de Furukawa (Furukawa, 1999) y el método de Kraemer (Kraemer y Kupfer, 2006). El método de Furukawa permite una predicción más precisa del *NNT* (Furukawa y Leucht, 2011).

Finalmente, Grissom y Kim (2012) señalan que los índices del tamaño del efecto tales como *NNT* y la diferencia de riesgo (los cuales se han expuesto en el apartado de índices del tamaño del efecto de la familia de riesgo) que se aplican en las tablas de contingencia 2x2 también pueden extenderse a las tablas de contingencia 2xc con c (columnas) categorías de resultados ordinales.

2.3.5. Índices de la "familia de asociación"

Cuantifican la fuerza de la asociación entre dos variables categóricas (Agresti y Finaly, 2009). Generalmente los datos se organizan en tablas de contingencia. Ejemplos de estos índices son: (1) Coeficiente de contingencia C de Pearson, (2) *V* de Cramer, (3) coeficiente *phi*, y (4) Lambda de Goodman y de Kruskal.

(1) **Coeficiente de contingencia *C* de Pearson**: es una versión ajustada del coeficiente *phi* que se usa para tests con más de un grado de libertad (Ellis, 2010). Generalmente, se utiliza para representar la fuerza de asociación en los contrastes de hipótesis ejecutados mediante el test de Chi-cuadrado en las tablas de contingencia más grandes de 2x2.

(2) ***V* de Cramer:** se puede utilizar para medir la fuerza de la asociación para tablas de contingencia de cualquier tamaño y se considera superior al coeficiente de contingencia *C* de Pearson (Ellis, 2010; Grissom y Kim, 2012). Generalmente, se utiliza para

representar la fuerza de asociación en los contrastes de hipótesis ejecutados mediante el test de Chi-cuadrado (Fritz y cols., 2012; Grissom, y Kim, 2012; Kline, 2013).

Sus valores oscilan entre 0 y 1 (Grissom y Kim, 2012; Kline, 2013), donde 0 equivale a independencia entre las variables y 1 es la máxima asociación entre las variables. De acuerdo a Kline (2013), la *V* de Cramer no es un coeficiente de correlación.

El intervalo de confianza para la *V* de Cramer se construye a partir de la distribución no central de X^2 (Grissom y Kim, 2012), lo cual requiere de programas estadísticos para su estimación.

(3) **Coeficiente *phi* (φ):** se utiliza para representar la fuerza de asociación cuando ambas variables son dicotómicas (Grissom y Kim, 2012) o han sido dicotomizadas (Sánchez-Meca y cols., 2003), por lo que los datos se organizan en tablas de contingencia 2x2.

Grissom y Kim (2012) señalan que el coeficiente *phi* es un caso limitado de *V* de Cramer, en tablas de 2x2. Por lo tanto, los valores del coeficiente *phi*, al igual que la *V* de Cramer, deberían oscilar entre 0 y 1, donde 0 equivale a independencia entre las variables y 1 es la máxima asociación entre las variables. Y, por el mismo razonamiento, dado que la *V* de Cramer no es un estadístico de correlación, el coeficiente *phi* tampoco debería ser considerado como tal, en tanto que es un supuesto limitado de la *V* de Cramer. Sin embargo, Grissom y Kim (*op. cit*.) consideran al coeficiente *phi* como una índice de correlación de Pearson aplicado a variables dicotómicas (p. 247) cuyos valores oscilan teóricamente entre -1 y +1, si bien, reconocen que debido a que la ecuación para estimar *phi* a partir del test de Chi cuadrado contempla una raíz cuadrada "*puede que no esté claro si phi es positivo o negativo. Sin embargo, el signo de phi es un resultado trivial del orden en que están dispuestas las dos columnas o las dos filas*" (p. 249).

Igual que en la *V* de Cramer, el intervalo de confianza para el coeficiente *phi* se construye a partir de la distribución no central de X^2 (Grissom y Kim, 2012), por lo que se requiere de programas estadísticos para su estimación.

(4) **Lambda de Goodman y de Kruskal (λ):** se utiliza cuando ambos X e Y se miden en escalas nominales y mide el porcentaje de mejora en la predicción del valor de la variable dependiente dado el valor de la variable independiente (Ellis, 2010; Fritz y

cols., 2012). De acuerdo a Fritz y cols. (2012) se utilizan cuando en la tabla de contingencia 2x2 las filas y columnas representan a una variable predictora y una variable predicha.

En general, como los índices de asociación expuestos, no son coeficientes de correlación, por lo que su cuadrado (V^2, phi^2, λ^2) no es un estimador de la proporción de varianza explicada (Fritz y cols., 2012; Kline, 2013).

2.3.6. Índices de la familia de la "probabilidad de superioridad o dominancia"

La diferencia de medias y la diferencia de medias estandarizadas no son los únicos estadísticos del tamaño del efecto que reflejan una diferencia entre dos grupos con respecto a una variable de resultado o dependiente. Existen otros estadísticos, menos conocidos, pero más intuitivos que estiman la diferencia entre dos poblaciones en términos de grado de solapamiento entre distribuciones de los dos grupos.

Dentro de estos índices se encuentran: (1) Lenguaje Común de McGraw y Wong (*Common Language, CL*), (2) probabilidad de superioridad de Grissom y Kim (*Probability of Superiority*, *PS*), (3) Probabilidad estocástica de Vargha y Delaney (*stochastic superiority, Â*), (4) *d* de Cliff, y (5) Área bajo la curva de ROC (AUC o ROC).

El *CL*, *PS* y *AUC* estiman la misma probabilidad de superioridad de la población, siendo el *CL* un estadístico paramétrico y el *PS* un estadístico no paramétrico. El estadístico AUC es un caso especial debido a que algunos autores utilizan esta terminología para referirse al *CL* y otros autores para referirse al *PS* (Grissom y Kim, 2012).

El estadístico $\hat{A}$ y *d* de Cliff son índices no paramétricos que estiman diferentes probabilidades de superioridad. Concretamente, $\hat{A}$ estima la superioridad teniendo en cuenta la probabilidad de las colas de los datos, mientras que la *d* de Cliff estima la superioridad teniendo en cuenta la superioridad en ambas direcciones (Pen y Cheng, 2014).

A continuación se expondrán cada uno de estos índices.

(1) Lenguaje Común (*Common Language, CL*): el estadístico *CL* fue propuesto por McGraw y Wong (1992) para un diseño de dos grupos independientes con una variable dependiente continua para estimar el parámetro $\Delta = Pr\ (Y_1 > Y_2)$ o la "*probabilidad de que una puntuación seleccionada al azar de una población será mayor que otra puntuación seleccionada al azar de otra población*" (*op. cit.*, p. 361).

Dado que el estadístico *CL* es paramétrico, requiere el cumplimento de los supuestos de normalidad e igualdad de varianzas (Li, 2015; McGraw y Wong, 1992; Peng y Chen, 2014; Ruscio, 2008a; Vargha y Delaney, 2000). Sin embargo, bajo condiciones de no normalidad, el rendimiento *CL* es adecuado; tan solo se deteriora bajo la violación de ambos supuestos, esto es, del supuesto de normalidad y de igualdad de la varianza (McGraw y Wong, 1992).

De acuerdo a Lakens (2013), matemáticamente el *CL* es "*la probabilidad de que una puntuación z sea mayor que el valor que corresponde a una diferencia entre los grupos de 0 en una curva de distribución normal"* (p. 4). Por lo tanto, la estimación de Δ a partir del estadístico *CL* se puede obtener a partir de un ecuación matemática basada en las puntuaciones *z* (ver Frías-Navarro, 2011b; Grissom y Kim, 2012; McGraw y Wong, 1992).

Una vez obtenido el valor de *z* se puede buscar en las tablas de la distribución normal tipificada la probabilidad asignada a dicho valor. Y esta probabilidad corresponde a *CL*. Existen programas estadísticos para estimar el *CL*, por ejemplo, Dunlop (1999) desarrolló un programa estadístico para calcular el *CL* y realizar transformaciones desde el *CL* a *d* de Cohen, al coeficiente de correlación y al valor *t* de Student.

Cuando se cumple el supuesto de normalidad y los tamaños de las muestras son iguales, los criterios para un tamaño del efecto pequeño, moderado y grande para *CL* son .56, .64, y .71, respectivamente, que equivalen a valores de 0.20, 0.50 y 0.80 en *d* de Cohen (Grissom, 1994; Li, 2015). Además el efecto nulo en CL es .50 (*CL* = .50) el cual equivale a una diferencia de medias en términos de *d* de Cohen igual a 0 (*d* de Cohen = 0).

(2) Probabilidad de superioridad (*Probability of Superiority, PS*): el índice de probabilidad de superioridad (*Probability of Superiority, PS*) es una extensión del estadístico *CL* paramétrico a las variables en escala ordinal. Por lo tanto, al igual que el

CL, el *PS* estima "*la probabilidad de que un sujeto seleccionado al azar de una población A tendrá una puntuación (Y_a) mayor que la puntuación (Y_b) alcanzada por otro sujeto seleccionado al azar de otra población B*" (Grissom y Kim, 2012, p. 149).

El *PS* es un estadístico no paramétrico, por lo que no requiere la asunción de los supuestos de normalidad de la variable dependiente ni homogeneidad de las varianzas en los grupos. A este respecto, el estudio de simulación realizado por Li (2015) ha mostrado su robustez ante las violaciones de estos supuestos.

Los valores de *PS* oscilan entre 0 y 1, puesto que *PS* es una probabilidad. Esto significa que los resultados más extremos cuando se comparan dos grupos son *PS* = 1 (donde todos los miembros de la población *a* tienen puntuaciones mayores que los miembros de la población *b*) y *PS* = 0 (donde todos los miembros de la población *a* son superados por los miembros de la población *b*). El valor del efecto nulo es *PS* = .5, donde los miembros de ambas poblaciones tienen puntuaciones iguales (Grissom y Kim, 2012). Por ejemplo, si se comparan los hombres y las mujeres en términos de peso, si *PS* = .60, significa que la probabilidad de que un hombre extraído al azar de la muestra de hombres tenga más peso que una mujer extraída al azar de la muestra de mujeres es 0.60.

Algunos índices del tamaño del efecto para tablas de contingencia se pueden transformar en *PS* (e.g., diferencia de riesgo, *NNT*). Además, es posible transformar los valores de *PS* en valores de *d* de Cohen, al igual que en el caso de *CL*, y su interpretación es similar a ésta. A este respecto, Grissom (1994) señala que cuando no existen diferencias entre los grupos en términos de *d* de Cohen (*d* = 0) la *PS* tiene un valor de .50; un tamaño del efecto pequeño (*d* = 0.20) se corresponde con un valor de *PS* de.56; un tamaño del efecto medio (*d* = 0.50) equivale a un valor de.64 de *PS*, y un tamaño del efecto grande (*d* = 0.80) se corresponde con un valor de .71 de *PS*.

(3) Probabilidad estocástica ($\hat{A}$): es un índice de superioridad propuesto por Vargha y Delaney (2000). La probabilidad estocástica hace referencia a la probabilidad de que una puntuación elegida al azar de un grupo 1 sea mayor a otra puntuación elegida al azar del grupo 2, más .5 veces la probabilidad de que una puntuación elegida al azar del grupo 1 sea igual a una puntuación elegida al azar del grupo 2. De acuerdo con Li (2015), $\hat{A}$ es considerado el estimador robusto de *CL*. De acuerdo a Vargha y Delaney (2000), $\hat{A}$ se aplica a cualquier variable que esté al menos en escala ordinal.

(4) *d* de Cliff (*d*): fue propuesto y desarrollado por Cliff (1993), también es conocido como "estadístico de dominancia" (Grissom y Kim, 2012). Es un índice no paramétrico, por lo que se aplica en variables con escala de medida ordinal, o en variables cuantitativas que no siguen una distribución normal.

La *d* de Cliff se puede calcular a partir del test *U* de Mann-Withney (Peng y Cheng, 2014). Sus valores oscilan entre -1 y +1, siendo el valor 0 el efecto nulo (o completo solapamiento entre las distribuciones de los dos grupos). Cuando d_{Cliff} = -1, todas las puntuación del grupo 1 están por debajo de todas las puntuaciones del grupo 2. Cuando d_{Cliff} = +1 todas las puntuaciones del grupo 1 están por encima de todas las puntuaciones del grupo 2. Cuando el valor de d_{Cliff} se sitúa entre -1 y +1, existe un grado de solapamiento entre las distribuciones del grupo 1 y 2.

(5) Área bajo la curva (*Area Under the Curve, AUC*): también conocido como el área bajo la característica del funcionamiento del receptor (*Receiver Operating Characteristic, ROC*)**.** De acuerdo con Frías-Navarro (2011b), el estadístico *AUC* o *ROC* es conceptual y matemáticamente similar al tamaño del efecto *CL* de McGraw y Wong (1992). Por lo tanto, el AUC es "*igual a la probabilidad de que una puntuación obtenida aleatoriamente de una población sea superior a la obtenida también de forma aleatoria de una segunda población*" (Frías-Navarro, 2011b, p. 155). Por ello, como Grissom y Kim (2012) indican, algunos autores hacen referencia al estadístico *CL* (en términos de McGraw y Wong, 1992) y *PS* (en términos de Grissom y Kim, 2012) como AUC o ROC.

De acuerdo a Grissom y Kim (2012), cualquier método adecuado para construir el intervalo de confianza para el estadístico *PS* y para el *CL* (se recuerda que en términos de estos autores, PS y CL son sinónimos), también es aplicable a la *d* de Cliff, puesto que la *d* de Cliff es una función lineal del *PS* (*d* = 2 *PS*-1).

De acuerdo con Ruscio y Mullen (2012), existen al menos siete procedimientos para estimar el intervalo de confianza de *CL* y $\hat{A}$, y por tanto, para el *PS* y *d* de Cliff. Entre ellos, el procedimiento para la estimación de intervalos de confianza de Cliff (1993) es considerado como uno de los métodos más populares (Li, 2015).

2.4. Software para la estimación de los tamaños del efecto y sus intervalos de confianza

La mayoría de los paquetes estadísticos estándares, como el SPSS o el SAS, no estiman de forma directa los índices del tamaño del efecto y sus intervalos de confianza, por lo que hay que hacer uso de macros, *scripts* o paquetes estadísticos específicos (Frías-Navarro, 2011b).

Algunos de estos paquetes estadísticos están disponibles como programas de software gratuito a través de Internet. Por ejemplo, el programa "*The Practical Meta-Analysis Effect Size Calculator*" desarrollado por David Wilson, disponible en la página web http://cebcp.org/practical-meta-analysis-effect-size-calculator/, el cual ha sido recientemente actualizado. Este programa puede ser utilizado para estimar una gran variedad de estadísticos del tamaño del efecto y sus intervalos de confianza al 95%, tales como los estadísticos de la familia de la diferencia estandarizada de medias, los de la familia del coeficiente de correlación, *odds ratio* y el riesgo relativo. También en Internet está disponible libremente un programa estadístico para estimar los tamaños del efecto robustos (estandarizados y no estandarizados) y sus intervalos de confianza propuesto por Algina y colaboradores desde la página web http://plaza.ufl.edu/algina/index.programs.html. También se pueden estimar los estadísticos del tamaño del efecto tales como *NNT* y sus intervalos de confianza en la página http://medcalc3000.com/BayesianAnalysis_1.htm y en la página web http://araw.mede.uic.edu/cgi-bin/nntcalc.pl. Por otro lado, desde la página web http://vassarstats.net/rho.html desarrollada por Richard Lowr se pueden estimar los intervalos de confianza para los coeficientes de correlación aplicando la transformación de la Z de Fisher.

Desde la página web http://www.thenewstatistics.com se puede descargar un software basado en Excel para el cálculo de los intervalos de confianza IC para la diferencia de medias estandarizadas elaborado por Cumming (2012). Asimismo, desde la página http://dl.dropbox.com/u/1857674/CIstuff/CI.html se pueden descargar *scripts* para SPSS, SAS, SPlus y R para calcular los intervalos de confianza para los tamaños del efecto asociados con los análisis de la prueba *t* de Student, ANOVA, análisis de regresión y Chi Cuadrado, desarrollados por Smithson (2003). Además esta página web facilita enlaces a otras páginas que pueden ser de utilidad. También se pueden descargar

scripts en la página web http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Programs.htm desarrollados por Wuensch.

También se puede descargar libremente el programa estadístico "*Effect Size Generator-Pro"* para la estimación de la diferencia de medias directa y estandarizada paramétricas y sus intervalos de confianza desde la página web del *Centre for Evaluation & Monitoring* http://www.cem.org/effect-size-calculator. Y, finalmente, hay disponibles en el mercado paquetes estadísticos comerciales que permiten calcular los tamaños del efecto y su intervalos de confianza como *Comprehensive Meta-analyis* (Borenstein, Hedges, Higgins, y Rothstein, 2014) y el DSTAT (Johnson, 1993).

2.5. Uso de los tamaños del efecto y sus intervalos de confianza en la literatura

Como se ha visto en el capítulo anterior, desde la década de 1990, los estadísticos habían sido conscientes de que las pruebas de la NHST eran, en muchos aspectos, insuficientes para interpretar los resultados de las investigaciones (e.g., Berkson, 1938; Cohen, 1994; Fidler y Loftus, 2009; Meehl, 1978). Posteriormente, el Grupo de Trabajo de Wilkinson (Wilkinson y TFSI, 1999) recomendaron informar sobre los tamaños del efecto y sus intervalos de confianza. No obstante, el uso de los índices del tamaño del efecto ha sido inconsistente durante muchos años (e.g., Fidler y cols., 2005), al menos hasta la publicación de la sexta edición del Manual de la APA (2010a), a partir del cual se ha observado un incremento en el uso de los estadísticos del tamaño del efecto, si bien este incremento no ha afectado al reporte de los intervalos de confianza de los tamaños del efecto.

Por ejemplo, Mathews y cols. (2008) analizaron 101 artículos publicados en cinco revistas de psicología (*Gifted and Talented International*, *Journal for the Education of the Gifted*, *Journal of Secondary Gifted Education, and Roeper Review*, y *High Ability Studies*) durante el período comprendido entre 1996 y 2005 y encontraron que, en general, el número de artículos que reportaron en los resultados algún índice del tamaño del efecto aumentó un 25.6% desde 1996 a 2000 y un 45.9% entre 2001 y 2005. A pesar de este incremento, la tasa de reporte del tamaño del efecto se mantuvo por debajo del 60%. Los índices del tamaño del efecto más reportados fueron los índices de la familia de la fuerza de las relaciones (r o R^2). En este sentido, de los estudios que reportaron índices del tamaño del efecto, aproximadamente el 65% informó sobre

medidas de correlación, mientras que el 35% restante de los estudios utilizó alguna medida de los índices de diferencias de medias estandarizadas. Finalmente, estos autores no registraron el uso de los intervalos de confianza de los tamaños del efecto.

Fidler y cols. (2005) analizaron 239 artículos publicados en el *Journal of Consulting and Clinical Psychology* desde el año 1993 hasta el año 2001, y encontraron que el reporte del tamaño del efecto se incrementó desde el 20% en el año 1993 hasta el 46% en el año 2001, si bien el informe de los intervalos de confianza solo alcanzó el 17% en 2001.

Odgaard y Fowler (2010) también revisaron los estudios de intervención publicados en los años 2003, 2004, 2007 y 2008 en el *Journal of Consulting and Clinical Psychology*, que fue la primera revista de la APA en requerir el reporte de los tamaños del efecto y sus intervalos de confianza (La Greca, 2005), y encontraron que en general el 75% de los estudios informaron de algún índice del tamaño del efecto, pero sólo en el 40% de los casos el reporte del tamaño del efecto iba acompañado por su intervalo de confianza. Por años, en 2003 el 69% de los estudios reportaron algún estadístico del tamaño del efecto, en 2004 el 64%, en 2007 el 74%, y en 2008 el 94%. Se observa pues un incremento en el reporte de los estadísticos del tamaño del efecto en el transcurso de los años, de 69% en el año 2003 hasta un 94% en el año 2008, si bien no sucedió lo mismo con los intervalos de confianza.

Sun y cols. (2010) analizaron los artículos publicados entre los años 2005 y 2007 en cinco revistas de Psicología (*Journal of Educational Psychology, Journal of Experimental Psychology: Applied, Journal of Experimental: Psychology Human Perception and Performance, Journal of Experimental Psychology: Learning, Memory & Cognition,* y *School Psychology Quarterly*) y encontraron que solo el 40% de los mismos reportaron algún índice del tamaño del efecto (estos autores no registraron el uso de los intervalos de confianza de los tamaños del efecto) De éstos, el 57% también llevaron a cabo su interpretación, si bien, la misma estuvo basada en la guía de Cohen (1988) del tamaño del efecto pequeño, mediano y grande. Los estadísticos del tamaño del efecto más frecuentemente reportados fueron los índices de la familia de la fuerza de la asociación o relación (62%) frente a las medidas de diferencia de medias (26%) y otros estadísticos (12%). Los autores piensan que esto puede ser debido al "*hecho de que el 76% de los 1,243 artículos (n = 938) usaron el modelo lineal general como el*

método principal para la prueba de NHST, y el 75% de los 610 artículos que informaron del tamaño del efecto (n=455) usaron el modelo lineal general" (p. 7).

Por otro lado, McMillan y Foley (2011) analizaron 417 artículos publicados entre los años 2008 y 2010 en cuatro revistas especializadas de educación y psicología (*Journal of Educational Psychology, Journal of Experimental Education, Journal of Educational Research,* y *Contemporary Educational Psychology*) y encontraron que el 74% de los estudios informaron alguna medida del tamaño del efecto. Los estadísticos del tamaño del efecto más reportados fueron la *d* de Cohen y la η^2, con un menor reporte de la *g* de Hedges, *odds ratio*, *f* de Cohen, y ω^2. Además, el reporte de la *d* de Cohen fue seguido de su interpretación basada en la guía que Cohen (1988) facilitó para la interpretación de los resultados (0.2 = "efecto pequeño", 0.5 = "efecto medio", y 0.8 o más alto = "efecto grande"). Estos autores concluyeron que, si bien se había incrementado el uso de los índices del tamaño del efecto en los informes de investigación, los debates sobre el significado sustantivo del tamaño del efecto seguían siendo deficientes. Estos autores tampoco registraron el uso de los intervalos de confianza de los tamaños del efecto.

Por su parte, Sesé y Palmer (2012) analizaron el uso de estadísticos en los artículos publicados en el año 2010 en ocho revistas de Psicología Clínica y de la Salud con índice de impacto ISI (*Journal of Behavioural Medicine, Behaviour, Research and Therapy, Depression and Anxiety, Behavior Therapy, Journal of Anxiety Disorders, International Journal of Clinical and Health Psychology, British Journal of Clinical Psychology,* y *British Journal of Health Psychology*), entre las que se encuentra la revista española *International Journal of Clinical and Health Psychology*. Estos autores encontraron que los índices del tamaño del efecto fueron reportados en el 61.04% de los estudios. Los estadísticos más usados fueron la R^2 seguido de la η^2 y *d* de Cohen. Sin embargo, sólo el 18.87% de los artículos informaron sobre los intervalos de confianza.

Fritz y cols. (2012) revisaron 71 artículos publicados en el *Journal of Experimental Psychology General* entre los años 2009 y 2010 y encontraron que en general el 58% de los estudios reportaron en los resultados algún índice del tamaño del efecto. Estos autores no registraron el uso de los intervalos de confianza de los tamaños del efecto. En cuanto a los tamaños del efecto reportados, observaron que había una dependencia de la técnica de análisis estadístico utilizada. Es decir, cuando el estudio realizaba el análisis de datos a través de un análisis de la varianza (ANOVA) o

covarianza (ANCOVA), solo el 56% de los estudios reportaron índices del tamaño del efecto, en este caso, η_p^2 fue la más comúnmente informada. Cuando el estudio utilizaba la prueba de la *t* de Student para el análisis de datos, solo el 30% de los estudios informaron de estadísticos del tamaño del efecto, en este supuesto, la *d* de Cohen fue la más utilizada. Cuando se ejecutaron análisis de regresión, el 35.3% de los estudios reportaron algún índice del tamaño del efecto, en concreto, todos los estudios informaron sobre la R^2. Sin embargo, Peng y cols. (2013) encontraron una mayor frecuencia de reporte de estadísticos del tamaño del efecto durante este mismo periodo. En concreto, estos autores analizaron 451 artículos publicados entre 2009 y 2010 en doce revistas, dos de ellas adscritas a la normativa de la *American Educational Research Association* (AERA), otras dos adscritas a la normativa de la APA (*Journal of Counselling Psychology* y *Journal of Educational Psychology*), y ocho adscritas a organizaciones profesionales. Respecto de las revistas de Psicología, estos autores encontraron que el 72.7% de los artículos publicados informaron algún índice del tamaño del efecto. Finalmente, estos autores tampoco registraron el uso de los intervalos de confianza de los tamaños del efecto.

En España, tras más de tres décadas en las que se ha recomendado informar sobre alguna medida del tamaño del efecto y sus intervalos de confianza junto a los resultados de las prueba de la NHST (e.g., Cohen, 1988; APA, 1996, 2001, 2010a; Wilkinson y TFSI, 1999), el seguimiento de tales recomendaciones ha sido más bien escaso. Por ejemplo, recientemente, Caperos y Pardo (2013) analizaron los artículos publicados en 2011 en cuatro revistas españolas de Psicología multidisciplinar (*Anales de Psicología, Psicológica, Psicothema,* y *Spanish Journal of Psychology*) indexadas en la base de datos JCR. Sus resultados indican que sólo el 24.3% de las pruebas NHST ejecutadas se acompañaron de un estadístico del tamaño del efecto (en la misma línea, Badenes-Ribera, Frías-Navarro, Monterde-i-Bort y Pascual Soler, 2013, en su revisión de los artículos publicados en la revista *Psicothema* y *Revista de Educación* durante el año 2011). Por otra parte, Badenes-Ribera y cols. (2013) encontraron que los índices del tamaño del efecto más reportados fueron el R^2, *d* de Cohen, y η^2. Finalmente, estos autores encontraron que sólo el 9.5% de los índices del tamaño del efecto fueron acompañados por sus intervalos de confianza (en el mismo sentido, Frías-Navarro, Monterde-i-Bort, Pascual-Soler, Pascual-Llobell, y Badenes-Ribera, 2012).

En conjunto, de los estudios analizados se observa que los índices del tamaño del efecto más frecuentemente reportados son el R^2, *d* de Cohen, y η^2 (Badenes-Ribera y cols., 2013; Peng y cols., 2013; Sesé y Palmer, 2012; Sun y cols. 2010). Sin embargo, como se ha comentado, estos estadísticos han sido ampliamente criticados por su sesgo (es decir, tienden a estar positivamente sesgados), su falta de robustez a los valores atípicos, y su inestabilidad bajo las violaciones de los supuestos estadísticos (Fritz y cols, 2012; Grissom y Kim, 2012; Kline, 2013; Thompson, 2002b; Wang y Thompson, 2007).

2.6. Replicación

La replicación, entendida como la confirmación de los resultados y conclusiones obtenidos en un estudio a partir de otro estudio realizado de forma independiente, se considera el estándar de oro científico (Jasny, Chin, Chong y Vignieri, 2011) y la piedra angular de la ciencia acumulativa (Asendorpf y cols., 2013; Cumming, 2012; Johnson, 1999; Kline, 2013).

De acuerdo con Kehle y cols. Kawano (2007), "*La ciencia se basa en la replicación y extensión, lo que permite la acumulación y la evolución del conocimiento y su aplicación*" (p. 419). Del mismo modo, Thompson (2006) observó que "*la ciencia es el negocio de descubrir las leyes (relaciones) acerca de los efectos que se producen (y se vuelvan a producir) bajo las condiciones establecidas*" (p. 252).

Los estudios de replicación permiten separar las conclusiones que son dignas de confianza de los resultados que son poco fiables (Koole y Lakens, 2012). En palabras de Shaver (1993) "*la cuestión de interés es si un tamaño del efecto de una magnitud considerada importante se ha obtenido consistentemente a través de válidas replicaciones. Si alguno o todos los resultados son estadísticamente significativos es irrelevante*" (p. 304). Por ello, las replicaciones son útiles para la prevención de la invalidez de la conclusión estadística (Error de Tipo I o falso positivo y Error de Tipo II o falso negativo). Es decir, la replicación es necesaria para la validación de los resultados positivos y para invalidar los falsos positivos (Jasny y cols., 2011). Por lo tanto, si los resultados de un estudio no pueden ser replicados, no tienen ninguna credibilidad (Cohen, 1994; Francis, 2012b; Roediger, 2012). Así pues, para tener carácter general, los resultados deben ser consistentes en una amplia variedad de circunstancias (Johnson, 1999). En consecuencia, los estudios de replicación son

esenciales para el desarrollo teórico a través de la confirmación y desconfirmación de los resultados (Brandt y cols., 2014). Sin replicación de los resultados, las teorías científicas y las leyes no tienen ninguna base (Harrison y cols., 2009).

Uno de los puntos más débiles en la investigación en Psicología ha sido durante muchos años la falta de énfasis en la replicación de los resultados (Henson, 2006; Koole y Lakens, 2012; Pashler y Wagenmakers, 2012; Schmidt, 2009). Por ejemplo, Makel, Plucker y Hegarty (2012) analizaron la prevalencia de los estudios de replicación en 100 revistas de alto impacto en el campo de la Psicología y encontraron que aproximadamente solo el 1.6% de todas las publicaciones utilizaron el término replicación en el texto. Sin embargo, un análisis más a fondo de 500 artículos seleccionados al azar de entre los artículos que utilizaron el término replicación reveló que sólo el 68% de los mismos fueron réplicas reales.

Esta falta de énfasis en la replicación de los resultados está relacionada con una excesiva confianza en los valores *p*. Por ejemplo, Stangor y Lemay (2016) señalan que solamente el 18% de los estudios con un valor de *p* mayor de .04 fueron replicados mientras que el 63% de los estudios que tenían un valor de *p* menor a .001 sí fueron replicados. Como Kline (2013) apunta las falacias sobre el valor *p* pueden hacer creer a los investigadores que no es necesario replicar los estudios. Del mismo modo, los editores de las revistas pueden albergar alguna de estas falsas creencias acerca del valor *p* que dificulte la publicación de los estudios de replicación. De hecho, los editores prefieren investigaciones originales con resultados estadísticamente positivos a estudios de replicación (Kline, 2013), si bien la tasa de estudios de replicación publicado se ha incrementado en las recientes décadas (Makel y cols., 2012). En consecuencia, los valores *p* y el sesgo de publicación han ido dirigiendo los resultados de la ciencia, fabricando un mundo de creencias y fantasías y perjudicando el avance acumulativo y válido del conocimiento científico.

Sin embargo, recientemente la ciencia psicológica ha empezado a reconsiderar la importancia de los estudios de replicación como la base de la construcción de un conocimiento científico acumulativo (Brandt y cols., 2014). Prueba de ello son los números especiales de revistas prestigiosas dedicados a los estudios de replicación como *Perspectives on Psychological Science* del año 2012 y *Journal of Experimental Social Psychology* del año 2016, y el hecho de que prestigiosas revistas del ámbito de la Psicología como *Journal of Experimental Social Psychology*, *Journal of Personality*

and Social Psychology, o *Psychological Science* han empezado a publicar estudios de replicación, e incluso a crear una sección para los estudios de replicación (Brandt y cols., 2014).

Los investigadores han propuesto muchas soluciones al problema de la actual crisis de la replicación. Por ejemplo, dejar disponibles en internet los métodos y datos utilizados en los estudios originales para consulta de los investigadores, hacer un llamamiento para llevar a cabo estudios de replicación, realizar estudios exploratorios que posteriormente sean confirmados con estudios de replicación con muestras de gran tamaño, etc. (Asendorf y cols., 2013; Sakaluk, 2016). En este sentido, se han dispuesto varias plataformas para registrar los protocolos de investigación tanto de estudios individuales como de trabajos de meta-análisis con el propósito de controlar decisiones a posteriori del investigador o prácticas cuestionables de investigación. El objetivo principal de todas estas actuaciones es aumentar la credibilidad de la ciencia y superar la crisis de confianza en la fiabilidad de los datos publicados (Hales, 2016).

2.7. Conclusión

Desde los inicios de la aplicación de la prueba de la NHST, se han sucedido los debates sobre la inadecuación de la misma para interpretar los resultados. En parte debido a estas críticas, la APA desde 1999 recomendó completar la información proporcionada por las pruebas de inferencia estadística (valor *p*) con algún índice del tamaño del efecto y su intervalo de confianza.

La presencia de los tamaños del efecto y su intervalo de confianza es una buena práctica estadística porque permite conocer la magnitud del efecto detectado, la precisión de la estimación puntual y tomar la decisión de mantener o rechazar la hipótesis nula. Utilizar una métrica estandarizada de los tamaños del efecto permite que los investigadores comparen las magnitudes de diferentes estudios, aportando conclusiones más sustantivas que las basadas únicamente en el valor p de probabilidad. Además, los tamaños del efecto facilitan la planificación de los nuevos estudios aportando el valor de un tamaño del efecto conocido y potencian el desarrollo del pensamiento meta-analítico mucho más integrador de los hallazgos de la literatura que la toma de decisión dicotómica propia del procedimiento clásico de comprobación de la hipótesis nula (NHST). Y, por supuesto, facilitan las tareas propias de los estudios de meta-análisis.

Sin embargo, como se ha visto, el uso de los índices del tamaño del efecto y sus intervalos de confianza ha sido inconsistente durante muchos años, si bien, a partir de la sexta edición del Manual de la APA (2010) se ha producido un incremento en el uso de los estadísticos del tamaño del efecto, pero no así de sus intervalos de confianza.

Por otra parte, existen docenas de índices del tamaño del efecto, como se ha expuesto en este capítulo. Sin embargo, los estudios publicados sobre el análisis de los usos de los estadísticos del tamaño del efecto sugieren que los índices más conocidos, por ser los más utilizados, son los índices como la *d* de Cohen, eta cuadrado y R^2. Sin embargo, se sabe que estos índices están positivamente sesgados en muestras de tamaño pequeño, y además, no son resistentes a las violaciones de los supuestos de las pruebas paramétricas como son la asunción de normalidad y homogeneidad de varianza, así como a la presencia de valores extremos (*outliers*). Existen otros índices del tamaño del efecto robustos frente a tales violaciones, que puede que sean menos conocidos, algunos de los cuales han sido expuestos en este capítulo.

Así las cosas es preciso estudiar el uso y conocimiento que los profesionales de la Psicología tienen y hacen de los estadísticos del tamaño del efecto, pues la utilización de estadísticos inadecuados puede entorpecer la acumulación de un conocimiento científico valido y, por otro lado, la aplicación de una práctica profesional basada en la evidencia.

3. MÁS ALLÁ DE LA PRUEBA NHST (II): META-ANÁLISIS

Como se ha comentado en el capítulo anterior, para evitar los problemas planteados por la prueba de la NHST respecto de su ineficacia para alcanzar el objetivo de toda ciencia como es la acumulación del conocimiento científico válido, se han propuesto diferentes métodos alternativos de análisis de datos. Entre ellos se ha abogado por el uso de revisiones sistemáticas de tipo cuantitativo o meta-análisis (APA, 2010a; Borenstein y cols., 2009; Bustamante y Delgado, 1994; Cooper, 1989; Cumming, 2012; Hedges y Olkin, 1985; Kline, 2013; Schmidt, 1996).

3.1. Revisiones sistemáticas: el Meta-análisis

De acuerdo con Sánchez-Meca, Marín-Martínez y López-López (2011, p. 96), "*Una revisión sistemática es una revisión objetiva de una pregunta formulada de forma clara para cuya respuesta es preciso integrar los estudios empíricos que se han llevado a cabo sobre ella*". Su objetivo es acumular de forma sistemática y objetiva las evidencias obtenidas en los estudios empíricos sobre un determinado problema (Huedo-Medina y Johnson, 2010). Por tanto, las revisiones sistemáticas son estudios secundarios que sintetizan la información científica disponible a través de métodos rigurosos y explícitos, lo que permite ahorrar tiempo a los profesionales de la salud y les ofrece una panorámica de lo que las evidencias científicas dicen respecto de ese problema (Sánchez-Meca y Botella, 2010).

Existen dos tipos de revisiones sistemáticas (Botella y Sánchez-Meca, 2015; Frías-Navarro y Monterde-i-Bort, 2014; Harris, Quatman, Manring, Siston, y Flanigan, 2014, Littell, Corcoran y Pillai, 2008; Sánchez-Meca y Botella, 2010; Sánchez-Meca y cols., 2011):

(1) ***Revisiones sistemáticas de tipo cualitativo*:** resumen o sintetizan de forma descriptiva (narrativa) la evidencia científica sobre una temática.
(2) ***Revisiones sistemáticas de tipo cuantitativo o meta-análisis***: revisiones en las que se utilizan métodos estadísticos para integrar cuantitativamente los resultados de los estudios empíricos primarios.

Las revisiones sistemáticas en general, y los trabajos de meta-análisis en particular, se consideran actualmente como las mejores herramientas para sintetizar las pruebas científicas respecto a qué tratamientos, intervenciones, programas de prevención y técnicas de diagnóstico deberían aplicarse para un determinado problema psicológico (Botella y Sánchez-Meca, 2015; Frías-Navarro y Pascual-Llobell, 2003; Harris y cols., 2014; Sánchez-Meca, 2010; Sánchez-Meca y Botella, 2010; Urra-Medina y Barría-Pailaquilén, 2010), sobre todo los trabajos de meta-análisis basados en estudios con buena calidad metodológica o estudios experimentales (Botella y Sánchez,-Meca, 2015; Mulrow y Cook, 1998; Sánchez-Meca, Boruch, Petrosino y Rosa-Alcázar, 2002). En este sentido, los estudios de meta-análisis permiten determinar qué programas son más efectivos, bajo qué condiciones y para qué tipos de personas (Sánchez-Meca y cols., 2011).

3.1.1. Meta-análisis: definición y principales características

El meta-análisis es una herramienta metodológica que permite integrar o sintetizar de forma cuantitativa los resultados obtenidos a partir de un conjunto de investigaciones realizadas sobre una temática concreta. De este modo, los resultados de cada estudio individual, expresados en términos de tamaño del efecto, son combinados para obtener conclusiones más generales (Botella y Sánchez-Meca, 2015; Catalá-López, y Tobías, 2014; Catalá-López, Tobías, y Roqué, 2014; Frías-Navarro, 2011b; Sánchez-Meca, Marín-Martínez y López-López, 2013).

Además, los estudios de meta-análisis permiten considerar aquellas variables que difieren entre los estudios primarios incluidos y que pueden contribuir a explicar la diferencia entre los resultados obtenidos en cada investigación particular (Borenstein y cols., 2009; Botella y Sánchez-Meca, 2015; Cooper, 1989; Glass, 1976; Glass, McGraw y Smith, 1981).

Los estudios de meta-análisis tienen entre sus características la objetividad, la replicabilidad y la precisión (Botella y Sánchez-Meca, 2015). La *objetividad* se refiere a que las decisiones que se toman y las operaciones que se realizan en el estudio meta-analítico están especificadas y dejan escaso margen a la discrecionalidad de quien lo realiza. La *replicabilidad* tiene que ver en el hecho de que todas las decisiones tomadas y operaciones ejecutadas quedan expresadas en el informe de investigación, lo que favorece que otros investigadores puedan llevar a cabo el mismo estudio y comprobar si

se dan o no los mismos resultados. Y, finalmente, la *precisión* hace referencia al formato de las respuestas, ya que éstas no sólo se expresan en palabras, sino que se reportan también con números que representan magnitudes.

Por ello, el meta-análisis favorece la acumulación del conocimiento científico en un determinado campo de investigación de una forma objetiva, sistemática y rigurosa, siempre que se utilice con buen juicio y siendo consciente de sus limitaciones (Cumming, 2014, 2012; Cumming y cols., 2012; Ellis, 2010; Huedo-Medina y Johnson, 2010; Kline, 2013; Marín-Martínez, Sánchez-Meca, y López-López, 2009; Sánchez-Meca, 2008; Sánchez-Meca y Botella, 2015; Urra-Medina y Barría-Pailaquilén 2010).

3.2. Revisiones sistemáticas versus revisiones narrativas

A diferencia de una revisión narrativa (también llamada revisión tradicional, cualitativa o subjetiva), una revisión sistemática implica un proceso de investigación científico basado en el rigor y la transparencia de cada una de las decisiones adoptadas durante la elaboración de dicha revisión tanto si es una revisión sistemática cualitativa como si se trata de una revisión sistemática cuantitativa tipo meta-análisis. Es decir, una revisión sistemática aplica el método científico en todo el proceso de revisión.

Por lo tanto, las revisiones sistemáticas difieren de las revisiones narrativas en una serie de características (Borenstein y cols., 2009; Cooper, 1998; Ellis, 2010; Field y Gillett, 2010; Hedges y Olkin, 1985; Huedo-Medina y Johnson, 2010; Marín-Martínez, Sánchez-Meca, Huedo-Medina, y Fernández-Guzmán, 2007; Marín-Martínez y cols., 2009; Rosenthal, 1995; Sánchez-Meca, 1986, 2008; Sánchez-Meca y Ato, 1989; Sánchez-Meca y Marín-Martínez, 2010):

(1) Rigor científico. Las revisiones sistemáticas cumplen con todas las normas del método científico que se exigen en las investigaciones primarias, posibilitando la replicación del proceso, es decir, que otros investigadores puedan repetir el meta-análisis en las mismas condiciones, comprobando si se dan o no los mismos resultados.

(2) Selección objetiva de los estudios. Los trabajos de investigación que formarán la unidad de análisis de las revisiones sistemáticas son seleccionados con criterios de búsqueda objetiva, sistemática y replicable de acuerdo a unas directrices previamente especificadas en el protocolo de revisión en función de la calidad o validez de las pruebas o evidencia aportada por los estudios científicos.

(3) Identificación de los estudios primarios. Los estudios primarios que forman parte del estudio de revisión sistemática son identificados de forma clara, facilitando la replicación de los estudios secundarios (revisiones sistemáticas). Generalmente dichos trabajos se añaden a las referencias bibliográficas del trabajo de meta-análisis con un asterisco.

(4) Seguridad. Al basarse en una metodología sistemática, objetiva y rigurosa, las conclusiones a las que se llegan con las revisiones sistemáticas son más fiables y seguras que las alcanzadas en revisiones cualitativas o narrativas de la investigación (Sánchez-Meca, 2008). Además, las principales conclusiones de los meta-análisis provienen del análisis estadístico de los resultados cuantitativos de los estudios primarios por lo que los estudios de meta-análisis constituyen una información precisa, objetiva y contrastable (Marín-Martínez y cols., 2009).

(5) Favorecen la Práctica Basada en la Evidencia (PBE): Las revisiones sistemáticas promueven la Practica Basada en la Evidencia en Psicología y otras Ciencias de la Salud, una nueva aproximación metodológica que anima a los profesionales a basar su práctica en la mejor evidencia obtenida de la investigación.

Además, las revisiones sistemáticas cuantitativas (estudios de meta-análisis) difieren de las revisiones narrativas en:

(**1**) **Eficiencia**. El meta-análisis tiene una mayor capacidad para tratar grandes cantidades de información que las revisiones narrativas, gracias a sus posibilidades de cuantificar y de codificar objetivamente las variables implicadas en los resultados de los estudios.

(**2) Detección de efectos pequeños**. Al centrarse en la magnitud de los efectos en lugar de en los resultados de las pruebas de significación estadística, el meta-análisis tiene una mayor capacidad para detectar efectos pequeños que pueden tener relevancia práctica o real. Las revisiones narrativas tienen más dificultades a la hora de detectar estos efectos. Sin embargo, los efectos pequeños son los más habituales en Psicología.

(3) Potencia estadística. Al manejar tamaños muestrales muy elevados como consecuencia de acumular las muestras de todos los estudios primarios analizados, los procedimientos de análisis estadístico propios del meta-análisis tienen mayor potencia estadística que las pruebas de significación estadística para detectar los efectos y las relaciones entre las variables implicadas.

(4) Énfasis en el tamaño del efecto. Otra de las grandes aportaciones del meta-análisis al proceso de investigación es el énfasis que pone en el tamaño del efecto, relegando las pruebas de significación estadística a un segundo plano (Borenstein y cols., 2009; Cumming, 2012, 2014; Frías-Navarro, 2011b; Kline, 2013).

(5) Análisis de la heterogeneidad. El meta-análisis permite analizar las fuentes de heterogeneidad en los resultados de los estudios primarios y detectar las características diferenciales de estos estudios que pueden explicar parte de esa heterogeneidad. Con ello, se consiguen explicar las posibles contradicciones entre los resultados de diferentes estudios sobre una misma temática (Marín-Martínez y cols., 2009). Es decir, el meta-análisis dispone de procedimientos estadísticos que permiten analizar el efecto de posibles variables moderadoras responsables de los resultados heterogéneos y contradictorios que pueden encontrarse en un determinado campo de investigación.

En definitiva, a diferencia de las revisiones narrativas, las revisiones sistemáticas no se basan en la selección parcial de la literatura ni utilizan criterios subjetivos de selección. La calidad de las revisiones sistemáticas está relacionada con la captación de todos los trabajos de investigación (publicados y no publicados) sobre una determinada temática. Dichos trabajos son valorados previamente desde el punto de vista de la validez de sus hallazgos para posteriormente formar parte de la revisión (de acuerdo con los criterios de inclusión y exclusión) donde se integran todos los hallazgos o evidencias, ofreciendo un resumen cualitativo sobre la situación de la temática analizada (revisión sistemática) o un resumen cuantitativo a través del cómputo del tamaño del efecto medio (en el caso del meta-análisis). Por todo ello, las revisiones sistemáticas se consideran fuentes valiosas de información dentro del modelo de la Práctica Basada en la Evidencia (Frías-Navarro y Pascual-Llobell, 2003; Pascual-Llobell, Frías-Navarro, y Monterde-i-Bort, 2004).

3.3. Proceso de revisión sistemática

La *Cochrane Collaboration*[4], la *Campbell Collaboration*[5], y la *PRISMA Statement*[6] señalan algunas consideraciones metodológicas a tener en cuenta y los pasos a seguir a la hora de elaborar revisiones sistemáticas de tipo cualitativo (o narrativas) y de tipo cuantitativo (o estudios de meta-análisis).

[4] Cochrane Collaboration (www.cochrane.org; www.cochrane.es)
[5] Campbell Collaboration (www.campbellcollaboration.org).
[6] PRISMA Statement (www.prisma-statement.org).

Estos organismos recomiendan la elaboración de un Protocolo de revisión previo al proceso de investigación que guie todo el proceso de manera explícita y replicable (en el mismo sentido ver Wright, Brand, Dunn, y Spindler, 2007).

3.3.1. Protocolo de revisión.

En el Protocolo de revisión se diseña el plan de actuación que se seguirá posteriormente en la revisión sistemática, incluyendo los argumentos que apoyen el desarrollo de la revisión (la necesidad de realizar el estudio de revisión), sus objetivos y los métodos que se utilizarán para localizar, seleccionar, valorar críticamente los estudios y analizar los datos.

El principal objetivo del Protocolo de revisión es prevenir el sesgo que podría ser introducido con cambios importantes en la cuestión de investigación original, como por ejemplo el sesgo en el proceso de selección de los estudios primarios o en el análisis de los datos. Gracias a la elaboración del protocolo, si se detectan problemas, se podrán corregir antes de iniciar el estudio de revisión propiamente dicho.

La identificación de la evidencia científica (los estudios empíricos primarios) requiere un proceso sistemático en el que se definan a priori, en un Protocolo de revisión, cuáles serán los criterios de inclusión y exclusión de los estudios, se diseñe la estrategia de localización de los mismos y su calidad metodológica.

Como Sánchez-Meca (2010) señala, los criterios de inclusión y exclusión dependen del objetivo de la revisión sistemática. Sin embargo, en un estudio de meta-análisis nunca pueden faltar los siguientes criterios de inclusión: (a) identificar los diseños de los estudios admisibles para el meta-análisis (experimentales, cuasi-experimentales, no experimentales); (b) definir los tipos de programas, tratamientos o intervenciones que se pretenden investigar; (c) definir las características de los participantes en los estudios; (d) determinar los datos estadísticos que deben aportar los estudios empíricos para poder calcular los tamaños del efecto (e.g., medias, desviaciones típicas, proporciones, pruebas t, pruebas F de ANOVA, etc.); (e) identificar cómo han de venir medidas las variables de resultado (e.g., escalas psicológicas, pruebas de rendimiento debidamente baremadas, medidas de autoinforme, etc.); (f) idioma en que tiene que estar escrito el estudio, y (g) rango temporal que se pretende examinar.

Además, el Protocolo de revisión debe incluir información detallada sobre el método de análisis estadístico a aplicar, como por ejemplo el método de estimación del *tamaño* del efecto y su varianza cuando las investigaciones primarias no lo proporcionan de forma directa, o el tipo de modelo utilizado para resumir la evidencia cuantitativa (efectos fijos/efectos aleatorios/efectos de coeficientes variables).

En definitiva, con el Protocolo de revisión se facilita la transparencia del proceso de investigación, se evitan las decisiones a posteriori adoptadas en función de los resultados y también se facilita la valoración externa del proceso de revisión, permitiendo detectar errores u omisiones antes de finalizar el trabajo de revisión sistemática. Además permite controlar la validez de los resultados y garantizar el rigor ético de la investigación.

3.3.2 Fases de un estudio de meta-análisis

La realización de un estudio de meta-análisis implica unas fases similares a las de cualquier investigación, si bien presenta algunas diferencias con éstas. Por ejemplo, en los estudios de meta-análisis, en cuanto se trata de una investigación secundaria, no se generan datos propios sino que se recogen las evidencia de informes de investigaciones primarias realizadas sobre el objeto de estudio, como artículos de revistas científicas, tesis doctorales, capítulos de libros, estudios técnicos, etc. (Botella y Sánchez-Meca, 2015).

Las principales ***fases*** de un estudio de meta-análisis son (Badenes-Ribera, 2013a, 2013b; Botella y Gambara, 2002, 2006; Botella y Sánchez-Meca, 2015; Conn y Rantz, 2003; Cooper, 1989; Cumming, 2012; Field y Gillett, 2010; Frías-Navarro y Monterde-i-Bort, 2014; Harris, y cols., 2014, Kline, 2013; Marín-Martínez y cols., 2009; Perestelo-Pérez, 2013; Sánchez-Meca, 2008, 2010; Sánchez-Meca y Botella, 2010; Sánchez-Meca y cols., 2011, 2013; Kline, 2013; Stroup y cols., 2000; Urra-Medina y Barría-Pailaquilén, 2010, Wright y cols., 2007):

1. Formulación del problema.
2. Búsqueda de los estudios empíricos primarios.
3. Selección de los estudios empíricos primarios.
4. Valoración la calidad metodológica de los estudios empíricos primarios.
5. Codificación de los datos de los estudios empíricos primarios.

6. Análisis estadísticos.

7. Publicación.

A. Formulación del problema

La formulación del problema supone plantear la pregunta de investigación de forma clara y precisa, lo que implica definir de forma teórica y operativa los constructos psicológicos objeto de estudio.

Una pregunta bien formulada responde al acrónimo PICOS (por sus siglas en inglés), es decir, que implica concretar los Participantes de estudio (*Participants*), la Intervención/tratamiento/exposición de los participantes (*Interventions*), las Comparaciones (*Comparisons*), las medidas de Resultados de la intervención (*Outcomes*) y el tipo de diseño del Estudio (*Study design*) (se puede encontrar más información en Harris y cols., 2014; Perestelo-Pérez, 2013).

Hay que tener en cuenta que cuanto más restringida sea una pregunta de investigación, más específico y focalizado será el estudio de meta-análisis, con lo que se podría tener dificultad en encontrar estudios empíricos que respondan a la cuestión planteada (Perestelo-Pérez, 2013; Wright y cols., 2007).

B. Búsqueda de los estudios empíricos primarios

Una vez definida de forma clara y precisa la pregunta de investigación, el siguiente paso es identificar la evidencia científica disponible que permita responder a esta pregunta. La búsqueda se debe basar en los criterios de inclusión y exclusión de los estudios empíricos primarios establecidos en el protocolo de revisión. El objetivo de esta fase es localizar *todos* los estudios con información relevante para que la revisión sistemática sea lo más comprehensiva y exhaustiva posible.

Se recomienda utilizar distintas estrategias de búsqueda. Algunos autores distinguen entre procedimientos de búsqueda formales e informales (Marín-Martínez y cols., 2009; Sánchez-Meca, 2008, 2010; Sánchez-Meca y Botella, 2010; Sánchez-Meca y cols., 2011, 2013).

→Dentro de los procedimientos formales de búsqueda destacan:

-búsquedas electrónicas en al menos dos bases de datos bibliográficas: Medline, PsycInfo, Embase, ERIC, WOS (web of Science), Scopus, CINAHL (Cumulative Index to Nursing and Allied Health Literature), PEDro (Physiotherapy Evidence Database),

CDSR (Cochrane Database of Systematic Reviews), etc., dependiendo del ámbito de estudio, siendo éstas de particular interés en el caso de las Ciencias de la Salud.

-*búsqueda manual en revistas científicas especializadas*, como mínimo los 6 meses anteriores a llevar a cabo la revisión sistemática.

-*búsqueda manual en el listado de referencias* de previas revisiones, estudios relevantes y de los estudios empíricos incluidos en la revisión.

→En los *procedimientos informales* de búsqueda destacan:

-*contacto con expertos e investigadores* en el ámbito de investigación para la identificación de o acceso a estudios que no han sido publicados o de difícil localización.

-búsqueda manual en los libros de actas de congresos, Tesis Doctorales, libros, capítulos de libros.

No obstante, aunque se utilizan distintas fuentes de información para identificar todos los estudios relevantes, resulta muy difícil (si no imposible) recuperar todos los estudios empíricos primarios existentes en un determinado campo. Lo ideal es recuperar el mayor número posible de estudios empíricos que cumplan con los criterios de inclusión.

C. Selección de los estudios empíricos primarios

La selección de los estudios se debe llevar a cabo al menos por dos revisores de manera independiente. Los desacuerdos entre los revisores se pueden resolver por consenso o a través de un tercer revisor. Algunos autores señalan que se debe comprobar el grado de acuerdo entre los revisores (Marín-Martínez y cols., 2009; Sánchez-Meca, 2008, 2010; Sánchez-Meca y Botella, 2010; Sánchez-Meca y cols., 2011, 2013). La comprobación del grado de acuerdo se puede realizar mediante la obtención de índices de acuerdo, como Kappa de Cohen para las variables cualitativas, o la correlación intra-clase para las variables continuas (Sánchez-Meca y cols., 2011).

Siguiendo a Perestelo-Pérez (2013), la selección de los estudios se desarrolla en dos fases: pre-selección y selección.

En la *fase de pre-selección* se identifican los estudios empíricos que cumplen los criterios de inclusión a partir de la lectura de los *abstracts* y el título del estudio. Si hay dudas sobre la inclusión o no de un estudio, se procede a su lectura a texto completo.

En la *fase de selección*, los estudios pre-seleccionados se analizan a texto completo y se decide si se incluyen o no en la revisión sistemática. Se debe justificar la exclusión de los estudios.

La clave en esta fase radica en seleccionar estudios empíricos homogéneos que aporten evidencias sobre una misma cuestión (Botella y Gambara, 2006; Botella y Sánchez-Meca, 2015; Kline, 2013). Es decir, seleccionar estudios que compartan características respecto del diseño de la investigación, modo en el que se han medido las variables de resultado, las características de los participantes y las características de los tratamientos (Botella y Sánchez-Meca, 2010).

D. Valoración de la calidad metodológica de los estudios primarios

El objetivo de esta fase es controlar la calidad metodológica de los estudios primarios que podría afectar a la calidad (es decir, la ausencia de sesgos) de las estimaciones de los efectos del trabajo de meta-análisis. Por tanto, el énfasis en la calidad metodológica de los estudios primarios es consistente con los objetivos de la ciencia de producir un conocimiento científico válido (Conn y Rantz, 2003).

La calidad metodológica del estudio primario hace referencia al grado en que un estudio se ha diseñado e implementado de una forma metodológica correcta, es decir, protegiéndose de las amenazas contra la validez interna y externa de los resultados (Botella y Sánchez-Meca, 2015).

Todo estudio de meta-análisis debe incorporar algún sistema de valoración de la calidad metodológica de los estudios primarios (Aguinis, Pierce, Bosco, Dalton y Dalton, 2011; Botella y Sánchez-Meca, 2015; Perestelo-Pérez, 2013; Sánchez-Meca, 2010; Wright y cols., 2007).

En cuanto a cómo valorar la calidad metodológica de los estudios primarios, no hay un único procedimiento. Se puede utilizar una plantilla o un listado de revisión previamente elaborado en la fase de desarrollo del protocolo de la revisión sistemática, o bien las escalas y listados de comprobación (Checklists) que ya existen en la literatura (donde se cuenta con más de 100 instrumentos), buena parte de las cuales se han desarrollado en la literatura médica (Jüni, Altman y Egger, 2001; Moher y cols., 2010; Moher, Jones y Lepage for the CONSORT group, 2001; Tritchler, 1999; Wortman, 1994).

Como Botella y Sánchez-Meca (2015) señalan, las escalas están compuestas por ítems referidos a aspectos específicos de la calidad metodológica. Los valores *asignados* a estos ítems se agregan para alcanzar una única puntuación global que mide un constructo general de calidad (aunque se puede desagregar en varias subescalas). Mientras que las listas de comprobación también son listas de ítems de verificación de cumplimiento, ya sea dicotómicamente (Sí/No) o con varias categorías, pero sin sumar las puntuaciones de los ítems en una puntuación total.

Una vez obtenidos los indicadores de calidad metodológica de los estudios primarios, se puede actuar de tres formas (Aguinis y cols., 2011; Botella y Sánchez-Meca, 2015; Conn y Rantz, 2003; Ellis, 2010; Kline, 2013; Sutton y Higgins, 2008; Wright y cols., 2007):

(1) *La* ***calidad metodológica como criterio de inclusión de los estudios primarios*** supone excluir del trabajo de meta-análisis aquellos estudios que no cumplan los criterios de calidad metodológica establecidos, por ejemplo, un punto de corte en las escalas de calidad o el cumplimiento de determinados ítems en las listas de *c*omprobación (Crowe y Shepard, 2011; Wright y cols., 2007*).* Sin embargo, Ellis (2010) señala diversas razones por las que se debe dudar de excluir estudios sobre la base de su calidad metodológica: 1) hacer juicios de calidad introduce un sesgo del revisor, pues el significado de "calidad metodológica" puede ser distinto para distintas personas; 2) prácticamente no hay estudios libres de fallos, por tanto, la exclusión de los estudios sin excelencia metodológica conduciría al rechazo de un gran número de evidencia sobre un tema, con lo que se desperdicia la investigación (además de que, como Botella y Sánchez-Meca (2015) señalan, la reducción drástica de estudios admitidos reduce la potencia de los análisis estadísticos y la validez de las conclusiones); 3) los estudios de baja calidad pueden proporcionar información que se puede combinar de manera significativa con los estudios de alta calidad. Después de todo, como Ellis (2010) señala "si los estudios estiman un efecto común, entonces la evidencia obtenida a partir de diferentes estudios debe converger" (p. 124).

(2) La ***calidad metodológica como criterio de ponderación en la estimación del tamaño del efecto medio***. Parte de la premisa de que los estudios con deficiencias metodológicas son menos informativos y, por tanto, deben tener menos "peso" en los resultados. Ponderar por las puntuaciones de calidad metodológica ofrece al menos dos ventajas (Conn y Rantz, 2003): (a) permite que todos los estudios sean incluidos en el

estudio de meta-análisis, lo cual evita los posibles sesgos en la selección de los estudios primarios cuando la calidad se erige como criterio de inclusión; y (b) pone *más énfasis* en los estudios de mayor rigor a fin de que estos estudios afecten los resultados en mayor medida.

(3) La ***calidad metodológica como cuestión empírica a analizar en el propio meta-análisis***. Supone examinar la relación que pueda existir entre la calidad metodológica de los estudios y los resultados del meta-análisis (tamaño del efecto medio). Es decir, considerar la calidad metodológica de los estudios primarios como variable moderadora de la heterogeneidad entre los tamaños del efecto de los estudios.

Finalmente, la valoración de la calidad metodológica de los estudios primarios debe llevarse a cabo forma independiente por al menos dos revisores. Los desacuerdos en la valoración de la calidad pueden ser superados por consenso entre los dos revisores o por un tercer revisor.

E. Extracción y codificación de los estudios primarios

La extracción de los datos supone obtener información relevante de cada una de los estudios primarios seleccionados con el objetivo de elaborar la revisión sistemática. El tipo de información que será recogida de cada estudio debe ser definido previamente en el Protocolo de la revisión sistemática, asegurando que los revisores extraerán toda la información necesaria de cada estudio de una manera uniforme.

Los datos a extraer no son iguales en todas las revisiones sistemáticas. Sin embargo, se suelen registrar las características de los estudios denominadas metodológicas, sustantivas y extrínsecas (Botella y Sánchez-Meca, 2015), como por ejemplo autor y año de publicación, tipo y características del diseño del estudio, número y características de los participantes, ámbito del estudio, medidas de resultado, fuentes de financiación, conflicto de intereses, etc. (Perestelo-Pérez, 2013). El objetivo es comprobar cuál de ellas está afectando o moderando los resultados, esto es, explicar por qué los estudios sobre un mismo tema alcanzan resultados diferentes, e incluso en ocasiones contradictorios (Botella y Sánchez-Meca, 2015; Lipsey y Wilson, 2001; Sánchez-Meca, 2010).

En un estudio de meta-análisis es muy importante evitar la inclusión de estudios que se han publicado varias veces utilizando los mismos datos. La inclusión de datos duplicados podría conducir a un sesgo en la estimación del efecto medio, sobrevalorando su tamaño.

Para que la extracción de los datos sea objetiva y sistemática se requiere un control externo mediante una plantilla de revisión. Además, también en esta fase se recomienda que al menos dos revisores efectúen de forma independiente la extracción y codificación de los datos de los estudios (o de una muestra aleatoria de éstos). Se debe comprobar la fiabilidad del proceso de codificación verificando el grado de acuerdo entre revisores mediante la obtención de índices de acuerdo como Kappa de Cohen para las variables cualitativas, o la correlación intra-clase para las variables continuas (Botella y Sánchez-Meca, 2015; Marín-Martínez y cols., 2009; Orwin y Vevea, 2010; Sánchez-Meca, 2010; Sánchez-Meca y Botella, 2010; Sánchez-Meca y cols., 2011). Al igual que antes se ha señalado, los desacuerdos se pueden resolver por consenso o por la participación de un tercer revisor.

F. Análisis estadístico e interpretación de los datos

El objetivo final de una revisión sistemática es ofrecer un resumen de los efectos detectados en los estudios empíricos primarios analizados. El resumen puede ser descriptivo o puede ofrecer una síntesis cuantitativa de los datos de las publicaciones cuando se trata de un trabajo de meta-análisis.

El análisis estadístico típico de un meta-análisis pasa por tres fases (Borenstein y cols., 2009; Field y Gillett, 2010; Lipsey y Wilson, 2001; Marín-Martínez y cols., 2009; Sánchez-Meca, 2010): (1) cálculo del tamaño del efecto medio con su intervalo de confianza y valoración de su significación estadística; (2) análisis de la heterogeneidad de los tamaños del efecto, y (c) si los tamaños del efecto son heterogéneos, búsqueda de variables moderadoras de tal variabilidad.

Por tanto, el objetivo de la mayoría de los análisis estadísticos es responder a las siguientes cuestiones (Botella y Sánchez-Meca, 2015; Huedo-Medina y Johnson, 2010; Sánchez-Meca, Marín-Martínez, y Huedo-Medina, 2006):

(1) ¿Cuál es la estimación combinada del tamaño del efecto? O lo que es lo mismo ¿es estadísticamente significativo el tamaño del efecto medio?
(2) ¿Son homogéneos los tamaños del efecto de los estudios primarios?

(3) En caso de ser heterogéneos los tamaños del efecto de los estudios primarios, ¿qué características de los estudios pueden explicar la heterogeneidad observada?

(4) ¿Es posible formular un modelo explicativo de la heterogeneidad de los tamaños del efecto a partir de las variables moderadoras codificadas?

La gran mayoría de los meta-análisis se realizan computando el tamaño del efecto medio de las diferencias estandarizadas de medias, *odds ratio*, o bien de los coeficientes de correlación (Sánchez-Meca, 2010). Los estudios de meta-análisis también pueden ser útiles para computar una estimación más precisa de datos de frecuencia como proporciones o prevalencias (Barendregt, Doi, Lee, Norman y Vos, 2013).

En los trabajos de meta-análisis el tamaño del efecto medio se acompaña de su intervalo de confianza para estimar el parámetro μ_θ. La amplitud entre el límite superior y el límite inferior del intervalo de confianza ofrece información sobre el grado de precisión de la estimación del efecto en la población. Cuanto más estrecho es el intervalo de confianza, mayor es la precisión de la estimación puntual del efecto en la población. Además, el intervalo de confianza permite realizar un contraste de hipótesis. Es decir, contrastar la hipótesis nula de que el efecto en la población es nulo, verificando si el valor 0 se encuentra dentro de los límites del intervalo de confianza.

La H_0 (hipótesis nula) también se puede contrastar mediante el estadístico de contraste *z* que bajo la H_0 se distribuye aproximadamente según una ley normal tipificada, N (0; 1). Los programas estadísticos utilizados para ejecutar los meta-análisis informan sobre el nivel crítico de probabilidad (el valor *p*) asociado al valor del estadístico de contraste *z*. Así, para un determinado nivel de significación alfa (e.g., alfa = .05) la H_0 será rechazada si el valor de p es menor al nivel alfa establecido (e.g., $p <$ .05). También, en casos de alta heterogeneidad, el contraste se puede realizar mediante el estadístico de contraste *t* que asume una distribución *t* con k-1 grados de libertad (Botella y Sánchez-Meca, 2015).

G. Publicación

La última etapa en la realización de un trabajo de meta-análisis es su publicación. Al tratarse de una investigación empírica, las secciones que debe incluir el informe escrito del meta-análisis son las típicas de un estudio empírico: introducción, método,

resultados y discusión y conclusiones (Botella y Gambara, 2006; Botella y Sánchez-Meca, 2015; Rosenthal, 1995; Sánchez-Meca y Botella, 2010).

Muy sucintamente, en la sección de *Introducción* se revisa el tema objeto de estudio, se definen los constructos implicados y se formulan los objetivos del trabajo del de meta-análisis y, en su caso, las hipótesis.

La sección *Método* debe contener todos los datos y decisiones tomadas en el meta-análisis para que pueda ser replicado por otros investigadores. Así, se describen los criterios de inclusión y exclusión de los estudios, las estrategias de búsqueda de los estudios utilizadas, el proceso de codificación de las características de los estudios y una descripción del índice del tamaño del efecto junto con las técnicas de análisis estadístico aplicadas. Como señala Sánchez-Meca (2010), de la precisión y meticulosidad con que se reporte la sección del método del informe de investigación dependerá el grado en que el lector podrá hacer una lectura crítica del trabajo de meta-análisis y valorar sus posibles deficiencias.

En la sección de *Resultados* se presentan los tamaños del efecto, el efecto medio y los datos estadísticos pertinentes para valorar su significación estadística y, en su caso, los análisis de variables moderadoras realizados para comprobar el influjo de algunas características de los estudios sobre la variabilidad de los tamaños del efecto individuales.

Finalmente, en la sección de *Discusión y Conclusiones*, los resultados del meta-análisis se ponen en relación con la literatura previa sobre el tema de investigación, se discute la relevancia práctica de los hallazgos, sus implicaciones para la práctica profesional y se apuntan líneas futuras de investigación.

3.4. Análisis estadísticos en el meta-análisis

Como se decía, el análisis estadístico típico de un trabajo de meta-análisis pasa por tres fases: (1) cálculo del tamaño del efecto medio con su intervalo de confianza y valoración de su significación estadística; (2) análisis de la heterogeneidad de los tamaños del efecto, y (c) si los tamaños del efecto son heterogéneos, búsqueda de variables moderadoras de tal variabilidad.

3.4.1- Modelos de estimación del tamaño del efecto medio

La estimación del tamaño del efecto medio en los estudios de meta-análisis puede realizarse mediante tres modelos estadísticos:

(1) el ***modelo de efecto fijo*** (Borenstein y cols., 2009, Borenstein, Hedges, Higgins, y Rothstein, 2010; Botella y Gambara, 2002, 2006; Botella y Sánchez-Meca, 2015; Cumming, 2012; Field y Gillett, 2010; Hedges y Vevea, 1998; Huedo-Medina y Johnson, 2010; Hunter y Schmidt, 2000; Sánchez-Meca y cols., 2006),

(2) el ***modelo de efectos aleatorios*** (Borenstein y cols., 2009, 2010; Botella y Gambara, 2002, 2006; Botella y Sánchez-Meca, 2015; Catalá-López y Tobías, 2014; Cumming, 2012; Field y Gillett, 2010; Hedges y Vevea, 1998; Huedo-Medina y Johnson, 2010; Hunter y Schmidt, 2000; Sánchez-Meca y cols., 2006), y

(3) el ***modelo de coeficientes variables*** (Boella y Sánchez-Meca, 2015; Bonnet, 2002, 2009, 2010; Sánchez-Meca, López-López y López-Pina, 2013).

Estos tres modelos difieren en cuanto a los supuestos de partida, las condiciones de aplicación y grado de generalización de los resultados.

Modelo de Efecto Fijo (EF)

El *"**modelo de efecto fijo"*** asume que los estudios incluidos en el trabajo de meta-análisis son homogéneos (muestras idénticas en composición y variabilidad), provienen de una misma población, y estiman a un mismo y único tamaño del efecto paramétrico o poblacional (θ) (Borenstein y cols., 2009, 2010; Botella y Sánchez-Meca, 2015; Frías-Navarro y Monterde-i-Bort, 2014; Huedo-Medina y Johnson, 2010). En consecuencia, la variabilidad que se observa entre los tamaños del efecto de los estudios primarios se debe al error de muestreo aleatorio inherente en cada estudio (variabilidad intraestudio), es decir, se debe al hecho de que los estudios primarios han utilizado muestras de sujetos diferentes, por lo que, las muestras están compuestas por diferentes sujetos. Por ello, el modelo matemático del modelo de efectos fijos es

$$T_i = \theta + u_i$$

donde u_i es el error de muestreo intraestudio al que están sometidas las estimaciones del tamaño del efecto, T_i, única fuente de variabilidad en este modelo.

El propósito del meta-analista es generalizar los resultados a una población de estudios con características idénticas a las incluidas en el estudio de meta-análisis (Borenstein y cols., 2010).

El estimador óptimo del efecto poblacional es la media de los valores de los tamaños del efecto de los estudios primarios ponderada por la inversa de sus respectivas varianzas intraestudio. De tal modo que, a los estudios con una estimación más precisa del tamaño del efecto de la población (mínima varianza) se les asigna más peso en la estimación combinada del tamaño del efecto, mientras que a los estudios con una estimación menos precisa del tamaño del efecto de la población (alta varianza) se les asigna un menor peso. En otras palabras, en la estimación del tamaño del efecto combinado, los estudios con muestras pequeñas tendrán un menor peso y, por el contrario, los estudios con muestras grandes tendrán un mayor peso (Borenstein y cols., 2009, 2010).

Finalmente, a partir del tamaño del efecto medio se construye un intervalo de confianza en torno a éste para estimar el parámetro μ_θ.

Modelo de efectos aleatorios (EA)

En el *"**modelo de efectos aleatorios***" se asume que cada uno de los estudios incluidos en un trabajo de meta-análisis estiman a un tamaño del efecto paramétrico o poblacional propio (θ), diferente en cada uno de los estudios y, a su vez, los efectos paramétricos estimados en los estudios son una muestra aleatoria de una población de efectos paramétricos (Botella y Sánchez-Meca, 2015). Es decir, se asume que los estudios primarios estiman una distribución de tamaños del efecto paramétricos en la población (Frías-Navarro y Monterde-i-Bort, 2014), una distribución que suele asumirse Normal, $\theta \sim N(\mu_0, \tau^2)$ (Botella y Sánchez-Meca, 2015; Sánchez-Meca y cols., 2006). El objetivo del meta-análisis es estimar la media de esta distribución (Borenstein y cols., 2009, 2010). De tal manera que, si se pudiese realizar un número infinito de estudios, los tamaños del efecto verdaderos de dichos estudios se distribuirían en torno a una media. Por lo tanto, los tamaños del efecto de los estudios incluidos en el trabajo de meta-análisis representan una muestra aleatoria de esos tamaños del efecto y de ahí el término de efectos aleatorios (Frías-Navarro y Monterde-i-Bort, 2014).

En consecuencia, los resultados pueden generalizarse a una población mayor de posibles resultados o estudios con características similares, aunque no necesariamente idénticas.

El modelo matemático incorpora dos términos de error: (1) variabilidad intraestudio, que coincide con la del modelo de efectos fijos (u_i) y (2) variabilidad interestudios (e_i), como la desviación de cada estudio respecto del tamaño del efecto medio, que hace referencia a las diferentes características de los estudios (e.g., intensidad, duración de una intervención, edad de los sujetos pueden haber variado de un estudio a otro).

$$T_i = \mu_0 + u_i + e_i$$

Al haber dos términos de error que se consideran independientes entre sí, la varianza de los tamaños del efecto individuales es la suma de los dos tipos de variabilidad: intraestudio e interestudio (Borenstein y cols., 2009, 2010; Huedo-Medina y Johnson, 2010; Sánchez-Meca y cols., 2006).

Por tanto, el modelo de efectos aleatorios implica ponderar cada estimación del tamaño del efecto por la inversa de la suma de las varianzas intraestudio e interestudio. De este modo, a cada estudio individual le corresponde una varianza intraestudio propia mientras que la varianza interestudio es común a todos los estudios, es decir, es una constante (Botella y Sánchez-Meca, 2015). Debido a que τ^2 (tau cuadrado) es una constante, se reducen las diferencias relativas entre los pesos de los estudios en función de su precisión (tamaño de la muestra), lo que significa que el peso relativo asignado a cada estudio está más equilibrado en el marco del modelo de efectos aleatorios de lo que está en el marco del modelo de efecto fijo (Borenstein y cols., 2009, 2010). Es decir, la estimación combinada del tamaño del efecto no está demasiado influenciada por el tamaño de la muestra de los estudios, en la forma en que lo haría el modelo de análisis de efecto fijo (asignar un peso muy pequeño a los estudios con alta varianza o poca precisión –tamaño muestral pequeño- y dar demasiada importancia a un estudio con baja varianza o alta precisión –tamaño muestral muy grande-).

Finalmente, en torno al tamaño del efecto se construye un intervalo de confianza tomando en consideración las dos fuentes de variabilidad (intraestudio e interestudio). Por ello, la amplitud del intervalo de confianza es más grande en el modelo de efectos aleatorios que en el modelo de efecto fijo.

Modelo de coeficientes variables

El *"**modelo de coeficientes variables"*** es una clase de modelo de efecto fijo en el que se asume que cada uno de los estudios incluidos en un trabajo de meta-análisis estima a un tamaño del efecto paramétrico o poblacional propio (θ). Por lo tanto, el modelo matemático es

$$T_i = \theta + u_i$$

Al igual que en el modelo de efectos fijos, el propósito del meta-analista es generalizar los resultados a una población de estudios con características idénticas a las incluidas en el estudio de meta-análisis (Bonnet, 2002, 2009, 2010; Sánchez-Meca y cols., 2013).

Por otro lado, al igual que el modelo de efectos aleatorios, se parte de que los estudios primarios estiman su propio tamaño del efecto paramétrico. Es decir, ambos modelos asumen que los parámetros poblacionales son heterogéneos. Sin embargo, el modelo de efectos aleatorios asume que los estudios primarios se han seleccionado al azar de una superpoblación de potenciales estudios y, como consecuencia, es posible estimar el tamaño del efecto medio de la superpoblación de estudios. Por el contrario, el modelo de coeficiente variable sostiene que los estudios no han sido seleccionados al azar de una superpoblación de estudios, por lo que los tamaños del efecto incluidos en el meta-análisis sólo representan a sus propios tamaños del efecto poblacionales, y el promedio de estos tamaños del efecto es el parámetro que debe ser estimado. En consecuencia, como se ha dicho, los resultados solo pueden generalizarse a una población de estudios con características idénticas a las incluidas en el estudio de meta-análisis (Sánchez-Meca y cols., 2013).

Elección del modelo estadístico

Los tres modelos expuestos difieren en cuanto a los supuestos de partida, las condiciones de aplicación y grado de generalización de los resultados. Por tanto, las consecuencias de asumir un modelo estadístico u otro afectan al grado en que los resultados se pueden generalizar. En este sentido, los modelos de efecto fijo y de coeficientes variables solo permiten generalizar los resultados meta-analíticos a los estudios de características similares a los incluidos en el trabajo de meta-análisis, mientras que el modelo de efectos aleatorios permite generalizar a una superpoblación más amplia de estudios.

Además, los diferentes modelos de análisis estadísticos propuestos varían con respecto a las transformaciones de los índices del tamaño del efecto y/o el peso de los tamaños del efecto cuando son estadísticamente combinados.

Sin embargo, decidir qué modelo estadístico es el más adecuado para realizar el trabajo de meta-análisis no es sencillo, pues se desconocen los parámetros poblacionales que se pretende estimar y por lo tanto no se conoce qué modelo se ajustará mejor a los datos. Se trata de una cuestión conceptual y empírica. A nivel conceptual, el investigador debe valorar el grado de generalización que pretende otorgar a sus resultados, y a nivel empírico, el meta-analista debe analizar el grado de heterogeneidad de los tamaños del efecto de los estudios primarios, si la distribución de los tamaños del efecto se aproxima a la distribución normal y el número de estudios primarios (Borenstein y cols., 2009,2010; Botella y Sánchez-Meca, 2015; Sánchez-Meca y cols., 2006).

Botella y Sánchez-Meca (2015) señalan que como guía orientativa, si se tiene un número razonable de estudios (e.g., $k > 30$), con una distribución de tamaños del efecto aproximadamente normal y elevada heterogeneidad, se puede asumir un modelo de efectos aleatorios. Si no se cumplen estas condiciones, estos autores aconsejan adoptar el modelo de efecto fijo si no existe heterogeneidad entre los tamaños del efecto. Finalmente, señalan que si además de no cumplirse los supuestos del modelo de efectos aleatorios, los tamaños del efecto individuales son heterogéneos, el modelo más apropiado sería el modelo de coeficientes variables.

No obstante, actualmente, el modelo de efectos aleatorios se considera como la opción más realista puesto que no asume la homogeneidad entre los estudios (Borenstein y cols., 2009; Cooper, Hedges, y Valentine, 2009; Field y Gillett, 2010; Kisamore y Brannick; 2008; Huedo-Medina y Johnson, 2010; National Research Council, 1992), lo cual permite generalizar los resultados más allá de los estudios incluidos en el meta-análisis (Botella y Sánchez-Meca, 2015; Borenstein y cols., 2010; Hedges y Vevea, 1998; Sánchez-Meca y cols., 2006).

Sin embargo, el modelo de efectos aleatorios no está exento de críticas. Como Bonnet (2002, 2009, 2010) señala, el modelo de efectos aleatorios requiere que los estudios incluidos en el trabajo de meta-análisis se hayan seleccionado de forma aleatoria de una población de estudios. No obstante, esta exigencia nunca será satisfecha

por el hecho de que los estudios incluidos en el trabajo de meta-análisis son los que se encuentran en la literatura, es decir, no se seleccionan aleatoriamente. Como Botella y Sánchez-Meca (2015) señalan, esta crítica es muy exigente, ya que podría extenderse a las investigaciones primarias, en cuyo caso se tendría que "condenar" el uso de los métodos de inferencia estadística en cualquier investigación en la que no se hubieran seleccionado aleatoriamente a los sujetos que conforman la muestra del estudio. En consecuencia, esta crítica invalidaría casi la totalidad de las investigaciones en Psicología y en las ciencias empíricas en general.

3.4.2.-Evaluación de la heterogeneidad

El segundo objetivo del meta-análisis es comprobar si los tamaños del efecto de los estudios primarios son heterogéneos entre sí. Por ello, la estimación del tamaño del efecto medio se acompaña de estadísticos que informan del grado de heterogeneidad estadística que se detecta entre los estudios o variación entre los resultados de los estudios más allá del error aleatorio (Song, Sheldon, Sutton, Abrams y Jones, 2001). El tamaño del efecto medio será representativo en la medida en que los tamaños del efecto individuales no sean muy heterogéneos entre sí (Botella y Sánchez-Meca, 2015).

La prueba de heterogeneidad *Q* de Cochran permite comprobar la heterogeneidad y su grado de significación estadística (Cochran, 1954). El estadístico *Q* tiene una distribución Chi Cuadrado con *k*-1 grados de libertad, siendo k el número de estudios, facilitando una decisión estadística acerca de la hipótesis de homogeneidad de los tamaños del efecto computados. Valores de $Q > k-1$ sugieren heterogeneidad estadística (Frías-Navarro y Monterde-i-Bort, 2014). El estadístico *Q* se acompaña de un valor *p* de probabilidad que, si es menor al nivel de alfa establecido (por ejemplo alfa = .05), sugiere presencia de heterogeneidad entre los tamaños del efecto y, en consecuencia, el tamaño del efecto medio no los representa bien.

El contraste de hipótesis nula de homogeneidad con el estadístico *Q* solo informa de si existe o no heterogeneidad estadísticamente significativa entre los tamaños del efecto individuales, pero no indica el grado de heterogeneidad exhibida por los tamaños del efecto (Borenstein y cols., 2009; Botella y Sánchez-Meca, 2015; Sánchez-Meca y Marín-Martínez, 2010). Por ello, la evaluación de la heterogeneidad se suele completar con el estadístico I^2 que informa del porcentaje de variación que se detecta entre los tamaños del efecto de los estudios que componen el trabajo de meta-

análisis (Takkouche, Cadarso-Sures y Spiegelman, 1999). El estadístico I^2 puede interpretarse en términos de porcentaje indicando el grado de variabilidad total de los tamaños del efecto que se debe a la variabilidad entre los estudios y no al error de muestreo aleatorio. Una ventaja del estadístico I^2 es que no está afectado por el tamaño de la muestra mientras que la prueba *Q* sí se ve afectada y además tiene baja potencia estadística (Higgins y Thompson, 2002; Huedo-Medina, Sánchez-Meca, Marín-Martínez y Botella, 2006). Higgins y Thompson (2002) proponen interpretar el valor de I^2 como heterogeneidad baja (25%), heterogeneidad media (50%) y heterogeneidad alta (75%).

3.4.3.-Evaluación de variables moderadoras

Si existe heterogeneidad entre los tamaños del efecto de los estudios primarios, entonces se hace preciso examinar el influjo de características de los estudios previamente codificadas. Las variables moderadoras actúan como variables predictoras, independientes o explicativas mientras que los tamaños del efecto actúan en el análisis como la variable dependiente, de resultado o de criterio.

Si la variable moderadora es cualitativa (por ejemplo, tipo de tratamiento, población de referencia, tipo de grupo control, etc.), se utilizan modelos de análisis de varianza (ANOVA), y si las variables moderadoras son cuantitativas (por ejemplo, edad media de los sujetos codificada en años, duración del tratamiento, porcentaje de varones o mujeres en la muestra, etc.,), se utilizan modelos de regresión, que en el campo del meta-análisis se llaman "modelos de meta-regresión".

Por tanto, mediante técnicas de ANOVA y de análisis de regresión, ambas ponderadas, es posible analizar el influjo de variables moderadoras sobre los tamaños del efecto de los estudios primarios (Huedo-Medina y Johnson, 2010; Sánchez-Meca, 2010).

Así pues, con las técnicas de ANOVA se analiza la "homogeneidad intercategorías" (mediante el estadístico de contraste Q_B) esto es, si existen diferencias –heterogeneidad- estadísticamente significativas entre los efectos medios de las distintas categorías de la variable moderadora cualitativa, y la "homogeneidad intracategoría" (mediante el estadístico de contraste Q_W), esto es, si existe heterogeneidad o diferencias entre los tamaños del efectos dentro de cada categoría. Además, el estadístico *Q*W se puede descomponer en tanta categorías como tenga la variable moderadora.

Finalmente, señalar que los estadísticos Q_B y Q_W se corresponden con las sumas de cuadrados intercategorías e intracategoría, respectivamente. Por tanto, la suma de ambos constituye la partición de la suma de cuadrado total ponderada que se corresponde con el estadístico Q de heterogeneidad ($Q = Q_B + Q_W$).

Al igual que las técnicas de ANOVA aplicables en los estudios primarios, el estadístico de contraste Q_B del ANOVA en el meta-análisis solo indica si existen diferencias estadísticamente significativas entre los efectos medios de las categorías de la variable moderadora, pero no aporta ninguna información sobre entre qué categorías de la variable se encuentran estas diferencias. Para dar respuesta a esta cuestión, es necesario realizar contrastes o comparaciones a posteriori (*post hoc*) entre los efectos medios, por ejemplo, mediante el procedimiento de Bonferroni y de Scheffé (Botella y Sánchez-Meca, 2015).

Los análisis de meta-regresión, al igual que en las investigaciones primarias, permiten analizar el rol de una variable moderadora continua (meta-regresión lineal simple) o varias variables moderadoras continuas y cualquier combinación de variables moderadoras continuas y cualitativas (meta-regresión lineal múltiple). Es decir, permiten analizar la existencia de una regresión lineal entre las variables moderadoras y los tamaños del efecto (Borenstein y cols., 2009). Las variables moderadas son los predictores del modelo y se consideran de efecto fijo y el tamaño del efecto es la variable independiente que se considera una variable continua distribuida según una ley normal.

En los análisis de meta-regresión, para comprobar si las variables moderadoras están relacionadas con los tamaños del efecto se utiliza el estadístico de contraste Q_R (Borenstein y cols., 2009).

Finalmente, mencionar que en el contexto del meta-análisis, las técnicas del ANOVA y de la meta-regresión se pueden ejecutar desde diferentes modelos estadísticos: (1) modelos de efectos fijos (*fixed-effects model*), y (2) modelos de efectos mixtos (*mixed-effects model*). El ***modelo de efectos fijos*** (EF) es una extensión del modelo de efecto fijo (EF), mientras que el ***modelo de efectos mixtos*** (EM) lo es del modelo de efectos aleatorios (EA). En ambos casos, se aplican métodos de ponderación que tienen en cuenta la precisión de las estimaciones de los efectos. El análisis de ambos modelos excede las pretensiones de este trabajo. Se puede profundizar en los modelos

estadísticos en las siguientes referencias (e.g., Borenstein y cols., 2009; Borenstein y Higgins, 2013; Botella y Sánchez-Meca, 2015; Hedges y Olkin, 1985; Sánchez-Meca y Marín-Martínez, 2010).

La elección del modelo estadístico para el análisis de las variables moderadoras (EF o EM) es una decisión que debe tomar el meta-analista en función del grado en que desea generalizar los resultados del ANOVA y/o de la meta-regresión (Borenstein y cols., 2010). Por ejemplo, si desea generalizar los resultados a una población de estudios con características idénticas a los estudios incluidos en el estudio de meta-análisis, el modelo estadístico de elección sería el modelo de efectos fijos (EF). Sin embargo, si el meta-analista desea generalizar los resultados a un mayor número de estudios con características parecidas pero no idénticas, el modelo de elección será del modelo de efectos mixtos (EM).

3.5.-Representacion gráfica: el *forest plot*

Las visualizaciones gráficas se han convertido en la principal herramienta para la presentación de los resultados en los estudios de meta-análisis. Así, el reporte de los resultados, esto es, informar del valor del tamaño del efecto combinado y su intervalo de confianza, se suele acompañar de una gráfica llamada *forest plot* donde se presentan los tamaños del efecto de los estudios individuales junto con el tamaño del efecto medio computado en el estudio de meta-análisis y sus intervalos de confianza, tanto de forma numérica como de forma gráfica. Por tanto, la lectura de los *forest plots* es una habilidad requerida para su lectura crítica (Anzures-Cabrera y Higgins, 2010; Borenstein y cols., 2009; Cumming, 2012; Ellis, 2010).

El *forest plot* se construye dibujando tantas filas como estudios incluidos en el trabajo de meta-análisis. En cada línea se representa el intervalo de confianza de cada estudio mediante una línea cuyos extremos se corresponden con los límites del intervalo de confianza asociado al tamaño del efecto estimado en cada estudio. La estimación puntual del tamaño del efecto se representa con algún símbolo convencional (cuadrados, círculos, rectángulos). También se presenta una línea vertical que hace referencia a la ausencia de efectos o efecto nulo. En la parte inferior del *forest plot* se presenta el tamaño del efecto medio junto con su intervalo de confianza obtenido en el meta-análisis.

Como ejemplo, en la Figura 2 se presenta el *forest plot* con los resultados del trabajo de meta-análisis realizado por Badenes-Ribera, Frías-Navarro, Bonilla-Campos, Monterde-i-Bort y Pons-Salvador (2015), en una revisión de estudios sobre la prevalencia de la violencia de pareja sufrida por mujeres lesbianas auto-identificadas en relaciones de pareja del mismo sexo. El *forest plot* permite obtener una visión global de los resultados y el grado de heterogeneidad existente entre los tamaños del efecto individuales.

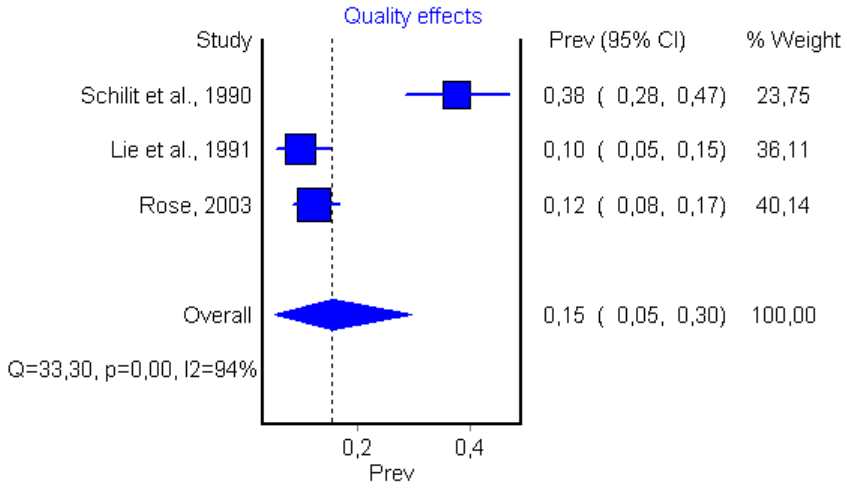


Figura 2. Forest plot sobre la prevalencia de la violencia de pareja sufrida por mujeres lesbianas auto-identificadas en relaciones de pareja del mismo sexo (tomado del estudio de Badenes-Ribera y cols., 2015). Prev: prevalencia. 95% CI: intervalo de confianza al 95% en torno a la prevalencia. Weight: peso del estudio en el meta-análisis. Q: estadístico de heterogeneidad *Q*. I^2: estadístico de heterogeneidad I^2.

En la parte izquierda de la imagen se detallan los estudios primarios que han formado parte del trabajo de meta-análisis. En la parte derecha se detallan los resultados en prevalencia de cada uno de los estudios primarios junto con su intervalo de confianza y su peso en el cálculo de la prevalencia media. Los cuadrados de la representación gráfica muestran las prevalencias obtenidas en cada uno de los estudios primarios, mientras que el diamante representa la prevalencia media obtenida en el meta-análisis (0.15 o 15%).

En este estudio, para combinar estadísticamente los resultados de los estudios primarios, se optó por un modelo de efectos aleatorios donde a priori no se asume homogeneidad en las magnitudes de las prevalencias, atribuyendo la variabilidad a error de muestreo y a la variabilidad entre los estudios. Para el cómputo de la prevalencia

media se utilizó el programa MetaXL que incluye la posibilidad de ejecutar un modelo de efectos aleatorios ponderando por la calidad de los efectos estimados de los estudios primarios, aspecto recomendable cuando se detecta heterogeneidad en los tamaños del efecto individuales (Barendregt y cols., 2013; Doi y Thalib, 2008).

Además, los *forest plots* elaborados por algunos de los programas estadísticos específicos para ejecutar meta-análisis, como por ejemplo el programa *Review Managener* (RevMan, 2008) y el programa MetaXL (Barendregt y cols., 2013) incorporan los valores de los estadísticos de heterogeneidad (Q y I^2). Sin embargo, otros programas estadísticos no lo hacen, como por ejemplo el *Comprehensive Meta-analysis version* 3.0 (Borenstein y cols., 2014).

En la Figura 1 mostrada anteriormente, la prueba Q de heterogeneidad fue $Q(2) = 33.30$, $p \leq .001$, lo que resulta estadísticamente significativa, y el índice $I^2 = 94\%\%$, reflejando ambos estadísticos una alta heterogeneidad entre los tamaños del efecto de los estudios.

3.6. Limitaciones del meta-análisis

Las principales limitaciones de los estudios de meta-análisis son (Aguinis y cols., 2011; Botella y Sánchez-Meca, 2015; Borenstein y cols., 2009; Egger, Dickersin y Smith 2001; Ellis, 2010; Field y Gillett, 2010; Hopewell, McDonald, Clarke y Egger, 2007; Ioannidis, 2011; Wright y cols., 2007):

(1) Deficiencias de los estudios primarios. La calidad de los datos de un meta-análisis, y por tanto, la fiabilidad de sus conclusiones, puede verse afectada por los datos e información defectuosa de los estudios primarios.

(2) Heterogeneidad entre los estudios primarios incluidos en el meta-análisis. Es una de las críticas más comunes y severas que ha recibido el meta-análisis, conocido como la "mezcla de manzanas y naranjas" ('*mixing apples and oranges*'). Esta crítica se basa en la idea de que el tamaño del efecto medio estimado en trabajos de meta-análisis con estudios primarios muy dispares entre sí posiblemente ignorará las diferencias importantes entre los estudios y será poco informativo.

Sin embargo, como Borenstein y cols. (2009) señalan, un trabajo de meta-análisis puede ser diseñado con el objetivo de analizar la fruta y el efecto diferencial

entre los distintos tipos de fruta. Por tanto, las manzanas, naranjas, peras, melones y otro tipo de frutas aportarían información valiosa.

(3) Calidad de los datos. La mezcla de estudios primarios de buena calidad metodológica con estudios de baja calidad puede afectar a las estimaciones de los efectos y, por tanto, dar lugar a estimaciones sesgadas. Por ello, como señala Sánchez-Meca (2008), la calidad metodológica de los estudios primarios se debe codificar como una variable moderadora más y analizar su posible relación con los tamaños del efecto detectados. Otra opción sería establecer desde el principio normas estrictas de calidad metodológica que deben cumplir los estudios para ser incluidos en el meta-análisis. Sin embargo, como se apuntó más arriba, Ellis (2010) señala diversas razones por las que se debe dudar de excluir estudios sobre la base de su calidad metodológica.

(4) Dependencia estadística. Las técnicas meta-analíticas de análisis de datos asumen la independencia de los datos, por tanto, la inclusión de más de un índice del tamaño del efecto calculado sobre la misma muestra de sujetos atenta contra ese supuesto de independencia de los datos. Este problema surge cuando un mismo estudio primario presenta diferentes índices del tamaño del efecto sobre diferentes variables dependientes. Si se incorporan todas las estimaciones del tamaño del efecto al trabajo de meta-análisis, se incurre en un problema de dependencia que afecta a la validez de la conclusión estadística.

Se han propuesto varias soluciones para resolver este problema (Sánchez-Meca, 2008): (1) obtener un promedio de los tamaños del efecto correspondientes a un mismo estudio, (2) realizar diferentes meta-análisis para cada variable dependiente, y (3) modelar la estructura correlacional entre las variables dependientes.

(5) Sesgo de disponibilidad. Idealmente un meta-análisis debe incluir todos los estudios primarios (publicados en distintas lenguas y no publicados) realizados sobre la temática objeto de estudio. No obstante, como ya se ha apuntado más arriba, por muy exhaustiva que sea la búsqueda de la literatura, nunca será posible localizar todos los estudios potencialmente seleccionables. En consecuencia, la exclusión (no intencionada) por parte del meta-analista de algunas investigaciones pertinentes puede conducir a un sesgo en la estimación del tamaño del efecto, y por tanto, amenazar la validez de los resultados del meta-análisis.

El sesgo de disponibilidad surge cuando las estimaciones del tamaño del efecto obtenidas a partir de los estudios primarios que están fácilmente disponibles difieren de las estimaciones reportadas en estudios que son menos accesibles. El sesgo de disponibilidad se presenta generalmente como resultado de un sesgo de reporte (*reporting bias*), el problema del archivador (*the file drawer problem*), el sesgo de publicación (*publication bias*), y el sesgo de la torre de Babel (*tower of babel bias*) o sesgo del lenguaje (*language bias*). El sesgo de publicación está muy relacionado con el problema del archivador. De hecho, el problema del archivador conduce a un sesgo de publicación.

(a) Sesgo de la Torre de Babel (o sesgo del lenguaje): Los criterios de exclusión de estudios basados en la lengua en la que están escritos los estudios pueden conducir a estimaciones del efecto sesgado y dar lugar al llamado sesgo de la Torre de Babel o sesgo del lenguaje (Borenstein y cols., 2009; Cumming, 2012; Ellis, 2010; Grégoire, Derderian y LeLorier, 1995). Por ejemplo, Grégoire y cols. (1995) revisaron dieciséis estudios de meta-análisis que habían excluido explícitamente los estudios que no estaban publicados en lengua inglesa. Estos autores buscaron estudios publicados en lengua no inglesa que eran relevantes para los distintos meta-análisis, encontrando un estudio escrito en alemán y publicado en una revista suiza que, de haber sido incluido en el trabajo de meta-análisis, habría arrojado un resultado no estadísticamente significativo en el meta-análisis en lugar de una conclusión estadísticamente significativa.

(b) Sesgo de publicación: Se refiriere a la tendencia por parte de investigadores, revisores y editores a presentar o aceptar manuscritos para su publicación en base a la dirección o la fuerza de los hallazgos del estudio. El sesgo de publicación se materializa en una publicación selectiva de estudios que obtienen resultados estadísticamente significativos y/o en la dirección esperada, lo que se traduce en una menor publicación de los estudios cuyos resultados no son estadísticamente significativos y/o no van en la dirección esperada. Por tanto, los estudios no publicados podrían ser sistemáticamente diferentes de los estudios publicados.

Existe evidencia de que el sesgo de publicación existe en Medicina, Biología, Ciencias Sociales y en Psicología (Fanelli, 2012; Francis, 2012a; Ioannidis y Trikalinos, 2007). Por ejemplo, Song y cols. (2010) encontraron "*evidencia empírica de que es más probable que ocurra la publicación de un estudio que muestra resultados*

estadísticamente significativos o "importantes" que la publicación de un estudio que no muestra estos resultados" (p. 19).

Los estudios con resultados estadísticamente significativos suelen ser de dos tipos (Banks, Kepes y Banks, 2012): (1) estudios con grandes tamaños muestrales, por lo que alcanzan la significación estadística independientemente de la magnitud del tamaño del efecto), y (2) estudios con tamaños de muestra pequeños, pero con tamaños del efecto grandes en magnitud y que, por lo tanto, logran la significación estadística a pesar del pequeño tamaño muestral.

Existen múltiples causas que pueden dar lugar al sesgo de publicación Banks, Kepes y Banks, 2012). Sin embargo, el estudio de Song y cols. (2010) señala que el sesgo de publicación se produce principalmente antes de la presentación de los resultados en las conferencias y de la presentación de manuscritos a las revistas, en el sentido de que los investigadores no presentan o envían estudios con resultados no estadísticamente significativos, estando su decisión afectada por la presión de publicar resultados estadísticamente significativos.

El sesgo de publicación de los estudios primarios afecta a la validez de los resultados de los estudios de meta-análisis (Cumming, 2012; Ioannidis, 2011), en tanto en cuanto los trabajos de meta-análisis son revisiones secundarias de estudios primarios, normalmente de los estudios publicados y, los estudios con resultados estadísticamente significativos tienen más probabilidad de ser publicados. Por tanto, un meta-análisis basado en estudios publicados puede estar sesgado (Johnson, 199; Sutton y Higgins, 2008). De hecho, Ioannidis y Trikalinosa (2007) analizaron 8 estudios de meta-análisis de ensayos clínicos con 50 estudios cada uno y encontraron evidencia de una mayor incidencia de estudios primarios con resultados estadísticamente significativos en 6 de 8 estudios de meta-análisis revisados. Por su parte, Ferguson y Brannick (2011) analizaron 91 meta-análisis publicados en la *American Psychological Association* y en la *Association for Psychological Science Journal* y encontraron que el 41% de los estudios de meta-análisis revisados aportaron evidencia del sesgo de publicación.

El sesgo de publicación afecta a la precisión del tamaño del efecto medio estimado, produciendo un tamaño del efecto sobreestimado. Por ello es imprescindible evaluar el posible impacto del sesgo de publicación en los resultados.

Existen diversos métodos para valorar el sesgo de publicación en los estudios de meta-análisis, como por ejemplo el *funnel plot*, el método de Trim-and-fill de Duval y Tweedie, el número de seguridad (*fail-safe number*), etc. (para una revisión más profunda de los diferentes métodos de evaluación del sesgo de publicación y sus limitaciones, el lector interesado puede consultar Rothstein, Sutton, y Borenstein, 2005a).

(I) El ***funnel plot*** (gráfico de embudo) es un método de detección del sesgo de publicación de uso frecuente en las ciencias de la salud (Sterne y cols., 2005). El *funnel plot* es una buena herramienta gráfica de carácter exploratorio que muestra la distribución de los tamaños del efecto de los estudios (Cumming, 2012; Sutton y Higgins, 2008). En concreto es un diagrama de dispersión de las estimaciones del tamaño del efecto, representada en el eje de abscisas (eje X), en función de una medida de la variabilidad del estudio (varianza o error típico) o el tamaño muestral (relacionado con la variabilidad) representada en el eje de ordenadas (eje Y).

En circunstancias normales, es decir, sin sesgo de publicación, la dispersión de los resultados (los puntos en el gráfico) describirá una forma de embudo simétrico en torno al valor poblacional, de ahí que el diagrama de dispersión se llame *funnel plot*. Sin embargo, en presencia de un sesgo de publicación, el gráfico será asimétrico: el lado contrario al efecto esperado mostrará una ausencia de puntos o una densidad menor de la esperada (en contrastes unilaterales que son los más frecuentes). Así pues, el grado de asimetría es una medida del sesgo de publicación en contrastes unilaterales (Botella y Sánchez-Meca, 2015).

En la Figura 3 aparecen los resultados de una simulación que representan 18 estudios con un tamaño del efecto estimado para cada uno de ellos (correlación de Pearson transformada a Z de Fisher).

Los círculos blancos de la figura representan los tamaños del efecto de los estudios primarios incluidos en el meta-análisis (k = 18) y el diamante blanco (en la base) representa el tamaño del efecto medio estimado. Observando los círculos blancos se detecta una asimetría en la forma del *funnel plot* que se interpreta como un indicio de sesgo de publicación.

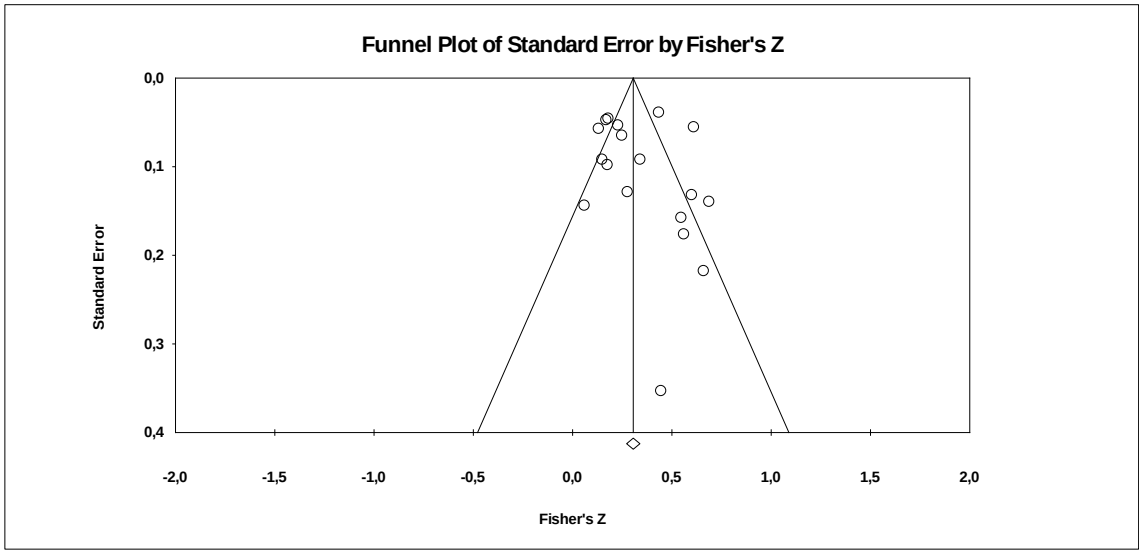


Figura 3. Funnel plot de 18 estudios (estudio de simulación a efectos ilustrativos)

(II) El ***método de Trim-and-fill***: introducido por primera vez por Duval y Tweedie (2000), está diseñado evaluar la simetría del *funnel plot* y ajustar los resultados del meta-análisis por la influencia potencial del sesgo de publicación. Esta técnica corrige la estimación del tamaño del efecto medio calculando un nuevo tamaño del efecto a partir de los tamaños del efecto observados sin tener en cuenta los valores más extremos de la cola asociada al efecto (generalmente la derecha), esto es, recortando los valores extremos. Después rellena el *funnel plot* imputando tamaños del efecto similares a los observados (originales) pero de signo contario (valores espejo). Por último, calcula una estimación del tamaño del efecto medio ajustada, teniendo en cuenta los tamaños del efecto originales (observados) y los tamaños del efecto imputados.

La figura 4 muestra el *funnel plot* corregido por el método de Duval y Tweedie. Los círculos blancos de la figura representan los tamaños del efecto observados (originales, de los estudios primarios), mientras que los círculos negros hacen referencia a los tamaños del efecto imputados que son similares a los valores extremos observados pero de signo contrario (valores espejo). El diamante blanco (en la base) representa el efecto estimado de los estudios originales mientras que el diamante negro representa el efecto medio estimado ajustado, teniendo en cuenta los tamaños del efecto originales (círculos blancos) y los cinco tamaños del efecto imputados (círculos negros). Se observa pues, que el tamaño del efecto medio original (diamante blanco) ha sido ajustado a la baja (diamante negro) por el método *de Trim-and-fill.*

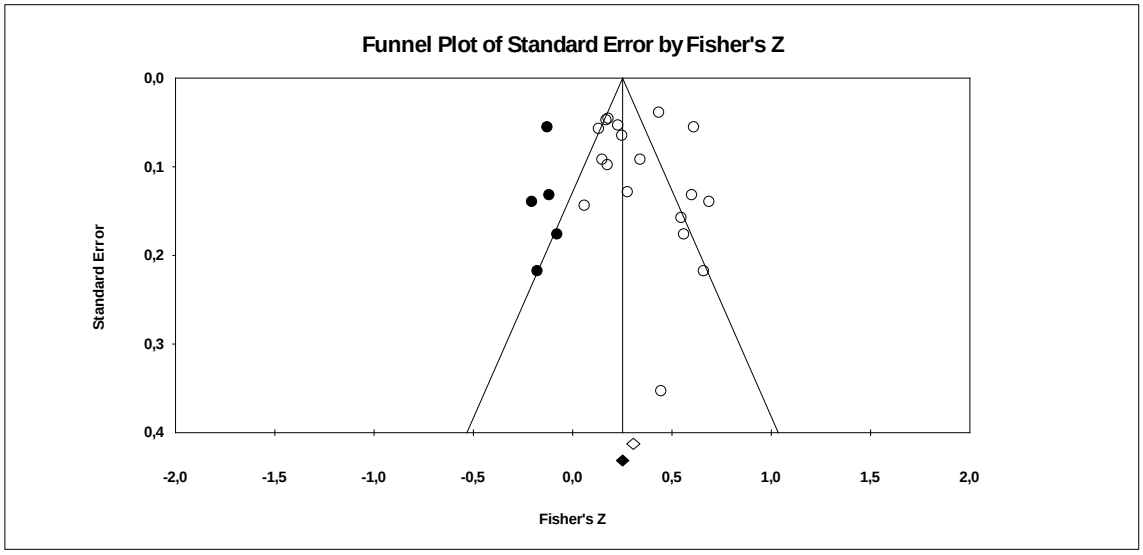


Figura 4. Funnel plot corregido por el método del test de Duval y Tweedie (estudio de simulación a efectos ilustrativos).

La valoración del impacto del sesgo de publicación se basa en la comparación entre el valor del tamaño del efecto medio original y su intervalo de confianza frente al tamaño del efecto medio ajustado y su intervalo de confianza, obtenido con la aplicación del método de Trimm-and-fill (test de Duval y Tweedie, 2000). Por tanto, la diferencia entre el valor del tamaño del efecto medio original y sus intervalos de confianza y el valor del tamaño del efecto medio ajustado, tras la imputación de los tamaños del efecto, provee información sobre la magnitud del sesgo de publicación. Así, una ventaja de este método es que proporciona una estimación de la magnitud del sesgo de publicación (Banks, Kepes y Banks, 2012).

De acuerdo a Rothstein, Sutton y Borenstein (2005b), es probable que no exista sesgo de publicación o que este sea insignificante cuando el tamaño del efecto medio observado y el tamaño del efecto medio ajustado no difieren mucho. Por el contrario, si la diferencia en las estimaciones del tamaño del efecto es notable, pero la conclusión final no cambia sustancialmente, se dice que el sesgo de publicación es moderado. Finalmente, si la diferencia entre las estimaciones del tamaño del efecto medio es tan grande que las conclusiones de la investigación cambian, se dice que el sesgo de publicación es grave.

Kepes, Banks, McDaniel y Whetzel (2012) ofrecieron también una guía para interpretar la magnitud del sesgo de publicación como "insignificante", "moderado" y "grave". En este sentido, Kepes y cols. señalan que se puede considerar ausente o insignificante el sesgo de publicación cuando la diferencia entre el tamaño del efecto medio observado y el tamaño del efecto medio ajustado es inferior al 20%. Por ejemplo, si el tamaño del efecto medio observado es grande (d = 0.80) y se ajusta hacia abajo pasando a ser un tamaño del efecto moderado (d = 0.50) (Banks, Kepes y Banks, 2012). Si la diferencia entre las dos estimaciones del tamaño del efecto medio está entre 20% y 40%, se considera que existe sesgo de publicación moderado. Por último, una diferencia entre los tamaños del efecto medio igual o superior al 40% denota un sesgo de publicación grave. Por ejemplo, cuando un tamaño del efecto medio original es d = 0.20 y se ajusta hacia abajo pasando a ser un tamaño del efecto irrelevante (Banks, Kepes y Banks, 2012).

En el ejemplo anterior, el tamaño del efecto medio original y su intervalo de confianza es r_+ = .32, IC 95% [.24, .40] mientras que el tamaño del efecto medio ajustado tras la imputación de los cinco estudios y su intervalo de confianza es r_+ = .24, IC 95% [.15, 33]. Por lo tanto, siguiendo las indicaciones de Kepes y cols. (2012), el sesgo de publicación es insignificante, pues tanto el tamaño del efecto medio original estimado como el ajustado, hacen referencia a tamaños del efecto moderados. Por lo que, en ambos casos, la conclusiones del estudio no cambian (Rothstein y cols., 2005b).

(III) El **número de seguridad** (***fail-safe number***) se define como el número de estudios que deberían haber quedado sin publicar, guardados en los archivadores, con resultados no estadísticamente significativos y que junto a los estudios encontrados harían que el efecto medio dejase de ser estadísticamente significativo. En general, se considera que existe sesgo de publicación si el número de seguridad es pequeño. Existen diferentes métodos para el cálculo del número de seguridad. En el ejemplo anterior, el número de seguridad es de 1,266 estudios, estimado con el método clásico, lo que significa que para cambiar el sentido de la conclusión se necesita que existan 1,266 estudios no publicados con resultados negativos o con un efecto medio igual a cero. Por otra parte, Rosenthal (1979) señala que los resultados de un meta-análisis son probablemente robustos si el número de seguridad no es más de cinco veces el número de estudios revisados más diez (FSN = 5 * k + 10, donde k es el número de estudios revisados). Esto es, por cada estudio publicado, existen 5 estudios no publicados con resultados

negativos, más un mínimo de 10 (Botella y Sánchez-Meca, 2015). Según esto, el proceso de censura cuando k = 18 podría implicar un total de 100 estudios perdidos (5*18+10=100). Como haría falta que hubieran 1,266 estudios no publicados para cambiar las conclusiones del meta-análisis, se puede establecer que el tamaño del efecto estimado es robusto respecto de la amenaza del sesgo de publicación.

Finalmente, señalar que debido a las limitaciones de los diferentes métodos de evaluación del sesgo de publicación, es aconsejable utilizar múltiples métodos para tener un mayor grado de confianza en la evaluación del impacto de este tipo de sesgo en los resultados del meta-análisis ejecutado (Banks, Kepes, y McDaniel, 2012).

3.7. Valoración de la calidad del meta-análisis

Los autores y lectores de revisiones sistemáticas deben ser conscientes de que es necesario realizar una lectura crítica de los trabajos de meta-análisis ya que las revisiones sistemáticas no están exentas de sesgo. Como Ellis (2010) señala, un buen trabajo de meta-análisis es aquel donde se han identificado las posibles fuentes de sesgo, se han medido sus consecuencias y se han adoptado las estrategias para su mitigación. Por tanto, como se ha dicho, los trabajos de meta-análisis deben ser sometidos a una lectura crítica que permita valorar el rigor del proceso de investigación secundaria realizada y la calidad de sus pruebas. Se ha comprobado que las deficiencias en la ejecución de los trabajos de meta-análisis podrían explicar las discrepancias encontradas en algunas áreas de investigación, cuando se comparan sus resultados con ensayos controlados aleatorios que utilizan tamaños de muestra grandes (LeLorier, Gregoire, Benhaddad, Lapierre y Derderian, 1997).

Existen protocolos de revisión de la calidad metodológica de las revisiones sistemáticas y de los trabajos de meta-análisis, a partir de escalas o listados de comprobación (*Checklist*) ya elaborados, como por ejemplo el AMSTAR *(Assessment of Multiple Systematic Review)* compuesto por 11 ítems (Shea, Grimshaw y cols., 2007), el MOOSE (*Meta-Analysis of Observational Studies in Epidemiology*) formado por 35 ítems (Stroup y cols., 2000); el PRISMA (*Preferred Reporting Items for Systematic reviews and Meta-analysis*) formada por 27 ítems (Moher, Liberati, Tetzlaff, Altman y el grupo PRIMSA, 2009) constituye una actualización del QUORUM (*Quality of Reporting of Meta-analyses*) compuesto por 17 ítems y centrado en los meta-análisis de ensayos clínicos aleatorizados (Moher y cols., 1999), el Protocolo de revisión de las

revisiones sistemáticas y meta-análisis compuesto por 10 ítems (Sánchez-Meca y Botella, 2010), el MARS (*Meta-analysis Reporting Standards*) (APA, 2010a, y el MECIR[7] (*Methodological Expectations of Cochrane Intervention Reviews*) compuesto por 80 ítems (Chandler, Churchill, Higgins, Lasserson y Tovey, 2013) cuyos ítems son conformes a los estándares PRISMA con la excepción del ítem 1 "'Título: identificar el informe como una revisión sistemática, meta-análisis, o ambos" que no está recogida en el MECIR.

El listado de comprobación PRISMA[8] es el instrumento más utilizado actualmente para valorar la calidad del reporte de revisiones sistemáticas cualitativas como de estudios de meta-análisis (Botella y Sánchez-Meca, 2015). PRISMA se centra en la presentación de informes de revisiones sistemáticas que evalúan los ensayos clínicos aleatorizados, pero también se puede utilizar como base para la presentación de informes de revisiones sistemáticas de otros tipos de investigación, en particular las evaluaciones de las intervenciones. Sus ítems hacen referencia a la adecuación del título del estudio, del *abstract*, de la introducción, de la metodología (e.g., criterios de selección, procesos de búsqueda, extracción de datos, definición y cálculo del tamaño del efecto, modelo estadístico asumido para la estimación del tamaño del efecto combinado, análisis del sesgo de publicación, etc.), presentación de resultados, y declaración de conflictos de intereses derivados de la posible financiación del estudio de meta-análisis.

Por su parte, el listado de comprobación AMSTAR ha demostrado tener buenas propiedades psicométricas en términos de validez aparente y de constructo, y de fiabilidad (Shea, Bouter, Peterson y cols., 2007; Shea, Hamel, Wells y cols., 2009). El AMSTAR está dirigido a valorar la adecuación de los métodos utilizados en las diferentes fases del estudio de meta-análisis (ver Tabla 2): búsqueda y selección de estudios, inclusión de estudios no publicados, extracción de las características de los estudios primarios, valoración de calidad de los estudios primarios, métodos estadísticos utilizados para la estimación del tamaño del efecto combinado, comprobación del sesgo de publicación y declaración sobre posibles conflictos de intereses.

[7] Puede obtenerse libremente en el sitio web: http://editorial-unit.cochrane.org/mecir
[8] Puede obtenerse libremente en el sitio web: http://www.prisma-statement.org/

Tabla 2: Lista de ítems del AMSTAR (tomado de Shea y cols., 2007)

1.- ¿Se proporcionó/facilitó un diseño 'a priori'? Los criterios de inclusión y la pregunta de investigación deben establecerse antes de la realización de la revisión
2.- ¿Se hizo por dos revisores independientes la selección de estudios y la extracción de los datos? La selección de estudios y la extracción de datos se debe realizar al menos por dos revisores independientes. En caso de desacuerdo, se debe iniciar un proceso de consenso
3.-¿Se realizó una búsqueda exhaustiva de la literatura? Al menos dos fuentes electrónicas deben ser utilizadas. El informe debe incluir los años y las bases de datos utilizadas (e.g., CENTRAL, EMBASE, y MEDLINE). Deben proporcionarse las palabras clave y/o términos MESH y la estrategia de búsqueda. Todas las búsquedas deben complementarse mediante la consulta de contenidos actuales, opiniones, libros de texto, registros especializados o expertos en el campo particular de estudio, y mediante la revisión de las referencias en los estudios encontrados
4. ¿Se utilizó como criterio de inclusión el estado de la publicación (e.g., literatura gris)? Los autores deben indicar que buscaron informes independientemente de su tipo de publicación. Los autores deben indicar si excluyeron informes (de la revisión sistemática), en función de su estado de publicación, idioma, etc.
5. ¿Se proporcionó una lista de estudios (incluidos y excluidos)? Se debe proporcionar una lista de los estudios incluidos y excluidos.
6. ¿Se proporcionaron las características de los estudios incluidos? En una forma agregada, como una tabla, los datos de los estudios originales deben proveerse de los participantes, las intervenciones y los resultados. Deben reportarse los rangos de las características en todos los estudios analizados e.g., edad, raza, sexo datos socioeconómicos relevantes, estado de la enfermedad, duración, gravedad, u otras enfermedades

(Continuación Tabla 2)

7.- ¿Se evaluó y documentó la calidad científica de los estudios incluidos? Se deben proporcionar métodos de evaluación 'a priori' (p.e. para estudios de efectividad si el/los autor(es) optó por incluir sólo estudios aleatorizados, doble-ciego, de placebo controlados, o encubrimiento/ocultación de la asignación como criterio de inclusión); para otro tipo de estudios serían relevantes ítems alternativos
8.- ¿Se utilizó adecuadamente calidad científica de los estudios incluidos en la formulación de las conclusiones? Los resultados del rigor metodológico y la calidad científica deben ser considerados en el análisis y las conclusiones de la revisión, y explícitamente en la formulación de las recomendaciones.
9.- ¿Fueron apropiados los métodos utilizados para combinar los resultados de los estudios? Para los resultados combinados, una prueba se debe hacer para asegurar que los estudios eran combinables, para evaluar su homogeneidad (es decir, la prueba de Chi cuadrado de homogeneidad, I^2). Se debe utilizar un modelo de efectos aleatorios si existe heterogeneidad y / o la idoneidad clínica de la combinación debe ser tenida en cuenta (es decir, ¿es sensato combinar los datos?).
10. ¿Se evaluó la probabilidad de sesgo de publicación? Una evaluación del sesgo de publicación debe incluir una combinación de ayudas gráficas (e.g., *funnel plot*) y/o pruebas estadísticas (e.g., test de regresión de Egger)
11.- ¿Se indicó la existencia de conflictos de intereses? Las fuentes potenciales de financiación deben ser claramente reconocidas tanto en la revisión sistemática como en los estudios incluidos.

Finalmente, el listado de comprobación MARS incluye recomendaciones sobre el título, redacción del *abstract*, de la introducción, del método, de los resultados y de la discusión. También hace referencia a los conflictos de intereses.

Todos los listados de comprobación citados se pueden utilizar como guía para la realización de estudios de meta-análisis o como guía para valorar la calidad de estudios de meta-análisis realizados por otros autores.

3.8. El meta-análisis en red

Una extensión de la metodología del meta-análisis es el llamado meta-análisis en red (*network meta-analysis*), también conocido como meta-análisis con comparaciones múltiples (*multiple-treatment comparisons*). Los meta-análisis en red permiten obtener estimaciones del efecto relativo (eficacia o seguridad comparada) de los distintos tratamientos a partir de comparaciones indirectas, teniendo en cuenta la «red completa» de estudios disponible (Català-López y Tobías, 2013; Cipriani, Higgins, Geddes, y Salanti, 2013). Para ello, se combinan las comparaciones directas (propias del meta-análisis tradicional) e indirectas entre diversos tratamientos, lo cual aumenta el poder estadístico de las estimaciones generadas (Català-López y cols., 2014).

La figura 5 muestra las comparaciones directas e indirectas de un estudio de meta-análisis en red.

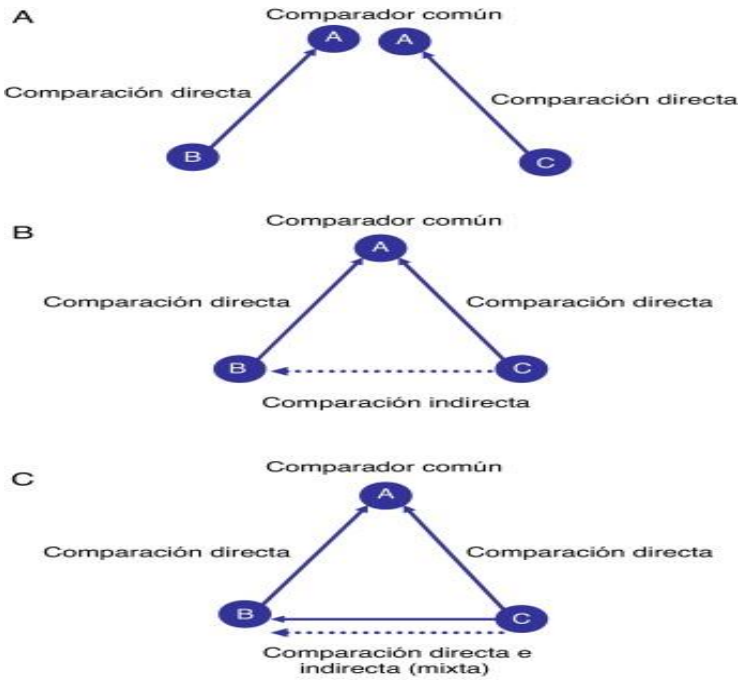


Figura 5 Comparaciones directas e indirectas (tomado de Català-López y cols., 2014, p. 574)

En la sección 1A se muestran dos comparaciones directas, una compara el tratamiento B frente A y la otra compara el tratamiento C frente a A. La sección B de la figura 1 muestra las dos comparaciones directas anteriores y una comparación indirecta de C frente a B, a partir de un comparador común A. Finalmente, la sección C de la figura 1 presenta tres comparaciones directas (comparación de B frente a A, comparación de C frente a A y comparación de C frente a B) y una comparación indirecta (de C frente a B). La combinación de una comparación indirecta y una comparación directa permite obtener una comparación mixta (Català-López y cols., 2014).

Actualmente, los estudios de meta-análisis en red están ganando popularidad entre investigadores, clínicos, editores de revistas, agencias de evaluación y planificadores sanitarios, como fuente de información de interés para conocer la eficacia y seguridad comparada de nuevas intervenciones. Sin embargo, la aplicación de la evidencia proporcionada por las comparaciones indirectas para guiar la toma de decisiones clínicas es, cuando menos, controvertida debido a que la metodología del meta-análisis en red todavía está en sus etapas iniciales (Català-López y cols., 2014).

Por otra parte, ya existen a disposición de los profesionales y consumidores de los estudios de meta-análisis de red listados de comprobación de su calidad, por ejemplo el PRISMA-NMA *(PRISMA for Network Meta-Analyses*), publicado en 2015 (Hutton y cols., 2015) y su versión española en 2016 (Hutton, Catalá-López, y Moher, 2016), el cual constituye una extensión del PRISMA (Moher y cols., 2009). La verificación de la calidad es esencial debido a que los meta-análisis en red están sujetos a las mismas limitaciones que los estudios de meta- análisis tradicionales (e.g., calidad de los estudios primarios, heterogeneidad y sesgos de disponibilidad) y además, las propias de este procedimiento.

En definitiva, los meta-análisis en red suponen una herramienta metodológica interesante, pues como Català-López y cols. (2014) señalan, proporcionan estimaciones del efecto de los tratamientos respecto a las múltiples alternativas disponibles desde un enfoque más completo que el que se venía utilizando tradicionalmente. Sin embargo, todavía están en fase de desarrollo por lo que hay que ser cautos en la interpretación de sus resultados, para lo que también se debe chequear su calidad metodológica.

3.9. Conclusión

El meta-análisis proporciona las herramientas para combinar información de los estudios "repetidos" sobre una temática y puede reducir la dependencia de las pruebas de significación estadística mediante el examen de los estudios replicados.

Por tanto, a diferencia de las pruebas NHST, los estudios de meta-análisis favorecen la acumulación del conocimiento al integrar de forma cuantitativa los resultados obtenidos en un conjunto de investigaciones primarias realizadas (estudios de replicación) sobre una determinada temática, permitiendo la generalización de los resultados. Además, al tener mayor potencia estadística que el procedimiento de la NHST debido al gran tamaño muestral sobre el que trabajan los estudios de meta-análisis, estos pueden detectar efectos pequeños que pueden tener relevancia práctica o clínica, donde la prueba NHST no es capaz.

Finalmente, las pruebas NHST al estar basadas en la significación estadística de los resultados (valor *p*) indican si un tratamiento o intervención es eficaz, es decir, si tiene un efecto distinto de cero, pero no cuantifican dicha magnitud. Los estudios de meta-análisis al estar centrados en los tamaños del efecto y sus intervalos de confianza, sintetizan la magnitud de la eficacia de los tratamientos aportando información sobre la precisión de dicha estimación. Además, el análisis de la variabilidad o heterogeneidad entre los tamaños del efecto de los estudios primarios permite determinar bajo qué condiciones es efectiva la intervención y para qué tipos de personas funciona.

Los profesionales de la salud no solo desean saber si una intervención es eficaz, sino el grado de su eficacia y las condiciones bajo las cuales es eficaz. Los estudios de meta-análisis responden a estas cuestiones. Por ello, se consideran una fuente valiosa de información dentro del modelo de la PBE.

Sin embargo, los trabajos de meta-análisis, tanto tradicionales como en red, están sujetos a deficiencias y a sesgos en sus estimaciones, por lo que es fundamental saber hacer una lectura crítica de un trabajo de meta-análisis, siendo capaz de depurar y valorar su calidad metodológica. Para ello, son útiles las guías como por ejemplo el AMSTAR, a través de la cual se puede hacer una reflexión crítica sobre el alcance de los resultados del meta-análisis tradicional en función de la calidad metodológica del mismo.

De especial importancia es la valoración del sesgo de publicación. El sesgo de publicación afecta tanto a los estudios de meta-análisis tradicionales como en red, a las revisiones sistemáticas cualitativas y a las revisiones narrativas o tradicionales. Respecto a los estudios de meta-análisis, el sesgo de publicación puede afectar a la validez de la estimación del tamaño del efecto medio, produciendo una sobreestimación del mismo, lo que puede cuestionar la eficacia de las intervenciones psicológicas determinada en estudios de meta-análisis con sesgo de publicación.

En este punto, es crucial tener un adecuado conocimiento de las técnicas gráficas y estadísticas diseñadas para valorar su impacto. En consecuencia, conocer e interpretar el gráfico *funnel plot*, método de evaluación del sesgo de publicación más utilizado en las Ciencias de la Salud, es indispensable para implementar correctamente en la práctica profesional los hallazgos de las investigaciones.

Teniendo en cuenta estas cuestiones, disponer de meta-análisis que incluyen las pruebas con mejor calidad supone trabajar con documentación válida, facilitando la toma de decisiones de profesionales e investigadores. De este modo, las actuaciones de los profesionales estarán siempre basadas en la mejor evidencia que exista sobre una cuestión, actuando desde los principios de la PBE.

4. STUDIES ON MISCONCEPTIONS OF THE *P* VALUE

4.1. Justification and purpose

The misconceptions of the *p* value are made based on certain beliefs and attributions about the significance of the results. These beliefs and attributions could influence the methodological behavior of the researcher, affect the decisions of the professional and jeopardize the quality of interventions and the accumulation of valid scientific knowledge. Consequently, knowing how to interpret *p* probability values is a core competence of the professional in Psychology and any discipline where statistical inference is applied. In this way, knowing the prevalence and category of the misinterpretations of the *p* value is important for deciding and planning statistical education strategies designed to rectify, amend incorrect interpretations.

For these reasons, the purpose of this chapter is to detect the statistical reasoning errors that Spanish academic psychologists and Spanish practitioner psychologists make when they are faced with the results of a statistical inference test. To this end, two questions have been analyzed. The first is the extension of the most common misconceptions of the *p* value and the second is the extent to which *p* values are correctly interpreted.

Furthermore, it was carried out a replication study with a sample of Chilean and Italian academic psychologists. Replication is the most objective method for checking if the result of a study is reliable (Asendorpt et al., 2013; Carver 1978; Cumming, 2008; Earp & Trafimow, 2015; Hubbard, 2004; Hubbard & Lindsay, 2008; Kline, 2013; Nickerson, 2000; Stroebe & Strack, 2014; Wilkinson & TFSI, 1999).

To study the prevalence of misconceptions and correct interpretations of the *p* value it was prepared a structured questionnaire that included a set of 10 questions that analyze the interpretations of the *p* value (see Table 3). Specifically, the questions evaluate the fallacies of inverse probability, replication, effect size and clinical or practical significance fallacy, the correct interpretation of the *p* value, and the correct decision in response to a result considered statistically significant.

Table 3 Questionnaire on interpretations of the p-value

Let's suppose that a research article indicates a value of p=0.001 in the results section (alpha=0.05). Mark which of the following statements are true (T) or false (F).

A.-Inverse probability fallacy:

1. The null hypothesis has been shown to be true.

2. The null hypothesis has been shown to be false.

3. The probability of the null hypothesis has been determined (p = 0.001).

4. The probability of the experimental hypothesis has been deduced (p = 0.001).

5. The probability that the null hypothesis is true, given the data obtained, is 0.01.

B.-Replication fallacy:

6. A later replication would have a probability of 0.999 (1-0.001) of being significant.

C.-Effect size fallacy:

7. The value $p < 0.001$ directly confirms that the effect size was large.

D.-Clinical or practical significance fallacy:

8. Obtaining a statistically significant result indirectly implies that the effect detected is important.

D.-Correct interpretation and decision made:

9. The probability of the result of the statistical test is known, assuming that the null hypothesis is true.

10. Given that p = 0.001, the result obtained makes it possible to conclude that the differences are not due to chance.

4.2. Study 1: Sample of Spanish academic psychologists[9]

4.2.1. Method

4.2.1.1. Design and Procedure

It was carried out a cross-sectional study through on-line survey. For this end, the e-mail addresses of academic psychologists were recorded after consulting publicly accessed sources, obtaining a sample framework consisting of 4,066 academics. Potential participants were invited to complete a survey through the use of a CAWI (Computer Assisted Web Interviewing) system. A follow-up message was sent one month later to non-respondents. The data collection was carried out during the 2013 and 2014 academic years.

4.2.1.2. Participants

It was used a non-probabilistic (convenience) sample. The sample initially comprised 472 academic psychologists. Of these 472 participants, 11.44% belonged to private university. Most of them did not respond to questions about interpretation of *p* values and they were eliminated from the analysis (n = 54) Therefore, the final sample was composed of 418 academic psychologists. The mean number of years of the professors at the University was 14.16 years (SD = 9.39, min = 1, max = 40). Men represented 48.56% and women 51.44%. Regarding university departments, 23.44% of the university professors (n = 98) belonged to the area of Personality, Evaluation and Psychological Treatments, 16.03% to the area of Behavioral Sciences Methodology (n = 67), 13.40% to the area of Basic Psychology (n = 56), 17.70% to the area of Social Psychology (n = 74), 6.94% to the area of Psychobiology (n = 29) and 22.49% to the area of Developmental and Educational Psychology (n = 94).

4.2.1.3. Instrument

The survey consisted of two sections. The first one included items related to information about sex and years of experience as an academic psychologist, Psychology knowledge areas, kind of university (public/private). The second section included the set of 10 questions that analyze the interpretations of the *p* value (see Table 3).

[9] This study is published as: Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from Spain. *Psicothema, 27,* 290-295. doi: 10.7334/psicothema2014.283

Finally, the instrument evaluated other questions, such as use and level of knowledge about the statistical terms (e.g., effect size, confidence intervals, meta-analysis), which are analyzed in the following chapter.

4.2.1.4. Data analysis

The analysis included descriptive statistics for the variables under evaluation. To calculate the confidence interval for percentages, we used score methods based on the works of Newcombe (2012). These methods perform better than traditional approaches when calculating the confidence intervals for percentages. In addition, it can be noted that in the published research, confidence intervals for percentages were not reported. These analyses were performed with the statistical program IBM SPSS v. 20 for Windows.

4.2.2. Results

Of the 4,066 academic psychologists who were sent an e-mail with the link to access the survey, 418 filled it out (10.26%). Therefore, the results must be qualified by the low response rate. However, it is possible that the participants who responded to the survey felt more confident about their statistical knowledge than those who did not respond. In this case, these results could underestimate the extension of the fallacies about the *p* value among Spanish academic psychologists in public universities.

Table 4 shows the percentage of responses by participants who endorsed the false statements about the *p* value, according to the Psychology knowledge areas. Regarding the "inverse probability fallacy", the table shows that the majority of the academic psychologists perceived some of the false statements about the *p* value to be true. The participants in the area of Methodology had fewer incorrect interpretations of the *p* value than the rest of the participants.

The false statements that received the most support were "the null hypothesis has been shown to be false" and "the probability of the null hypothesis has been determined ($p = 0.001$)". The percentage of those who rated the 5 statements correctly ranged from 0% for the participants from the area of Psychobiology to 19.57% for the participants from the area of Methodology.

Concerning the "replication fallacy" it can be noted that the majority of the participants correctly evaluated the false statement.

Table 4 Percentage of the misconceptions of the p-value by Psychology knowledge area [and 95% Confidence Intervals]

Ítem	1 *n* = 98	2 *n* = 67	3 *n* = 56	4 *n* = 74	5 *n* = 29	6 *n* = 94	Total *n* = 418
Inverse probability fallacy							
1. The null hypothesis has been shown to be true	8.16	1.49	7.14	5.41	6.90	12.77	7.42
	[4.19, 15.29]	[0.26, 7.98]	[2.81, 16.98]	[2.12, 13.09]	[1.91, 21.96]	[7.46, 21]	[5.27, 10.33]
2. The null hypothesis has been shown to be false	65.31	35.82	60.71	66.22	55.17	61.70	58.61
	[55.47, 73.99]	[25.40, 47.78]	[47.63, 72.42]	[54.88, 75.95]	[37.55, 71.59]	[51.60, 70.89]	[53.83, 63.23]
3. The probability of the null hypothesis has been determined (p = 0.001)	51.02	58.21	67.86	62.16	62.07	56.38	58.37
	[41.27, 60.69]	[46.27, 69.26]	[54.82, 78.60]	[50.77, 72.35]	[44, 77.31]	[46.30, 65.96]	[53.59, 63]

Note. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology.

Table 4 (Continued)

Ítem	1 *n* = 98	2 *n* = 67	3 *n* = 56	4 *n* = 74	5 *n* = 29	6 *n* = 94	Total *n* = 418
Inverse probability fallacy							
4. The probability of the experimental hypothesis has been deduced (p = 0.001)	40.82 [31.61, 50.71]	13.43 [7.23, 23.60]	23.21 [14.10, 35.77]	36.49 [26.44, 47.87]	37.93 [22.69, 56]	43.62 [34.04, 53.70]	33.73 [29.37, 38.39]
5. The probability that the null hypothesis is true, given the data obtained, is 0.01	32.65 [24.18, 42.44]	19.40 [11.71, 30.42]	25 [15.52, 37.69]	31.08 [21.69, 42.34]	41.38 [25.51, 59.26]	36.17 [27.18, 46.25]	30.62 [26.40, 35.20]
% Participants who correctly rate the 5 statements as false	4.08 [1.60, 10.03]	19.40 [11.71, 30.42]	5.36 [1.84, 14.61]	2.70 [0.74, 9.33]	0 [0, 11,70]	4.26 [1.67, 10.44]	6.2

Note. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

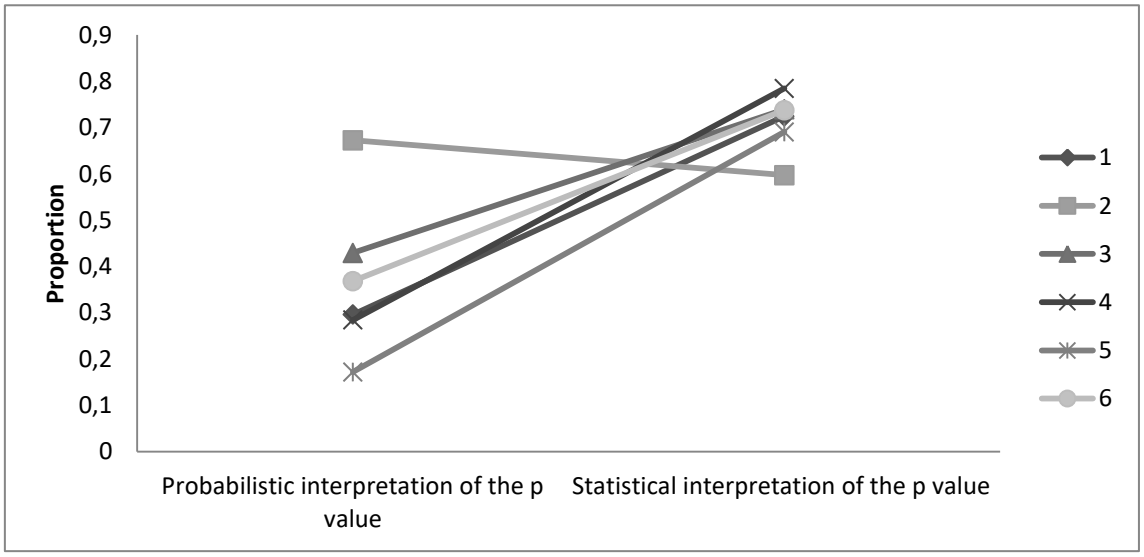
Table 4 (Continued)

Ítem	1 *n* = 98	2 *n* = 67	3 *n* = 56	4 *n* = 74	5 *n* = 29	6 *n* = 94	Total *n* = 418
Replication fallacy							
6. A later replication would have a probability of 0.999 (1-0.001) of being significant.	34.69 [26.01, 44.53]	16.42 [9.42, 27.06]	35.71 [24.46, 48.81]	39.19 [28.86, 50.58]	27.59 [14.70, 45.72]	45.74 [36.04, 55.78]	34.69 [30.28, 39.37]
Effect size fallacy							
7. The value $p = 0.001$ directly confirms that the effect size was large	12.24 [7.15, 20.19]	2.99 [0.82, 10.25]	8.93 [3.87, 19.26]	16.22 [9.53, 26.24]	24.14 [12.22, 42.11]	18.09 [11.61, 27.07]	13.16 [10.25, 16.74]
Clinical/practical significance fallacy							
8. Obtaining a statistically significant result indirectly implies that the effect detected is important	39.80 [30.67, 49.70]	22.39 [14.06, 33.71]	28.57 [18.42, 41.48]	35.14 [25.24, 46.50]	27.59 [14.70, 45.72]	45.74 [36.04, 55.78]	35.17 [30.74, 39.86]

Note. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology.

With regard to "effect size fallacy" and "clinical or practical fallacy", the false statement that received the most support was the one related to the clinical or practical significance of the findings. The percentage of participants who rated both statements correctly ranged from 48.9% in the area of Developmental and Educational Psychology to 76.1% in the area of Methodology.

Finally, Figure 6 shows the proportion of the different groups of participants endorsing each of the two statements referred to correct interpretation and statistical decision adopted on the *p* value. The majority of the participants in the different knowledge areas had problems with the probabilistic interpretation of the *p* value.



Note. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

Figure 6. Proportion of correct interpretation and statistical decision adopted by knowledge area

The interpretation of the *p* value improved when performed in terms of the statistical conclusion, compared to the probabilistic interpretation, except in the academic psychologists from the area of Methodology, where this improvement was not observed. In this case, the professors presented greater problems with the statistical interpretation of the *p* value than with the probabilistic interpretation, but with a smaller difference between the two, compared to the other knowledge areas.

4.2.3. Discussion

The results indicate that the comprehension and correct application of many statistical concepts continue to be problematic among Spanish academic psychologists.

The "inverse probability fallacy" is the most frequently observed misinterpretation. This means that participants confuse the probability of obtaining a result or a more extreme result if the null hypothesis is true ($Pr(Data|H_0)$) with the probability that the null hypothesis is true given some data ($Pr(H_0|Data)$).

The results also indicate that academic psychologists from the area of Methodology are not immune to erroneous interpretations. However, they show fewer problems than their colleagues from other areas. These data are consistent with previous studies (Haller & Krauss, 2002; Lecoutre et al., 2003; Monterde-i-Bort et al., 2010).

The differences between the psychologists from the area of Methodology and those from the rest of the areas in the correct appraisal of the p value can be due to the fact that the probabilistic interpretation requires thinking about the significance of the p value as a conditional probability and a random variable, while the statistical interpretation is only based on the valuation of the p value compared to the alpha value. The results of the statistical programs include the p value and only require the researcher to routinely apply the $p < alpha$ rule. By contrast, the probabilistic interpretation involves statistical reasoning; that is, it means reflecting on the statistical processes involved in the behavior of the p value when the null hypothesis is not rejected (Pfannkuch, & Wild, 2004).

It must be acknowledged several limitations in this study. The low response rate might affect the representativity of the sample and, therefore, the generalization of the results. Furthermore, it should be kept in mind that this study is descriptive. Nonetheless, our results agree with the findings of previous studies (e. g., Gordon, 2001; Haller & Kraus, 2002; Lecoutre et al., 2003; Mittag & Thompson, 2000; Oakes, 1986) indicating the need to adequately train academic psychologists in order to produce valid scientific knowledge and improve professional practice. This training should lead to a better understanding of the p value and correcting the misconceptions committed by academics psychologists when they face the results of significance test.

4.3. Study 2: Sample of Spanish practitioner psychologists[10].

4.3.1. Method

4.3.1.1. Design and Procedure

The data were collected from a cross-sectional on-line survey of Spanish psychologists. We send an e-mail to Spanish Psychological Associations inviting them to participate in the on-line survey on professional practice in Psychology. Potential participants were invited to complete a survey through the use of a CAWI (Computer Assisted Web Interviewing) system. A follow-up message was sent three weeks later. The data collection was performed from May to September 2015.

4.3.1.2. Participants

It was used a non-probabilistic (convenience) sample. The sample was initially made up of 113 Spanish psychologists. Of these, 68.1% were practitioner psychologists, 28.3% were academic psychologists, 0.9% were researchers, and 2.7% reported other role. Since the objective of the study was to analyze the barriers to evidence-based practice, participants who were not practitioners psychologists were eliminated from the sample (n= 36).

The final sample consisted of 77 Spanish psychologists with an average age of 41.44 years (*SD* = 9.42, min = 25, max = 64). Of the 77 participants, 31.2% were men and 68.8% were women. The mean number of years as member of Spanish Psychological Associations was 13.73 years (*SD* = 9.30, *min* = 0, max = 33). The mean degree of familiarity with EBP was 4.65 (*SD* = 2.18, *min* = 1, *max* = 7). With regard to education degree, 27.27% of the participants had bachelor's degree, 46.75% Master's degree, and 25.97% Ph.D. Regarding the clinical setting where the participants worked, 38.96% of them worked in public setting (61.04% private setting).

Finally, Mann Whitney U test indicated that there were not statistically significant differences between men and women for the variable number of years as member as psychologist (z = -1.29, p = .196, r = -.15, 95% CI [-.35, .08]) and for the

[10] This study is under review as: Badenes-Ribera, L. Bonilla-Campos, A. & Frías-Navarro, D. (2016). Barriers to Evidence Based Practice in Spanish practitioner psychologists: An exploratory study.

variable degree of familiarity with Evidence Based Practice approach (z = -0.07, p = .942, r = -.01, 95% CI [-.23, .21]).

Furthermore, the results of a oneway ANOVA showed that there were not statistically significant differences on mean scores for the degree of familiarity with evidence-based practice approach by education level of the participants (F(2,74) = 2.16, p = .123, Cohen's d = 0.4 , 95% CI [-0.13, 0.98]).

4.3.1.4. Data analysis

The analysis included descriptive statistics for the variables under evaluation. To calculate the confidence interval for percentages, we used score methods based on the works of Newcombe (2012). These methods perform better than traditional approaches when calculating the confidence intervals of percentages. These analyses were performed with the statistical program IBM SPSS v. 20 for Windows.

4.3.2. Results

Table 5 presents the percentage of responses by participants who endorse the eight false statements about the p values. It is noteworthy that more than a third of the participants had problems with "clinical or practical significance fallacy". These participants stated that p-value indicate the importance of the findings. Consequently, they linked the statistical significance of the findings with their clinical or practical significance.

Regarding the "inverse probability fallacy", the majority of the psychologists perceived some of the false statements about the p value to be true.

Finally, the majority of the participants had problems with the probabilistic interpretation of the p value like the Spanish academic psychologists, except those from the area of Methodology who presented greater problems with the statistical interpretation of the p value than with the probabilistic interpretation. That is, the interpretation by practitioner psychologists improved when performed in terms of the statistical conclusion, compared to the probabilistic interpretation of the p value. There was not overlapping between confidence intervals; therefore, the difference between percentages was statistically significant.

Table 5 Percentage of participants who endorsed the statements

Ítem	*n*	%	95% CI
Inverse probability Fallacy			
1. The null hypothesis has been shown to be true	15	19.48	[12.18, 29.69]
2. The null hypothesis has been shown to be false.	15	19.48	[12.18, 29.69]
3. The probability of the null hypothesis has been determined (p =0.001))	8	10.39	[5.36, 19.18]
4. The probability of the experimental hypothesis has been deduced (p = 0.001).)	8	10.39	[5.36, 19.18]
5. The probability that the null hypothesis is true, given the data obtained, is 0.01.	15	19.48	[12.18, 29.69]
% Participants who not endorsed the five false statements	0	0	[0, 4.75]
Replication fallacy			
6. A later replication would have a probability of 0.999 (1-0.001) of being statistically significant.	3	3.90	[1.33, 10.84]
Effect size fallacy			
7. The value p = 0.001 directly confirms that the effect size was large.	8	10.39	[5.36, 19.18]
Clinical/practical significance fallacy			
8. Obtaining a statistically significant result indirectly implies that the effect detected is important.	28	36.36	[26.51, 47.52]
Correct interpretation of the *p* value			
9. The probability of the result of the statistical test is known, assuming that the null hypothesis is true	6	7.79	[3.62, 15.98]
10. Given that p = 0.001, the result obtained makes it possible to conclude that the differences are not due to chance	36	46.75	[36.03, 57.78]
% Participants who endorsed the two correct statements	1	1.30	[0.02, 7]

Note. CI = confidence interval.

4.3.3. Discussion

The findings indicate that the comprehension of many statistical concepts continues to be problematic among Spanish practitioner psychologists. Interpreting a statistically significant result as important or useful, confusing the alpha's significance level with the probability that the null hypothesis is true, relating p-value to effect size, and believing that the probability of replicating a result is 1-*p* are erroneous or false interpretations that continue to exist among Spanish practitioner psychologists.

The "clinical or practical significance fallacy" was the most frequently observed misinterpretation. Nevertheless, a statistically significant result does not indicate that the result is important, in the same way that a non-statistically significant result might still be important (Nickerson, 2000). Clinical significance refers to the practical or applied value or importance of the effect of an intervention. That is, whether it makes any real (e.g., genuine, palpable, practical, noticeable) difference to the clients or to others with whom they interact in everyday life (Kazdin, 1999; 2008).

The correct interpretation of the *p* value was improved when performed in terms of statistical conclusion, compared to the probabilistic interpretation of the *p* value. This might be due to the fact that the probabilistic interpretation requires thinking about the significance of the *p* value as a conditional probability and a random variable, which means reflecting on the statistical processes involved in the behavior of the p value when the null hypothesis is not rejected. While the statistical interpretation is only based on the valuation of the *p* value, provided by the outputs of statistical programs, compared to the alpha value and, thus, only requires the researcher to routinely apply the $p <$ alpha rule (Pfannkuch & Wild, 2004).

The methodological errors and the poor methodological knowledge have been and continue to be a source of direct threat to properly implement the EBP in professional practice. Statistical significance tests have a purpose, and respond to some problems and not to others (Perezgonzalez, 2015).

A statistical significance test does not talk about result importance, replicability, or even the probability that a result was due to chance (Carver, 1978). *P*-value informs us whether an effect exists, but it does not reveal the size of the effect, or the clinical/practical significance of the effect (Sullivan & Feinn, 2012). The effect size can

only be determined by directly estimating its value with the appropriate statistic and its confidence interval (Cohen, 1994; Cumming, 2012; Kline, 2013).

Finally, several limitations should be acknowledged in the present study. The low response rate might affect the representativity of the sample and, therefore, the generalizability of the findings among practitioner psychologists. Nevertheless, it is possible that the participants who responded to the survey felt more confident about their statistical knowledge than those who did not respond. Should this be the case, the results might underestimate the barriers to EBP. In addition, the results agree with the findings of previous studies on misconceptions of the *p*-value in samples of academic psychologists and undergraduates of Psychology degree (Badenes-Ribera et al., 2015; Badenes-Ribera, Frías-Navarro; Bryan, Bonilla-Campos, & Longobardi, 2016; Badenes-Ribera, Frías-Navarro & Pascual-Soler, 2015; Falk & Greenbaum, 1995; Haller & Krauss, 2002, Kühberger et al., 2015; Monterde-i-Bort et al., 2010; Oakes, 1986).

All of this leads to point out the need to adequately train Psychology professionals in order to improve the professional practice. EBP requires professionals to critically evaluate the findings of psychological research (Daset & Cracco, 2013). To be able to do so, training is necessary in statistical concepts, research design methodology, and results of statistical inference tests and meta-analytic studies.

4.4. Study 3: A replication study from Chile and Italy[11]

4.4.1. Justification and purpose

As it was said before, replication is the most objective method for checking if the result of a study is reliable and this concept plays an essential role in the advance of scientific knowledge (Asendorpt et al., 2013; Carver 1978; Cumming, 2008; Earp & Trafimow, 2015; Hubbard, 2004; Hubbard & Lindsay, 2008; Kline, 2013; Nickerson, 2000; Stroebe & Strack, 2014; Wilkinson & TFSI, 1999). "*Replication includes repetitions by different researchers in different places with incidental or deliberate changes from the original experiment*" (Cumming, 2008, p. 287).

Previous studies have shown that the fallacies about p values are common among academic psychologists and undergraduates students majoring in Psychology from several countries (Germany, USA, Spain, Israel). However, nothing is known about the extension of these misinterpretations in Chile (a Latin-American country) and Italy (another country from Europe Union). Consequently, research on misconceptions of the *p* values in these countries is useful to improve our current knowledge about the extension of the fallacies among academic psychologists. Furthermore, the present study is part of a cross-cultural research project between Spain and Italy about statistical cognition, and it is framed within the line of research on cognition and statistical education that our research group has been developing for many years.

4.4.2. Method

4.4.2.1. Design

This work is a replication of the study by Badenes-Ribera et al. (2015) which analyzed the extent of the most common misconceptions of *p* value and two correct interpretations of this among Spanish academic psychologists. We modified tree aspects from the original research: the answer scale format of the instrument for measuring misconceptions of *p* values, the geographical areas of the participants (Italian and Chilean academic psychologists) and the moment in time.

[11] This study is published as: Badenes-Ribera, L., Frías-Navarro, D., Bryan, I., Bonilla-Campos, A. & Longobardi, C. (2016). Misconceptions of the p value among Chilean and Italian academic psychologists. *Frontiers in Psychology, 7*, doi: 10.3389/fpsyg.2016.01247

In the original study, the instrument included a set of 10 questions that analyzed the interpretations of *p* value. The questions were posed using the following format in the heading: "Suppose that a research article indicates a value of $p = 0.001$ in the results section (alpha = 0.05) […]" and the participants had to mark which of the statements were true or false. Therefore, the response scale format was dichotomous. In the present study, we changed the response scale format, and the participants could only indicate which answers were true instead of being forced to choose between a true or false option. This response scale format ensures that answers given by the subjects have a higher level of confidence, but on the other hand it does not allow knowing if the items that were not chosen are considered false or "does not know".

The second modification was the geographic area and the moment in time. The original research was conducted in Spain during 2013-2014, while the present study was carried out in Chile and Italy in 2015.

4.4.2.2. Procedure

It was conducted a cross-sectional study through on-line survey. In order to do this, the e-mail addresses of academic psychologists were found by consulting the websites of Chilean and Italian universities, resulting in 2,321 potential participants (1,824 Italians, 497 Chilean). Potential participants were invited to complete a survey through the use of a CAWI (Computer Assisted Web Interviewing) system. A follow-up message was sent two weeks later to non-respondents. The data collection was performed from March to May 2015.

Individual informed consent was also collected from academics along with written consent describing the nature and objective of the study according to the ethical code of the Italian Association for Psychology (AIP). The consent stated that data confidentiality would be assured and that participation was voluntary.

The questions were administered in Italian and Spanish language respectively. Original items were in Spanish, and therefore all of them were translated into Italian by applying the standard back-translation procedure, which implied translations from Spanish to Italian and vice versa (Balluerka, Gorostiaga, Alonso-Arbiol, & Haranburu, 2007).

Thirty participants (25 Italian and 5 Chilean) did not respond to questions about *p* value misconceptions and were therefore removed from the study. The response rate was 7.07% (Italian 7.35%, Chilean 6.04%).

4.4.2.3. Participants

It was used a non-probabilistic (convenience) sample. The sample initially comprised 194 academic psychologists from Chile and Italy. Of these 194 participants, thirty did not respond to questions about misconceptions of *p* values and were removed from the analysis. Consequently, the final sample consisted of 164 academic psychologists; 134 of them were Italian and 30 were Chilean. Table 6 presents a description of the participants.

Of the 134 Italian participants, 46.3% were men and 53.7% were women, with a mean age of 48.35 years (*SD* = 10.65, min = 28, max = 83). The mean number of years that the professors had spent in academia was 13.28 years (*SD* = 10.52, min = 1, max = 46).

Of the 30 Chilean academic psychologists, men represented 50% of the sample. In addition, the mean age of the participants was 44.50 years (*SD* = 9.23, min = 30, max = 69). The mean number of years that the professors had spent in academia was 15.53 years (*SD* = 8.69, min = 4, max = 41).

4.4.2.4. Data analysis

The analysis included descriptive statistics for the variables under evaluation. To calculate the confidence interval for percentages, we used score methods based on the works of Newcombe (2012). These methods perform better than traditional approaches when calculating the confidence intervals of percentages.

These analyses were performed with the statistical program IBM SPSS v. 20 for Windows.

Table 6. Description of the participants

	Chile (*n* = 30)		Italy (*n* = 134)	
	n	%	*n*	%
Sex				
Men	15	50	62	46.27
Women	15	50	72	53.73
Psychology knowledge areas				
Development and Educational Psychology	7	23.33	25	18.66
Clinical and Dynamic Psychology	7	23.33	23	17.16
Social Psychology	5	16.67	22	16.42
Methodology	5	16.67	13	9.7
Neuropsychology	1	3.33	14	10.45
Work and Organizational Psychology	3	10	10	7.46
General Psychology	2	6.67	27	20.15
Type of University				
Public	13	43.33	116	86.57
Private	17	56.67	18	13.43
Have you been reviewers for scientific journals in the last year?				
Yes	17	56.67	115	85.82
No	13	43.33	19	14.18

4.4.3. Results

Table 7 presents the percentage of responses by participants who endorse the eight false statements about the *p* values, according to the Psychology knowledge areas and nationality of the participants.

Regarding the "inverse probability fallacy", the majority of the Italian and Chilean academic psychologists perceived some of the false statements about the *p* value to be true, like in the study of Badenes-Ribera et al. (2015).

Overall, the false statement that received the most support was "The null hypothesis has been shown to be false". By sample, Italian participants encountered the biggest problems with the false statement "The probability of the null hypothesis has been determined ($p = 0.001$)", while Chilean participants with "The null hypothesis has been shown to be false".

The participants in the area of Methodology made fewer incorrect interpretations of *p* values than the rest of the participants. There were, however, overlaps among the confidence intervals; therefore, the differences between percentages were not statistically significant. In addition, Italian methodologists presented more problems than their Chilean peers recognizing the false statements "The probability of the null hypothesis has been determined ($p = 0.001$)" and "The probability that the null hypothesis is true, given the data obtained, is 0.001", although there were overlaps among the confidence intervals. Consequently, the differences among the percentages were not statistically significant.

Concerning the "replication fallacy", as Table 7 shows, the majority of the participants (87.80%) correctly evaluated the false statement, like in the studies of Badenes-Ribera et al. (2015) with 65.3% of correct answers and Haller and Krauss (2002) with 51% of correct answers. The participants in the area of Methodology had fewer incorrect interpretations of the *p* value than the rest also in this case, but there were overlaps among the confidence intervals; therefore, the differences between percentages were not statistically significant.

Table 7. Percentage of participants who endorsed the false statements [and 95% Confidence Intervals]

Item	Chile (*n* = 30)				Italy (*n* = 134)				Total (N = 164)					
	Methodology (*n* = 5)		Other knowledge areas (*n* = 25)		Methodology (*n* = 13)		Other knowledge areas (*n* = 121)		Methodology (*n* = 18)		Other knowledge areas (*n* = 146)		Total (N = 164)	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Inverse probability fallacy														
1. The null hypothesis has been shown to be true	0	0 [0, 43.45]	1	4 [0.71, 19.54]	0	0 [0, 22.81]	5	4.13 [1.78, 9.31]	0	0 [0, 17.59]	6	4.11 [1.90, 8.68]	6	3.66 [1.69, 7.75]
2. The null hypothesis has been shown to be false	2	40 [11.76, 76.93]	15	60 [40.74, 76.60]	3	23.08 [8.18, 50.26]	34	28.10 [20.86, 36.69]	5	27.78 [12.50, 50.87]	49	33.56 [26.41, 42.56]	54	32.93 [26.20, 40.44]
3. The probability of the null hypothesis has been determined (p = 0.001)	1	20 [3.62, 62.45]	3	12 [4.17, 29.96]	4	30.77 [12.68, 57.63]	31	25.62 [18.68, 34.06]	5	27.78 [12.50, 50.87]	34	23.29 [17.17, 30.77]	39	23.78 [17.91, 30.85]

Table 7 (Continued)

Item	Chile (*n* = 30)				Italy (*n* = 134)				Total (N = 164)					
	Methodology (*n* = 5)		Other knowledge areas (*n* = 25)		Methodology (*n* = 13)		Other knowledge areas (*n* = 121)		Methodology (*n* = 5)		Other knowledge areas (*n* = 25)		Methodology (*n* = 13)	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
4. The probability of the experimental hypothesis has been deduced (*p* = 0.001)	0	0 [0, 43.45]	4	16 [6.40, 34.65]	1	7.69 [1.37, 33.31]	15	12.40 [7.66, 19.45]	1	5.56 [0.99, 25.76]	19	13.01 [8.49, 19.43]	20	12.20 [8.03, 18.09]
5. The probability that the null hypothesis is true, given the data obtained, is 0.01	0	0 [0, 43.45	2	8 [2.22, 24.97]	3	23.08 [8.18, 50.26]	17	14.05 [8.96, 21.35]	3	16.67 [5.84, 39.22]	19	13.01 [8.49, 19.43]	22	13.41 [9.03, 19.48]
Participants who not endorse the 5 false statements	3	60 [23.07, 88.24]	7	28 [14.28, 47.58]	5	38.46 [17.71, 64.48]	48	39.67 [31.40, 48.57]	8	44.44 [24.56, 66.28]	55	37.67 [30.22, 45.75]	63	38.41 [31.32, 46.04]

Table 7 (Continued)

Item	Chile (*n* = 30)				Italy (*n* = 134)				Total (N = 164)			
	Methodology (*n* = 5)		Other knowledge areas (*n* = 25)		Methodology (*n* = 13)		Other knowledge areas (*n* = 121)		Methodology (*n* = 5)		Other knowledge areas (*n* = 25)		Methodology (*n* = 13)	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Replication fallacy														
6. A later replication would have a probability of 0.999 (1-0.001) of being significant.	0	0 [0, 43.45]	5	20 [8.86, 39.13]	1	7.69 [1.37, 33.31]	14	11.57 [7.02, 18.49]	1	5.56 [0.99, 25.76]	19	13.01 [8.49, 19.43]	20	12.20 [8.03, 18.09]
Effect size fallacy														
7. The value *p* = 0.001 directly confirms that the effect size was large	0	0 [0, 43.45]	0	0 [0, 13.32]	1	7.69 [1.37, 33.31]	7	5.79 [2.83, 11.46]	1	5.56 [0.99, 25.76]	7	4.79 [2.34, 9.57]	8	4.88 [2.49, 9.33]

Table 7 (Continued)

Item	Chile (*n* = 30)				Italy (*n* = 134)				Total (N = 164)			
	Methodology (*n* = 5)		Other knowledge areas (*n* = 25)		Methodology (*n* = 13)		Other knowledge areas (*n* = 121)		Methodology (*n* = 5)		Other knowledge areas (*n* = 25)		Methodology (*n* = 13)	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Clinical or practical fallacy														
8. Obtaining a statistically significant result indirectly implies that the effect detected is important	0	0 [0, 43.45]	2	8 [2.22, 24.97]	1	7.69 [1.37, 33.31]	11	9.09 [5.15, 15.55]	1	5.56 [0.99, 25.76]	13	8.90 [5.28, 14.64]	14	8.54 [5.15, 13.82]

Regarding the percentage of participant responses that endorsed the false statements about the *p* value as an effect size indicator and as having clinical or practical significance, it is noteworthy that only a limited percentage of the participants believed that small *p* values indicates that the results are important and that the *p* values indicate effect size. These findings are in line with the results of the study of Badenes-Ribera et al. (2015). By sample, Chilean participants presented fewer misconceptions than Italian participants, however, there were overlaps among the confidence intervals; thus, the differences between percentages were not statistically significant.

Figure 7 shows the percentage of participants in each group who endorse some of the false statements in comparison to the studies of Oakes (1986), Haller and Krauss (2002), and Badenes-Ribera et al. (2015). The number of statements (and the statements themselves) posed to the participants differed across studies, and this should be borne in mind. The study by Oakes and that of Haller and Krauss presented the same six wrong statements to the participants. In the study of Badenes-Ribera et al. and the current study, the same eight false questions were presented.

Overall it is noteworthy that despite the fact that 30 years have passed since the Oakes' original study (1986) and 14 years since the study of Haller and Krauss (2002), and despite publication of numerous articles on the misconceptions of *p* values, most of the Italian and Chilean academic psychologists do not know how to correctly interpret *p* values.

Finally, Figure 8 shows the percentage of participants who endorsed each of the two correct statements about *p* value interpretation. It can be noted that the majority of academic psychologists, including participants from Methodology area, had problems with the probabilistic interpretation of the *p* value, unlike the study of Badenes-Ribera (2015) where the Methodology instructors showed more problems with the interpretation of *p*-value in terms of the statistical conclusions (results not shown).

In addition, in all cases the interpretation of *p* values improved when performed in terms of the statistical conclusions, compared to their probabilistic interpretation.

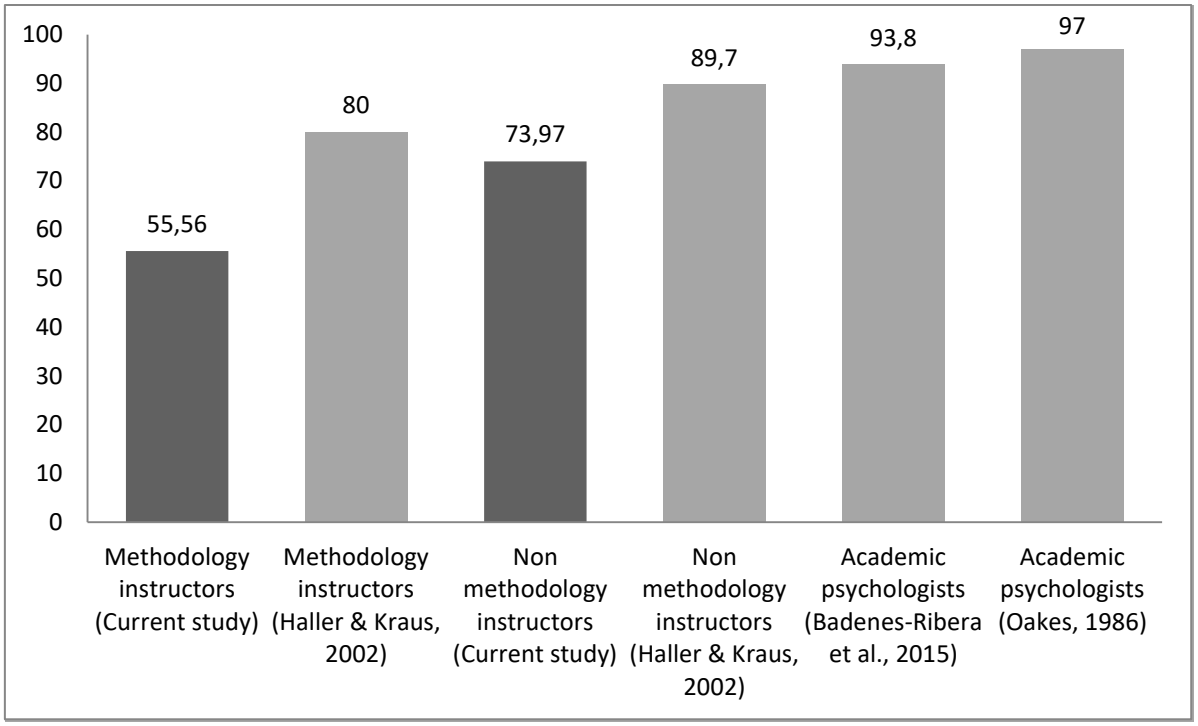


Figure 7. Percentages of participants in each group who endorse at least one of the false statements in comparison to the studies of Badenes-Ribera et al. (2015), Haller and Krauss (2002), and Oakes (1986).

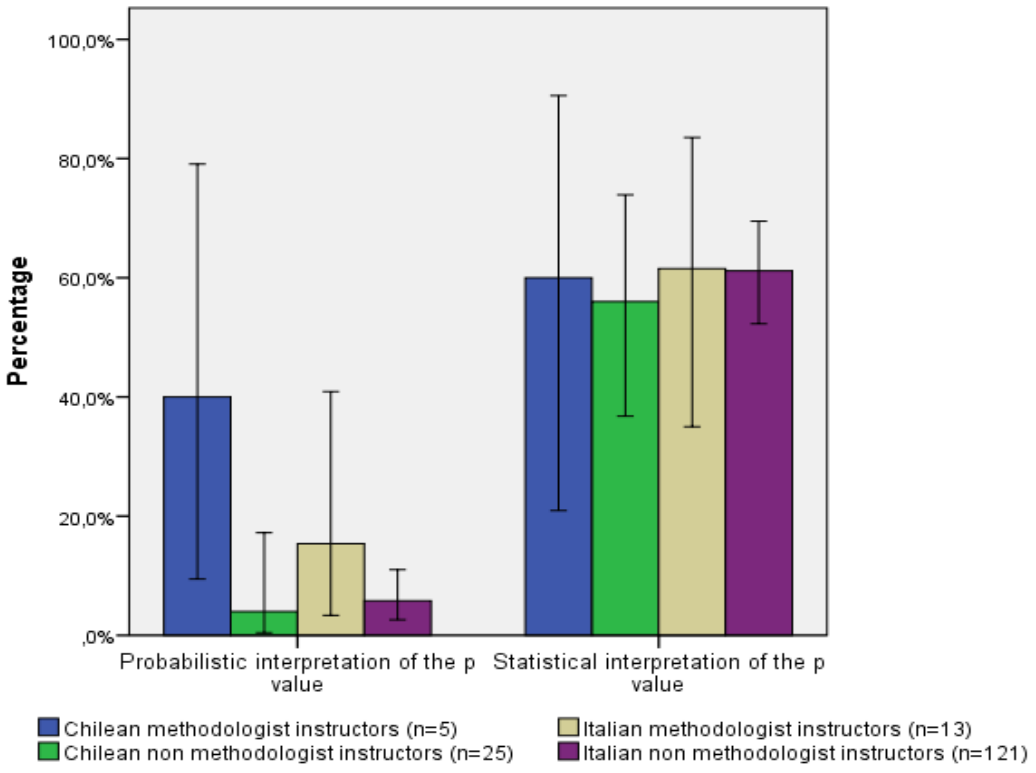


Figure 8. Percentage of correct interpretation and statistical decision adopted broken down by knowledge area and nationality

4.4.4. Discussion

This work is a replication of the study by Badenes-Ribera et al. (2015) which modified aspects of the original research design. In particular, we changed the answer scale format of the instrument for identifying misconceptions of *p* values, the geographical area of the participants and the moment in time. The geographical area of the participants helps generalize the findings of the previous study to other countries.

Firstly, it is noteworthy that our answer scale format obtained a lower rate of misinterpretation of *p* values than the original. Nevertheless, the difference between the studies might also be caused by a difference between countries.

In addition, the findings indicate that the comprehension and correct application of many statistical concepts continue to be problematic among Chilean and Italian academic psychologists.

As in the original research, the "inverse probability fallacy" was the most frequently observed misinterpretation, followed by the "replication fallacy" in both samples. This means that some Chilean and Italian participants confuse the probability of obtaining a result or a more extreme result if the null hypothesis is true ($Pr(Data|H_0)$) with the probability that the null hypothesis is true given some data ($Pr(H_0|Data)$). In addition, not rejecting the null hypothesis does not imply its truthfulness. Thus, it should never be stated that the null hypothesis is "accepted" when the *p* value is greater than alpha; the null hypothesis is either "rejected" or "not rejected" (Badenes-Ribera et al., 2015; Wilkinson & TFSI, 1999).

On the other hand, the results also indicate that academic psychologists from the area of Methodology are not immune to erroneous interpretations, and this can hinder the statistical training of students and facilitate the transmission of these false beliefs, as well as their perpetuation (Haller & Krauss, 2002; Kirk, 2001; Kline, 2013; Krishnan & Idris, 2014). This data is consistent with previous studies (Haller & Krauss, 2002; Lecoutre et al., 2003; Monterde-i-Bort et al., 2010) and with the original research from Spain (Badenes-Ribera et al., 2015). Nevertheless, we have not found differences between academic psychologists from the area of Methodology and those from the other areas in the correct evaluation of p values. It should however be borne in mind that lack of statistically significant differences does not imply evidence of equivalence. Furthermore, sample sizes in some sub-groups are very small (e.g. n = 5), yielding

confidence intervals that are quite large. Thus, considerable overlaps of CIs for percentages are unsurprising (and not very informative).

The interpretation of p values improved in all cases when performed in terms of the statistical conclusion, compared to the probabilistic interpretation alone. These results are different from those presented in the original study (Badenes-Ribera et al., 2015), but again the difference between the studies might be due to a difference between countries. In Spain, for example, literature about misconceptions of p values has been available for 15 years (Monterde-i-Bort et al., 2006; Monterde et al., 2010; Pascual et al., 2000), and this research can possibly be known by Methodology instructors. Therefore, it is possible that Spanish Methodologists are more familiar with these probabilistic concepts than their Italian counterparts. In this sense, language barriers may have posed a problem. Nevertheless, all literature published in Spanish was readable by Chilean researchers as well. Thus, lack of available literature cannot explain potential differences found between Spanish and Chilean researchers. In this last case, it is possible that differences exist in academic training between Spanish and Chilean Methodologists.

On the other hand, it can be noted that if participants consider the statement about probabilistic interpretation of *p* values unclear, that may explain, at least partially, why it was endorsed by fewer lecturers than the statement on statistical interpretation of the *p* value, both in the original and current samples. Future research should expand on this question using other definitions of *p* value, such as adding further questions like "the probability of witnessing the observed result or a more extreme value if the null hypothesis is true".

It must be acknowledged some limitations of this study that need to be mentioned. Firstly, the results must be qualified by the low response rate, because of the 2,321 academic psychologists who were sent an e-mail with the link to access the survey, only 164 took part (7.07%). The low response rate could affect the representativity of the sample and, therefore, the generalizability of the results. Moreover, it is possible that the participants who responded to the survey had higher levels of statistical knowledge than those who did not respond, particularly in the Chilean sample. Should this be the case, the results might underestimate the extension of the misconceptions of *p* values among academic psychologists from Chile and Italy. Furthermore, it must also be acknowledged that some participants do not use

quantitative methods at all in their research. These individuals may have been less likely to respond, as well.

Another limitation of our study is the response format. By not asking to explicitly classify statements as either true of false, it is not possible to differentiate omissions from items identified as false. A three-response format (True/False/Don't know) would have been far more informative since this would have also allowed to identify omissions as such.

Nevertheless, the results of the present study agree with the findings of previous studies in samples of academic psychologists (Badenes-Ribera et al., 2015; Haller & Krauss, 2002; Monterde-i-Bort et al., 2010; Oakes, 1986), statistics professionals (Lecoutre et al., 2003), psychology undergraduate students (Badenes-Ribera, Frías-Navarro & Pascual-Soler, 2015; Falk & Greenbaum, 1995; Haller & Krauss, 2002) and members of the American Educational Research Association (Gordon, 2001; Mittag & Thompson, 2000).

4.5. Overall discussion, conclusion and methodological recommendations

Overall, the results of three studies indicate that the comprehension and correct application of many statistical concepts continue to be problematic among Spanish, Chilean and Italian academic psychologists and Spanish practitioner psychologists.

Among academic psychologists, the "inverse probability fallacy" is the most frequently observed misinterpretation. This means that academic psychologists confuse the probability of obtaining a result or a more extreme result if the null hypothesis is true ($Pr(Data|H_0)$) with the probability that the null hypothesis is true given some data ($Pr(H_0|Data)$). As Kirk (1996) points out, statistical inference tests do not respond to what researchers want to know. When researchers perform a statistical inference test, they want to find out the probability that the Null Hypothesis (H_0) is true, given certain data (D), that is, $Pr(H_0|Data)$. However, the statistical inference tests indicate the probability of obtaining a result or a more extreme result if the null hypothesis is true (*op. cit*., p. 747). Not rejecting the null hypothesis does not imply the truth of the null hypothesis. For this reason, it should never be stated that the null hypothesis is

"accepted" when the *p* value is greater than alpha; the null hypothesis is rejected or not rejected (Palmer & Sesé, 2013; Wilkinson & TFSI, 1999).

The findings also indicate that academic psychologists from the area of Methodology are not immune to erroneous interpretations. However, they show fewer problems than their colleagues from other areas. These data are consistent with previous studies (Haller & Krauss, 2002; Lecoutre et al., 2003). The fact that academic psychologists from the area of Methodology erroneously interpret the *p* value hinders the students' statistical training and facilitates the transmission of these false beliefs, as well as their perpetuation (Haller & Krauss, 2002; Kirk, 2001; Kline, 2013). It is, therefore, necessary to improve the statistical education or training of academic psychologists and the content of statistics textbooks in order to guarantee high quality training of future professionals (Cumming, 2012; Gliner et al., 2002; Kline, 2013; Hallan & Krauss, 2002; Wasserstein & Lazar, 2016).

Among Spanish practitioner psychologist, the "clinical or practical significance fallacy" was the most frequently observed misinterpretation. Nevertheless, a statistically significant result does not indicate that the result is important, in the same way that a non-statistically significant result might still be important (Nickerson, 2000; Wasserstein & Lazar, 2016). Clinical significance refers to the practical or applied value or importance of the effect of an intervention. That is, whether it makes any real (e.g., genuine, palpable, practical, noticeable) difference to the clients or to others with whom they interact in everyday life (Kazdin, 1999; 2008). The presentation of a lot of asterisks along with the *p* value of probability or very small *p* values only highlights that in this design the null hypothesis is not very plausible, but from there it cannot be inferred that the effect found is important (Gliner et al., 2001). As Fethney (2010) noted,

> In many people's minds, the word 'significant' means 'important', but in the world of statistics, it is a statement about the likelihood of a result being due to chance, or the amount of uncertainty we are prepared to accept, not its importance (p. 93).

Thus, to distinguish between the importance or practical significance of the findings and their statistical significance, the term "statistically significant" should be used to describe the results linked to a value of $p <$ alpha (Cumming, 2012; Gliner et al., 2001; Kline, 2013; Frías et al., 2000; Monterde-i-Bort et al., 2010; Thompson, 1996).

In summary, as Verdam et al. (2014) point out, the *p* value is not the probability of the null hypothesis; rejecting the null hypothesis does not prove that the alternative hypothesis is true; not rejecting the null hypothesis does not prove that the alternative hypothesis is false, and the *p* value does not give any indication of the effect size. Furthermore, the *p* value does not indicate replicability of results. Therefore, NHST only tells us about the probability of obtaining data which are equally or more discrepant than those obtained in the event that H_0 is true (Balluerka et al., 2009; Cohen, 1994; Kline, 2013; Nickerson, 2000).

It is noteworthy that these misconceptions are interpretation problems originating from the researcher and they are not a problem of NHST itself (Leek, 2014). Behind these erroneous interpretations are some beliefs and attributions about the significance of the results. Therefore, it is necessary to improve the statistical education or training of researchers and the content of statistics textbooks in order to guarantee high quality training of future professionals (Cumming, 2012; Kline, 2013; Haller & Krauss, 2002).

Problems in understanding the *p* value influence the conclusions that professionals draw from their data (Hoekstra et al., 2014), jeopardizing the quality of the results of psychological research (Frías-Navarro, 2011a). The value of the evidence depends on the quality of the statistical analyses and their interpretation (Faulkner et al., 2008).

Reporting the effect size and its confidence intervals (CIs) could help avoid these erroneous interpretations (APA, 2010a; Balluerka et al., 2009; Coulson et al., 2010; Fidler & Loftus, 2009; Hoekstra, Johnson et al., 2012; Gliner et al., 2001; Kirk, 1996; Maher, Markey, & Ebert-May, 2013; Newcombe, 2012; Monterde-i-Bort et al., 2010; Palmer & Strelan, 2015¸ Savalei & Dunn, 2015; Téllez et al., 2015; Wilkinson & TFSI, 1999). Empirical studies have provided evidence that CIs avoid some of the incorrect interpretations of *p* values (Fidler & Loftus, 2009; Hoekstra, Johnson et al., 2012). This statistical practice would enhance the body of scientific knowledge (Frías-Navarro, 2011b; Lakens, 2013). However, CIs are not immune to incorrect interpretations either (Belia, Fidler, Williams, & Cumming, 2005; Hoekstra et al., 2014; Miller & Ulrich, 2016).

On the other hand, the use of effect size statistics and its CIs facilitates the development of "meta-analytic thinking" among researchers. "Meta-analytic thinking" redirects the design, analysis and interpretation of the results towards the effect size and, in addition, contextualizes its value within a specific area of investigation of the findings (Coulson et al., 2010; Cumming, 2012; Frías-Navarro, 2011b; Kline, 2013; Peng et al., 2013). This knowledge enriches the interpretation of the findings, as it is possible to contextualize the effect, rate the precision of its estimation, and aid in the interpretation of the clinical and practical significance of the data. Finally, the real focus for many applied studies is not only finding proof that the therapy or whatever intervention worked, but also quantifying its effectiveness. Nevertheless, reporting the effect size and its confidence intervals continues to be uncommon (Frías-Navarro et al., 2012; Fritz et al., 2012; Peng et al., 2013). As several authors point out, the "effect size fallacy" and the "clinical or practical significance fallacy" could underlie deficiencies in scientific reports published in high-impact journals when reporting effect size statistics (Fidler, 2005; Kirk, 2001; Kline, 2013).

It has been acknowledged several limitations in this series of studies. The low response rate might affect the representativity of the sample and, therefore, the generalizability of the findings among Chilean, Italian and Spanish academic psychologists and Spanish practitioner psychologists. Nevertheless, as it was said before, it is possible that the participants who responded to the survey felt more confident about their statistical knowledge than those who did not respond. Should this be the case, the results might underestimate the barriers to EBP.

In addition, the results agree with the findings of previous studies on misconceptions of the *p*-value in samples of academic psychologists and undergraduates of Psychology (Badenes-Ribera, Frías-Navarro & Pascual-Soler, 2015; Falk & Greenbaum, 1995; Haller & Krauss, 2002, Kühberger et al., 2015; Monterde-i-Bort et al., 2010; Oakes, 1986), statistics professionals (Lecoutre et al., 2003) and members of the American Educational Research Association (Gordon, 2001; Mittag & Thompson, 2000).

All of this leads to indicate the need to adequately train Psychology professionals to produce valid scientific knowledge and improve the professional practice.

In conclusion, this work provides more evidence of the need for better statistical education, given the problems related to adequately interpreting the results obtained with the null hypothesis significance procedure. As Falk and Greenbaum (1995) point out, "*unless strong measures in teaching statistics are taken, the chances of overcoming this misconception appear low at present*" (p. 93). The results of these studies report the prevalence on different misconceptions about *p*-value among academic psychologists, university students and psychologists. This information is fundamental for approaching and planning statistical education strategies designed to intervene in order to address incorrect interpretations. Future research in this field should be directed toward intervention measures against the fallacies or interpretation errors related to the *p* value of probability.

5. STUDIES ON KNOWLEDGE LEVEL OF EFFECT SIZE, CONFIDENCE INTERVALS AND META-ANALYSES

5.1. Justification and purpose

The "statistical reform" movement recommends an important change in researcher behavior, asking academics to change their perspective from "how probable or improbable the sample result is" (e.g., to apply the traditional statistical significance tests and dichotomous statistical decisions based on the comparison of the *p* value and the alpha value combined with power or confidence intervals) to new analytic strategies that estimate the effect size (ES) and its confidence intervals (CIs), and favor the replication of the findings, as well as their practical/clinical significance (Wilkinson & TFSI, 1999).

These recommendations were incorporated into the revised fifth edition of the Publication Manual of the American Psychological Association (APA, 2001), and they were again included in the sixth edition (APA, 2010a).

The sixth edition of the Manual of the American Psychological Association (APA, 2010a) reinforces the use of CIs, the ES and their CIs, and confirms meta-analysis as part of the mainstream in the use of statistics for a better research practice. In this way, the APA (2010a) states that "*For the reader to appreciate the magnitude or importance of a study's finding, it is almost always necessary to include some measure of effect size in the results section*" (p. 34), and that "*The inclusion of confidence intervals (for estimation of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results* […]. *The use of confidence intervals is therefore strongly recommended*" (p. 34). In addition, the APA (2010a) recommends that researchers should "*Whenever possible, provide a confidence interval for each ES reported to indicate the precision of estimation of the ES*" (p. 34).

An effect size represents the strength or magnitude of a relationship between the variables in the population, or a sample-based estimate of that quantity (Cohen, 1988).

There are dozens of ES measures available (Henson, 2006; Kirk, 1996). In general, they can be classified into two broad groups: measures of mean differences and measures of strength of relations (Frías-Navarro, 2011b; Henson, 2006; Huberty, 2002; Kirk, 1996; Rosnow & Rosenthal, 2009).The former is based on the standardized group mean difference (e. g. Cohen's *d*, Glass's *g*, Hedges' g_u, Cohen's *f*); the latter is based on the proportion of variance accounted for or correlation between two variables (e. g., R^2/r^2, η^2, w^2).

The most frequently reported ES measures are the unadjusted R^2, Cohen's *d*, and η^2 (Badenes-Ribera et al., 2013; Lakens, 2013; Peng et al., 2013; Sesé & Palmer, 2012; Sun et al., 2010). These statistics have been criticized for bias (i.e., they tend to be positively biased), lack of robustness to outliers, and instability under violations of statistical assumptions (Fritz et al., 2012; Grissom & Kim, 2012; Kline, 2013; Thompson, 2002b, 2007; Wang & Thompson, 2007).

CIs as an interval estimator, indicate the precision of the parameter estimate (e.g., population mean or population standard deviation). In this case, the width of a confidence interval represents the precision of the point estimate of a statistic. The smaller the confidence interval is, the more precise the estimation.

CIs for effect size statistics are not the same as CIs for other sample statistics (e.g., means, standard deviation) because they are not computed in the same manner. There are several methods to assist researchers in designing confidence intervals for ES estimates (e.g., Grissom & Kim, 2012).

Meta-analysis "*is a research methodology that aims to quantitatively integrate the results of a set empirical studies about a given topic*" (Sánchez-Meca & Marín-Martínez, 2010, p. 151). It facilitates more precise effect size estimations, it allows researchers to rate the stability of the effects, and it helps them to contextualize the effect size values obtained in their studies. Moreover, the results of a meta-analytic study help to plan future sample sizes by providing the value of the estimated effect size in a specific research context. Graphical displays have become the principal tool for presenting the results of multiple studies on the same research question (Anzures-Cabrera & Higgins, 2010; Frías-Navarro & Monterde-i-Bort, 2014).

The main purpose of this chapter is to analyze what Spanish academic psychologists and Spanish practitioner psychologists know about ES, their CIs, and

meta-analyses, given that this is one of the main recommendations proposed by the APA (2010a) to improve statistical practice and favor the accumulation of knowledge and the replication of findings. For this purpose, the participants were asked about their statistical knowledge and statistical analyses performed by them.

In addition, it was carried out a direct replication study with a sample of Chilean and Italian academic psychologists, who were also asked about their statistical knowledge and statistical analyses performed by them.

5.2. Study 1: Sample of Spanish academic psychologists[12]

5.2.1 Method

5.2.1.1. Design and Procedure

It has been carried out a cross-sectional study through on-line survey. For this purpose, the e-mail addresses of academic psychologists were found by consulting the webs of the Spanish universities, resulting in 4,463 potential participants. Potential participants were invited to complete a survey through the use of a CAWI (Computer Assisted Web Interviewing) system. A follow-up message was sent two weeks later to non-respondents. The data collection was performed during the 2013-2014 school year. The response rate was 10.58%.

5.2.1.2. Participants

It was used a non-probabilistic (convenience) sample consisted of 472 academic psychologists. The mean number of years of the professors in the University was 13.56 years (*SD* = 9.27, min = 1, max = 40). Men represented 45.76% (*n* = 216) and women 54.24% (*n* = 256).

Regarding university departments, 23.94% of the university professors (*n* = 113) belonged to the area of Personality, Evaluation and Psychological Treatments, 14.83% to the area of Behavioral Sciences Methodology (*n* = 70), 16.10% to the area of Basic Psychology (*n* = 76), 16.31% to the area of Social Psychology (*n* = 77), 6.78% to the area of Psychobiology (*n* = 32) and 22.03% to the area of Developmental and Educational Psychology (*n* = 104). Regarding kind of university, 87.92% belonged to

[12] This study is published as: Badenes-Ribera, L., Frías-Navarro, D., Pascual-Soler, M., & Monterde-i-Bort, H., (2016). Knowledge level on effect size statistics, confidence intervals and meta-analysis in Spanish academic psychologists. *Psicothema, 28, 448-456.* doi: 10.7334/psicothema2016.24

public university (12.08% private university). Finally, 64.83% of the participants have been reviewer of scientific journals last year.

5.2.1.3. Instrument

The survey consisted of two sections. The first one included items related to information about sex and years of experience as an academic psychologist, Psychology knowledge areas, kind of university (public/private).

The second section included items related to statistical knowledge and statistical practice of the researcher. They were the following:

1.-Knowledge and use of statistical terms, evaluated with 4 questions.

A.-*"What terms from the following list do you know sufficiently: standard deviation, sedimentation graph, forest plot, ANOVA, funnel plot, correlation, meta-analysis, regression analysis, effect size*". On this item, more than one response can be chosen.

B.- "*Can you give the name of an effect size statistic?*".

C.- "*If your answer is Yes, please specify its name*" (open-ended question).

D.-"*In your reports, what type of statistics do you use more often*?". Likert-type response scale with 5 response ratings that range from 0=not at all, to 4=used often.

2.-Opinions about meta-analysis, evaluated with 1 question.

A.-"*What type of review do you think has the most credibility and objectivity?*" (select only one response):

a) The narrative review carried out by experts (such as those performed in the "Annual Review").

b) The quantitative review or meta-analysis.

c) The qualitative review.

3.- Use of meta-analytic study, evaluated with 1 question: "*Have you read or used a meta-analytic study?*"

a) I have never read or used one.

c) Yes: I have read or used 1 -2 meta-analytic studies.

d) Yes, I have read or used more than 2 meta-analytic studies.

4.- Researcher's behavior, evaluated with 11 questions related to research design (e. g., estimate a priori sample size, strategies used to do it, and so on), reporting on p value, and interpretation of *p* value (see Table 5).

5.2.1.4. Data analysis

The analysis included descriptive statistics for the variables under evaluation such as frequencies and percentage. To calculate the confidence interval for percentages, we used score methods based on the works of Newcombe (2012). These methods perform better than traditional approaches when calculating the confidence intervals for percentages. It can be noted that in the published research, confidence intervals for percentages were not reported. These analyses were performed with the statistical program IBM SPSS v. 20 for Windows.

5.2.2. Results

Table 8 shows the participants' responses by psychology knowledge areas to the item that rates their knowledge about statistical terms. It can be noted that more than 90% of the participants said they have adequately known about standard deviation, correlation, analysis of variance and regression analysis.

In addition, more of 80% of them adequately know the statistical terms of effect size and meta-analysis. However, it is noteworthy that the graphics that usually accompany meta-analytic studies (forest plot and funnel plot) were rated as sufficiently known by a very low percentage of the participants, especially the funnel plot. Funnel plot graphic presents the heterogeneity among effect sizes and it is an often used publication bias detection method in the health sciences. And, forest plot graphic presents the mean effect size and its confidence interval along with the effect sizes and CIs of the primary studies.

Regarding their knowledge about ES statistics, 72.3% of the participants (n = 41) stated to know some effect size statistic. However, only 68.4% of them indicated the name of an effect size statistic (n = 323).

Table 8. Statistical terms the participants know sufficiently (%) [and 95% Confidence Intervals]

Statistical terms	1 (*n* = 113)	2 (*n* = 70)	3 (*n* = 76)	4 (*n* = 77)	5 (*n* = 32)	6 (*n* = 104)	Total (*n* = 472)
Standard deviation	100	100	100	98.70	93.75	98.08	98.94
	[96.71, 100]	[94.80, 100.]	[95.19, 100]	[93, 99,77]	[79.85, 98.27]	[93.26, 99.47]	[97.54, 99.55]
Correlation	98.23	100	100	98.70	96.88	97.12	98.52
	[93.78, 99.51]	[94.80, 100.]	[95.19, 100]	[93, 99,77]	[84.26, 99.45]	[91.86, 99.01]	[96.97, 99.28]
ANOVA	97.35	100	98.68	94.81	100	96.15	97.46
	[92.48, 99.09]	[94.80, 100.]	[92.92, 99.77]	[87.39, 97.96]	[89.28, 100]	[90.53, 98.49]	[95.61, 98.54]
Regression analysis	96.46	98.57	93.42	96.10	90.63	90.38	94.49
	[91.25, 98.61]	[92.34, 99.75]	[85.51, 97.16]	[89.16, 98.67]	[75.78, 96.76]	[83.20, 94.69]	[92.05, 96.21]

Note. More than one response could be selected. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

Table 8 (Continued)

Statistical terms	1 (*n* = 113)	2 (*n* = 70)	3 (*n* = 76)	4 (*n* = 77)	5 (*n* = 32)	6 (*n* = 104)	Total (*n* = 472)
Effect size	94.69	92.86	89.47	84.42	78.13	77.88	87.08
	[88.90, 97.54]	[84.34, 96.91]	[80.58, 94.57]	[74.71, 90.85]	[61.25, 88.98]	[69, 84.79]	[83.75, 89.81]
Meta-analysis	92.92	91.43	81.58	85.71	71.88	86.54	86.86
	[86.65, 96.37]	[82.53, 96.01]	[71.42, 88.70]	[76.20, 91.83]	[54.63, 84.44]	[78.66, 91.81]	[83.52, 89.62]
Sedimentation graphic	57.52	67.14	27.63	45.45	15.63	38.46	45.13
	[48.31, 66.24]	[55.50, 77]	[18.84, 38.58]	[34.81, 56.53]	[6.86, 31.75]	[29.68, 48.06]	[40.70, 49.64]
Forest plot	12.39	27.14	9.21	6.49	6.25	4.81	11.02
	[7.53, 19.73]	[18.12, 38.54]	[4.53, 17.81]	[2.81, 14.32]	[1.73, 20.15]	[2.07, 10.76]	[8.50, 14.16]
Funnel plot	6.19	22.86	2.63	6.49	6.25	0.96	6.99
	[3.03, 12.24]	[14.59, 33.95]	[0.72, 9.10]	[2.81, 14.32	[1.73, 20.15]	[0.17, 5.25]	[5.02, 9.66]

Note. More than one response could be selected. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

By Psychology knowledge areas, the percentage of those who stated to know some effect size statistic were 78.8% in the area of Personality, Evaluation and Psychological Treatments, 97.1% in Methodology, 65.8% in Basic Psychology, 70.1% in Social Psychology, 40.6% in Psychobiology and 64.4% in Developmental and Educational Psychology. And the percentage of those who gave the name of an effect size statistic were 75.2% in the area of Personality, Evaluation and Psychological Treatments, 91.4% in Methodology, 61.8% in Basic Psychology, 64.9% in Social Psychology, 37.5% in Psychobiology and 62.5% in Developmental and Educational Psychology. Consequently, there is greater knowledge about the term of effect size than about statistics of effect size.

The most familiar statistics to the participants were those that evaluate differences between the means of the groups analyzed (standardized mean differences), followed by the proportion of variance explained (η^2) and correlation coefficients (Table 9).

Concerning the use of ES statistics in research reports (Table 10), 40.7% of the participants stated that they use the ES a lot in their studies, but only 24.4% of them estimated the confidence interval around the ES. Approximately 36.8% of the participants said they use the ES little or not at all in their statistical reports. Furthermore, most of participants (57.8%) recognized that they use effect sizes and their CIs very little or not at all (not utilized, scarcely utilized and somewhat utilized).

Table 9. Known effect size statistics (responses of 323 participants)

Effect size statistics	*n*	%	95% CI
Cohen's *d*	228	70.59	[65.40, 75.29]
η^2	142	43.96	[38.65, 49.42]
Correlation/Association coefficient	80	24.77	[20.38, 29.75]
(Pearson, Spearman, biserial, phi, Cramer's V)			
Hedges' *g*	35	10.84	[7.90, 14.70]
R^2	32	9.91	[7.11, 13.65]
Omega/Omega^2	26	8.05	[5.55, 11.53]
Odds Ratio	19	5.88	[3.80, 9]
Cohen's *f*/Cohen's f^2)	9	2.79	[1.47, 5.21]
Relative Risk	8	2.48	[1.26, 4.81]
Glass' delta	6	1.86	[0.85, 3.99]
Beta	3	0.93	[0.32, 2.69]
Number Needed to Treat (NNT)	3	0.93	[0.32, 2.69]
Wilk's Lambda	2	0.62	[0.17, 2.23]
Epsilon/Epsilon^2	2	0.62	[0.17, 2.23]
Cliff's delta	1	0.31	[0.05, 1.73]
Common Language (CL)	1	0.31	[0.05, 1.73]

Note. The majority of participants reported knowing more than one effect size statistic. CI = confidence interval.

Table 10. Use of the statistics (%) [and 95% Confidence Intervals] (N=472)

Statistics	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized
ANOVA	65.25 [60.85, 69.41]	25.42 [21.70, 29.54]	6.14 [4.31, 8.68]	2.54 [1.46, 4.39]	0.64 [0.22, 1.85]
Correlation	55.72 [51.21, 60.14]	26.06 [22.30, 30.20]	12.92 [10.19, 16.25]	4.66 [3.10, 6.96]	0.64 [0.22, 1.85]
T tests	44.70 [40.28, 49.21]	29.24 [25.31, 3350]	17.58 [14.42, 21.28]	7.84 [5.74, 10.62]	0.64 [0.22, 1.85]
Regression	44.49 [40.07, 49]	27.54 [23.71, 31.74]	18.43 [15.19, 22.18]	8.26 [6.10, 11.10]	1.27 [0.58, 2.75]
Effect size	40.68 [36.34, 45.17]	22.46 [18.92, 26.44]	14.83 [11.91, 18.32]	14.19 [11.33, 17.63]	7.84 [5.74, 10.62]
Confidence intervals	26.06 [22.30, 30.20]	21.82 [18.33, 25.77]	22.46 [18.92, 26.4]4	22.88 [19.32, 26.88]	6.36 [4.49, 8.93]
Exploratory factorial analysis	24.79 [21.11, 28.88]	23.94 [20.31, 27.99]	21.19 [17.74, 25.10]	22.46 [18.92, 26.44]	7.63 [5.56, 10.38]
Effect Size and Confidence interval	24.36 [20.71, 28.43]	17.80 [14.61, 21.50]	18.01 [14.80, 21.73]	15.68 [12.68, 19.23]	24.15 [20.51, 28.21]

Table 10 (Continued)

Statistics	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized
MANOVA	21.61	22.46	20.76	28.18	6.99
	[18.13, 25.54]	[18.92, 26.44]	[17.35, 24.65]	[24.31, 32.40]	[5.02, 9.66]
Confirmatory factor analysis	19.92	22.25	18.01	28.39	11.44
	[16.56, 23.75]	[18.73, 26.21]	[14.80, 21.73]	[24.51, 32.62]	[8.88, 14.63]
Structural equations	13.35	18.01	15.04	31.78	21.82
	[10.57, 16.71]	[14.80, 21.73]	[12.10, 18.55]	[27.74, 36.11]	[18.33, 25.77]
Discriminant analysis	5.08	10.17	26.06	40.25	18.43
	[3.44, 7.45]	[7.76, 13.23]	[22.30, 30.20]	[35.93, 44.74]	[15.19, 22.18]

Table 11 shows that the most of participants (57.4%) pointed out that meta-analytics study are the type of review with the most credibility and objectivity. Nevertheless, 42.6% said they give more importance to narrative reviews carried out by experts and/or qualitative reviews. By Psychology knowledge areas, the percentage of participants who gave more importance to narrative reviews carried out by experts and/or qualitative reviews ranged from 24.3% in the area of Methodology to 59.4% in the area of Psychobiology. In addition, the majority of the participants said they have used or read a meta-analytic study for their research. By knowledge areas, the percentage of participants who stated utilizes meta-analytic studies ranged from 75% in the area of Psychobiology to 92.2% in the area of Social Psychology.

In addition, Table 11 shows that the majority of the participants said they have used or read a meta-analytic study for their research. By knowledge areas, the percentage of participants who stated utilize meta-analytic studies ranged from 75% in the area of Psychobiology to 92.2% in the area of Social Psychology.

Table 11. Opinions about the review with most credibility and objectivity and use of meta-analytic studies (%) [and 95% Confidence intervals]

	1 (*n* = 113)	2 (*n* = 70)	3 (*n* = 76)	4 (*n* = 77)	5 (*n* = 32)	6 (*n* = 104)	Total (*N* = 472)
Opinions about the review with most credibility and objectivity							
The quantitative review or meta-analysis	64.60 [55.44, 72.81]	75.71 [64.50, 84.25]	52.63 [41.55, 63.46]	51.95 [40.96, 62.75]	46.88 [30.87, 63.55]	48.08 [38.72, 57.58]	57.42 [52.91, 61.80]
The narrative review carried out by experts	27.47 [20.05, 36.30]	20 [12.30, 30.82]	42.11 [31.65, 53.32]	38.96 [28.84, 50.13]	40.63 [25.52, 57.74]	40.38 [31.46, 49.99]	34.32 [30.18, 38.72]
The qualitative review	7.96 [4.25, 14.45]	4.29 [14.7, 11.86]	5.26 [2.07, 12.77]	9.09 [4.47, 17.60]	12.50 [4.97, 28.07]	11.54 [6.72, 19.09]	8.26 [6.10, 11.10]

Note. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

Table 11 (Continued)

	1 (*n* = 113)	2 (*n* = 70)	3 (*n* = 76)	4 (*n* = 77)	5 (*n* = 32)	6 (*n* = 104)	Total (*n* = 472)
Reading or use of meta-analytic studies e							
I have never read or used one	9.73 [5.52, 16.59]	12.86 [6.91, 22.66]	13.16 [7.31, 22.55]	7.79 [3.62, 15.98]	25 [13.25, 42.11]	23.08 [16.03, 32.05]	14.41 [11.53, 17.86]
I have read or used 1-2 meta-analytic studies	21.24 [14.71, 29.66]	27.14 [18.12, 38.54]	34.21 [24.54, 45.40]	31.17 [21.93, 42.20]	34.38 [20.41, 51.69]	36.54 [27.92, 46.12]	30.08 [26.12, 34.37]
I have read or used more than 2 meta-analytic studies	69.03 [59.99, 76.81]	60 [48.29, 70.67]	52.63 [41.55, 63.46]	61.04 [49.87, 71.16]	40.63 [25.52, 57.74]	40.38 [31.46, 49.99]	55.51 [51, 59.93]

Note. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

Finally, it was analyzed the profile of researchers based on whether they could or could not indicate the name of an effect size statistic. The results indicated that academics who could name an effect size statistic have a behavior more close to good statistical practices and of research design (see Table 12 and Table 13).

In this way, as Table 12 shows academics who could name an effect size statistic compared to participants who could not had higher proportion of participants who had read or used meta-analysis studies, had been reviewers for scientific journals, had published articles in journals with impact factor JCR (Journal Citation Reports of WoS) and thought that meta-analysis studies are the type of review with the most credibility.

Regarding their behavior when they plan or prepare a study (Table 13), academics who could name an effect size statistic perform better methodological practices than the rest of the participants, since a larger proportion of them estimate a priori sample size (both groups have a high proportion), plan the number of participants, and use statistical criteria seeking that the sample represents the characteristics of the population. It is noteworthy that academics who named an effect size statistic confuse in a lesser extent planning the statistical power a priori as a strategy to adjust the significance level or alpha value, and also make in a lesser extent the clinical or practical size fallacy, where the statistical significance of the effect is related to the importance of effect, although both groups of academics exceed 30% of subjects who believe in that association. However, it should be noted that a statistically significant effect can be found, but could not have any clinical importance, and vice versa. The clinical or practical importance of the findings should be described by an expert in the field, and not placed in the statistics alone.

In addition, they follow the APA recommendations avoiding expressions of *p* value as $p <$ alpha or $p >$ alpha and using its exact value in a higher proportion than the rest of the participants.

Finally, both groups of academics said in a high proportion that they do not know any checklist to assess the design quality of a study (91.9% of academics who could not name a statistical effect size and 78% of academics who could do it) and that do not know that currently there is some kind of open debate on statistical issues or research design (79.9% in the group of academics who could not name a statistical effect size and 53.9% in the group of academics who did it).

Table 12. Researcher's behavior and opinion according to knowing or not knowing the name of effect size statistics

Item	Not Knowing (n = 149)		Knowing (n = 323)	
	%	95% CI	%	95% CI
1. Have you read or used a meta-analytic study?				
I have never read or used one	28.86	[22.19, 36.59]	7.74	[5.30, 11.18]
Yes: I have read or used 1 -2 meta-analytic studies	**36.91**	**[29.58 44.90]**	26.93	[22.39, 32.02]
Yes, I have read or used more than 2 meta-analytic studies	34.23	[27.09, 42.16]	**65.33**	**[59.98, 70.31]**
2. Have you been reviewer for scientific journals in the last year?				
No	**48.32**	**[40.44, 56.29]**	29.10	[24.42, 34.28]
Yes: 1-2 reviewed articles	38.26	[30.84, 46.26]	33.44	[28.51, 38.75]
Yes: more than 2 reviewed articles	13.42	[8.86, 19.82]	**37.46**	**[32.36, 42.86]**

Note. CI = confidence interval.

Table 12 (Continued)

Item	Not Knowing (n = 149)		Knowing (n = 323)	
	%	95% CI	%	95% CI
3. Have you published an article in a journal indexed in the WoS with JCR impact factor in the last year?				
No	**39.60**	[32.10, 47.62]	21.67	[17.53, 26.48]
Yes: 1-2 published articles	**39.60**	[32.10, 47.62]	**43.03**	**[37.75, 48.48]**
Yes: more than 2 published articles	20.81	[15.06, 28.02]	35.29	[30.28, 40.65]
4. Do you know checklist for assessing research design of a study?				
No	**91.95**	**[86.45, 95.33]**	**78.02**	**[73.19, 82.19]**
Sí	8.05	[4.67, 13.55]	21.98	[17.81, 26.81]

Note. CI = confidence interval.

Table 12 (Continued)

Item	Not Knowing (n = 149)		Knowing (n = 323)	
	%	95% CI	%	95% CI
5. What type of review do you think has the most credibility and objectivity?				
The narrative review carried out by experts	**40.94**	**[33.37, 48.97]**	31.27	[26.46, 36.52]
The quantitative review or meta-analysis	**41.61**	**[34, 49.64]**	**64.71**	**[59.35, 69.72]**
The qualitative review	17.45	[12.20, 24.34]	4.02	[2.37, 6.76]
6. In your opinion, what statistical questions or issues related to the study design are currently being debated?				
I don't know	**79.87**	**[72.71, 85.52]**	**53.87**	**[48.42, 59.23]**
I don't think there are any debates open	2.01	[0.69, 5.75]	2.17	[1.05, 4.41]
There is some debate	18.12	[12.76, 25.08]	43.96	[38.65, 49.42]

Note. CI = confidence interval.

Table 13. Researcher's methodological behavior according to knowing or not knowing the name of effect size statistics

Item	Not Knowing (n = 149)		Knowing (n = 323)	
	%	95% CI	%	95% CI
1. When you plan a study, do you estimate a priori the sample size you will need?				
No	21.48	[15.64, 28.74]	14.55	[11.12, 18.81]
Yes	**78.52**	**[71.26, 84.36]**	**85.45**	**[81.19, 88.88]**
2. What kind of strategy do you use when you want to plan the sample size of a study?				
You try to achieve the greatest number of participants possible	33.10	[26.05, 41]	25.08	[20.66, 30.08]
You use software or tables to estimate the sample size according to the statistical criteria	25.20	[18.91, 32.74]	34.67	[29.69, 40.02]
You try to make the sample represent the characteristics of the population	33.80	[26.70, 41.72]	37.46	[32.36, 42.86]
You do not use any strategy because it isn't part of your research interests.	7.90	[4.55, 13.36]	2.79	[1.47, 5.21]

Note. CI = confidence interval.

Table 13 (Continued)

Item	Not Knowing (n = 149)		Knowing (n = 323)	
	%	95% CI	%	95% CI
3. In your opinion, what is the purpose of calculating the statistical power a priori?				
To adjust the significance level or alpha value	46.98	[39.14, 54.97]	33.13	[28.22, 38.43]
To explore the reliability of the scales	13.42	[8.86, 19.82]	4.95	[3.07, 7.89]
To estimate the sample size	**39.60**	**[32.10, 47.62]**	**61.92**	**[56.51, 67.05]**
4. In your opinion, obtaining a statistically significant result implies indirectly that the detected effect is important				
No	45.64	[37.85, 53.64]	**69.66**	**[64.44, 74.42]**
Yes	**54.36**	**[46.36, 62.15]**	30.34	[25.58, 35.56]

Note. CI = confidence interval.

Table 13 (Continued)

Item	Not Knowing (n = 149)		Knowing (n = 323)	
	%	95% CI	%	95% CI
5. When you perform a statistical test, do you consider a priority to always report the statistical significance obtained?				
No	5.37	[2.75, 10.24]	3.72	[2.14, 6.38]
Yes, and using expressions like $p < .05$, $p > .05$	**59.73**	**[51.71, 67.27]**	41.80	[36.54, 47.24]
Yes, and using expressions with the p value of exact probability	34.90	[27.71, 42.85]	**54.49**	**[49.04, 59.84]**

Note. CI = confidence interval.

5.2.3. Discussion

The results of this study are novel because, until now, there were no self-report data about the following of the statistical reform and the APA Manual recommendations among Spanish researchers, even though these recommendations have to be followed in almost all of the psychological journals.

The results indicate that the emphasis the statistical reform places on the use of the ES and its confidence interval has also had an impact on participants, especially the estimation of the effect size. The majority of the interviewees state that they use effect size statistics (63.2%) in a fair amount or a lot. Moreover, 42.2% of them say that they use effect sizes and their confidence intervals in a fair amount or a lot as well. Therefore our results point to a higher self-reported use of the ES and CIs than prior studies that analyze the actual use of the ES and CIs in articles published at Spanish psychology journals (e.g., Badenes-Ribera et al., 2013; Caperos & Pardo, 2013; Frías-Navarro et al., 2012; Sesé & Palmer, 2012). For example, Caperos and Pardo (2013) reviewed 186 articles published in four Spanish psychology journals indexed at the Journal Citation Reports from 2009 (Social Science Edition) (*Anales de Psicología, Psicológica, Psicothema, and Spanish Journal of Psychology*) and they found that only 24.3% of the statistical inference tests were accompanied by an ES statistic. It could be a sign of the change in the analytic behavior of the researcher.

This discrepancy between self-reported use and the actual use of effect size statistics and confidence intervals in scientific reports might be explained in part by social desirability, as it is possible that the participants stated that they used effect size statistics and confidence intervals in their reports higher than what they actually used them. In fact, the findings of the present study show that the percentage of academic psychologists who stated to know effect size statistics was higher than the percentage of them who actually could give a name.

However, self-reported use of CIs were not nearly as frequently as effect size point estimate (along the same lines several studies that review articles published in Spanish and international Psychology journals (such as Badenes-Ribera et al., 2013; Fritz et al., 2012; Peng et al., 2013; Sesé & Palmer, 2012). This result goes against the APA recommendation that "*Whenever possible, provide a confidence interval for each*

effect size reported to indicate the precision of estimation of the effect size" (APA, 2010a, p. 34). It could be expected that it will improve in future studies, since the change in statistical practices takes time.

Regarding the type of effect size statistic they know, the participants mention to a greater degree the effect size statistics from the family of standardized mean differences and η^2 (parametric effect size statistics). These findings are in line with previous researches that analyze the use of effect size statistics in journals. For example, Peng et al. (2013) found that the most frequently reported ES measures were R^2, Cohen's *d*. Nevertheless, effect size statistics from the family of standardized differences in means (e.g. Cohen's *d*, Glass' delta, Hedges' *g*,) and from the family of correlation (Pearson correlation, R^2, η^2, omega^2, and so on) have been criticized for lack of robustness against outliers or departure from normality, and instability under violations of statistical assumptions (Algina et al., 2005; Grissom & Kim, 2012; Kline, 2013; Peng & Chen, 2014; Wang & Thompson, 2007).

The results suggest that the modern robust statistical methods are not known by most of participants, or at least, majority of participants did not give the name of robust effect size statistics. In fact, only 0.9% of the participants (n=3) gave the name of a robust effect size statistic (e.g., Number Needed to Treat, NNT).

Regarding the knowledge of meta-analytic studies, the majority of the participants give more credibility and objectivity to systematic reviews and meta-analytic studies than to other types of literature reviews. Also, they have an adequate knowledge of meta-analyses. However, they have a poor knowledge of graphical displays for meta-analyses (i.e., forest plots and funnel plots) which can become in a misinterpretation of results. The graphical presentation of results is an important part of a meta-analysis and it has become the primary tool for presenting the results of multiple studies on the same research question (Anzures-Cabrera & Higgins 2010; Borenstein, et al. 2009; Ellis, 2010; Sánchez-Meca & Marín-Martínez, 2010).

Finally, the analysis of the researcher's behavior associated with methodological practices point out that academics who know some effect size statistics present a profile more close to good statistical practices and design research, participate more actively in the process of peer review, and publish in journals with impact.

It must be acknowledged some limitations in this study. Firstly, the low response rate could affect the representativity of the sample and, therefore, the generalizability of the results.

Moreover, it is possible that the participants who responded to the survey had higher level of statistical knowledge than those who did not respond. Should this be the case, the results might overestimate the extension of the impact of the statistical reform in Spanish academic psychologists.

Furthermore, it must also be acknowledged that some participants do not use quantitative methods at all. These individuals may have been less likely to respond, as well. Nevertheless, our findings are in line with previous researches that analyze the use of effect size statistics in journals. For example, Sesé & Palmer (2012) found that the most frequently reported ES measures were R^2 and Cohen's *d*. In addition, Peng et al. (2013) pointed out that robust effect size statistics were reported fewer than non-robust statistics, such as standardized differences in means.

In addition, it is possible that there has been an effect of social desirability as it may always happen when data are collected through self-report questionnaires. In this way, the percentage of participants who stated that could give the name of an effect size statistic was higher than the percentage of them who actually did it. A way of control this bias in future research would be formulate the questions (e.g., what is the correct interpretation of a specific forest plot, funnel plot, effect size or regression analysis) with three or four-response format, or with open-end question. These response formats would permit us to assess the level of knowledge of the statistical terms, thus, they would have been far more informative.

5.3. Study 2: Sample of Spanish Practitioner Psychologists[13]

5.3.1. Method

5.3.1.1. Design and Procedure

The data were collected from a cross-sectional on-line survey of Spanish psychologists. We send an e-mail to Spanish Psychological Associations inviting them to participate in the on-line survey on professional practice in Psychology. Potential participants were invited to complete a survey through the use of a CAWI (Computer Assisted Web Interviewing) system. A follow-up message was sent three weeks later. The data collection was performed from May to September 2015.

5.3.1.2. Participants

It was used a non-probabilistic (convenience) sample. The sample was initially made up of 113 Spanish psychologists. Of these, 68.1% were practitioner psychologists, 28.3% were academic psychologists, 0.9% were researchers, and 2.7% reported other role. Since the objective of the study was to analyze the barriers to evidence-based practice, participants who were not practitioner psychologists were eliminated from the sample (*n* = 36).

The final sample consisted of 77 Spanish psychologists with an average age of 41.44 years (*SD* = 9.42, min = 25, max = 64). Of the 77 participants, 31.2% were men and 68.8% were women. The mean number of years as member of Spanish Psychological Associations was 13.73 years (*SD* = 9.30, *min* = 0, max = 33). The mean degree of familiarity with EBP was 4.65 (*SD* = 2.18, *min* = 1, *max* = 7). With regard to education degree, 27.27% of the participants had bachelor's degree, 46.75% Master's degree, and 25.97% Ph.D. Regarding the clinical setting where the participants worked, 38.96% of them worked in public setting (61.04% private setting).

Finally, Mann Whitney U test indicated that statistically significant differences between men and women were not observed for the variable number of years as member as psychologist (z = -1.29, p = .196, r = -.15, 95% CI [-.35, .08]) and for the variable degree of familiarity with Evidence Based Practice approach (z = -0.07, p = .942, r = -.01, 95%CI [-.23, .21]).

[13] This study is under review as: Badenes-Ribera, L. Bonilla-Campos, A. & Frías-Navarro, D. (2016). Barriers to Evidence Based Practice in Spanish practitioner psychologists: An exploratory study.

Furthermore, the results of a one-way ANOVA showed that there were not statistically significant differences on mean scores for the degree of familiarity with evidence-based practice approach by education level of the participants ($F(2,74) = 2.16$, $p = .123$, Cohen's $d = 0.4$, 95% CI [-0.13, 0.98]).

5.3.1.3. Instrument

It was prepared a structured questionnaire that included items related to information about sex, age, education degree, years of experience as practitioner psychologist, clinical setting (public or private), degree of familiarity with EBP movement, as well as items related to knowledge on methodological issues associated with EBP, such as effect size statistics, meta-analysis studies, and methodological quality checklist of the studies.

1.-Knowledge of the statistical terms, evaluated with 3 questions:

A.-*"What terms from the following list do you know sufficiently: standard deviation, sedimentation graph, forest plot, ANOVA, funnel plot, correlation, meta-analysis, regression analysis, effect size*". On this item, more than one response can be chosen.

B.- "*Can you give the name of an effect size statistic?*"

C.- "*If your answer is Yes, please specify its name*" (open-ended question).

2. Knowledge of the methodological quality checklist of the studies evaluated with 1 question: *Do you know methodological quality checklist of the studies*? (Yes/no).

3. Use of meta-analytic study in professional practice, evaluated with 1 question: "*Have you used a meta-analytic study in your professional practice?*" (Yes/no)

5.3.1.4. Data analysis

The analysis included descriptive statistics for the variables under evaluation such as frequencies and percentage. To calculate the confidence interval for percentages, we used score methods based on the works of Newcombe (2012). These methods perform better than traditional approaches when calculating the confidence intervals of percentages. These analyses were performed with the statistical program IBM SPSS v. 20 for Windows.

5.3.2. Results

Table 14 shows the participants' responses to the item that rates their knowledge about statistical terms. It can be noted that more than 80% of the participants said to have adequately known about standard deviation, correlation and confidence intervals.

In addition, most of the participants (67.53%) stated they adequately know meta-analysis studies. Nevertheless, it is noteworthy that only 31.7% said they adequately know effect size statistics. Furthermore, the graphics that usually accompany meta-analytic studies, such as, forest plot and funnel plot, were rated as sufficiently known by a very low percentage of the participants (1.3%), like the study with Spanish academic pasychologists. Forest plot graphic presents the mean effect size and its confidence interval along with the effect sizes and CIs of the primary studies. Funnel plot graphic presents the heterogeneity among effect sizes and it is an often used publication bias detection method in the health sciences.

Table 14. Statistical terms the participants know sufficiently (%)

	n	%	95% CI
Standard deviation	69	89.61	[80.82, 94.64]
Correlation	65	84.42	[74.71, 90.85]
Confidence interval	63	81.82	[71.76, 88.85]
Meta-analysis	52	67.53	[56.46, 76.94]
ANOVA	46	59.74	[48.58, 69.98]
Regression analysis	40	51.95	[40.96, 62.75]
Effect size	24	31.17	[21.93, 42.20]
Sedimentation graphic	5	6.49	[2.81, 14.32]
Forest Plot	1	1.30	[0.23, 0.70]
Funnel Plot	1	1.30	[0.23, 0.70]

Note. More than one response could be selected. CI = confidence interval.

Regarding their knowledge about ES statistics, although 31.7% of the participants stated they know effect size, only 10.39% of them indicated the name of an effect size statistic (n = 9). Consequently, there is greater knowledge about the term of effect size than of statistics of effect size, like the study with Spanish academic psycholigists. As Table 15 shows, the statistics most familiar to the participants were those that evaluate the differences between the means of the groups analyzed (standardized mean differences), specifically, Cohen's *d*, followed by the proportion of variance explained (η^2) and the Glass' *g*.

Table 15. Known effect size statistics (responses of 9 participants)

Effect size statistics	*n*	%	95% CI
Cohen's *d*	7	77.78	[45.26, 93.68]
Glass' *g*	2	22.22	[6.32, 54.74]
η^2	2	22.22	[6.32, 54.74]
Biserial correlation	1	11.11	[1.99, 43.50]
Cox's *d*	1	11.11	[1.99, 43.50]

Note. 50% of the participants reported knowing more than one effect size statistic. CI = confidence interval.

With regard to level of knowledge on checklists, only 24.68% of the participants, 95% CI [16.40, 35.35] stated knowing any checklist.

Finally, the majority of the participants said they have used a meta-analytic study in their clinical practice (51.95, 95% CI [40.96, 62.75]).

5.3.3. Discussion

The findings of this study indicate that the comprehension of many statistical concepts continues to be problematic. The poor methodological knowledge has been and continues to be a source of direct threat to properly implement the EBP in professional practice. Most of the participants reported they have used meta-analytic studies in their professional practice and have adequate knowledge about them. Nevertheless, they acknowledged having a poor knowledge of effect size and graphical displays for meta-

analyses, such as forest plot and funnel plot, which may become in a misinterpretation of results and, therefore, lead to bad practice, taking into account that most of the participants said that they used meta-analytic studies in their professional practice, like in the study with Spanish academic psychologists (Badenes-Ribera et al., 2016). As several authors point out, the graphical presentation of results is an important part of a meta-analysis and it has become the primary tool for presenting the results of multiple studies on the same research question (Anzures-Cabrera & Higgins, 2010; Borenstein et al., 2009; Sánchez-Meca & Marín-Martínez, 2010). In this way, forest plot and funnel plot are graphics used in meta-analytic studies to present pooled effect size estimates and publication bias and/or heterogeneity, respectively.

With regard to type of effect size statistic they know, the participants mentioned to a greater degree the effect size statistics from the family of standardized mean differences and η^2 (parametric effect size statistics), like the study with Spanish academic psychologists (Badenes-Ribera, Frías-Navarro, Pascual-Soler et al., 2016). The findings suggest that the participants do not know the alternatives for parametric effect size statistics, such as non-parametric statistics (e.g., Spearman's correlation), the robust standardized mean difference (trimmed mean and winsorized variance) or the probability of superiority (PS), the number needed to treat (NNT) (Erceg-Hurn & Mirosevich, 2008; Ferguson, 2009; Grissom & Kim, 2012; Keselman et al., 2008; Kraemer & Kupfer, 2006; Wilcox & Keselman, 2003).

Concerning the methodological quality checklists, again most of the participants said not having knowledge about them. In relation to the above, it should be clarified that there are checklists for primary studies (e.g., CONSORT) and for meta-analytic studies (e.g., AMSTAR or PRISMA-NMA).

Finally, several limitations should be acknowledged in this study. The low response rate might affect the representativity of the sample and, therefore, the generalizability of the findings among practitioner psychologists. Nevertheless, it is possible that the participants who responded to the survey felt more confident about their statistical knowledge than those who did not respond. Should this be the case, the results might underestimate the barriers to EBP. In addition, our results agree with prior researches on knowledge level of effect size and meta-analytic studies in Spanish academic psychologists (Badenes-Ribera, Frías-Navarro, Pascual-Soler et al., 2016).

5.4. Study 3: A replication study from Chile and Italy[14]

5.4.1. Justification and purpose

Replication is the most objective method for checking if the result of a study is reliable and this concept plays an essential role in the advance of scientific knowledge (Asendorpt et al., 2013; Carver 1978; Cumming, 2008; Earp & Trafimow, 2015; Hubbard, 2004; Hubbard & Lindsay, 2008; Kline, 2013; Nickerson, 2000; Stroebe & Strack, 2014; Wilkinson & TFSI, 1999).

For this reason, that is, to check whether the results of the study by Badenes-Ribera, Frías-Navarro, Pascual-Soler et al. (2016) on the level of knowledge of the effect sizes, confidence intervals and meta-analysis conducted in Spanish academic psychologists are reliable, we carried out a replication study with a sample of Chilean and Italian academic psychologists.

The "*replication includes repetitions by different researchers in different places with incidental or deliberate changes to the experiment*" (Cumming, 2008, p. 287). As Stroebe & Strack (2014) state, "*exact replications are replications of an experiment that operationalize both the independent and the dependent variable in exactly the same way as the original study*" (p. 60). Therefore, in exact and direct replication, the only differences between the original and replication studies would be the participants, location and moment in time.

5.4.2. Method

5.4.2.1 Design

It has been conducted an exact or direct replication study of the original study with Spanish academics psychologists. Consequently, we modified the geographic area of the participants and the moment in time. The original research was conducted in Spain in 2013-2014, while the present study was carried out in Chile and Italy in 2015.

[14] This study is under review as: Badenes-Ribera, L., Frías-Navarro, D., Bryan, N., Bonilla-Campos, A. & Longobardi, C. (2016). Survey on (mis)use of effect size, confidence intervals and meta-analysis in Chilean and Italian academic psychologists.

5.4.2.2. Procedure

It was carried out a cross-sectional study through on-line survey. For this purpose, the e-mail addresses of academic psychologists were found by consulting the websites of Chilean and Italian universities, resulting in 2,321 potential participants (1,824 Italians, 497 Chilean). The data collection was performed from March to May 2015. Potential participants were invited to complete a survey through the use of a CAWI (Computer Assisted Web Interviewing) system. A follow-up message was sent two weeks later to non-respondents. Individual informed consent was also collected from academics along with written consent describing the nature and objective of the study according to the ethical code of the Italian Association for Psychology (AIP). The consent stated that data confidentiality would be assured and that participation was voluntary. The mean response rate was 8.36% (Italian 8.72%, Chilean 7.04%).

The questions were administered in Italian and Spanish respectively. Original items were in Spanish, and therefore all of them were translated into Italian by applying the standard back-translation procedure, which implied translations from Spanish to Italian and vice versa (Balluerka et al., 2007).

5.4.2.3. Participants

It has been used a non-probabilistic (convenience) sample. The sample comprised 194 academic psychologists from Chile and Italy. Of these 194 participants, 159 were Italian and 35 were Chilean.

Of the 159 Italians participants, 45.91% were men and 54.09% were women, with a mean age of 47.65 years (*SD* = 10.47, min = 28, max = 83). The mean number of years that the professors had spent in academia was 12.90 years (*SD* = 10.21, min = 0, max = 46).

Of the 30 Chilean academic psychologists, men represented 45.71% of the sample and women represented 54.29%. In addition, the mean age of the participants was 43.60 years (*SD* = 9.17, min = 30, max = 69). The mean number of years that the professors had spent in academia was 15 years (*SD* = 8.61, min = 1, max = 41). Table 16 presents a description of the participants.

Table 16. Description of the participants

	Chile (*n*= 35)		Italy (*n* = 159)	
	n	%	*n*	%
Sex				
Men	16	45.71	73	45.91
Women	19	54.29	86	54.09
Psychology knowledge areas				
Development & Educational Psychology	8	22.86	29	18.24
Clinical and Dynamic Psychology	8	22.86	36	22.64
Social Psychology	7	20	28	17.61
Methodology	6	17.14	17	10.69
Neuropsychology	1	2.86	16	10.06
Work and Organizational Psychology	3	8.57	13	8.18
General Psychology	2	5.71	28	17.61
Type of University				
Public	16	45.71	137	86.16
Private	19	54.29	22	13.84
Have you been reviewer for scientific journals in the last year?				
Yes	20	57.14	135	84.91
No	15	42.86	24	15.09

5.4.2.4 Data analysis

The analysis included descriptive statistics for the variables under evaluation such as frequencies and percentage. To calculate the confidence interval for percentages, we used score methods based on the works of Newcombe (2012). These methods perform better than traditional approaches when calculating the confidence intervals of percentages. These analyses were performed with the statistical program IBM SPSS v. 20 for Windows.

5.4.3 Results

Table 17 presents the percentage of responses by participants about their level of knowledge of statistical terms, according to the nationality of the participants. It can be noted that more than 90% of Chilean and Italian participants said they have an adequate knowledge of CIs, analysis of variance, regression analysis and standard deviation. Additionally, more than 80% of them have an adequate knowledge of the statistical terms of correlation, effect size and meta-analyses.

Finally, the statistical terms of forest plot and funnel plot (i.e., graphics that usually accompany meta-analytic studies) are rated as being sufficiently known by a very low percentage of the Chilean and Italian participants, especially the funnel plot graphics that are used primarily as a visual aid for detecting publication bias and heterogeneity, like we observed in Spanish academic psychologists, like the study of Badenes-Ribera et al. (2016).

Concerning their knowledge about effect size statistics, 82% of the participants state that they have an adequate knowledge of effect sizes. However, only 54.29% (95% CI 38.19, 69.53) of Chilean participants (n = 19) and 44.65% (95% CI 37.14, 52.42) of Italian participants (n = 71) state that they know some effect size statistic. Therefore, there is greater knowledge of the term “effect size” than of the actual statistics of effect sizes, again, like we observed in the study with Spanish academic psychologists, like the study of Badenes-Ribera et al. (2016).

Table 17. Statistical terms the participants know sufficiently (%)

	Chile (n = 35)		Italy (n = 159)	
	%	95% CI	%	95% CI
Confidence Intervals	97.14	[85.47, 99.49]	93.71	[88.81, 96.55]
ANOVA	94.29	[81.39, 98.42]	98.74	[95.53, 99.65]
Regression analysis	94.29	[81.39, 98.42]	98.11	[94.60, 99.36]
Standard deviation	91.43	[77.62, 97.04]	99.37	[96.52, 99.89]
Correlation	80	[64.11, 89.96]	96.86	[92.85, 98.65]
Effect size	80	[64.11, 89.96]	81.76	[75.03, 86.99]
Meta-analysis	80	[64.11, 89.96]	92.45	[87.27, 95.63]
Sedimentation graphic	60	[43.57, 74.45]	8.81	[5.32, 14.24]
Forest plot	20	[10.04, 35.89]	17.61	[12.47, 24.27]
Funnel plot	8.57	[2.96, 22.38]	13.84	[9.32, 20.06]

Note. More than one answer could be selected.

Table 18 shows the effect size statistics known by participants according to their nationality. The effect size statistics most familiar to the participants (Chilean and Italian academics) were those that evaluate the differences between the means of the groups analyzed (standardized mean difference), followed by the proportion of variance explained(η^2) and correlation coefficients.

Table 19 presents the use of statistics in research reports by nationality of the participants. Overall, it can be noted that analysis of variance (ANOVA) was the most widely-used statistic in research reports. It is noteworthy as well that the majority of the Italian participants (52.9%) and 42.9% of the Chilean participants said they use the effect sizes and CIs little (17 % and 18.9% Italian, 8.6% and 8.6% Chilean) or not at all (17% Italian, 25.7% Chilean) in their statistical reports.

In addition, the majority of the Italian participants state that they use MANOVA (57.8%) a fair amount or a lot compared to 28.6% of Chilean participants. Finally, discriminate analysis is the least utilized by Italian and Chilean participants.

Table 18. Known effect size statistics (responses of 20 Chilean and 73 Italian participants)

	Chile		Italy	
	%	95% CI	%	95% CI
Cohen's *d*	75	[53.13, 88.81]	64.38	[52.93, 74.40]
η^2	65	[43.29, 81.88]	58.90	[47.45, 69.47]
Correlation/Association coefficient (Pearson, Spearman, biserial, Phi, Cramer's *V*)	25	[11.19, 46.87]	34.25	[24.39, 45.67]
Cohen's *f*	15	[5.24, 36.04]	4.11	[1.41, 11.40]
Hedge's *g*	10	[2.79, 30.10]	10.96	[5.66, 20.16]
R^2	5	[0.89, 23.61]	13.70	[7.61, 23.41]
w^2	5	[0.89, 23.61]	15.07	[8.63, 25]
Odds Ratio	10	[2.79, 30.10]	8.22	[3.82, 16.79]
Relative Risk	5	[0.89, 23.61]	5.48	[2.15, 13.26]
Glass's delta	5	[0.89, 23.61]	0	[0, 5]
Cohen's f^2	0	[0, 16.1]	6.85	[2.96, 15.05]
Beta	0	[0, 16.1]	2.74	[0.75, 9.45]
Cohen's *q*	0	[0, 16.1]	4.1	[1.41, 11.40]

Note. The majority of participants reported knowing more than one effect size statistic.

Table 19. Use of statistics [95% Confidence Interval]

	Chile (*n* = 35)					Italy (*n* = 159)				
	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized
ANOVA	51.43 [35.57, 67.01]	17.14 [8.10, 32.68]	17.14 [8.10, 32.68]	0 [0, 9.89]	14.29 [6.26, 29.38]	62.89 [55.16, 70.02]	20.75 [15.18, 27.71]	10.06 [6.29, 15.72]	3.14 [1.35, 7.15]	3.14 [1.35, 7.15]
Correlation	42.86 [27.98, 59.14]	28.57 [16.33, 45.05]	11.43 [4.54, 25.95]	0 [0, 9.89]	17.14 [8.10, 32.68]	54.72 [46.96, 62.25]	25.16 [19.05, 32.43]	14.47 [9.84, 20.77]	1.89 [0.64, 5.40]	3.77 [1.74, 7.99]
T tests	37.14 [23.17, 53.66]	31.43 [18.55, 47.98]	14.29 [6.26, 29.38]	2.86 [0.51, 14.53]	14.29 [6.26, 29.38]	27.67 [21.31, 35.09]	37.74 [30.57, 45.48]	22.64 [16.83, 29.75]	7.55 [4.37, 12.73]	4.40 [2.15, 8.81]

Table 19 (Continued)

	Chile (*n* = 35)					Italy (*n* = 159)				
	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized
Regression	40 [25.55, 56.43]	20 [10.04, 35.89]	22.86 [12.07, 39.02]	2.86 [0.51, 14.53]	14.29 [6.26, 29.38]	47.80 [40.18, 55.52]	27.04 [20.74, 34.43]	16.98 [11.94, 23.58]	5.03 [2.57, 9.61]	3.14 [1.35, 7.15]
Effect size	40 [25.55, 56.43]	20 [10.04, 35.89]	5.71 [1.58, 18.61]	14.29 [6.26, 29.38]	20 [10.04, 35.89]	35.85 [28.81, 43.56]	18.24 [13.55, 25.66]	13.84 [9.32, 20.06]	16.98 [11.94, 23.58]	15.09 [10.36, 21.48]
Confidence intervals	37.14 [23.17, 53.66]	14.29 [6.26, 29.38]	14.29 [6.26, 29.38]	17.14 [8.10, 32.68]	17.14 [8.10, 32.68]	27.67 [21.31, 35.09]	27.67 [21.31, 35.09]	17.61 [12.47, 24.27]	16.98 [11.94, 23.58]	10.06 [6.29, 15.72]
Size effect and CIs	25.71 [14.16, 42.07]	31.43 [18.55, 47.98]	8.57 [2.96, 22.38]	8.57 [2.96, 22.38]	25.71 [14.16, 42.07]	25.79 [19.61, 33.10]	21.38 [15.73, 28.39]	16.98 [11.94, 23.58]	18.87 [13.55, 25.66]	16.98 [11.94, 23.58]

Table 19 (Continued)

	Chile (n = 35)					Italy (n = 159)				
	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized
Exploratory factorial analysis	20 [10.04, 35.89]	37.14 [23.17, 53.66]	8.57 [2.96, 22.38]	8.57 [2.96, 22.38]	25.71 [14.16, 42.07]	26.42 [20.18, 33.77]	23.90 [17.94, 31.09]	19.50 [14.09, 26.34]	16.98 [11.94, 23.58]	13.21 [8.80, 19.35]
Confirmatory factorial analysis	14.29 [6.26, 29.38]	25.71 [14.16, 42.07]	17.14 [8.10, 32.68]	11.43 [4.54, 25.95]	31.43 [18.55, 47.98]	23.27 [17.38, 30.42]	18.87 [13.55, 25.66]	21.38 [15.73, 28.39]	24.53 [18.49, 31.76]	11.95 [7.79, 17.91]
Structural equations	14.29 [6.26, 29.38]	17.14 [8.10, 32.68]	11.43 [4.54, 25.95]	25.71 [14.16, 42.07]	31.43 [18.55, 47.98]	17.61 [12.47, 24.27]	16.98 [11.94, 23.58]	16.35 [11.41, 22.88]	22.01 [16.27, 29.07]	27.04 [20.74, 34.43]
MANOVA	8.57 [2.96, 22.38]	20 [10.04, 35.89]	20 [10.04, 35.89]	34.29 [20.83, 50.85]	17.14 [8.10, 32.68]]	24.53 [18.49, 31.76]	33.33 [26.48, 40.98]	21.38 [15.73, 28.39]	14.47 [9.84, 20.77]	6.29 [3.45, 11.19]

	Chile (n = 35)					Italy (n = 159)				
	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized
Exploratory factorial analysis	20 [10.04, 35.89]	37.14 [23.17, 53.66]	8.57 [2.96, 22.38]	8.57 [2.96, 22.38]	25.71 [14.16, 42.07]	26.42 [20.18, 33.77]	23.90 [17.94, 31.09]	19.50 [14.09, 26.34]	16.98 [11.94, 23.58]	13.21 [8.80, 19.35]
Confirmatory factorial analysis	14.29 [6.26, 29.38]	25.71 [14.16, 42.07]	17.14 [8.10, 32.68]	11.43 [4.54, 25.95]	31.43 [18.55, 47.98]	23.27 [17.38, 30.42]	18.87 [13.55, 25.66]	21.38 [15.73, 28.39]	24.53 [18.49, 31.76]	11.95 [7.79, 17.91]
Structural equations	14.29 [6.26, 29.38]	17.14 [8.10, 32.68]	11.43 [4.54, 25.95]	25.71 [14.16, 42.07]	31.43 [18.55, 47.98]	17.61 [12.47, 24.27]	16.98 [11.94, 23.58]	16.35 [11.41, 22.88]	22.01 [16.27, 29.07]	27.04 [20.74, 34.43]
Discriminate analysis	5.71 [1.58, 18.61]	5.71 [1.58, 18.61]	40 [25.55, 56.43]	20 [10.04, 35.89]	28.57 [16.33, 45.05]	3.14 [1.35, 7.15]	11.32 [7.28, 17.18]	22.64 [16.83, 29.75]	37.74 [30.57, 45.48]	24.53 [18.49, 31.76]

Table 20 shows that most of the participants (54.95% of Chilean academics and 66.23% of Italian academics) pointed out that meta-analytic studies are the type of review with the most credibility and objectivity. Nevertheless, 45.05 % of Chilean academics and 33.77% of Italian participants said they give more importance to narrative reviews carried out by experts and/or qualitative reviews. Furthermore, the majority of participants said they have used or read a meta-analytic study for their research.

Table 20. Reading or use of meta-analytic studies (%)

	Chile (*n* = 35)			Italy (*n* = 159)		
	n	%	95% CI	*n*	%	95% IC
Opinions about the review with most credibility and objectivity						
The quantitative review or meta-analysis	25	71.43	[54.95, 83.67]	117	73.58	[66.23, 79.82]
The narrative review carried out by experts	5	14.29	[6.26, 29.38]	26	16.35	[11.41, 22.88]
The qualitative review	5	14.29	[6.26, 29.38]	16	10.06	[6.29, 15.72]
Reading or use of meta-analytic studies						
I have never read or used one	6	17.14	[8.10, 32.68]	44	27.67	[21.31, 35.09]
I have read or used 1-2 meta-analytic studies	12	34.29	[20.83, 50.85]	98	61.64	[53.89, 68.83]
I have read or used more than 2 meta-analytic studies	17	48.57	[32.99, 64.43]	17	10.69	[6.78, 16.45]

Note: CI = Confidence Interval

Finally, Table 21 and Table 22 show the profile of researchers according to be able or not to indicate the name of an effect size statistic and nationality of the participants. Overall and like in the study with Spanish academic psychologists, the academics who gave the name of an effect size statistic have a behavior more close to good statistical practices and of research design. In this way, academics who gave the name of an effect size statistic compared to participants who did not give it had higher proportion of participants who had read or used meta-analysis studies, had been reviewers for scientific journals, had published an article in journals with impact factor JCR (Journal Citation Reports of WoS), and thought that meta-analysis studies are the type of review with the most credibility.

Furthermore, academics who named an effect size statistic perform better methodological practices than the rest of the participants, since a larger proportion of them estimate a priori sample size (both groups have a high proportion), plan the number of participants, and use statistical criteria seeking that the sample represents the characteristics of the population and follow the APA recommendations avoiding expressions of *p* value as $p <$ alpha or $p >$ alpha and using its exact value in a higher proportion than the rest of the participants.

It can be noted that academics who named an effect size statistic confuse in a lesser extent planning the statistical power a priori as a strategy to adjust the significance level or alpha value, and also make in a lesser extent the clinical or practical size fallacy where the statistical significance of the effect is related to the importance of effect, like in the study with Spanish academic Psychologists. Also, they said they knew that currently there is some kind of open debate on statistical issues or research design, which do not agree with the findings of the study with Spanish academic psychologists, where most of the participants (who called an effect size statistics and who did not so) said they did not know that currently there is some kind of open debate on statistical issues or research design.

Finally, the majority of the participants said that they did not know any checklist to assess the design quality of a study, like we observed in the study with Spanish academic psychologists.

Table 21. Researcher's behavior and opinion according to knowing or not knowing the name of effect size statistics (%) [and 95% Confidence Intervals]

	Chile		Italy	
Item	Not Knowing (*n* = 16)	Knowing (*n* =19)	Not Knowing (*n* = 88)	Knowing (*n* =71)
1 Have you read or used a meta-analytic study?				
I have never read or used one	31.25	5.26	35.23	18.31
	[14.16, 55.60]	[0.94, 24.64]	[26.06, 45.63]	[11.02, 28.85]
Yes: I have read or used 1 -2 meta-analytic studies	**43.75**	26.32	**52.27**	**73.24**
	[23.10, 66.82]	[11.81, 48.79]	**[41.96, 62.39]**	**[61.95, 82.15]**
Yes, I have read or used more than 2 meta-analytic studies	25 [	**68.42**	12.50	8.45
	10.18, 49.50]	**[46.01, 84.64]**	[7.13, 21.01]	[3.93, 17.24]

Table 21 (Continued)

	Chile		Italy	
Item	Not Knowing (*n* = 16)	Knowing (*n* =19)	Not Knowing (*n* = 88)	Knowing (*n* =71)
2 Have you been reviewer for scientific journals in the last year?				
No	**68.75**	21.05	21.59	7.04
	[44.40, 85.84]	[8.51, 43.33]	[14.28, 31.28]	[3.05, 15.45]
Yes: 1-2 reviewed articles	25	47.37	25	32.39
	[10.18, 49.50]	[27.33, 68.29]	[17.13, 34.96]	[22.66, 43.94]
Yes: more than 2 reviewed articles	6.25	**31.58**	**53.41**	**60.56**
	[1.11, 28.33]	**[15.36, 53.99]**	**[43.06, 63.47]**	**[48.94, 71.11]**

Table 21 (Continued)

	Chile		Italy	
Item	Not Knowing (n = 16)	Knowing (n =19)	Not Knowing (n = 88)	Knowing (n =71)
3 Have you published an article in a journal indexed in the WoS with JCR impact factor in the last year?				
No	**50**	36.84	22.73	11.27
	[28, 72]	[19.15, 58.96]	[15.22, 32.51]	[5.82, 20.69]
Yes: 1-2 published articles	**43.75**	**47.37**	31.82	33.80
	[23.10, 66.82]	**[27.33,68.29]**	[23.02, 42.13]	[23.88, 45.38]
Yes: more than 2 published articles	6.25	15.79	**45.45**	**54.93**
	[1.11, 28.33]	[5.52,37.57]	**[35.46, 55.83]**	**[43.40, 65.95**]

Table 21(Continued)

	Chile		Italy	
Item	Not Knowing (*n* = 16)	Knowing (*n* =19)	Not Knowing (*n* = 88)	Knowing (*n* =71)
4. What type of review do you think has the most credibility and objectivity?				
The narrative review carried out by experts	12.50	15.79	17.05	15.49
	[3.50, 36.02]	[5.52, 37.57]	[10.61, 26.24]	[8.88, 25.65]
The quantitative review or meta-analysis	**62.50**	**78.95**	**65.91**	**83.10**
	[38.64, 81.52]	**[56.67, 91.49]**	**[55.53, 74.96]**	**[72.74, 90.06]**
The qualitative review	25	5.26	17.05	1.41
	[10.18, 49.50]	[0.94, 24.64]	[10.61, 26.24]	[0.25, 7.56]
5. Do you know checklist for assessing research design of a study?				
No	**68.75**	**52.63**	**87.50**	**83.10**
	[44.40, 85.84]	**[31.71, 72.67]**	**[78.99, 92.87]**	**[72.74, 90.06]**
Sí	31.25	47.37	12.50	16.90
	[14.16, 55.60]	[27.33, 68.29]	[7.13, 21.01]	[9.94, 27.26]

Table 21 (Continued)

	Chile		Italy	
Item	Not Knowing (n = 16)	Knowing (n =19)	Not Knowing (n = 88)	Knowing (n =71)
6. In your opinion, what statistical questions or issues related to the study design are currently being debated?				
I don't know	**62.50**	36.84	**61.36**	26.76
	[38.64, 81.52]	[19.15, 58.96**]**	**[50.92, 70.86]**	[17.85, 38.05]
I don't think there are any debates open	12.50	0	9.09	4.23
	[3.50, 36.02]	[0, 16.82]	[4.68, 16.93]	[1.45, 11.70]
There is some debate	25 [10.18, 49.50]	**63.13 [41.04, 80.85**]	29.55 [21.03, 39.77]	**69.01 [57.52, 78.56]**

Tabla 22. Researcher's methodological behavior according to knowing or not knowing the name of effect size statistics (%) [and 95% Confidence Intervals]

	Chile		Italy	
Item	Not Knowing (*n* = 16)	Knowing (*n* =19)	Not Knowing (*n* = 88)	Knowing (*n* =71)
1. In your opinion, obtaining a statistically significant result implies indirectly that the detected effect is important				
No	**83.33**	**100**	**90**	**92.19**
	[55.20, 95.30]	**[82.41, 100]**	**[80.77, 95.07]**	**[82.98, 96.62]**
Yes	16.67	0	10	7.81
	[4.70, 44.80**]**	[0, 16.82]	[4.93, 19.23]	[3.38, 17.02]
2. When you plan a study, do you estimate a priori the sample size you will need?				
No	25	10.53	15.91	16.90
	[10.18, 49.50]	[2.94, 31.39]	[9.72. 24.95]	[9.94, 27.26]
Yes	**75**	**89.47**	**84.09**	**83.10**
	[50.50, 89.82]	**[68.61, 97.06]**	**[75.05, 90.28]**	**[72.74, 90.06]**

Table 22 (Continued)

	Chile		Italy	
Item	Not Knowing (n = 16)	Knowing (n =19)	Not Knowing (n = 88)	Knowing (n =71)
3. What kind of strategy do you use when you want to plan the sample size of a study?				
You try to achieve the greatest number of participants possible	12.50	10.53	25	22.54
	[3.50, 36.02]	[2.94, 31.39]	[17.13, 34.96]	[14.37, 33.52]
You use software or tables to estimate the sample size according to the statistical criteria	**56.25**	**68.42**	9.09	**32.39**
	[33.18, 76.90]	**[46.01, 84.64]**	[4.68, 16.93]	**[22.66, 4394**]
You try to make the sample represent the characteristics of the population	0	5.26	**56.82**	29.58
	[0, 19.36]	[0.94, 24.64]	**[46.40, 66.67]**	[20.23, 41.02]
You do not use any strategy because it isn't part of your research interests.	31.25	15.79	9.09	15.49
	[14.16, 55.60]	[5.52, 37.57]	[4.68, 16.93]	[8.88, 25.65]

Table 22 (Continued)

	Chile		Italy	
Item	Not Knowing (n = 16)	Knowing (n =19)	Not Knowing (n = 88)	Knowing (n =71)
4. In your opinion, what is the purpose of calculating the statistical power a priori?				
To adjust the significance level or alpha value	25	0	28.41	4.23
	[10.18, 49.50]	[0, 16.82]	[20.04, 38.58]	[1.45, 11.70]
To explore the reliability of the scales	12.50	0	10.23	4.23
	[3.50, 36.02]	[0, 16.82]	[5.47,18.31]	[1.45, 11.70]
To estimate the sample size	**37.50**	**89.47**	**42.05**	**76.06**
	[18.48, 61.36]	**[68.61, 97.06]**	**[32.28, 52.48]**	**[64.96, 84.48]**
I don’t know/don’t respond	25	10.53	19.32	15.49
	[10.18, 49.50]	[2.94, 31.39]	[12.43, 28378]	[8.88, 25.65]

Table 22 (Continued)

	Chile		Italy	
Item	Not Knowing (*n* = 16)	Knowing (*n* =19)	Not Knowing (*n* = 88)	Knowing (*n* =71)
5. When you perform a statistical test, do you consider it a priority to always report the statistical significance obtained?				
No	31.25	0	9.09	22.54
	[14.16, 55.60]	[0, 16.82]	[4.68, 16.93]	[14.37, 33.52]
Yes, and using expressions like p<0.05, p>0.05	**56.25**	42.11	**68.18**	**40.85**
	[33.18, 76.90]	[23.14, 63.72]	**[57.87, 76.98]**	**[30.17,52.46]**
Yes, and using expressions with the p value of exact probability	12.50	**57.89**	22.73	**36.62**
	[3.50, 36.02]	**[36.28, 76.86]**	[15.22, 32.51]	**[26.37, 48.24]**

5.4.4. Discussion

Like in the original study with Spanish academic psychologists, the findings at the present study indicate that the emphasis the statistical reform places on the use of the ES and its confidence interval has also had an impact on Chilean and Italian academic psychologists, especially the estimation of the effect size. The majority of the participants stated that they use effect size statistics and effect size and their confidence intervals a fair amount. Therefore the results point to a higher self-reported use of the ES and CI than prior studies, but CIs were reported not nearly as frequently as effect size point estimate (Fritz et al., 2012; Peng et al., 2013; Sesé & Palmer, 2012). Again, this results goes against the APA recommendation that "*Whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size*" (APA, 2010a, p. 34), like the study by Badenes-Ribera, Frías-Navarro, Pascual-Soler et al. (2016). It could be expected that they will improve in future studies, since the change in statistical practices takes time.

Regarding the type of effect size statistic they know, the participants mentioned to a greater degree the effect size statistics from the family of standardized mean differences and η^2 (parametric effect size statistics). These findings are in line with previous researches that analyze the use of effect size statistics in journals. For example, Peng et al. (2013) found that the most frequently reported ES measures were R^2, Cohen's *d*. Nevertheless, standardized differences in means (e.g. Cohen's *d*, Glass' delta, Hedges' *g*,) and from the family of correlation (Pearson's correlation, R^2, η^2, omega^2, and so on) have been criticized for lack of robustness against outliers or departure from normality, and instability under violations of statistical assumptions (Algina et al., 2005; Grissom & Kim, 2012; Kline, 2013, Peng & Chen, 2014; Wang & Thompson, 2007).

In addition, the findings suggest that the modern robust statistical methods are not known by most of the participants, or at least, majority of the participants did not give the name of robust effect size statistics, like in the study of Spanish academic psychologists (Badenes-Ribera, Frías-Navarro, Pascual-Soler et al., 2016).

Regarding the opinion about reviews, the majority of the participants give more credibility and objectivity to systematic reviews and meta-analytic studies than to other types of literature reviews. Also, they have an adequate knowledge of meta-analyses.

However, they have a poor knowledge of graphical displays for meta-analyses (i.e., forest plots and funnel plots) which can become misinterpretation of results. The graphical presentation of results is an important part of a meta-analysis and it has become the primary tool for presenting the results of multiple studies on the same research question (Anzures-Cabrera & Higgins, 2010; Borenstein et al., 2009; Ellis, 2010; Sánchez-Meca & Marín-Martínez, 2010).

Finally, like in the study by Badenes-Ribera, Frías-Navarro, Pascual-Soler et al. (2016) the analysis of researcher's behavior associated with methodological practices point out that academics who know some effect size statistics present a profile more close to good statistical practices and design research, participate more actively in the process of peer reviewing, and publish in journals with impact.

It must be acknowledged several limitations of this study. Firstly, the low response rate could affect the representativity of the sample and, consequently, the generalizability of the results. Moreover, it is possible that the participants who responded to the survey had higher level of statistical knowledge than those who did not respond. Should this be the case, the results might overestimate the extension of the impact of the statistical reform in Chilean and Italian academic psychologists. Furthermore, it must also be acknowledged that some participants do not use quantitative methods at all. These individuals may have been less likely to respond, as well. Nevertheless, the findings are in line with the study with Spanish academic psychologists (Badenes-Ribera, Frías-Navarro, Pascual-Soler et al., 2016), and prior researches that analyzed the use of effect size statistics in journals (Fritz et al., 2012; McMillan & Foley, 2011; Peng et al., 2013; Sesé & Palmer, 2012).

In addition, it is possible that there has been an effect of social desirability as it may always happen when data are collected through self-report questionnaires. For instance, like in the study by Badenes-Ribera, Frías-Navarro, Pascual-Soler et al. (2016), the percentage of Italian and Chilean academic psychologists who stated that could give the name of an effect size statistic was higher than the percentage of them who actually did it. A way of control this bias in future research would be formulate the questions (e.g., what is the correct interpretation of a specific forest plot, funnel plot, effect size or regression analysis) with three or four-response format, or with open-end

question. These response formats would permit us to assess the level of knowledge of statistical terms, thus they would have been far more informative.

5.5. Overall discussion, conclusion and methodological recommendations

As it has been acknowledged before, the studies conducted have several limitations (e.g., low response rate, social desirability and so on), however, the findings of the three studies are consistent among them.

Taking into account these limitations, the results are novel because, until now, there were no self-report data about the following of the statistical reform and the APA Manual recommendations among Spanish, Italian and Chilean researchers, even though these recommendations have to be followed in almost all of the psychological journals.

The results of these studies indicate that the emphasis the statistical reform places on the use of the ES and its confidence interval has also had an impact on participants, especially the estimation of the effect size. The majority of surveyed stated that they use effect size statistics a fair amount or a lot. And, nearly to majority of them said that they used effect sizes and their confidence intervals a fair amount or a lot, also. Therefore the results point to a higher self-reported use of the ES and CIs compared to previous studies (Badenes-Ribera et al., 2013; Caperos & Pardo, 2013; Frías-Navarro et al., 2012; Sesé & Palmer, 2012). However, CIs were reported not nearly as frequently as effect size point estimate (along the same lines, Badenes-Ribera et al., 2013; Fritz et al., 2012; Peng et al., 2013; Sesé & Palmer, 2012). These findings go against the APA recommendation that "*Whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size*" (APA, 2010a, p. 34). It could be expected that they will improve in future studies, since the change in statistical practices takes time.

Nevertheless, the change in statistical practice is slow, if we take into account that the recommendations about using the effect size and its confidence interval was introduced in the 1999 report by the statistical inference workgroup of the American Psychological Association (Wilkinson & TFSI, 1999). The elaboration of this report was the APA's response to a broad set of criticisms against the null hypothesis

statistical technique (NHST), proposing an improvement in the statistical practices (Balluerka et al., 2009; Nickerson, 2000; Monterde-i-Bort et al., 2010).

Regarding the type of effect size statistic they know, the participants mentioned to a greater degree the effect size statistics from the family of standardized mean differences and η^2 (parametric effect size statistics). These findings are in line with previous researches that analyze the use of effect size statistics in journals. For example, Peng et al. (2013) found that the most frequently reported ES measures were R^2 and Cohen's *d*. Nevertheless, standardized differences in means (e.g. Cohen's *d*, Glass' delta, Hedges' *g*) and from the family of correlation (Pearson's correlation, R^2, η^2, omega^2, and so on) have been criticized for lack of robustness against outliers or departure from normality, and instability under violations of statistical assumptions (Algina, et al., 2005; Grissom & Kim, 2012; Kline, 2013; Peng & Chen, 2014; Wang & Thompson, 2007; Wilcox, 2010).

There are theoretical reasons and empirical evidence that outliers and violations of assumptions are common in practice (Erceg-Hurn & Mirosevich, 2008; Grissom & Kim, 2001). Consequently, researchers should consider using effect size statistics that are more resistant to outliers and violations of statistical assumptions (Erceg-Hurn & Mirosevich, 2008; Grissom & Kim, 2012; Keselman et al., 2008; Kline, 2013). Moreover, confidence intervals are not immune to outliers or departure from normality and the violations of statistical assumptions.

There are some alternatives for parametric effect size statistics: on the one hand, non-parametric effect size statistics, such as Spearman's correlation (*rho*), Cliff's delta, and so on, and on the other hand, the modern robust effect sizes statistics, such as, the robust standardized mean differences based on robust estimators (trimmed means and winsorized variances), the probability of superiority (PS) which is defined as the probability that a randomly sampled score from one population is larger than a randomly sampled score from a second population, the number needed to treat (NNT), an effect size index appropriate for conveying information in psychotherapy outcome studies or other behavioral research that involves comparisons between treatments or between treatment and control or placebo conditions (e. g., Arnau, Bendayan, Blanca, & Bono, 2013; Erceg-Hurn & Mirosevich, 2008; Grissom & Kim, 2012; Keselman et al., 2008; Wilcox & Keselman, 2003; Wilcox, 2012).

The results suggest that the modern robust statistical methods are not known by most participants, or at least, the participants did not give the name of robust effect size statistics. In fact, only 0.9% of the Spanish academic psychologists called a robust effect size statistic (NNT). As Erceg-Hurn and Mirosevich (2008) pointed out this might be due to lack of exposure to these methods. In this way, "*the psychology statistics curriculum, journal articles, popular textbooks, and software are dominated by statistics developed before the 1960s*" (*op. cit*., p.593).

Regarding the knowledge of meta-analytic studies, the majority of the participants give more credibility and objectivity to systematic reviews and meta-analytic studies than to other types of literature reviews. Also, they have an adequate knowledge of meta-analyses. However, they have a poor knowledge of graphical displays for meta-analyses (e.g., forest plots and funnel plots) which can become in a misinterpretation of results. The graphical presentation of results is an important part of a meta-analysis and it has become the primary tool for presenting the results of multiple studies on the same research question (Anzures-Cabrera & Higgins, 2010; Borenstein et al., 2009; Ellis, 2010; Sánchez-Meca & Marín-Martínez, 2010).

In addition, it is known that publication bias is common in psychological meta-analytic studies. For example, Ferguson & Brannick (2011) reviewed 91 meta-analyses published in American Psychological Association and Association for Psychological Science Journal and found 41% of the meta-analyses reported finding evidence of publication bias. Publication bias is an important threat to the validity of meta-analytic studies, since meta-analytically derived estimates could be inaccurate, typically overestimated. In fact, as Kepes et al. (2014) point out, given the influence of meta-analytic studies on future research directions and evidence-based practice, publication bias has been referred to as ''the Achilles' heel of systematic reviews'' (Torgerson, 2006), ''*the kryptonite of evidence-based practice*'' (Banks & McDaniel, 2011), and the ''*antagonist of effective policy making*'' (Banks, Kepes & Banks, 2012). It should be noted that funnel plot is used as publication bias detection method in the health sciences (Sterne et al., 2005). Therefore, researchers, academics and practitioners must adequately know funnel plots, which is a basic tool of meta-analytic studies to detect bias publication and heterogeneity of effect sizes.

The analysis of researcher's behavior associated with its methodological practices point out that academics who know some effect size statistics present a profile more close to good statistical practices and design research, participate more actively in the process of peer review, and publish in journals with impact.

However, three issues alert on the knowledge that Spanish, Chilean and Italian academic psychologists have about effect size and validity of statistical conclusion in general: they associate wrongly effect size with the importance of a finding (clinical or practical significance fallacy), they continue to use in a high proportion *p*-value expressions that revolve around the oracle of the value of alpha and, they don't know the purpose of planning a priori statistical power.

Finally, two events that have allowed the science debate on statistical procedures, progress towards a statistical reform and greater transparency and quality of studies, such as, the open debate on the uses and abuses of statistical significance tests (which started almost since the beginning of its use) and the development of check tools such checklist (CONSORT, STROBE, PRISMA…), continue to be unknown in a high proportion by academic psychologists and Spanish practitioner psychologists.

The Evidence-based Practice requires professionals to critically evaluate the results of psychological research in order to decide whether their use is appropriate or not (Beyth-Maron et al., 2008; Frías-Navarro, 2011a; Sánchez-Meca & Botella, 2010M Vázquez & Nieto, 2003). The information provided by the studies depends on the statistical analyses performed; therefore, their value largely depends on the quality of the statistical analyses and the interpretation of the results (Cumming, 2012; Cumming et al., 2012; Kline, 2013; Palmer & Sesé, 2013; Wilkinson & the TFI, 1999).

Estimating effect sizes means contextualizing their value within a research area, and not only deciding whether an effect is statistically significant or not. The interpretation of the magnitude of the effect implies making a judgment within a specific research context, indicating whether it is a small, medium or big effect. To make this judgment, the researcher must pose questions of practical and/or clinical significance, abandoning the emphasis on whether the result was or was not statistically significant (Cumming et al., 2012).

Estimating effects and evaluating their magnitude within a specific context means that researchers' statistical practices must be complemented by their experience and judgment, fomenting Evidence-based Practice. This type of performance facilitates better comprehension of the study results, and it helps professionals (practitioners) in the true interpretation of the findings and their possible use in their clinical practice.

6. CONCLUSION

Currently there is an open scientific and social debate that could change the course of statistical practices among researchers. For example, during the last three years criticism against the classical statistical inference procedure based on the probability value *p* and the dichotomous decision to keep or reject the null hypothesis has been hardened (Allison et al., 2016; Nuzzo, 2014; Wasserstein & Lazar, 2016). In addition, the low proportion of replication studies, publication bias that lead to an overestimation of the magnitude of effects, questionable statistical practices (Questionable Research Practices, QRPs) leading to find statistically significant results (called *p*-hacking), such as recording many response variables and deciding which to report after the analysis, reporting only statistically significant results, remove outliers and increase sample size to get statistical significance, and fraud also are current issues of discussion (Earp & Trafimow, 2015; Ioannidis, 2005a, 2005b; Kepes et al., 2014).

The realization of this study has tried to contribute to this debate, providing evidence of the current state of affairs, in what refers to the knowledge and practices of academic and professional psychologists in relation to methodology and research designs.

The findings of this work are an empirical evidence of all the inappropriate behaviors surrounding the process of statistical inference and that for decades have been studied by researchers, such as misinterpretations and misuse of statistical inference techniques due to statistic and effect size fallacies that surround it. Academics, scientists and professionals are not immune to such beliefs. The problem has not been resolved despite the recommendations and alerts that have been permanently detailed in scientific publications. Statistical reeducation to correct the errors of interpretation of the various fallacies and incorporating an Evidence Based Statistical Practice oriented to the conscious and explicit use of all elements surrounding the process of statistical inference is essential to interpret critically the results of statistical inference.

Evidence-Based Practice requires professionals to critically evaluate the findings of psychological research. In order to do so, training is necessary in statistical concepts, research design methodology, and results of statistical inference tests.

In this sense, to improve teaching and the appropriate use of the NHST provide a better way for the future than abandon it. In addition, it should be noted that research problems are not only linked to the use of p values or the analysis of data; in fact they can affect any stage of research design from the literature review to the interpretation of results in a particular study.

Therefore, new programs and manuals of statistics are needed. For example, textbooks should include a section on the current debate and criticisms of the NHST procedure, in terms of whether statistical significance tests are or not the best way to advance the body of valid scientific knowledge. In so far as researchers are aware of the issues surrounding the use of p values, they will develop a critical or active reading when reviewing the literature of their research area. For instance, the proper use of statistical inference tests that are required to check the assumptions of the statistical model in a study is supposed to improve the quality of evidence. In addition, the statistical concepts should be presented correctly, and professors should be prepared to teach the concepts properly.

Furthermore, programs and manuals of statistics should include alternatives to traditional methods. They should consider statistics that are more resistant to outliers and robust to violations of the assumptions of population normality and homogeneity of variance (e.g., modern robust statistical methods, such as, trimmed mean and winsorized variance), and add information about how to calculate and report the effect size and its confidence intervals, both in statistically significant results and in the non-significant ones. And finally, the authors should give examples in order to decide whether the result has practical or clinical importance.

The teaching of statistical inference techniques also requires students to develop critical thinking that allows them to assess the quality of research design, acquiring knowledge on a wide variety of analytical techniques and not only on statistical significance test.

On the other hand, as Kirk (2001) points out, "*authors of statistical software packages have a responsibility to assist researchers in following the best practices in statistics*" (p. 216). In this ways, statistical software programs (e.g., SPSS) should also be updated to include in their menus other techniques such as the estimation of confidence intervals for parametric effect size statistics, and the estimation of effect size statistics more resistant to extreme values (outliers) and violations of the assumptions of the parametric tests (normal distribution and homogeneity of variance), such as modern robust effect size statistics and their confidence intervals. There are several websites that offer computing routines/programs for general or specific effect size estimators and confidence intervals of various effect sizes.

Finally, as Wasserstein & Lazar (2016) point out:

> Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning (p. 132).

Therefore, the debate should be focused on how to improve statistical practices and the improvement should include discussing the understanding of statistical methods but also other basic issues surrounding the process of research design, such as the review and critique of literature, hypothesis formulation, planning the study, collecting data, checking statistical assumptions, the development of the research report and replication of the findings.

In short, good statistical practice, good research design and correct interpretation of the results in a context are essential components of a good scientific practice for the accumulation of a valid scientific knowledge. There remains a clear need to raise awareness among professionals and academics in Psychology about these guidelines, but especially to promote the education and training in the use of these practices in order to build better science.

7. REFERENCIAS

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). En L. L. Harlow, S. A. Mulaik, y J. H. Steiger (Eds.), *What if there were no significance tests*? (pp. 117–141). Mahwah, NJ: Lawrence Erlbaum

Agresti, A., y Finaly, B. (2009). *Statistical methods for the social sciences* (4th. Edition). New Jersey: Pearson Prentice Hall.

Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., y Dalton, C. M. (2011). Debunking myths and urban legends about meta-analysis. *Organizational Research Methods, 14*, 306–331. doi:10.1177/1094428110375720.

Aguinis, H., Werner, S, Abbott, J, Angert, C, Park, J. H, y Kohlhausen, D. (2010). Customer-Centric Science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods 13*, 515-539. doi: 10.1177/1094428109333339

Algina, J., Keselman, H. J., y Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10*, 17–328. doi: 10.1037/1082-989X.10.3.317

Allison, D. B., Brown, A. W., George, B. J., y Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. Nature, 530, 27-29. doi: 10.1038/530027a.

American Psychological Association. (1952). *Publication Manual of the American Psychological Association* (1st ed.). Washington, DC: American Psychological Association

American Psychological Association. (1974). Publication Manual of the American Psychological Association (2nd ed.). Washington, DC: American Psychological Association.

American Psychological Association. (1983). *Publication Manual of the American Psychological Association (3rd ed.).* Washington, DC: American Psychological Association.

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

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

American Psychological Association. (2006). Evidence-based practice in psychology: APA Presidential Task Force on evidence-based practice. *American Psychologist, 61,* 271–285. doi:10.1037/0003-066X.61.4.271

American Psychological Association (2010a). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Psychological Association (2010b). *Ethical Principle of Psychologist and Code of Conduct*. Disponible en: www.apa.org/ethics/

Anderson, D. R., Burnham, K. P., y Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*, 912-923. doi: 10.2307/3803199

Anzures-Cabrera, J., y Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods, 1,* 66-80. doi:10.1002/jrsm.6

Armijo-Olivo, S, Warren, S., Fuentes, J., y Magee, D. J. (2011) Clinical relevance vs. statistical significance: Using neck outcomes in patients with temporomandibular disorders as an example. *Manual Therapy 16,* 563-572. doi:10.1016/j.math.2011.05.006

Arnau, J., Bendayan, R., Blanca, M. J., y Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*, 873-879. doi: 10.3758/s13428-012-0306-x.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … y Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108-119. doi: 10.1002/per.1919

Babione, J. M. (2010). Evidence-Based Practice in Psychology: An ethical framework for graduate education, clinical training, and maintaining professional competence. *Ethics y Behavior, 20*, 443-453. doi:10.1080/10508422.2010.521446

Badenes-Ribera, L. (2013a). *Prevalencia y correlatos de la violencia de pareja en mujeres lesbianas: Una revisión sistemática*. Trabajo fin de Máster. Facultad de Psicología. Universidad de Valencia. No publicado.

Badenes-Ribera, L. (2013b). *Violencia de pareja en mujeres lesbianas: Un meta-análisis de su prevalencia.* Trabajo fin de Máster. Facultad de Derecho. Universidad de Valencia. No publicado.

Badenes-Ribera, L. Bonilla-Campos, A. & Frías-Navarro, D. (2016). Barriers to Evidence Based Practice in Spanish practitioner psychologists: An exploratory study. (En revision)

Badenes-Ribera, L., y Frías-Navarro, D. (2014). Análisis del razonamiento inferencial en docentes universitarios de Psicología. En D. Frias-Navarro, M. Pascual-Soler, Badenes-Ribera, L., y H. Monterde-i-Bort., H. (Eds.). *Reforma estadistica en Psicología* (pp. 110-150). Valencia: Palmero Ediciones.

Badenes-Ribera, L., Frias-Navarro, D., Bonilla-Campos, A., Pons-Salvador, G., y Monterde-i-Bort., H. (2015). Intimate Partner Violence in Self-identified Lesbians: a Meta-analysis of its Prevalence. *Sexuality Research and Social Policy, 12,* 47-59. doi:10.1007/s13178-014-0164-7

Badenes-Ribera. L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., y Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian academic psychologists. *Frontiers in Psychology, 7,* 1247. doi: 10.3389/fpsyg.2016.01247

Badenes-Ribera, L., Frias-Navarro, D., Monterde-i-Bort, H., y Pascual-Soler, M. (2013, Febrero). *Informar e interpretar el tamaño del efecto en Psicología y Educación*. Trabajo presentado en el XIV Congreso Virtual de Psiquiatria.com. Interpsiquis, Palma de malloca, España.

Badenes-Ribera, L., Frias-Navarro, D., Monterde-i-Bort, H., y Pascual-Soler, M. (2015). Interpretation of the p value. A national survey study in academic psychologists from Spain. *Psicothema*, 27, 290-295. doi: 10.7334/psicothema2014.283

Badenes-Ribera, L., Frias-Navarro, D., y Pascual-Soler, M. (2015). Errors d'interpretació dels valor p en estudiants universitaris de Psicologia. *Anuari de Psicologia de la Societat Valenciana, 16*, 15-32. doi: 10.7203/anuari.Psicologia.16.2.15

Badenes-Ribera, L., Frias-Navarro, D., Pascual-Soler, M., y Monterde-i-Bort, H. (2016). Knowledge level on effect size statistics, confidence intervals and meta-analysis in Spanish academic psychologists. *Psicothema*, 28, 448-456. doi: 10.7334/psicothema2016.24

Baicus, C., y Cariol, S. (2009). Leeter a Editor: effect measure for quantiative endpoints: statsitical versus clinical significance, or "how large the scale is?". European *Journal of Internal Medicine, 20*, 124-125. doi: 10.1016/j.ejim.2008.10.002

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423–437. doi: 10.1037/h0020412

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37,* 379–384. doi: 10.3758/BF03192707

Bakker, M. (2014). *Good science, bad science.Questioning research practices in Psychological research*. Enschede (Netherland): Ipskamp Drukkers.

Bakker, M., y Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*, 666-678. doi:10.3758/s13428-011-0089-5

Balluerka, N., Gómez, J., y Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology, 1*, 55–70. doi:10.1027/1614-1881.1.2.55

Balluerka, N., Gorostiaga, A., Alonso-Arbiol, I., y Haranburu, M. (2007). La adaptación de instrumentos de medida de unas culturas a otras: una perspectiva práctica.

Psicothema, 19, 124-133. Disponible en: http://www.psicothema.com/psicothema.asp?id=3338

Balluerka, N., Vergara, A. I., y Arnau, J. (2009). Calculating the main alternatives to Null Hypothesis Significance testing in between subject experimental designs. *Psicothema, 21,* 141-151. Disponible en: http://www.psicothema.com/psicothema.asp?id=3607

Banks, G. C., Kepes, S., y Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis, 34*, 259–277. doi: 10.3102/0162373712446144

Banks, G. C., Kepes, S., y McDaniel, M. A. (2012). Publication Bias: A call for improved meta-analytic practice in the organizational sciences. *International Journal of Selection and Assessment, 20,* 182-196. doi: 10.1111/j.1468-2389.2012.00591.x

Banks, G. C., y McDaniel, M. A. (2011). The kryptonite of evidencebased I-O psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4,* 40–44. doi: 10.1111/j.1754-9434.2010.01292.x.

Barendregt, J. J., Doi, S. A., Lee, Y. Y., Norman, R. E., y Vos, T. (2013). Meta-analysis of prevalence. *Epidemiology Community and Health, 67,* 974-978. doi:10.1136/jech-2013-203104.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41,* 257-278. doi: 10.1111/j.2044-8317.1988.tb00901.x

Belia, S., Fidler, F., Williams, J., y Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389-396. doi:10.1037/1082-989X.10.4.389

Berben, L., Sereika, S. M., y Engberg, S. (2012). Effect size estimation: Methods and examples. *International Journal of Nursing Studies 49,* 1039–1047. doi:10.1016/j.ijnurstu.2012.01.015

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526–536. Recuperado 18/0/2016 desde https://www.jstor.org/stable/pdf/2279690.pdf

Beyth-Maron, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal, 7,* 20-39. Recuperado 18/08/2016 desde http://iase-web.org/documents/SERJ/SERJ7(2)_Beyth-Maron.pdf

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27,* 335–340. doi: 10.3102/10769986027004335

Bonett, D. G. (2009). Meta-analytic confidence intervals for standardized and unstandardized mean differences. *Psychological Methods, 14*, 225–238. doi: 10.1037/a0016619

Bonett, D. G. (2010). Varying coefficient meta-Analytic methods for alpha reliability. *Psychological Methods, 15*, 368-385. doi: 10.1037/a0020142

Borenstein, M., Hedges, L. V., Higgins, J. P. T. y Rothstein, H. (2009). *Introduction to Meta-analysi*s. Chichester : Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., y Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97–111. doi: 10.1002/jrsm.12

Borenstein, M. J., Hedges, L. V., Higgins, J., y Rothstein, H. (2014). *Comprehensive meta-analysis Version 3.0*. Englewood, NJ: Biostat.

Borenstein, M. J., y Higgins, J. (2013). Meta-analysis and subgroups. *Preventive Science, 14*, 134-143. doi 10.1007/s11121-013-0377-7

Borges, A. (1997) Algunos problemas frecuentes en la interpretación de los contrastes de hipótesis estadísticas en psicología. *Iberpsicología, 2, (3),* 7. Recuperado desde: https://www.researchgate.net/publication/28059708_Algunos_problemas_frecuentes_en_la_interpretacion_de_los_contrastes_de_hipotesis_estadisticas_en_psicologia

Borges, A., San Luis, C., Sánchez, J. A., y Cañadas, I. (2001). El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa. *Psicothema, 13*, 173-178. Disponible en: http://www.psicothema.com/psicothema.asp?id=430

Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin, 16,* 335–338. doi:10.1037/h0074554

Botella, J., y Gambara, H. (2002). *Qué es el meta-análisis*. Madrid: Biblioteca Nueva.

Botella, J., y Gambara, H. (2006). Doing and reporting a meta-analysis. *International Journal of Clinical and Health Psychology, 6,* 425-440.

Botella, J., y Sánchez-Meca, J. (2015). *Meta-análisis en ciencias sociales y de la salud*. Madrid: Ediciones Síntesis.

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online [serie en línea], 4(2).* Acceso 22/03/2016. Disponible en http://www.dgps.de/fachgruppen/methoden/mpr-online/issue7/art2/brandstaetter.pdf

Brandt, M., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217-224. doi:10.1016/j.jesp.2013.10.005

Bustamante, F. J., y Delgado, J. (1994). Las corrientes del meta-análisis: Líneas de convergencia. *Psicológica, 15*, 225-274.

Campitelli, G. (2015). Answering research questions without calculating the mean. *Frontiers in Psychology, 6,* 1379-1381. doi: 10.3389/fpsyg.2015.01379

Caperos, J. M., y Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*, 408-414. doi: 10.7334/psicothema2012.207

Carlson, K. D., y Schmidt, F. L. (1999). Impact of experimental design on effect sizes: Findings from the research literature on training. *Journal of Applied Psychology, 84,* 851-862. doi: 10.1037/0021-9010.84.6.851

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287–292. doi:10.1080/00220973.1993.10806591

Catalá-López, F., y Tobías, A., (2013). Síntesis de la evidencia clínica y metaanálisis en red con comparaciones indirectas. *Medicina Clínica, 140*, 2013, 182–187. doi:10.1016/j.medcli.2012.09.013

Catalá-López, F., y Tobías, A., (2014). Metaanálisis de ensayos clínicos aleatorizados, heterogeneidad e intervalos de predicción. *Medicina Clínica, 142*, 270-274. doi: 10.1016/j.medcli.2013.06.013

Catalá-López, F., Tobías, A., y Roqué, M. (2014). Conceptos básicos del meta-análisis en red. *Atención Primaria, 46*, 573-581. doi: 10.1016/j.aprim.2014.01.006

Chandler, J., Churchill, R., Higgins, J., Lasserson, T., y Tovey, T. (2013). Methodological standards for the conduct of new Cochrane intervention reviews (MECIR). Disponible en: http://editorial-unit.cochrane.org/mecir

Cipriani, A., Higgins, J. P. T., Geddes, J. R., y Salanti, G. (2013). Conceptual and technical challenges in network meta-analysis. *Annals of Internal Medicine: Journal, 159*, 130-137. doi:10.7326/0003-4819-159-2-201307160-00008

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494–509. doi:10.1037/00332909.114.3.494

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101-129.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312. doi: 10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003. doi:10.1037/0003-066X.49.12.997

Conn, V. S., y Rantz, M. J. (2003). Research methods: Managing primary study quality in meta-analyses. *Research in nursing & Health, 26*, 322-333. doi: 10.1002/nur.10092

Consejo General de Colegios Oficiales de Psicólogos (2010). Código Deontológico del Consejo General de Colegios Oficiales de Psicólogos. Recuperado desde:

https://www.cop.es/pdf/Codigo-Deontologico-Consejo-Adaptacion-Ley-Omnibus.pdf

Cooper, H. (1989). *Integrating Research: A Guide for Literature Reviews*. Beverly Hills, CA: Sage.

Cooper, H. (1998). Synthesizing research (3rd ed.) Thousand Oaks, CA: Sage.

Cooper, H., Hedges, L. V., y Valentine, J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., y Schuler, T.C. (2007). Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal, 7,* 541-546. doi: 10.1016/j.spinee.2007.01.008

Cortina, J. M., y Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*, 161–172. doi: 10.1037/1082-989X.2.2.161

Coulson, M., Healey, M., Fidler, F., y Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology, 1, 26.* doi: 10.3389/fpsyg.2010.00026

Crowe, M., y Shepard, L. (2011). A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64,* 79-89. doi: 10.1016/j.jclinepi.2010.02.008.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3,* 286-300. doi: 10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*, 7-29. doi: 10.1177/0956797613504966

Cumming, G., Fidler, F., Kalinowski, P, y Lai, J. (2012). The statistical recommendations of the American Psychological Association publication manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64,* 138-146. doi: 10.1111/j.1742-9536.2011.00037.x

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., …, y Wilson, S. (2007). Statistical reform in Psychology: Is anything changing? *Psychological Science, 18*, 230-232. doi: 10.1111/j.1467-9280.2007.01881.x

Cumming, G., y Finch, S. (2005). Inference by eye: Confidence intervals, overlap, and how to read pictures of data. *American Psychologist, 60*, 170-180. doi: 10.1037/0003-066X.60.2.170

Cumming, G., Williams, J., y Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 199-311. doi: 10.1207/s15328031us0304_5

Daset, L. R. y Cracco, C. (2013). Psicología Basada en la Evidencia: algunas cuestiones básicas y una aproximación a través de una revisión bibliográfica. *Ciencias Psicológicas; 7*, 209 – 220. Recuperado en 17 de agosto de 2016, de http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S1688-42212013000200009&lng=es&tlng=es.

Doi, S. A., y Thalib, L. (2008). A quality-effects model for meta-analysis. *Epidemiology, 19*, 94-100. doi:10.1097/EDE.0b013e31815c24e7

Dunlap, W. P. (1999). A program to compute McGraw and Wong's common language effect size indicator. *Behavioral Research Methods, Instruments and Computer, 31*, 706-709. doi: 10.3758/BF03200750

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., y Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170-177. doi: /10.1037/1082-989X.1.2.170

Durlak, J.A., (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34,* 917–928. doi: 10.1093/jpepsy/jsp004

Duval, S., y Tweedie, R. (2000). Trim and Fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455-63. doi: 10.1111/j.0006-341X.2000.00455.x

Earp, B. D., y Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6,* 621. doi: 10.3389/fpsyg.2015.00621

Egger M., Dickersin K., y Smith G. D. (2001). Problems and limitations in conducting systematic reviews. En M. Egger, G. D. Smith, & D. G. Altman, *Systematic reviews in health care* (pp. 43-68). London: BMJ Publishing Group, 2nd ed.

Ellis, P. D. (2010). *The essential guide to effect size. Statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge.

Erceg-Hurn, D. M. y Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63,* 591-601. doi: 10.1037/0003-066X.63.7.591

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist*, 53,798-799. doi: 10.1037/0003-066X.53.7.798

Falk, R., y Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75-98. doi: 10.1177/0959354395051004

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States data. *PLoS ONE 5(4*), e10271. doi:10.1371/journal.pone.0010271

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90,* 891–904. doi:10.1007/ s11192-011-0494-7

Faulkner, C., Fidler, F., y Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behavior Research and Therapy, 46,* 270-281. doi: 10.1016/j.brat.2007.12.001

Ferguson, C. J. (2009). An effect size primer: a guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40,* 532-538. doi: 10.1037/a0015808

Ferguson, C. J., y Brannick, M. T. (2011). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17,* 120–128. doi:10.1037/a0024445

Ferrill, M. J., Brown, D.A., y Kyle, J. A. (2010). Clinical versus statistical significance: interpreting p values and confidence intervals related to measures of association to

guide decision making. *Journal of Pharmacy Practice, 23,* 344-351. doi: 10.1177/0897190009358774

Fetheny, J. (2010). Statistical and clinical significance, and how to use confidence intervals to help interpret both. *Australian Critical Care, 23,* 93-97. doi:10.1016/j.aucc.2010.03.001

Fidler, F. (2002). The fifth edition of the APA publication manual: why its statistics recommendations are so controversial. *Educational & Psychological Measurement, 62,* 749-770. doi: 10.1177/001316402236876

Fidler, F. (2005). *From statistical signifi cance to effect estimation: Statistical reform in psychology, medicine and ecology* (Doctoral Thesis, University of Melbourne, Australia). Recuperado desde: https://www.researchgate.net/publication/267403673_From_statistical_significance_to_effect_estimation_Statistical_reform_in_psychology_medicine_and_ecology

Fidler, F. (2010). The american psychological association publication manual sixth edition: Implications for statistics education. *ICOTS8.* Disponible en: http://iase-web.org/documents/papers/icots8/ICOTS8_C156_FIDLER.pdf

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., . . . Schmitt, R. (2005). Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology, 73*, 136 –143. doi:10.1037/0022-006X.73.1.136

Fidler, F., y Loftus, G.R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology, 217,* 27-37. doi: 10.1027/0044-3409.217.1.27

Fidler, F., y Thompson, B. (2001). Computing correct confidence intervals for anova fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61,* 575-604. doi: 10.1177/0013164401614003

Field, A. P., y Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63,* 665-694. doi: 10.1348/000711010X502733.

Finch, S., Cumming, G., y Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181-210. doi: 10.1177/00131640121971167

Finch, S., Thomason, N., y Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology, 12*, 825-853. doi: 10.1177/0959354302126005

Francis, G. (2012a). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19,* 151-156. doi: 10.3758/s13423-012-0227-9

Francis, G. (2012b). The Psychology of replication and replication in Psychology. *Perspectives on Psychological Science, 7*, 585 –594. doi: 10.1177/1745691612459520

Frías-Navarro, D. (2011a). *Técnica estadística y diseño de investigación*. Valencia: Palmero Ediciones.

Frías-Navarro, D. (2011b). Reforma estadistica. Tamaño del efecto. En D. Frias-Navarro, D. *Técnica estadística y diseño de investigación* (pp. 123-168). Valencia: Palmero Ediciones.

Frías-Navarro, D., y Gómez-Frias, R. (2014). Metodología de investigación y contraste de hipótesis. En D. Frias-Navarro, M. Pascual-Soler, L. Badenes-Ribera y H. Monterde-i-Bort (Eds). *Reforma Estadística en Psicología* (pp. 50-70). Valencia: Ediciones Palmero.

Frías-Navarro, D., y Monterde-i-Bort, H. (2014). Revisión sistemática. Introducción al meta-análisis. En D. Frias-Navarro, M. Pascual-Soler, L. Badenes-Ribera y H. Monterde-i-Bort (Eds.). *Reforma Estadística en Psicología* (pp. 152-168). Valencia: Ediciones Palmero.

Frías-Navarro, D., Monterde-i-Bort, H., Pascual-Soler, M., Pascual-Llobell, J., y Badenes-Ribera, L. (2012, Julio). *Improving statistical practice in clinical production: A case Psicothema.* Poster presentado al V European Congress of Methodology, Santiago de Compostela, España.

Frías-Navarro, D. y Pascual-Llobell, J. (2003). Psicología clínica basada en pruebas: Efecto del tratamiento. *Papeles del Psicólogo, 85*, 11-118. Disponible en: http://www.papelesdelpsicologo.es/vernumero.asp?id=1074

Frías, M. D., Pascual, J., y García, J.F. (2000). Tamaño del efecto del tratamiento y signficación estadistica. *Psicothema, 12,* 236-240. Disponible en: http://www.psicothema.com/pdf/555.pdf

Frías, M. D., Pascual, J., y García, J.F. (2002). La hipótesis nula y la significación práctica. *Metodología de las Ciencias del Comportamiento*, 181-185. Recuperado 18/05/2016 desde http://www.valencia.edu/~garpe/C_/A_/C_A_0020.pdf

Frías-Navarro, D., Pascual-Soler, M., Badenes-Ribera, L., y Monterde-i-Bort, H. (2014). *Reforma Estadística en Psicología*. Valencia: Ediciones Palmero.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379–390. doi: 10.1037/1082-989X.1.4.379

Fritz, C. O., Morris, P. E., y Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*, 2–18. doi:10.1037/a0024338.

Furukawa, T. A. (1999). From effect size into number needed to treat. *Lancet, 353*, 1680. doi: 10.1016/S0140-6736(99)01163-0

Furukawa, T. A., y Leucht, S. (2011). How to obtain NNT from Cohen's d: Comparison of two methods. *PLoS ONE 6*(4), e19070. doi: 10.1371/ journal.pone.0019070

Gadbury, G. L, y Allison, D. B. (2014). Inappropriate fiddling with statistical analyses to obtain a desirable p-value: Tests to detect its presence in published literature. *PLoS ONE, 7*: e46363. 16. doi: 10.1371/journal.pone.0046363

Gardner, M. J., y Altman, D. G. (2000). Confidence intervals rather than p values. En D.G. Altman, D. Machin, T. Bryant y M. J. Gardner (eds.) (2nd edition) , *Statistics whith confidence* (pp. 15-27). London: British Medical Journal.

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., y Zieffler, A. (2008). *Developing students' statistical reasoning. Connecting research and teaching practice*. New York, NY: Springer Publishers

Garfield, J., y Franklin, C. (2011). Assessment of learning, for learning, and as learning in statistics education. En C. Batanero, G. Burrilll, C. Reading and A. Rossman (eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 133-145). New York, NY: Springer Publishers.

Garfield, J., Zieffler, A., Kaplan, D., Cobb, G., Chance, B., y Holcomb, J. P. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician, 65,* 1-10. doi: 10.1198/tast.2011.08241

Gelman, A., y Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant, *The American Statistician, 60*, 328-331, doi: 10.1198/000313006X152649

Gibbson, R. D., Hedeker, D. R., y Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational and Behavioral Statistics, 18,* 271-279. doi: 10.3102/10769986018003271

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. En L. Kruger, G. Gigerenzer, y M. Morgan (Eds.), *The probabilistic revolution*: Vol. 2. Ideas in the sciences (pp. 11-33). Cambridge, MA: MIT Press.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics 33,* 587-606. doi: :10.1016/j.socec.2004.09.033

Gigerenzer, G., Kraus, S., y Vitouch, O. (2004). The null ritual what you always wanted to know about significance testing but were afraid to ask. En: Kaplan, D. (Ed.), *Handbook on Quantitative Methods in the Social Sciences* (pp.389-406*)*. Sage, Thousand Oaks, CA.

Gigerenzer, G., y Marewski, J. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41*, 421-440. doi: 10.1177/0149206314547522

Gigerenzer, G., y Murray, D. J. (1987). *Cognition as Intuitive Statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly, 52*, 647-674. doi: 10.1177/106591299905200309

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8. doi: 10.3102/0013189X005010003

Glass, G. V., Mcgraw, B., y Smith, M. L. (1981). *Meta-analysis in Social Research*. Beverly Hills, CA: Sage.

Gliner, J. A., Vaske, J. J., y Morgan, G. A. (2001). Null hypothesis significance testing: Effect size matters. *Human Dimensions of Wildlife, 6*, 291-301. doi: 10.1080/108712001753473966

Gliner, J. A., Leech, N. L., y Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71*, 83-92. doi:10.1080/00220970209602058

Goodman, S. (1999). Toward evidence-based medical statistics 1: The p value fallacy. *Annals of Internal Medicine, 130*, 995-1004. doi:10.7326/0003-4819-130-12-199906150-00008

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology, 45*, 135-140. doi: 10.1053/j.seminhematol.2008.04.003

Gordon, H. R. D. (2001). American vocational education research association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Educational Research, 26*, 1-18. doi: 10.5328/JVER26.2.244

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics on investigating theoretical models. *Psychological Reviews, 69*, 54-61.

Greenfield, M. L., Kuhn, J. E., y Wojtys, E. M. (1996). A statistics primer. P values: probability and clinical significance. *The American Journal of Sports Medicine, 24*, 863–865.

Greenland, S., y Poole, C. (2011). Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. *Jurimetrics*, 51, 113–129. Disponible en: http://www.ph.ucla.edu/epi/faculty/greenland/Epi204/GreenlandPoole2011.InterpretingStats.pdf

Greenstein, G. (2003). Clinical versus statistical significance as they relate to the efficacy of periodontal therapy. *Journal of the American Dental Association*, 134, 583-91. doi: 10.14219/jada.archive.2003.0225

Grègoire, G., Derderian, F., y LeLorier, J. (1995). Selecting the language of the publications included in a meta-analysis: Is there a Tower of Babel bias?. *Journal of Clinical Epidemiology, 48,* 159–163. doi:10.1016/0895-4356(94)00098-B

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79*, 314–316. doi: 10.1037/0021-9010.79.2.314

Grissom, R. J., y Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135-146. doi: 10.1037/1082-989X.6.2.135

Grissom, R. J., y Kim, J. J. (2012). *Effect sizes for research*. New York, USA: Routledge

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: a methodological perspective. *Theory & Psychology, 23*, 251–270. doi: 10.1177/0959354312465483

Hales, A. H. (2016). Does the conclusion follow from the evidence? Recommendations for improving research. *Journal of Experimental Social Psychology, 66*, 39-46. doi: 10.1016/j.jesp.2015.09.011

Haller, H., y Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research Online [On-line serial], 7,* 120. Recuperado 30 Julio 2014, desde http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf

Halpin, P. F., y Stam, H. J. (2006). Inductive inference or inductive behavior: fisher and Neyman–Pearson approaches to statistical testing in psychological research (1940-1960). *American Journal of Psychology, 119,* 625–653.doi:10.2307/20445367

Harlow, L. L., Mulaik, S. A., y Steiger, J. H. (1997, Eds.). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, J. D., Quatman, C. E., Manring, M. M., Siston, R. A., y Flanigan, D. C. (2014). How to write a systematic review. *American Journal of Sports Medicine, 42,* 2761-2768. doi: 10.1177/0363546513497567

Harrison, J., Thompson, B., y Vannest, K. J. (2009). Interpreting the evidence for effective interventions to increase the academic performance of students with ADHD: Relevance of the statistical significance controversy. *Review of Educational Research, 79*, 740-775. doi: 10.3102/0034654309331516

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., y Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology 13*, e1002106. doi: 10.1371/journal.pbio.1002106

Hedges, L. V. (1981). Distribution theory for Glass's estimator effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*, 107-128. doi: 10.3102/10769986006002107

Hedges, L. V., y Olkin, I. (1985). *Statistical Methods for Meta-analysis*. New York: Academic Press.

Hedges, L. V., y Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods, 3*, 486-504. doi: 10.1037/1082-989X.3.4.486

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *Counseling Psychologist, 34*, 601-629. doi: 10.1177/0011000005283558

Higgins, J. P. T., y Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558. doi: 10.1002/sim.1186

Hoekstra, R., Finch, S., Kiers, H. A. L., y Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review, 13*, 1033-1037. doi: 10.3758/BF03213921

Hoekstra, R., Kiers, H.A.L. y Johnson, A., (2012). Are assumptions of well-known statistical techniques checked, and why (not)?. *Frontiers in Psychology, 5*, doi: 10.3389/fpsyg.2012.00137

Hoekstra, R., Johnson, A., y Kiers, H.A.L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement, 72*, 1039-1052. doi: 10.1177/0013164412450297

Hoekstra, R., Morey, R.D., Rouder, J.N., y Wagenmakers, E. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin& Review, 21*, 1157-1164. doi: 10.3758/s13423-013-0572-3

Hopewell, S., McDonald, S., Clarke, M., y Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database of Systematic Reviews, 2*. Art No: MR000010.

Hubbard, R. (2004). Alphabet soup. Blurring the distinctions between p's and α's in psychological research. *Theory & Psychology, 14*, 295-327. doi: 10.1177/0959354304043638

Hubbard, R., y Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*, 69-88. doi: 10.1177/0959354307086923

Hubbard, R., y Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology and its future prospects. *Educational and Psychological Measurement, 60*, 661–681. doi: 10.1177/0013164400605001

Huberty, C. J. (1993). Historical origins of statistical testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks. Journal of *Experimental Education, 61*, 317–333. doi: 10.1080/00220973.1993.10806593

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*, 227–240. doi: 10.1177/0013164402062002002

Huedo-Medina, T. B., y Johnson, B. T. (2010). *Modelos estadísticos en meta-análisis*. Oleiros, La Coruña: Netbiblio.

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., y Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods, 11,* 193-206. doi: 10.1037/1082-989X.11.2.193

Hunter, J. E., y Schmidt, F. L. (2000). Fixed effects vs random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection & Assessment, 8,* 275-292. doi: 10.1111/1468-2389.00156

Hutton, B., Catalá-López, F., y Moher, D. (2016). La extensión de la declaración PRISMA para revisiones sistemáticas que incorporan metaanálisis en red: PRISMA-NMA. *Medicina Clínica*. doi: 10.1016/j.medcli.2016.02.025

Hutton, B., Salanti, G., Caldwell, D. M., Chaimani, A., Schmid, C. H., Cameron, C., J.P, … y, Moher D. (2015). The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Annals of Internal Medicine: Journal. 162*, 777-784. doi: 10.7326/M14-2385.

Ioannidis, J. P. A. (2005a). Why most published research findings are false. *PLoS Medicine*, 2(8): e124. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2005b). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294*, 218–28. doi: 10.1001/jama.294.2.218.

Ioannidis, J. P. A. (2011). Meta-research: The art of getting it wrong. *Research Synthesis Methods, 1,* 169-184. doi: 10.1002/jrsm.19

Ioannidis, J. P. A., y Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4*, 245–253. doi: 10.1177/1740774507079441

Jacobsen, N. S., Follette,W.C., y Revenstorf, D. (1984). Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336-352. doi: 10.1016/S0005-7894(84)80002-7

Jasny, B. R., Chin, G., Chong, L., y Vignieri, S. (2011). Again, and Again, and Again … *Science, 334*, 1225. doi: 10.1126/science.334.6060.1225

John, L. K., Loewenstein, G., y Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives. *Psychological Science, 23,* 524-532. doi: 10.1177/0956797611430953

Johnson, B. T. (1993). DSTAT 1.10. *Software for the meta-analytic review of research literatures*. Mahwah, NJ:Erlbaum

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management, 63*, 763-772.

Jüni, P., Altman, D. G., y Egger, M. (2001). *Assessing the quality of randomised controlled trials*. En M. Egger, G.D. Smith y D.G. Altman (Eds.), Systematic reviews in health care: Meta-analysis in context (2ª ed.) (pp. 87-108). Londres: BMJ Books.

Kalinowski, P., y Fidler, F. (2010). Interpreting significance: the differences between statistical significance, effect size, and practical importance. *Newborn & Infant Nursing Reviews 10,* 50–54. doi: 10.1053/j.nainr.2009.12.007

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332-339. doi: 10.1037/0022-006X.67.3.332

Kazdin, A. E. (2001). Almost Clinically Significant ($p < .10$): Current measures may only approach clinical significance. *Clinical Psychology Science and Practice, 8*, 455-462. doi: 10.1093/clipsy.8.4.455

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*, 146-159. doi: 10.1037/0003-066X.63.3.146

Kelley, K., y Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137-152. doi: 10.1037/a0028086

Kehle, T. J., Bray, M. A., Chafouleas, S. M., y Kawano, T. (2007). Lack of statistical significance. *Psychology in the Schools, 44*, 417–422. doi: 10.1002/pits.20233

Kendall, P. C. (Ed.). (1999). Special section: Clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 283–339. doi: 10.1037/0022-006X.67.3.283

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., y Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 285–299. doi: 10.1037/0022-006X.67.3.285

Kepes, S., Banks, G. C., McDaniel, M., y Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 29,* 183-203. doi: 10.1177/1094428112452760.

Kepes, S., Banks, G. C., y Oh, I.-S. (2014). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology, 29*, 183-203. doi: 10.1007/s10869-012-9279-0

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., y Deerin, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*, 110–129. doi: 10.1037/1082-989X.13.2.110.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759. doi: 10.1177/0013164496056005002

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*, 213-218. doi: 10.1177/00131640121971185

Kisamore, J. L., y Brannick, M. T. (2008). An illustration of the consequences of meta-analysis model choice. *Organizational Research Methods, 11*, 35–53. doi: 10.1177/1094428106287393

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association: Washington, DC.

Kline, R. B. (2013). *Beyond significance testing: Statistic reform in the behavioral sciences*. Washington, DC: American Psychological Association: Washington, DC.

Koole, S. L., y Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7,* 608 – 614. doi: 10.1177/1745691612462586

Kraemer, H. C., y Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry, 59,* 990–996. doi: 10.1016/j.biopsych.2005.09.014

Krishnan, S., y Idris, N. (2014). Students' misconceptions about hypothesis test. *REDIMAT, 3*, 276-293. doi: 10.4471/redimat.2014.54

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *The American Psychologist, 56*, 16-26. doi: 10.1037/0003-066X.56.1.16

Kühberger, A., Fritz, A., Lermer, E. y Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC Research Notes*, *17*, 8, 84. doi: 10.1186/s13104-015-1020-4.

La Greca, A. M. (2005). Editorial. *Journal of Consulting and Clinical Psychology, 73*, 3–5. doi: 10.1037/0022-006X.73.1.3

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4,* 1-10. doi: 10.3389/fpsyg.2013.00863

Lambdin, C. (2012). Significance test as sorcery: Science is empirical – significance tests are not. *Theory & Psychology*, *22*, 67-90. doi: 10.1177/0959354311429854

Lecoutre, M. P., Poitevineau, J., y Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis tests. *International Journal of Psychology, 38,* 37-45. doi: 10.1080/00207590244000250

Leek, J. (2014). On the scalability of statistical procedures: Why the p-value bashers just don't get it. *Simply Statistics Blog*, Dispnible en: http://simplystatistics.org/2014/02/14/on-the-scalability-of-statisticalprocedures-why-the-p-value-bashers-just-dont-get-it/

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., y Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal Medical, 337*, 536-542. doi: 10.1056/NEJM199708213370806

Li, J. C H. (2015). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods*. doi: 10.3758/s13428-015-0667-z

Lipsey, M. W. y Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Littell, J. H., Corcoran, J., y Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford UK: Oxford University Press

Long, J. D., y Cliff, N. (1997). Confidence intervals for Kendall's tau. *Britrish Journal of Mathematical and Statsitical Psychology, 50,* 31-41. doi: 10.1111/j.2044-8317.1997.tb01100.x

Lovell, D. P. (2013). Biological importance and statistical significance. *Journal of Agricultural and Food Chemistry, 61,* 8340–8348. doi: 10.1021/jf401124y

Ludbrook, J. (2013). Should we use one-sided or two-sided p values in tests of significance? *Clinical and Experimental Pharmacology and Physiology, 40,* 357–361. doi: 10.1111/1440- 1681.12086

Macdonald, R. R. (1997). On statistical testing in psychology. *British Journal of Psychology, 88,* 333–347. doi: 10.1111/j.2044-8295.1997.tb02638.x

Maher, J. M., Markey, J. C., y Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. *CBE Life Sciences Education*, 12, 345–351. doi: 10.1187/cbe.13-04-0082

Makel, M. C., Plucker, J. A., y Hegarty, B. (2012). Replications in psychology research: How often do they really occur?. *Perspectives on Psychological Science, 7,* 537-542. doi: 10.1177/1745691612460688

Manriquez, J. J., Villouta, M. F., y Williams, H. C. (2007). Evidence-based dermatology: Number needed to treat and its relation to other risk measures. *Journal of the American Academy of Dermatology*, *56,* 664-671. doi: 10.1016/j.jaad.2006.08.024

Marín-Martínez, F., Sánchez-Meca, J., Huedo-Medina, T. B. y Fernández-Guzmán, I. (2007). Meta-análisis: Dónde estamos y hacia dónde vamos. En A. Borges y P. Prieto (Eds.), *Psicología y ciencias afines en los albores del siglo XXI (Homenaje al profesor Alfonso Sánchez Bruno*) (pp. 87-102). Grupo Editorial Universitario. Disponible en: http://www.um.es/metaanalysis/publications.php

Marín-Martínez, F., Sánchez-Meca, J., y López-López, J. A. (2009). El meta-análisis en el ámbito de las Ciencias de la Salud: Una metodología imprescindible para la eficiente acumulación del conocimiento. *Fisioterapia*, 31, 107-114. doi:10.1016/j.ft.2009.02.002

Matthews, M. S., Gentry, M., McCoach, D. B., Worrell, F. C., Matthews, D., y Dixon, F. (2008). Evaluating the state of a field: effect size reporting in gifted education. *The Journal of Experimental Education, 77*, 55–65. doi: 10.3200/JEXE.77.1.55-68

McGough, J. J., y Faraone, S. V. (2009). Estimating the size of treatment effects: Moving beyond p values. *Psychiatry (Edgmont), 6*, 21–29.

McGraw, K. O., y Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361–365. doi: 10.1037/0033-2909.111.2.361

McMillan, J. H., y Foley, J. (2011). Reporting and discussing effect size: Still the road less traveled. *Practical Assessment, Research & Evaluation, 16(*14). Recueperado desde: http:// pareonline.net/getvn.asp?v=16&n=14

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*, 806–834. doi: 10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195-244. doi: 10.2466/pr0.1990.66.1.195

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology*, *64*, 553-557.

Miller, J., y Ulrich, R. (2016). Interpreting confidence intervals: A comment on Hoekstra, Morey, Rouder and Wagenmakers (2014). *Psychonomic Bulletin & Review*, 23, 124-130. doi: 10.3758/s13423-015-0859-7

Mittag, K. C., y Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance test and others statistical issues. *Educational Researcher, 29*, 14-20.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., y Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomized controlled trials: the QUORUM statement. Quality of Reporting of Meta-analyses. Lancet, 354, 1896-1900. doi: 10.1016/S0140-6736(99)04149-5

Moher, D., Hopewell S., Schulz, K. F., Montori, V., Gøtzsche P.C., Devereaux, P. J., … Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: updated

guidelines for reporting parallel group randomised trials. *British Medical Journal, 340:*c869. doi: 10.1136/bmj.c869

Moher, D., Jones, A. y Lepage, L. for the CONSORT Group (2001). Use of the CONSORT statement and quality of reports for randomized trials: A comparative before-and-after evaluation. *Journal of the American Medical Association, 285*, 1992-1995. doi: 10.1001/jama.285.15.1992

Moher, D., Liberati, A., Tetzalaff, J., Altman, D. G., y the PRISMA group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *Plos Medicine, 6 (7): e1000097*. doi: 10.1371/journal.pmed.1000097

Monterde-i-Bort, H., Frias-Navarro, D., y Pascual-Llobell, J. (2010). Uses and abuses of statistical significance tests and other statistical resources: A comparative study. *European Journal of Psychology of Education, 25*, 429-447. doi: 10.1007/s10212-010-0021-x

Monterde-i-Bort, H., Pascual-Llobell, J., y Frias-Navarro, D. (2006). Errores de interpretación de los métodos estadísticos: importancia y recomendaciones. *Psicothema, 18,* 848-856. Recuperado el 18/08/2016 desde http://www.psicothema.com/psicothema.asp?id=3319

Morgan, P. (2003). Null Hypothesis Significance Testing: Philosophical and Practical Considerations of a Statistical Controversy. E*xceptionality: A special Education Journal, 11*, 209–221. doi: 10.1207/S15327035EX1104_2

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analyses or repeated measures. *British Journal of Mathematical and Statistical Psychology, 53,* 17-29. doi: 10.1348/000711000159150

Morris, S. B. (2008). Estimating effect from pretest-posttest-control group designs. *Organizational Research Methods, 11,* 364-386. doi: 10.1177/1094428106291059

Morris, S. B., y DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-group designs. *Psychological Methods, 7,* 105-125. doi: 10.1037/1082-989X.7.1.105

Morrison, D. E., y Henkel, R. E. (1970). *The significance test controversy: a reader*. Chicago, Illinois: Aldine Publishing.

Motulsky, H. J. (2015). Common misconceptions about data analysis and statistics. *British Journal of Pharmacology, 172,* 2126–2132. doi: 10.1111/bph.12884

Mulrow, C., y Cook, D. (Eds.) (1998). *Systematic reviews: Synthesis of best evidence for health care decisions*. Philadelphia, PA: American College of Physicians.

Musselman, K. E. (2007). Clinical significance testing in rehabilitation research: what, why, and how? *Physical Therapy Reviews*, 12, 287-296. doi: 10.1179/108331907X223128

Nakagawa, S., y Cutchill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews, 82,* 591-605. doi: 0.1111/j.1469-185X.2007.00027.x

National Research Council. (1992). Combining information: Statistical issues and opportunities for research. Washington, DC: National Academy Press.

Nelson, N., Rosenthal. R., y Rosnow, R .L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41*, 1299–1301. doi: 10.1037/0003-066X.41.11.1299

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics 45,* 401-410. doi: 10.2307/2986064

Newcombe, R.G. (2012). *Confidence intervals for proportions and related mesuares of effect size*. Boca Raton, FL: CRC Press

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301. doi: 10.1037/1082-989X.S.2.241

Nuzzo R. (2014). Statistical errors: P values, the “gold standard” of statistical validity, are not as reliable as many scientists assume. *Nature*, *130*, 150-152. Recuperado 18/05/2016 desde http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chicester: John Wiley & Sons.

Odgaard, E. C., y Fowler, R. L. (2010). Confidence intervals for effect sizes: compliance and clinical significance in the Journal of Consulting and Clinical

Psychology. *Journal of Consulting and Clinical Psychology, 78,* 287–297. doi: 10.1037/a0019294.

Ogles, B. M., Kirk, M. Lunnen, K. M., y Bonesteel, K. (2001). Clinical significance: history, application, and current practice. *Clinical Psychology Review, 21*, 421–446, doi: 10.1016/S0272-7358(99)00058-6.

Olejnik, S., y Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*, 434–447. doi: 10.1037/1082-989X.8.4.434

Orwin, R.G., y Vevea, J. L. (2010). Evaluating coding decisions. En H. Cooper, L.V. Hedges y J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* 2ª ed. (pp. 177-203). Nueva York: Russell Sage Foundation.

Page, P. (2014). Beyond statistical significance: clinical interpretation of rehabilitation research literature. *International Journal of Sports Physical Therapy*, 9, 726–736. Recuperado desde: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4197528/

Palmer, A., y Sesé, A. (2013). Recommendations for the use of statistics in clinical and health psychology. *Clínica y Salud, 24*, 47-54. doi: 10.5093/cl2013a6

Palmer, M. A., y Strelan, P. (2015). Commentary on Dutta and Pullig: Corrective action is more effective than downplaying harm for restoring brand equity. *Journal of Business Research*, 68, 1271-1272. doi: 10.1016/j.jbusres.2014.11.007

Pascual, J., Frías, D. y García, J. F. (2000). El procedimiento de significación estadística (NHST): Su trayectoria y actualidad. *Revista de Historia de la Psicología, 21*, 9-26.

Pascual-Llobell, J., Frias-Navarro, D. y Monterde-i-Bort, H. (2004). Tratamientos psicológicos con apoyo empírico y práctica clínica basada en la evidencia. *Papeles del psicólogo, 87*, 1-8. Recuperado 18/05/2016 desde http://www.papelesdelpsicologo.es/vernumero.asp?id=1134

Pashler, H., y Wagenmakers, E. J. (2012). Editors' Introduction to the special section on replicability in Psychological science: A crisis of confidence?. *Perspectives on Psychological Science, 7,* 528-530. doi: 10.1177/1745691612465253

Peng, C.-Y. J., y Chen, L.-T. (2014). Beyond Cohen's d: alternative effect size measures for between subject designs. *The Journal of Experimental Education, 82*, 22-50. doi: 10.1080/00220973.2012.745471

Peng, C.-Y J., Chen, L.-T, Chiang, H., y Chiang, Y. (2013). The Impact of APA and AERA Guidelines on effect size reporting. *Educational Psychology Review*, *25*, 157-209. doi: 10.1007/s10648-013-9218-2

Perestelo-Pérez, L. (2013). Standards on how to develop and report systematic reviews in Psychology and Health. I*nternational Journal of Clinical and Health Psychology, 13,* 49-57. doi: 10.1016/S1697-2600(13)70007-3

Perezgonzalez, J. D. (2014). A reconceptualization of significance testing. *Theory & Psychology, 24*, 852–859. doi: 10.1177/0959354314546157

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*, 223. doi: 10.3389/fpsyg.2015.000223

Pfannkuch, M., y Wild, C. (2004). Towards an understanding of statistical thinking. En D. Ben-Zvi, y J. Garfield, (Eds.). *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17-45). Dordrecht: Kluwer Academic Publishers.

Poitevineau, J., y Lecoutre, B. (2001). Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated. *Psychonomic Bulletin & Review, 8,* 847-850. doi: 10.3758/BF03196227

Preacher, K. J., y Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16,* 93–11. doi: 10.1037/a0022658

Review Manager (RevMan) (2008). [Computer Program]. Version 5.0. Copenhagen: he Nordic Cochrane, The Cochrane Collaboration.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. APS Observer, 25. Recuperado desde http://www .psychologicalscience.org/index.php/publications/observer/2012/ february-11-2012-observer-publications/psychology';s-woesand-a-partial-cure-the-value-of-replication.html

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641. doi: 10.1037/00332909.86.3.638.

Rosenthal, R. (1984). *Meta-Analytic procedures for social research*. BeverlyHills, CA: Sage.

Rosenthal, R. (1995).Writing meta-analytic reviews. *Psychological Bulletin, 18,* 183-192. doi: 10.1037/0033-2909.118.2.183

Rosenthal, R., y Gaito, J. (1963). The interpretation of level of significance by psychological researchers. *Journal of Psychology, 55:* 33-38. doi: 10.1080/00223980.1963.9916596

Rosnow, R. L., y Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284. doi: 10.1037/0003-066X.44.10.1276

Rosnow, R., y Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57,* 221–237. doi: doi.org/10.1037/h0087427

Rosnow, R. L., y Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Journal of Psychology, 217,* 6–14. doi: 10.1027/0044-3409.217.1.6

Rosnow, R., Rosenthal, R. y Rubin, D. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science 11,* 446-453. doi: 10.1111/1467-9280.00287

Rothstein, H. R., Sutton, A. J., y Borenstein, M. (2005a). Publication bias in meta-analysis: Prevention, assessment and adjustments. Chichester, UK: Wiley

Rothstein, H. R., Sutton, A. J., y Borenstein, M. (2005b). Publication bias in meta-analyses. En H. R. Rothstein, A. J. Sutton, y M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 1–7). West Sussex: Wiley

Ruscio, J. (2008a). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13,* 19–30. doi: 10.1037/1082-989X.13.1.19

Ruscio, J. (2008b). Constructing confidence intervals for Spearman's rank correlation with ordinal data: A simulation study comparing analytic and bootstrap methods. *Journal of Modern Applied Statistical Methods, 7,* 416-434. Disponible en: http://digitalcommons.wayne.edu/jmasm/vol7/iss2/7

Ruscio, J., y Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research, 47,* 201–223. doi: 10.1080/00273171.2012.658329

Sackett, D.L., Straus, S. L., Richardson, W.S., Rosenberg, W.M.C. y Haynes, R. B. (2000). *Evidence Based Medicine. How to practice y teach EBM (3rd Edition)*. Edinburgh, England & New York (NY): Churchill Livingstone.

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology, 66,* 47-54. doi: 10.1016/j.jesp.2015.09.013

Sánchez-Meca, J. (1986). La revisión cuantitativa: una alternativa a las revisiones tradicionales Anales *de Psicología, 3,* 79-107. Disponible en: http://www.um.es/metaanalysis/pdf/7004.pdf

Sánchez-Meca, J. (2008). Meta-análisis de la investigación. En M. A. Verdugo, M. Crespo, M. Badía, & B. Arias (Coords.), *Metodología en la investigación sobre discapacidad: Introducción al uso de las ecuaciones estructurales*. Salamanca: Publicaciones del INICO (Colección ACTAS, 5/2008). Disponible en: http://www.um.es/metaanalysis/pdf/5023.pdf

Sánchez-Meca, J. (2010). Cómo realizar una revisión sistemática y un meta-análisis . *Aula Abierta, 38, 53-64* Disponible en: http://www.um.es/metaanalysis/pdf/5030.pdf

Sánchez-Meca, J., y Ato-García, M. (1989). Meta-análisis: una alternativa metodológica a las revisiones tradicionales de la investigación . En J. Arnau & H. Carpintero (Eds.), Tratado de Psicología General (pp. 617-669). Madrid. Alhambra. Disponible en: http://www.um.es/metaanalysis/pdf/6201.pdf

Sánchez-Meca, J., Boruch, R. F., Petrosino, A., y Rosa-Alcázar, A. I. (2002). La Colaboración Campbell y la Práctica basada en la Evidencia. *Papeles del Psicólogo, 83,* 44-48. Disponible en: http://www.papelesdelpsicologo.es/vernumero.asp?id=896

Sánchez-Meca, J., y Botella, J. (2010). Revisiones Sistemáticas y Meta-análisis: herramientas para la práctica profesional. *Papeles del Psicólogo, 31,* 7-17. Disponible en: http://www.papelesdelpsicologo.es/pdf/1792.pdf

Sánchez-Meca, J., López-López, J. A., y López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology, 66,* 402–425. doi: 10.1111/j.2044-8317.2012.02057.x

Sánchez-Meca, J., y Marín-Martínez, F. (2010). Meta-analysis in psychological research. *International Journal of Psychological Research, 3,* 150-162. doi: 10.21500/20112084.860

Sánchez-Meca, J., Marín-Martínez, F., y Chacón-Mosco, S. (2003).Effect-size índices for dicothomized outcomes in meta-analysis. *Psychologial Methods, 8,* 448-467. doi: 10.1037/1082-989X.8.4.448

Sánchez-Meca, J., Marín-Martínez, F., y Huedo-Medina, T. B. (2006). Modelo de efectos fijos y modelo de efectos aleatorios. En J. L. R. Martín, A. T. Garcés y T. Seoane. *Revisiones sistemáticas en las Ciencias de la Vida. El concepto de salud a través de la síntesis de la Evidencia Científica*. Castilla la Mancha: FISCAM, Fundación para la Investigación Sanitaria en Castilla-LaMancha. Disponible en: http://www.um.es/metaanalysis/pdf/5003.pdf

Sánchez-Meca, J., Marín-Martínez, F., y López-López, J. A. (2011). Meta-análisis e intervención psicosocial basada en la evidencia. *Psychosocial Intervention, 20,* 95-107. doi: 10.5093/in2011v20n1a8

Sánchez-Meca, J., Marín-Martínez, F., y López-López, J. A. (2013). Metodología del meta-análisis. En J. F. J. Sarabia (Coord.), *Métodos de investigación social y de la empresa* (pp. 447-470). Madrid: Pirámide.

Santana-Cárdenas, S. (2006). *Manual de estilo de publicaciones de la APA*. Acceso 21/03/206. Disponible en: http://www.cusur.udg.mx/induccion-alumnos/sites/default/files/manual_de_estilo_para_publicaciones_apa_comentado.pdf

Savalei, V., y Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis?. *Frontiers in Psychology*, 6, 245. doi: 10.3389/fpsyg.2015.00245

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129. doi: 10.1037/1082-989X.1.2.115

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100. doi:10.1037/a0015108

Schmidt, F. L., y Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. En L. L. Harlow, S. A. Mulaik, y J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Lawrence Erlbaum.

Schulz, R., O'Brien, A., Czaja, S., Ory, M., Norris, R., Martire, L.M., … B Stevens, A., (2002). Dementia caregiver intervention research: in search of clinical significance. *The Gerontologist, 42*, 589–602. Recuperado desde: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2579772/pdf/nihms72032.pdf

Sesé, A., y Palmer, A. (2012). El uso de la estadística en psicología clínica y de la salud a revisión. *Clínica y Salud, 23*, 97-108.

Shaver, J. P. (1993). What statistical significance testing is, and what is not. *The Journal of Experimental Education, 61,* 293-316. Recuperado 18/08/2016 desde http://www.jstor.org/stable/20152383?seq=1#page_scan_tab_contents

Shea, B., Bouter, L. M., Peterson, J., Boers, M., Andersson, N., Ortiz, Z. … Grimshaw, J. (2007). External validation of a mesuarement tool to assess systemativ reviews (AMSTAR). Plos One/ www.plosone.org., December 2007, Issue 12, e1350. doi: 10.1371/journal.pone.0001350

Shea, B. J., Grimshaw, J. M, Wells, G.A., Boers, M., Andersson, N., Hamel, C., … Bouter, L. M. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*, 10-16. doi: 10.1186/1471-2288-7-10

Shea, B. J., Hamel, C., Wells, G. A., Bouters, L. M., Kristjansson, E., Grimshaw, J. … Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology 62*, 1013-1020. doi: 10.1016/j.jclinepi.2008.10.009

Simonsohn, U., Nelson, L., y Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534–547. doi: 10.1037/a0033242

Simonsohn, U., Nelson, L., y Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias: Using only significant results. *Perspectives on Psychological, 9,* 666-681. doi: 10.1177/1745691614553988

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage

Smithson, M. (2011). Confidence interval. En M. Lovric (Ed.) *International Encyclopedia of Statistical Science* (pp. 283-284). Berlin:Springer

Song, F., Sheldon, T., Sutton, A., Abrams, K., y Jones, D. R. (2001). Methods for exploring heterogeneity in meta-analysis. *Evaluation and the Health Professions, 24*, 126-151. doi: 10.1177/016327870102400203

Song, F., Parekh, S., Hooper, L., Loke, Y. K , Ryder, J., Sutton, A. J., … Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment, 14*, 1–220. doi: 10.3310/hta14080.

Spielman, S. (1978). Statistical dogma and the logic of significance testing. *Philosophy of Science, 45*, 120–135. doi: 10.1086/288784

Stangor, C., y Lemay, E. P. (2016). Editorial. Introduction to the special issue on methodological rigor and replicability. *Journal of Experimental Social Psychology, 66*, 1-3. doi: 10.1016/j.jesp.2016.02.006

Sterne, J. A., Gavaghan, D., y Egger, M. (2005). The funnel plot. En H. R. Rothstein, A. J. Sutton, y M. Borenstein (Eds.), *Publication bias in metaanalysis*: Prevention, assessment and adjustments (pp. 75–98). Chichester, UK: Wiley

Stroebe, W. (2016). Are most published social psychological findings false?. *Journal of Experimental Social Psychology, 66,* 134-144. doi: 10.1016/j.jesp.2015.09.017

Stroebe, W., y Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59-71. doi: 10.1177/1745691613514450

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., ... Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology. A proposal for reporting. *JAMA: The Journal of the American Medical Association, 283*, 2008-2012. doi:10.1001/jama.283.15.2008

Sullivan, G. M., y Feinn, R. (2012). Using effect sSize—or why the p value is not enough. *Journal of Graduate Medical Education, 4*, 279–282. doi: 10.4300/JGME-D-12-00156.1

Sun, S., Pan, W., y Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 10*, 989-1004. doi: 10.1037/a0019507

Sutton, A. J., y Higgins, J. (2008). Recent developments in meta-analysis. *Statistics in medicine, 27*, 625-650. doi: 10.1002/sim.2934

Takkouche, B., Cadarso-Sures, C., y Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology, 150*, 206-215. Disponible en: http://www.biostat.jhsph.edu/~fdominic/teaching/bio656/references/Takkouche.1999.pdf

Taylor, M. J., y White, K. R. (1992). An evaluation of alternative methods for computing standardized mean difference effect size. *Journal of Experimental Education, 61*, 63-72. doi: 10.1080/00220973.1992.9943850

Téllez, A., García, C. H., y Corral-Verdugo, V. (2015). Effect size, confidence intervals and statistical power in psychological research. *Psychology in Russia: State of the Art, 8*, 27-46. doi: 10.11621/pir.2015.0303

Terwee, C. B., Roorda, L. D., Knol, D. K., De Boer, M. R., y De Vel, H. C. W. (2009). Linking measurement error to minimal important change of patient-reported outcomes. *Journal of Clinical Epidemiology, 62*, 1062-1067. doi: 10.1016/j.jclinepi.2008.10.011

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70,* 434-438. doi: 10.1002/j.1556-6676.1992.tb01631.x

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61,* 361–377. doi: 10.1080/00220973.1993.10806596

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25,* 26-30. doi: 10.3102/0013189X025002026

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools, 5,* 33–38. Disponible en: http://www.personal.psu.edu/faculty/d/m/dmr/sigtest/4mspdf.pdf

Thompson, B. (1999). Journal editorial polices regarding statistical significance tests: heat is to fire as p is to importance. *Educational Psychology Review, 11*, 157-169. doi: 10.1023/A:1022028509820

Thompson, B. (2002a). "Statistical," "Practical," and "Clinical": How many kinds of significance do counselors need to consider. J*ournal of Counseling & Development, 80*, 64-71. doi: 10.1002/j.1556-6678.2002.tb00167.x

Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect size. *Educational Researcher, 31*, 25-32. doi: 10.3102/0013189X031003025

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Newbury Park, CA: Sage.

Thompson, B. (2006). Role of effect sizes in contemporary research in counseling. *Counseling and Values, 50*, 176-186. doi: 10.1002/j.2161-007X.2006.tb00054.x

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432. doi: 10.1002/pits.20234

Torgerson, C. J. (2006). Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies, 54*, 89–102. doi: 110.1111/j.1467-8527.2006.00332.x.

Trafimow, D., y Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*,1–2. doi: 10.1080/01973533.2015.1012991

Tressoldi, P. E., Giofré, D., Sella, F., y Cumming, G. (2013). High Impact = high statistical standards? Not necessarily so. *PloS One, 8(2),* e56180. doi: 10.1371/journal.pone.0056180

Trigo-Sánchez, M. E., y Martínez-Cervantes, R. J. (2016). Generalized eta squared for multiple comparisons on between-groups designs. *Psicothema, 28,* 340-345. doi: 10.7334/psicothema2015.124

Tritchler, D. (1999). Modelling study quality in meta-analysis. *Statistics in Medicine, 18,* 2135-2145. doi: 10.1002/(SICI)1097-0258(19990830)18:16<2135::AID-SIM183>3.0.CO;2-5

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53,* 796 doi: 10.1037/0003-066X.53.7.796.b

Turner, D., Schünemannb, H.J, Griffith, L:E., Beatone , D.E, Griffiths, AM., Critch, JN., y Guyatt, G.H. (2010). The minimal detectable change cannot reliably replace the minimal important difference. *Journal of Clinical Epidemiology, 63,* 28-36. doi: 10.1016/j.jclinepi.2009.01.024

Urra-Medina, E., y Barría-Pailaquilén, R. (2010). Systematic review and its relationship with Evidence-Based Practice in health. *Revista Latino-Americana de Enfermería, 18(4),* 824-831. Recuperado desde 18/08/2016 http://www.scielo.br/pdf/rlae/v18n4/23.pdf

Vacha-Haase, T. (2001) Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61,* 219-224. doi: 10.1177/00131640121971194

Vacha-Haase, T., y 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,* 104-105. doi: 10.1037/0735-7028.30.1.104

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T.S., y Thompson, B., (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory and Psychology, 10,* 413–425. doi: 10.1177/0959354300103006

Vacha-Haase, T., y Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51,* 473–481. doi: 10.1037/0022-0167.51.4.473

Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypotheses testing by university students. *Themes in Education, 3,* 183–198.

Vallecillos, A., y Batanero, C. (1997). Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios. *Recherches en Didactique des Mathematiques, 17,* 29-48

Valera-Espín, A., y Sánchez-Meca, J. (1997). Pruebas de significación y magnitud del efecto: Reflexiones y propuestas. *Anales de Psicología, 13,* 85-90. Disponible en: http://www.um.es/analesps/v13/v13_1/09-13-1.pdf

Valera-Espín, A., y Sánchez-Meca, J., y Marín-Martínez, F. (2000). Contraste de hipótesis e investigación psicológica española: análisis y propuestas. *Psicothema, 12,* 549-552. Disponible en: http://www.psicothema.com/pdf/623.pdf

Vargha, A., y Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132. doi: 10.3102/10769986025002101

Vázquez, C., y Nieto, M. (2003). Psicología (clínica) basada en la evidencia) (PBE): una revisión conceptual y metodológica. En J. L. Romero (Ed), Psicópolis: *Paradigmas actuales y alternativos en la psicología contemporánea.* Barcelona. Paidós.

Verdam, M. G. E., Oort, F. J., y Sprangers, M. A. G. (2014). Significance, truth and proof of p values: reminders about common misconceptions regarding null hypothesis significance testing. *Quality of Life Research, 23,* 5-7. doi: 10.1007/s11136-013-0437-2

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin y Review, 14,* 779-804. doi: 10.3758/BF03194105

Wang, Z., y Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? *Journal of Experimental Education, 75,* 109–125. doi: 10.3200/JEXE.75.2.109-125

Wasserstein., R. L., y Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician, 70*, 129-133. doi: 10.1080/00031305.2016.1154108

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods* (2nd editon). New York, NY: Springer.

Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing. (3rd edition). San Diego, CA: Elsevier

Wilcox, R. R., y Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254–274. doi: 10.1037/1082-989X.8.3.254

Wilkinson, L., y the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist*, 54, 594-604. doi: 10.1037/0003-066X.54.8.594

Wortman, P.M. (1994). Judging research quality. En H. Cooper y L.V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 97-109). Nueva York: Russell Sage Foundation.

Wright, R. W., Brand, R. A. Dunn, W., y Spindler, K. (2007). How to write a systematic review. *Clinical Orthopaedics and related research*, *455*, 23-29. doi:10.1097/BLO.0b013e31802c9098