DOI: https://doi.org/10.11144/Javeriana.upsy16-3.fvcp

Falacias sobre el valor *p* compartidas por profesores y estudiantes universitarios*

Fallacies about *p*-Value Shared by Professors and University Students

Recepción: 19 Junio 2016 | Aprobación: 08 Febrero 2017

Laura Badenes-Ribera[a]
Universitat de València, España
ORCID: http://orcid.org/0000-0002-4706-690X
Maria Dolores Frias Navarro
Universitat de València, España

[a] Autor de correspondencia. Correo electrónico: Laura.badenes@uv.es

Para citar este artículo: Bardenes-Ribera, L., & Frías Navarro, M. D. (2017). Falacias sobre el valor p compartidas por profesores y estudiantes universitarios. *Universitas Psychologica*, *16*(3), 1-10.
https://doi.org/10.11144/Javeriana.upsy16-3.fvcp

RESUMEN

La "Práctica Basada en la Evidencia" requiere a los profesionales valorar de forma crítica los resultados de las investigaciones psicológicas. Sin embargo, las interpretaciones incorrectas de los valores *p* de probabilidad son abundantes y repetitivas. Estas interpretaciones incorrectas pueden afectar las decisiones profesionales y poner en riesgo la calidad de las intervenciones y la acumulación de un conocimiento científico válido. Por lo tanto, identificar el tipo de falacia que subyace a las decisiones estadísticas es fundamental para abordar y planificar estrategias de educación estadística dirigidas a intervenir sobre las interpretaciones incorrectas. En consecuencia, el objetivo de este estudio es analizar la interpretación del valor *p* en estudiantes y profesores universitarios de psicología. La muestra estuvo formada por 161 participantes (43 profesores y 118 estudiantes). La antigüedad media como profesor fue de 16.7 años (*DE* = 10.07). La edad media de los estudiantes fue de 21.59 (*DE* = 1.3). Los hallazgos sugieren que los estudiantes y profesores universitarios no conocen la interpretación correcta del valor *p*. La falacia de la probabilidad inversa presentó mayores problemas de comprensión. Además, se confundieron la significación estadística y la significación práctica o clínica. Estos resultados destacan la necesidad de la educación estadística y reeducación estadística.

Palabras clave

practica basada en la evidencia; interpretación errónea; valores p; tamaño del efecto; educación estadística

ABSTRACT

The "Evidence Based Practice" requires professionals to critically assess the results of psychological research. However, incorrect interpretations of *p* values of probability are abundant and repetitive. These misconceptions may affect professional decisions and compromise the quality of interventions and the accumulation of a valid scientific knowledge. Therefore, identifying the types of fallacies that underlying statistical decisions is fundamental for approaching and planning statistical education strategies designed to intervene in incorrect interpretations. Consequently, the aim of this study is to analyze the interpretation of *p* value among university students of psychology and academic psychologists. The sample was composed of 161 participants (43 academics and 118 students). The mean number of years as academic was

16.7 (*SD* = 10.07). The mean age of university students was 21.59 years (*SD* = 1.3). The findings suggest that college students and academics do not know the correct interpretation of *p* values. The inverse probability fallacy presented major problems of comprehension. In addition, the participants confused statistical significance and practical significance or clinical or the findings. There is a need for statistical education and statistical re-education.

Keywords
evidence-based practice; misconception; p value; statistical education

Introducción

La Práctica Basada en la Evidencia (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], 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 (APA, 2005; Babione, 2010; Daset & Cracco, 2013; Frias-Navarro & Pascual-Llobell, 2003; Sánchez-Meca, Boruch, Petrosino, & Rosa-Alcázar, 2002; Vázquez & Nieto, 2003). Por tanto, el enfoque de la PBE requiere de los profesionales nuevas habilidades como la capacidad para evaluar y jerarquizar la calidad de las investigaciones psicológicas (Beyth-Maron, Fidler, & Cumming, 2008; Pascual-Llobell, Frias-Navarro, & Monterde-i-Bort, 2004).

Dentro de ese proceso de valoración crítica de la evidencia científica es crucial conocer y comprender el proceso de contraste de hipótesis estadísticas mediante la ejecución de la prueba de significación de la hipótesis nula (Null Hypothesis Significance Testing [NHST]), sobre todo teniendo en cuenta que en el ámbito de la psicología el procedimiento de la NHST es la técnica por excelencia en el análisis de datos (Cumming et al., 2007). Por tanto, saber interpretar los valores *p* de probabilidad es una competencia básica del profesional de la psicología y de todas aquellas disciplinas donde se aplica la inferencia estadística.

Sin embargo, el procedimiento de la NHST ha sido criticado desde el principio de su aplicación en psicología y otras ciencias (Berkson, 1938; Cohen, 1994). Una de las cuestiones que más controversia ha provocado es la interpretación correcta del valor *p* asociado al resultado de la prueba estadística (Frias-Navarro, Pascual-Soler, Badenes-Ribera, & Monterde-i-Bort, 2014; Wasserstein & Lazar, 2016).

El valor *p* es la probabilidad del resultado observado o un valor más extremo si la hipótesis nula es cierta (Hubbard & Lindsay, 2008; Johnson, 1999; Kline, 2013). La definición es clara y precisa, sin embargo, las interpretaciones incorrectas siguen siendo abundantes y repetitivas (Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, & Pascual-Soler, 2015; Perezgonzalez, 2015a, 2015b; Verdam, Oort, & Sprangers, 2014).

Las cuatro interpretaciones erróneas del valor *p* más comunes son: (1) Falacia de la probabilidad inversa; (2) Falacia de la probabilidad contra el azar; (3) Falacia del tamaño del efecto y (4) Falacia de la significación clínica o práctica (Badenes-Ribera et al., 2015; Balluerka, Gómez, & Hidalgo, 2005; Cohen, 1994; Cumming, 2012; Falk & Greenbaum, 1995; Kline, 2013; Nickerson, 2000; Téllez, García, & Corral-Verdugo, 2015).

La falacia de la "probabilidad inversa" es la falsa creencia de que el valor p hace referencia a la probabilidad de que la hipótesis nula (H_0) sea verdadera dados ciertos datos [$Pr(H_0|Datos)$]. Por su parte, la "falacia de las probabilidades contra el azar" señala que el valor *p* es la probabilidad de obtener el resultado por azar o la probabilidad de que el resultado ocurra como consecuencia del proceso de la selección de la muestra. Ambas están relacionadas con el mismo problema: confundir la probabilidad del resultado, asumiendo que la hipótesis nula es cierta [$Pr(Datos|H_0)$], con la probabilidad de la hipótesis nula, dados ciertos datos [$Pr(H_0|Datos)$]. Las pruebas de significación estadística no ofrecen información de la probabilidad condicional de la hipótesis nula dados los datos obtenidos en la investigación (Kirk, 1996; Shaver, 1993).

La "falacia del tamaño del efecto" vincula la significación estadística con la magnitud del efecto (Gliner, Vaske, & Morgan, 2001). De este modo, los valores pequeños de *p* son interpretados como efectos grandes (Kline, 2013). Sin embargo, el valor *p* no informa de la magnitud de un efecto (Cumming, 2012). El tamaño del efecto solo puede ser conocido estimando directamente su valor con el estadístico adecuado y su intervalo de confianza (Gliner et al., 2001; Wasserstein & Lazar, 2016; Wilkinson, 1999).

La "falacia de la significación clínica o práctica" asocia el valor *p* con la importancia práctica o clínica de un hallazgo. Sin embargo, un resultado estadísticamente significativo no indica que sea un resultado importante desde el punto de vista clínico, práctico o sustantivo y viceversa (Gliner, Leech, & Morgan, 2002; Kirk, 1996; Palmer & Sesé, 2013; Wasserstein & Lazar, 2016). Bajo esta falsa creencia es posible que efectos sin significación estadística pero con significación clínica o importancia práctica sean rechazados. Y, al contrario, efectos con significación estadística y poca significación clínica o importancia práctica se tomen como significativos o importantes (Frias-Navarro, 2011).

Detrás de estas interpretaciones erróneas hay unas creencias y atribuciones sobre el significado de los resultados. Por ello, es necesario comprender el razonamiento estadístico o la forma de razonar con ideas estadísticas y dar sentido a la información estadística que realizan las personas (Garfield, 2002; Leek, 2014).

Estudios previos han detectado la presencia de este tipo de falacias sobre el valor *p* de probabilidad en muestras de profesores y estudiantes universitarios de distintas disciplinas (p. ej., Castro-Sotos, Vanhoof, Van den Noortgate, & Onghena, 2009; Falk & Greenbaum, 1995; Frias-Navarro et al., 2014; Haller & Kraus, 2002; Monterde-i-Bort, Frias-Navarro, & Pascual-Llobel 2010; Oakes, 1986; Vallecillos, 2002; Vallecillos & Batanero, 1997). Por ejemplo, en el ámbito de la psicología, Oakes (1986) encontró que el 97 % de los profesores universitarios interpretaron de forma incorrecta el significado del valor *p*. Haller y Kraus (2002) replicaron el estudio de Oakes (1986) en una muestra de profesores y estudiantes universitarios. Sus resultados revelaron que el 80 % de profesores de Metodología, el 89.7 % de profesores que no enseñaban metodología y el 100 % de los estudiantes cometieron algún tipo de error de interpretación del valor *p*. Finalmente, en los recientes estudios de Badenes-Ribera et al. (2015) y Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos y Longobardi (2016), en sendas muestras de profesores universitarios, se observó que muchos de ellos, incluidos los profesores del área de Metodología, no sabían interpretar correctamente los valores de *p* asociados a los resultados de las pruebas de inferencia estadística. En ambos estudios, la "falacia de la probabilidad inversa" presentó los mayores problemas de comprensión. Además, el 35.2 % de los profesores cometieron la "falacia de la significación clínica o práctica" de los resultados, es decir, confundieron la significación estadística de los resultados con la significación clínica o práctica de los mismos (Badenes-Ribera et al., 2015).

El objetivo de la presente investigación es detectar los errores de razonamiento estadístico que estudiantes y profesores universitarios realizan ante los resultados que aporta una prueba de inferencia estadística, pues su visión e interpretación de los hallazgos son un filtro de calidad en su vida profesional que no puede estar sometido a falsas creencias del procedimiento estadístico que representa la herramienta fundamental para obtener conocimiento científico. La competencia de la 'lectura crítica' dentro del modelo de la PBE requiere conocer e interpretar adecuadamente la calidad metodológica de las pruebas o evidencias aportadas por la literatura. Del mismo modo, los investigadores deben producir evidencia o pruebas empíricas correctamente interpretadas no confundiendo el alcance de sus resultados. En consecuencia, identificar el tipo de falacia que subyace a las decisiones estadísticas es fundamental para abordar y planificar estrategias de educación estadística dirigidas a intervenir sobre las interpretaciones incorrectas.

Método

Muestra

Se utilizó una muestra no probabilística (conveniencia). La muestra estuvo formada por 161 participantes de la Universidad de Valencia. De ellos, el 26.7 % (n = 43) fueron profesores (n = 43) y el 73.3 %, estudiantes de cuarto curso del grado de psicología que ya habían cursado las asignaturas de Estadística y Psicometría (n = 118). El 39.5 % de los profesores fueron hombres (n = 17) y el 60.5 % mujeres (n = 26). La antigüedad media como profesor fue de 16.7 años (DE = 10.07). Respecto de la muestra de estudiantes, el 19.5 % fueron hombres (n = 23), el 78.8 % fueron mujeres (n = 93) y el 1.7 % no indicó su sexo (n = 2) con una edad media de 21.59 (DE = 1.3).

Instrumento

En la primera sección de la encuesta se recogió información sobre variables sociodemográficas: sexo, edad (estudiantes universitarios), antigüedad como profesor en la universidad.

La segunda sección incluyó la encuesta sobre interpretaciones del valor p de Badenes-Ribera et al. (2015). Este instrumento está compuesto por 10 ítems con una escala de respuesta dicotómica (verdadero o falso) pensados para detectar interpretaciones erróneas sobre el valor p de probabilidad asociadas a las pruebas de inferencia estadística y su interpretación correcta. En el presente estudio, se utilizaron los ítems referentes a falacia de la probabilidad inversa (5 ítems), falacia del tamaño del efecto (1 ítem), falacia de la significación clínica o práctica (1 ítem) y, finalmente, interpretaciones correctas del valor p del procedimiento de contraste de hipótesis (2 ítems). Las cuestiones se plantearon con el siguiente argumento:

"Supongamos que un artículo de investigación señala un valor de p = 0.001 en el apartado de resultados (alfa = 0.05). Señale si las siguientes afirmaciones son verdaderas o falsas".

A.-Falacia de la Probabilidad inversa:

1. Se ha probado que la hipótesis nula es verdadera.
2. Se ha probado que la hipótesis nula es falsa.
3. Se ha determinado la probabilidad de la hipótesis nula (p = 0.001).
4. Se ha deducido la probabilidad de la hipótesis experimental (p = 0.001).
5. La probabilidad de que la hipótesis nula sea verdadera, dados los datos obtenidos, es de 0.001.

B.-Falacia del tamaño del efecto:

6. El valor p = 0.001 confirma de forma directa que el tamaño del efecto ha sido grande.

C.- Falacia de la significación clínica o práctica:

7. Obtener un resultado estadísticamente significativo indica que el efecto detectado es importante.

D.- Interpretación correcta y decisión adoptada:

8. Se conoce la probabilidad del resultado de la prueba estadística, asumiendo que la hipótesis nula es cierta.
9. Dado que p = 0.001 entonces el resultado obtenido permite concluir que las diferencias no se deben al azar.

Procedimiento

La participación en el estudio fue voluntaria y completamente anónima. Los estudiantes respondieron a la encuesta en horas de clase y no recibieron ninguna compensación por ello. Por su parte, los profesores cumplimentaron las preguntas a través de internet, mediante encuesta online. Para ello, se registraron las direcciones de correo electrónico de los profesores y se les envió un mensaje invitándolos a participar en el estudio sobre cognición y práctica estadística.

En la realización del presente estudio, se siguieron los valores éticos requeridos en la investigación con seres humanos, respetando los principios fundamentales incluidos en la Declaración de Helsinki, en sus actualizaciones y normativas vigentes (consentimiento informado y derecho a la información, protección de datos

personales y garantías de confidencialidad, no discriminación y posibilidad de abandonar el estudio en cualquiera de sus fases).

Análisis de datos

Se utilizó el programa estadístico IBM SPSS v. 20 para Windows. Se analizaron las frecuencias y porcentajes de respuestas a los ítems. Además, los análisis incluyeron la estimación del intervalo de confianza para los porcentajes. Para el cálculo de los intervalos de confianza de los porcentajes se utilizó el método score de Wilson (Newcombe, 2012).

Resultados

La Tabla 1 muestra el porcentaje de participantes que estuvieron de acuerdo con las afirmaciones sobre el valor *p* de probabilidad.

TABLA 1
Porcentaje de participantes que están de acuerdo con las afirmaciones (intervalo de confianza 95 %)

Ítem	Estudiantes (n = 118)	Profesores (n = 43)
Falacia de la probabilidad inversa		
1. Se ha probado que la hipótesis nula es verdadera	11.02 [6.55, 17.94]	2.33 [0.41, 12.6]
2. Se ha probado que la hipótesis nula es falsa	68.64 [59.80, 76.32]	62.79 [47.86, 75.62]
3. Se ha determinado la probabilidad de la hipótesis nula (p = 0.001)	57.63 [48.61, 66.16]	60.47 [45.58, 73.63]
4. Se ha deducido la probabilidad de la hipótesis experimental (p = 0.001)	29.66 [22.17, 38.44]	39.53 [26.37, 54.42]
5. La probabilidad de que la hipótesis nula sea verdadera, dados los datos obtenidos, es de 0.001	44.92 [36.24, 53.91]	30.23 [18.6, 45.11]
% Participantes que han valorado correctamente las cinco afirmaciones como falsas	3.39 [1.33, 8.39]	0 [0, 8.2]
Falacia del tamaño del efecto y de la significación clínica		
6. El valor p = 0.001 confirma de forma directa que el tamaño del efecto ha sido grande	11.02 [6.55, 17.94]	13.95 [6.56, 27.26]
7. Obtener un resultado estadísticamente significativo indica que el efecto detectado es importante	19.49 [13.35, 27.55]	32.56 [20.49, 47.48]
% Participantes que han valorado correctamente las dos afirmaciones como falsas	67.8 [58.92, 75.55]	62.79 [47.86, 75.62]
Interpretaciones correctas y decisión adoptada por el investigador		
8. Se conoce la probabilidad del resultado de la prueba estadística, asumiendo que la hipótesis nula es cierta	54.24 [45.26, 62.95]	30.23 [18.60, 45.11]
9. Dado que p = 0.001 entonces el resultado obtenido permite concluir que las diferencias no se deben al azar	58.47 [49.45, 66.96]	74.42 [59.76, 85.07]
% Participantes que han valorado correctamente las dos afirmaciones como verdaderas	31.36 [23.68, 40.2]	18.6 [9.74, 32.62]

Fuente: elaboración propia.

Respecto de la falacia de la probabilidad inversa, se observa que gran parte de los participantes percibieron como verdadera alguna de las cinco interpretaciones erróneas del valor *p*. Las interpretaciones que mayor respaldo recibieron tanto por parte de los estudiantes como de los profesores fueron "se ha probado que la hipótesis nula es falsa" y "se ha determinado la probabilidad de la hipótesis nula (p = 0.001)". Además, ninguno de los profesores valoró correctamente las cinco afirmaciones sobre las interpretaciones del valor *p*, frente al 3.39 % (IC 95% [1.33, 8.39]) de los estudiantes que sí que lo hicieron.

Respecto de las interpretaciones erróneas del valor *p* relacionadas con la falacia del tamaño del efecto y la falacia de la significación clínica o práctica de los hallazgos, se observa que la mayoría de los participantes, tanto en los estudiantes como en los profesores, valoraron correctamente estas afirmaciones, es decir, no cometieron este tipo de interpretaciones erróneas. Además, en ambas muestras, la afirmación falsa (o interpretación errónea) que mayor respaldo recibió fue "obtener un resultado estadísticamente significativo indica que el efecto detectado es importante". En consecuencia, los participantes presentaron mayores problemas en discernir entre la significación estadística de los resultados obtenidos y la significación práctica o clínica de los mismos.

Finalmente, respecto de las interpretaciones correctas del valor *p* de probabilidad, se observa un patrón distinto entre los profesores y los estudiantes en la comprensión del valor *p* de probabilidad. Mientras que los profesores presentaron mayores problemas con la interpretación probabilística del valor *p* mejorando notablemente su interpretación cuando se hace en términos de decisión estadística. Los estudiantes presentaron mayores problemas de comprensión con la interpretación estadística del valor *p*, mejorando su interpretación cuando la misma se hace en términos de probabilidad. Si bien en el caso de los estudiantes, la diferencia entre ambas interpretaciones no es estadísticamente significativa puesto que existe un solapamiento

entre los intervalos de confianza de los porcentajes. Finalmente, solo un pequeño porcentaje de participantes en ambas muestras valoraron correctamente las dos afirmaciones como verdaderas, sobre todo en el caso de los profesores universitarios.

Discusión

Los resultados del estudio indican que la comprensión e interpretación de los valores *p* de las pruebas de inferencia estadística sigue siendo problemática entre los estudiantes universitarios y los profesores universitarios. Confundir el nivel de significación de alfa con la probabilidad de que la hipótesis nula sea cierta, interpretar un resultado estadísticamente significativo como un resultado importante o útil son interpretaciones erróneas o falsas creencias que continúan existiendo entre estudiantes y profesores universitarios de psicología (Falk & Greenbaum 1995; Haller & Kraus, 2002; Kühberger, Fritz, Lermer, & Scherndl, 2015; Oakes, 1986).

Estos resultados son consistentes con estudios previos en muestras de estudiantes universitarios (p. ej., Castro-Sotos et al., 2009; Falk & Greenbaum, 1995; Haller & Kraus, 2002; Kühberger et al., 2015; Vallecillos, 2002, Vallecillos & Batanero, 1997) en muestras de profesores universitarios de psicología (p. ej., Badenes-Ribera et al., 2015; Badenes-Ribera et al., 2016; Haller & Kraus, 2002; Monterde-i-Bort et al., 2010; Oakes, 1986), en muestras de miembros de la American Educational Research Association (AERA) (p. ej., Mittag & Thompson, 2000) y en profesionales de la Estadística (Lecoutre, Poitevineau, & Lecoutre, 2003).

La "falacia de la probabilidad inversa" es la que se observó con mayor frecuencia. Además, un gran número de estudiantes universitarios y profesores de psicología confundieron la significación estadística de los resultados obtenidos con la significación práctica de los mismos. Sin embargo, el valor *p* no ofrece información de la magnitud del efecto o importancia del resultado (Gliner et al., 2002; Rosenthal, 1993; Shaver, 1993). La significación clínica o sustantiva no se corresponde con el valor de ningún estadístico, ni del resultado de la prueba de inferencia estadística (valor *p*) ni de la magnitud del tamaño del efecto (APA, 2010; Cumming, 2012; Kline, 2013). 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). Por tanto, la presentación de muchos asteriscos junto al valor *p* de probabilidad o valores *p* muy pequeños solo señalan que en ese diseño la hipótesis nula es poco plausible, pero de ahí no se puede inferir que el efecto encontrado es importante, que la relación entre las variables es fuerte o que existe una relevancia sustantiva (Frias-Navarro, 2011; Gliner et al., 2001; Palmer & Sesé, 2013; Newcombe, 2012).

Las falacias del tamaño del efecto y de la significación clínica o práctica de los resultados representan una de las críticas más fuertes contra las pruebas de significación estadística y, en gran medida, han provocado el movimiento de la reforma estadística (Cumming, 2012; Kline, 2013; Wilkinson, 1999) que aboga por acompañar los valores *p* con información del tamaño del efecto y sus intervalos de confianza (Balluerka, Vergara, & Arnau, 2009; Cumming, 2012; Maher, Markey, & Ebert-May, 2013; Newcombe, 2012; Savalei & Dunn, 2015; Téllez et al., 2015; Valera-Espín, Sánchez-Meca, & Marín-Martínez, 2000), tal y como señala el manual de la APA (2010). Sin embargo, estudios previos han mostrado que los intervalos de confianza no están exentos de errores de interpretación (e. g., Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Perezgonzalez, 2015a).

Finalmente, los hallazgos del estudio deben interpretarse con ciertas limitaciones. El procedimiento de muestreo (muestra de conveniencia) limita la validez externa de nuestros resultados. Sin embargo, los resultados son consistentes con estudios previos tal y como ya se ha comentado. Así pues, la presencia

de interpretaciones erróneas del valor *p* entre estudiantes y profesores universitarios indica la necesidad de mejorar la formación o educación estadística de los profesores para evitar la perpetuación de estas falacias (Haller & Kraus, 2002; Kline, 2013) y producir un conocimiento científico válido. También se necesita mejorar el contenido de los libros de estadística para garantizar una formación de calidad a los futuros profesionales de la psicología (Cumming, 2012; Gliner et al., 2002; Kline, 2013). La enseñanza de la estadística no solo debe consistir en cálculos de enseñanza, procedimientos y fórmulas, deberían centrarse mucho más en el pensamiento y la comprensión de los métodos estadísticos (Haller & Kraus, 2002; Perezgonzalez, 2015b).

Para ello, se requiere un enfoque multifacético que implique a los autores de libros de texto, profesores, autores de paquetes estadísticos de software, editores de revistas y manuales de publicación (Balluerka et al., 2005; Kirk, 2001). Por ejemplo, como Gliner et al. (2002) señalan, los autores de libros de texto deberían incluir una sección sobre el debate y críticas del procedimiento NHST que incluyera un apartado sobre su origen. Pues, probablemente, la mayoría de los problemas de interpretación vinculados al valor *p* radican en que este procedimiento es una fusión entre la prueba de significación de Fisher y la prueba de la hipótesis estadística de Neyman y Pearson (Hager, 2013; Ivarsson, Andersen, Stenling, Johnson, & Lindwall, 2015; Perezgonzalez, 2015b). Además, los libros deberían incluir una sección sobre el cálculo del tamaño del efecto y sus intervalos de confianza y, poner ejemplos sobre la importancia práctica o clínica de un hallazgo. En los dos primeros libros de la colección Reforma Estadística editados en España, se detallan todas estas cuestiones en diferentes capítulos (Frias-Navarro, 2011; Frias-Navarro et al., 2014).

La PBE requiere de profesionales que valoren críticamente los hallazgos de los estudios o investigaciones psicológicas y, para ello, es necesaria una formación en conceptos estadísticos, en metodología de diseños de investigación y en resultados de pruebas de inferencia estadística (Badenes-Ribera et al., 2016; Beyth-Maron et al., 2008). Finalmente, la investigación futura debe ir dirigida ahora a la intervención sobre las falacias vinculadas a la interpretación del valor *p* de probabilidad.

Agradecimientos

Este estudio fue financiado por el programa VALI +d de Formación Pre-doctoral de Investigadores (ACIF/2013/167) de la Conselleria d'Educació, Cultura i Esport, Generalitat Valenciana (España), y parcialmente subvencionado por el Ministerio de Economía y Competitividad (I + D + i) (España) (EDU2011-22862).

Referencias

American Psychological Association. (2005). *Policy Statement on Evidence-Based Practice in Psychology*. Washington, DC: Autor.

American Psychological Association. (2006). Evidence-based practice in psychology: APA Presidential Task Force on evidence-based practice. *American Psychologist*, *61*, 271-285. http://dx.doi.org/10.1037/0003-066X.61.4.271

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th. ed.). Washington, DC: Autor.

Babione, J. M. (2010). Evidence-Based Practice in Psychology: An ethical framework for graduate education, clinical training, and maintaining professional competence. *Ethics & Behavior*, *20*, 443-453. http://dx.doi.org/10.1080/10508422.2010.521446

Badenes-Ribera. L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian academic psychologists. *Frontiers in Psychology*, *7*, 1247. http://dx.doi.org/10.3389/fpsyg.2016.01247

Badenes-Ribera, L., Frias-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. http://dx.doi.org/10.7334/psicothema2014.283

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology*, *1*, 55-70. http://dx.doi.org/10.1027/1614-1881.1.2.55

Balluerka, N., Vergara, A. I., & Arnau, J. (2009). Calculating the main alternatives to Null Hypothesis Significance testing in between subject experimental designs. *Psicothema*, *21*(1), 141-151.

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.

Beyth-Maron, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal*, *7*(2), 20-39.

Castro-Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education*, *17*(2). (Número de servicio de reproducción de documentos ERIC EJ856367). Recuperado de http://www.amstat.org/publications/jse/v17n2/castrosotos.html

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997-1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Nueva York: Routledge.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P, Christiansen, A., Kleinig, A.,… & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*, 230-232. http://dx.doi.org/10.1111/j.1467-9280.2007.01881.x

Daset, L. R., & Cracco, C. (2013). Psicología Basada en la Evidencia: algunas cuestiones básicas y una aproximación a través de una revisión bibliográfica. *Ciencias Psicológicas*, *7*(2), 209-220.

Falk, R., & Greenbaum, C. W. (1995). Significance tests Die Hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75-98. http://dx.doi.org/10.1177/0959354395051004

Frias-Navarro, D. (2011). *Técnica estadística y diseño de investigación*. Valencia: Palmero Ediciones.

Frias-Navarro, D., & Pascual-Llobell, J. (2003). Psicología clínica basada en pruebas: efecto del tratamiento. *Papeles del Psicólogo*, *24*(85), 11-18.

Frias-Navarro, D., Pascual-Soler, M., Badenes-Ribera, L., & Monterde-i-Bort, H. (2014). *Reforma estadística en psicología*. Valencia: Palmero Ediciones.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistic Education*, *10*. Recuperado de http://www.amstat.org/publications/jse/v10n3/garfield.html

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education*, *71*, 83-92. http://dx.doi.org/10.1080/00220970209602058

Gliner, J. A., Vaske, J. J., & Morgan, G. A. (2001). Null hypothesis significance testing: Effect size matters. *Human Dimensions of Wildlife*, *6*, 291-301. http://dx.doi.org/ 10.1080/108712001753473966

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: A methodological perspective. *Theory & Psychology*, *23*, 251-270. http://dx.doi.org/10.1177/0959354312465483

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online* [On-line serial], 7, 120. Recuperado

de http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157-1164. http://dx.doi.org/10.3758/s13423-013-0572-3

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*, 69-88. http://dx.doi.org/10.1177/0959354307086923

Ivarsson, A., Andersen, M. B., Stenling, A., Johnson, U., & Lindwall, M. (2015). Things we still haven't learned (so far). *Journal of Sport & Exercise Psychology*, *37*, 449-461. http://dx.doi.org/10.1123/jsep.2015-0015

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management*, *63*, 763-772.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 332-339. http://dx.doi.org/10.1037/0022-006X.67.3.332

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759. http://dx.doi.org/10.1177/0013164496056005002

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*, 213-218. http://dx.doi.org/ 10.1177/00131640121971185

Kline, R. B. (2013). *Beyond significance testing: Statistic reform in the behavioral sciences*. Washington, DC: APA.

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC Research Notes*, *17*(8), 84. http://dx.doi.org/10.1186/s13104-015-1020-4

Lecoutre, M. P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Tests. *International Journal of Psychology*, *38*, 37-45. http://dx.doi.org/10.1080/00207590244000250

Leek, J. (14 de febrero de 2014). On the scalability of statistical procedures: Why the p-value bashers just don't get it [Simply Statistics Blog]. Recuperado de http://simplystatistics.org/2014/02/14/on-the-scalability-of-statisticalprocedures-why-the-p-value-bashers-just-dont-get-it/

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. *CBE Life Sciences Education*, *12*, 345-351. http://dx.doi.org/10.1187/cbe.13-04-0082

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance test and others statistical issues. *Educational Researcher*, *29*, 14-20. http://dx.doi.org/10.3102/0013189X029004014

Monterde-i-Bort, H., Frias-Navarro, D., & Pascual-Llobel, 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. http://dx.doi.org/10.1007/s10212-010-0021-x

Newcombe, R. G. (2012). *Confidence intervals for proportions and related measures 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. http://dx.doi.org/10.1037/1082-989X.S.2.241

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chichester, England: Wiley.

Palmer, A., & Sesé, A. (2013). Recommendations for the use of statistics in clinical and health psychology. *Clínica y Salud*, *24*, 47-54. http://dx.doi.org/10.5093/cl2013a6

Pascual-Llobell, J., Frias-Navarro, D., & Monterde-i-Bort, H. (2004). Tratamientos psicológicos con apoyo empírico y práctica

clínica basada en la evidencia. *Papeles del Psicólogo*, *25*(87), 1-8.

Perezgonzalez, J. D. (2015a).Confidence intervals and tests are two sides of the same research question. *Frontiers in Psychology*, 6, 34. http://dx.doi.org/10.3389/fpsyg.2015.00034

Perezgonzalez, J. D. (2015b).Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, 6, 223. http://dx.doi.org/10.3389/fpsyg.2015.000223

Rosenthal, R. (1993). Cumulating evidence. En G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519-559). Hillsdale, NJ: Erlbaum.

Sánchez-Meca, J., Boruch, R. F., Petrosino, A., & Rosa-Alcázar, A. I. (2002). La Colaboración Campbell y la Práctica basada en la Evidencia. *Papeles del Psicólogo*, *83*, 44-48.

Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*, 6, 245. http://dx.doi.org/10.3389/fpsyg.2015.00245

Shaver, J. P. (1993). What statistical significance testing is, and what is not. *The Journal of Experimental Education*, *61*, 293-316.

Téllez, A., García, C. H., & Corral-Verdugo, V. (2015). Effect size, confidence intervals and statistical power in psychological research. *Psychology in Russia: State of the Art*, *8*, 27-46. http://dx.doi.org/10.11621/pir.2015.0303

Valera-Espín, A., Sánchez-Meca, J., & Marín-Martínez, F. (2000). Contraste de hipótesis e investigación psicológica española: análisis y propuestas. *Psicothema*, *12*(Supl. 2), 549-552.

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., & Batanero, C. (1997). Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios. *Recherches en Didactique des Mathématiques*, 17, 29-48.

Vázquez, C., & 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* (pp. 465-485). Barcelona: Paidós.

Verdam, M. G. E., Oort, F. J., & 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. http://dx.doi.org/10.1007/s11136-013-0437-2

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129-133. http://dx.doi.org/10.1080/00031305.2016.1154108

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist*, *54*, 594-604. http://dx.doi.org/10.1037/0003-066X.54.8.594

Notas

* Artículo de investigación.