Universitat d'Alacant
Universidad de Alicante

XIII JORNADES DE XARXES D´INVESTIGACIÓ EN DOCÈNCIA UNIVERSITÀRIA

Noves estratègies organitzatives i metodològiques en la formació universitària per a respondre a la necessitat d'adaptació i canvi

JORNADAS DE REDES DE INVESTIGACIÓN EN DOCENCIA UNIVERSITARIA XIII

Nuevas estrategias organizativas y metodológicas en la formación universitaria para responder a la necesidad de adaptación y cambio

ISBN: 978-84-606-8636-1

Coordinadores

María Teresa Tortosa Ybáñez
José Daniel Álvarez Teruel
Neus Pellín Buades

Universidad de Alicante
Vicerrectorado de Estudios, Formación y Calidad
Instituto de Ciencias de la Educación (ICE)

ISBN: 978-84-606-8636-1

Revisión y maquetación: Neus Pellín Buades

Publicación: Julio 2015

Necesidad de re-educación estadística en profesores universitarios: errores de interpretación valores p y tamaño del efecto

L. Badenes-Ribera*; D. Frias-Navarro; M. Pascual-Soler, H. Monterde-i-Bort; & O. Molina-Palomero

**Metodologia de les Ciències del Comportament. Universitat de València (España)*

*** ESIC Business & Marketing School (España)*

Universitat

RESUMEN

La estadística es una materia difícil de enseñar y aprender y hay amplia evidencia que señala que su aplicación a menudo es deficiente.

Las interpretaciones incorrectas de los valores *p* de probabilidad vinculados a una prueba de significación estadística afectan a las decisiones de los profesionales, ponen en peligro la calidad de las intervenciones y la acumulación de conocimiento científico válido. Nuestro estudio analiza los errores de razonamiento estadístico que los profesores universitarios realizan ante los resultados que aporta una prueba de inferencia estadística. La muestra está compuesta por 230 profesores del área de las Ciencias de la Educación. El 55.2% son hombres (n=127) y el 44.8% son mujeres (n=103). La media de años como profesor universitario es 12.22 (DT = 9.22). Nuestros hallazgos sugieren que muchos profesores no saben interpretar correctamente los valores de *p* de probabilidad. Los profesores adscritos al área de MIDE también cometen errores de interpretación. Estos resultados resaltan la importancia de la re-educación estadística de los profesores y la necesidad de una formación continuada. Finalmente se ofrecen recomendaciones para la mejora de la enseñanza y práctica estadística.

Palabras clave: falacias, valor p, prueba de significación, cognición estadística

1. INTRODUCCIÓN

1.1 Problema/cuestión

El valor *p* vinculado a una prueba de inferencia estadística es la probabilidad del resultado observado o un valor más extremo si la hipótesis nula es cierta (Fisher, 1925; Hubbard y 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 y, Pascual-Soler, 2015; Balluerka, Gomez y Hidalgo, 2005; Goodman, 1993, 2008; Wagenmakers, 2007). Estas interpretaciones incorrectas afectan a las decisiones de los profesionales, ponen en peligro la calidad de las intervenciones y la acumulación de conocimiento científico válido.

Nuestro estudio de encuesta analiza los errores de razonamiento estadístico que los profesores universitarios realizan ante los resultados que aporta una prueba de inferencia estadística. Este trabajo se enmarca dentro de la línea de investigación sobre cognición y educación estadística que nuestro equipo de investigación viene desarrollando desde hace años en el Departamento de Metodología de las Ciencias del Comportamiento de la Universidad de Valencia (Spain) (REME).

1.2 Revisión de la literatura.

Las cuatro interpretaciones erróneas del valor *p* más comunes son: (1) "Falacia de la probabilidad inversa"; (2) "Falacia de replicación"; (3) "Falacia del tamaño del efecto" y (4) Falacia de la significación clínica o práctica.

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 ($Pr(H_0|Datos)$). Sin embargo, 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; Sharver, 1993).

La "falacia de la replicación" es la falsa creencia de que el valor *p* indica el grado de replicabilidad del resultado. Es decir, una replicación del estudio tiene una probabilidad 1-*p* de obtener un resultado estadísticamente significativo (Carver, 1978). Bajo esta falsa creencia, dado un valor de $p<.05$, un investigador podría inferir que la probabilidad de replicación es >.95 (Kline, 2013).

La "falacia del tamaño" vincula la significación estadística con el tamaño del efecto y supone creer que valores pequeños de *p* son indicadores de efectos grandes

(Gliner, Vaske, y Morgan, 2001, Kline, 2013). Sin embargo, el tamaño del efecto sólo puede ser conocido estimando directamente su valor con el estadístico adecuado y su intervalo de confianza (Cumming, 2012; Gliner et al., 2001; Grissom y Kim, 2012).

La "falacia de la significación clínica o práctica" vincula el valor *p* con la importancia práctica o clínica del hallazgo (Gliner, Leech y Morgan, 2002; Kirk, 1996).

Estudios previos

Oakes (1986) realizó un estudio con profesores universitarios de Psicología a quienes les presentó una situación de investigación donde los resultados de la prueba t de Student tenían un valor de *p*=.01. Los participantes debían señalar como verdadera o falsa un conjunto de seis afirmaciones que realmente eran todas falsas (falacia de la probabilidad inversa, falacia de la replicación, etc). Sus resultados señalan que el 97% de los profesores percibieron como verdaderas al menos una de las seis opciones falsas del significado del valor *p*.

Haller y Krauss (2002) replicaron el estudio de Oakes (1986) encontrando que el 80% de profesores de Metodología, el 89.7% de profesores de Psicología que no enseñaban metodología cometieron algún tipo de error de interpretación del valor p.

Los estudios de Badenes-Ribera, et al., (2015) y Montender-i-Bort, Frias-Navarro y Pascual-Llobel (2010) encontraron que parte de los profesores universitarios interpretaban de forma incorrecta el valor *p* vinculándolo al concepto de tamaño del efecto y a la importancia del hallazgo.

1.3 Propópsito.

El propósito de nuestra investigación es detectar los errores de razonamiento estadístico que los profesores universitarios de Ciencias de la Educación realizan ante los resultados que aporta una prueba de inferencia estadística. En concreto, nuestro estudio se centra en el análisis de los errores de interpretación del valor *p* vinculados con la falacia de la probabilidad inversa, la falacia del tamaño del efecto, la falacia de la significación clínica o práctica así como la interpretación correcta del valor *p*.

2. METODOLOGIA

2.1. Descripción del contexto y de los participantes

La muestra está compuesta por 230 profesores del área de las Ciencias de la Educación. El 55.2% son hombres (n=127) y el 44.8% son mujeres (n=103). La media de años como profesor es 12.22 (DT = 9.22). El 23,9% de los profesores universitarios (n=55) pertenecen al área de Didactica y Organización Escolar, el 20.9% al área de

Metodos de Investigación y Diagnóstico en Educación (n=48), el 16.1% al área de Didáctica de la Expresión (n=37), el 8.3% al área de Didáctica de la Educación Física (n=19), el 8.3% al área de Teoría de la Educación (n=19), el 6.5% al área de Didáctica de la Matemática (n=15), el 6.5% al área de Didáctica de las Ciencias Experimentales (n=15), el 5.2% al área de Didáctica de las Ciencias Sociales (n=12), el 2.6% al área de Didáctica de la Lengua y la Literatura (n=6), y, el 1.7% al área de Didácticas Específicas (n=4).

2.2. Instrumentos

La encuesta recoge preguntas relacionadas con información sobre variables socio-demográficas: sexo, área de conocimiento, antigüedad como docente o investigador en la universidad (PDI). Además, incluye un conjunto de 10 preguntas que analizan las interpretaciones del valor *p* del procedimiento del contraste de hipótesis. Las preguntas se planteaban 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 0,01.

B.-Falacia de replicación:

6. Una replicación posterior tendría 0.999 de probabilidad (1-0.001) de ser significativa.

C.-Falacia del tamaño del efecto:

7. El valor $p<0,001$ confirma de forma directa que el tamaño del efecto ha sido grande.

D.- Falacia de la significación clínica o práctica

8. Obtener un resultado estadísticamente significativo implica de forma indirecta que el efecto detectado es importante.

E.-Interpretación correcta y decisión adoptada:

9. Se conoce la probabilidad del resultado de la prueba estadística, asumiendo que la hipótesis nula es cierta.

10. Dado que p=0.001 entonces el resultado obtenido permite concluir que las diferencias no se deben al azar.

Finalmente, el instrumento evalúa otras cuestiones como el conocimiento sobre la reforma estadística que no se analizan en este artículo.

2.3. Procedimento

Se registraron las direcciones de correo electrónico de docentes o investigadores universitarios del ámbito de la Ciencias de la Educación a partir de la consulta de fuentes de acceso público. Se obtuvo un marco muestral compuesto por 5,659 docentes o investigadores universitarios. La recogida de datos se realizó durante el curso académico de los años 2013 y 2014 mediante la utilización de un sistema CAWI (Computer Assisted Web Interviewing). Se recogieron un total de 230 cuestionarios válidos. La tasa de respuesta fue del 9.37%

3. RESULTADOS

Los resultados pueden calificarse por su baja tasa de respuesta. No obstante, es posible que los participantes que respondieron a la encuesta se sintieran más seguros de sus conocimientos estadísticos que los que no contestaron. Si esto es así, estos resultados pueden subestimar la extensión de las falacias sobre el valor *p* entre los profesores universitarios de Ciencias de la Educación de las universidades públicas de España.

A. -Falacia de la probabilidad inversa

La Tabla 1 muestra el porcentaje de respuestas de los participantes relacionadas con la falacia de la probabilidad inversa. Se observa que gran parte de los participantes perciben como verdadera alguna de las afirmaciones falsas sobre el valor *p*.

Los participantes de las áreas de Didácticas Específicas y de Métodos de Investigación y Diagnóstico en Educación (MIDE) poseen menos interpretaciones incorrectas del valor *p* comparado al resto de participantes.

Tabla 1. Falacia de la probabilidad inversa por área de conocimiento (%)

Ítem	1 n=56	2 n=52	3 n=55	4 n=48	5 n=19	Total n=230
1. Se ha probado que la hipótesis nula es verdadera	39.3	28.8	41.8	16.7	21.1	31.3
2. Se ha probado que la hipótesis nula	50	30.8	43.6	47.9	52.6	43.9

es falsa						
3. Se ha determinado la probabilidad de la hipótesis nula (p<0.001)	46.4	67.3	65.5	47.9	57.9	57
4. Se ha deducido la probabilidad de la hipótesis experimental (p<0.001)	53.6	36.5	45.5	43.8	21.1	43
5. La probabilidad de que la hipótesis nula sea verdadera, dados los datos obtenidos, es de 0.01	30.4	28.8	36.4	25	31.6	30.4
% Participantes que han valorado correctamente las 5 afirmaciones como falsas	5.36	13.46	3.64	12.5	5.26	8.26

Nota. 1= Didáctica de la Expresión (corporal, plástica, musical y educación física); 2= Didácticas Específicas (matemática, ciencias experimentales, ciencias sociales, lengua y literatura), 3= Didáctica y Organización Escolar, 4= MIDE, 5= Teoría e Historia de la Educación.

Las afirmaciones falsas que mayor respaldo han recibido son "se ha probado que la hipótesis nula es falsa" con una ratio de 30.8% para los participantes del área de Didácticas Específicas a 52.6% del área de Teoría e Historia de la Educación y, "se ha determinado la probabilidad de la hipótesis nula (p<0.001)" con una ratio del 46.4% para los participantes del área Didáctica de la Expresión hasta el 67.3% para los participantes del área de Didácticas Específicas. El rango de porcentaje de quienes han valorado correctamente las 5 afirmaciones como falsas (sería una respuesta correcta) oscila entre 5.26% para los participantes del área de Teoría e Historia de la Educación y el 13.46% para los participantes del área de Didáctica Específicas

B.-Falacia de la replicación

La Tabla 2 muestra las respuestas de los participantes por área de conocimiento que respaldan las afirmaciones falsas del valor *p* como grado de replicabilidad del resultado. Se observa que la mayoría de los participantes de las distintas áreas de conocimiento evalúan correctamente las afirmaciones con ratios de respuesta correcta de 52.6% en el área de Teoría e Historia de la Educación hasta el 67.9% en el área de Didáctica de la Expresión.

Tabla 2. Falacia de replicación por área de conocimiento (%)

Ítem	1 n=56	2 n=52	3 n=55	4 n=48	5 n=19	Total n=230
6. Una replicación posterior tendría 0.999 de probabilidad (1-0.001) de ser significativa	32.1	42.3	47.3	35.4	47.4	40
% Participantes que han valorado la afirmación como falsa	67.9	57.7	52.7	64.6	52.6	60

Nota. 1= Didáctica de la Expresión (corporal, plástica, musical y educación física); 2= Didácticas Específicas (matemática, ciencias experimentales, ciencias sociales, lengua y literatura), 3= Didáctica y Organización Escolar, 4= MIDE, 5= Teoría e Historia de la Educación.

C.-Falacia del tamaño del efecto y de la significación clínica/pratica

La Tabla 3 muestra el porcentaje de respuestas de los participantes por área de conocimiento que apoyan las afirmaciones falsas del valor de *p* como tamaño del efecto y significación clínica/práctica del resultado.

Tabla 3. Falacia del tamaño del efecto y de la significación clínica/práctica por área de conocimiento (%)

Ítem	1 n=56	2 n=52	3 n=55	4 n=48	5 n=19	Total n=230
7. El valor $p<0.001$ confirma de forma directa que el tamaño del efecto ha sido grande	37.5	42.3	45.5	22.9	47.4	38.3
8. Obtener un resultado estadísticamente significativo implica de forma indirecta que el efecto detectado es importante	71.4	53.8	60	50	63.2	59.6
% Participantes que valoran correctamente las dos afirmaciones	19.64	28.85	21.82	47.92	26.32	28.70

Nota. 1= Didáctica de la Expresión (corporal, plástica, musical y educación física); 2= Didácticas Específicas (matemática, ciencias experimentales, ciencias sociales, lengua y literatura), 3= Didáctica y Organización Escolar, 4= MIDE, 5= Teoría e Historia de la Educación.

La afirmación falsa que mayor respaldo ha recibido es "obtener un valor resultado estadísticamente significativo implica de forma indirecta que el efecto detectado es importante" con ratios que oscilan entre 50% en el área de MIDE y 71.4% en el área de Didáctica de la Expresión. En consecuencia los participantes presentan mayores problemas en discernir entre la significación estadística de los resultados obtenidos y la significación práctica de los mismos.

El rango de porcentaje de quienes han valorado correctamente las 2 afirmaciones como falsas (sería una respuesta correcta) oscila entre 19.64 en el área de Didáctica de la Expresión y el 47.92% en el área de MIDE.

D.-Interpretación correcta del valor p y decisión adoptada

La Tabla 4 muestra el porcentaje de participantes por area de conocimiento que apoyan cada una de las afirmaciones correctas del valor *p*. Se observa que la mayoría de los participantes en las distintas áreas de conocimiento tienen problemas con la interpretación probabilística del valor *p*.

Tabla 4. Interpretación correcta del valor *p* por área de conocimiento (%)

Ítem	1 n=56	2 n=52	3 n=55	4 n=48	5 n=19	Total n=230
9. Se conoce la probabilidad del resultado de la prueba estadística, asumiendo que la hipótesis nula es cierta	46.4	36.5	43.6	39.6	10.5	39.1
10. Dado que p=0.001 entonces el resultado obtenido permite concluir que las diferencias no se deben al azar	67.9	65.4	65.5	75	63.2	67.8
% Participantes que valoran correctamente las 2 afirmaciones como verdaderas	33.93	28.85	29.09	33.33	5.26	29.13

Nota. 1= Didáctica de la Expresión (corporal, plástica, musical y educación física); 2= Didácticas Específicas (matemática, ciencias experimentales, ciencias sociales, lengua y literatura), 3= Didáctica y Organización Escolar, 4= MIDE, 5= Teoría e Historia de la Educación.

Sin embargo, la interpretació mejora cuando se ejecuta en términos de conclusión estadistica.

El rango de porcentaje de quienes han valorado las 2 afirmaciones como verdaderas (sería una respuesta correcta) oscila entre 5.26% en el área Teoria e Historia de la Educacion y el 33.93% en el área de Didáctica de la Expresión.

4. CONCLUSIONES

Los resultados indican que la comprensión de los resultados de las pruebas de inferencia estadística continúa siendo problemática entre profesores universitarios de Ciencias de la Educación. Estos resultados son consistentes con investigaciones previas con muestras de profesores universitarios españoles (Badenes-Ribera, et al., 2015; Monterde-i-Bort, et al., 2010), en profesores universitarios de otros países (Haller y Krauss, 2002; Oakes, 1986), en muestras de miembros de la American Educational Research Association (AERA) (Mittag, y Thompson, 2000) y en profesionales de la Estadística (Lecoutre, Poitevineau, y Lecoutre, 2003).

Además, los resultados de esta investigación identifican el tipo de falacias que los PDI de Ciencias de la Educación realizan en torno a la interpretación del valor *p* de probabilidad que acompaña a una prueba de inferencia estadística

La "*falacia de la significación clínica o práctica*" es la que con mayor frecuencia se observa. Los participantes confunden la significación estadística de los resultados obtenidos con la significación práctica de los mismos. 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. Sin embargo, rechazar la hipótesis nula no indica la importancia de los hallazgos. No hay que confundir la significación estadística de los resultados con la significación clínica o práctica (Thompson, 1996).

Sin embargo, el valor *p* no ofrece información ni de la magnitud del efecto ni de la importancia del resultado (Gliner et al., 2002, Grant, 1962; Rosenthal, 1993; Shaver, 1993). El tamaño del efecto solo puede ser conocido mediante su estimación a través de un estadístico adecuado junto con su intervalo de confianza y, por otro lado, 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. Por tanto, valores *p* muy pequeños sólo 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 (Frías-Navarro, 2011; Gliner, et al., 2001).

La "falacia del tamaño del efecto" 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). De hecho, en el estudio de Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, y Pascual-Soler (2013) donde se analizaron los artículos publicados entre 2007 y 2012 en dos revistas españolas indexadas en el JCR (Psicothema y Revista de Educación) señala que la mayoría de los artículos revisados no incluyen ninguna estimación del tamaño del efecto. Sin embargo, un análisis evolutivo de los artículos publicados, es decir, por años de publicación, se observa que desde 2010 hasta 2012 existe una tendencia creciente a informar sobre el tamaño del efecto junto a los valores *p* de probabilidad. Lo que puede significar que los autores y editores de las revistas cada vez son más conscientes de que el valor *p* no indica la magnitud del efecto y, por tanto, se hace necesario informar sobre algún estadístico del mismo. Sin embargo, no existe consciencia de la necesidad de informar sobre los intervalos de confianza, pues, sólo el 9,5% de los artículos que informaron sobre el tamaño del efecto acompañaron dicho estadístico con su intervalo de confianza.

Los resultados también indican que los profesores universitarios adscritos al área de Métodos de Investigación y Diagnóstico en Educación no son inmunes a las interpretaciones erróneas del valor *p* de probabilidad, lo que compromete la formación académica de estudiantes universitarios y facilita la transmisión de estas falsas creencias del valor *p* así como su perpetuación (Badenes-Ribera et al., 2015; Haller y Krauss, 2002; Kline, 2013).

Además, la labor que realizan los docentes queda reflejada en sus publicaciones que de manera directa afectan a la acumulación del conocimiento Por lo tanto, su visión e interpretación de los hallazgos es un filtro de calidad que no puede estar sometido a creencias o interpretaciones erróneas del procedimiento estadístico que representa la herramienta fundamental para obtener conocimiento científico.

En definitiva, la presencia de interpretaciones erróneas del valor *p* resaltan la necesidad de mejorar la formación o educación estadística de docentes e investigadores

universitarios y el contenido de los libros de texto de estadística para garantizar una formación de calidad a los futuros profesionales (Cumming, 2012; Gliner, et al., 2002; Kline, 2013). La enseñanza de la estadística no sólo debe consistir en cálculos de enseñanza, procedimientos y fórmulas, deberían centrarse mucho más en el pensamiento y la comprensión de los métodos estadísticos (Haller y Krauss, 2002).

La literatura que se ha desarrollado sobre la cognición estadística y su educación tiene abierta toda una línea de investigación sobre esta problemática (Beyth-Maron, Fidler & Cumming, 2008; Garfield, Ben-Zvi, Chance, Medina, Roseth, & Zieffler , 2008; Garfield, & Franklin, 2011; Garfield, Zieffler, Kaplan, Cobb, Chance, & Holcomb, 2011).

Limitaciones

Una de las limitaciones de la presente investigación es la baja tasa de respuesta lo que afecta a la representatividad de la muestra, y por tanto, a la generalización de los resultados entre profesores universitarios de las Ciencias de la educación. Sin embargo, como se señalaba anteriormente, es posible que los profesores que han participado en este estudio sean aquellos que se sienten más confiados y seguros de sus conocimientos estadísticos que los profesores que decidieron no participar. Por ello, los resultados ofrecidos pueden infraestimar la extensión de las falacias sobre el valor p entre los miembros de dicha población.

No obstante, los resultados del presente estudio van en la línea de los hallazgos de estudios previos (e.g., Badenes-Ribera et al., 2015; Haller y Krauss, 2002; Lecoutre, et al., 2003; Monterde-i-Bort, et al., 2010; Oakes, 1986). Todo ello nos viene a indicar la necesidad de formar adecuadamente a los profesionales de la Ciencias de la Educación para producir un conocimiento científico y válido y mejorar la práctica profesional.

Este trabajo forma parte de un Estudio subvencionado por el Ministerio de Economía y Competitividad, Plan Nacional de I+D+i (EDU2011-22862) y por el Programa VALi+d para Personal Investigador en Formación de carácter Pre-doctoral (ACIF/2013/167). Conselleria d´Educació, Cultura i Esport, Generalitat Valenciana (España).

5. REFERENCIAS BIBLIOGRÁFICAS

Badenes-Ribera, L., Frias-Navarro, D., Monterde-i-Bort, H., i Pascual-Soler, M. (in press). Interpretation of the p value. A national survey study in academic psychologists from Spain. *Psicothema, 27*.

Badenes-Ribera, L., Frias-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2013). *Informar e interpretar el tamaño del efecto en Psicología y Educación*. XIV Congreso Virtual de Psiquiatria.com. Interpsiquis, 2013: 1-28 Febrero.

Balluerka, N., Gómez, J., y Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*, 55-70.

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

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48:* 378-399.

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

Fidler, F. (2005). From statistical significance to effect estimation: statistical reform in psychology, medicine and ecology. PhD Thesis History and Philosophy of Science. Melbourne, Australia. Department of History and Philosophy of Science. University of Melbourne.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, UK. Oliver and Boyd.

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

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. Springer.

Garfield, J. y Franklin, C. (2011). Assessment of learning, for learning, and as learning in statistics education. In C. Batanero, G. Burrilll, C. Reading and A. Rossman (eds.), T*eaching statistics in school mathematics—Challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 133–145). New York: 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

Gliner, J. A., Vaske, J. J., y Morgan, G. A. (2001). Nul Hypothesis Significance Testing: Effect Size Matters. Human Dimensions of Wildlife, 6, 291-301.

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

Goodman, S. (1993). P-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology, 137*, 485-496.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Semin Hematol 45*, 135-140.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics on investigating theoretical models. *Psychological Reviews, 69*, 54-61.

Grissom, R. J. & Kim, J. J. (2012). *Effect sizes for research*. New York, USA: Routledge

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. Retrieved July 30, 20134, from http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf

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

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

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

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*, 213–218

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

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.

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.

Monterde-i-Bort, H., Frías-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.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chicester: John Wiley & Sons.

Rosenthal, R. (1993). Cumulating evidence. En G. keren y C. Lewis (eds.), *A handbook for data anylisis in the behavioral sciencies: Methodological issues* (p. 519-559). Hillsdale, NJ:Erlbaum.

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

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*, 26–30

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779-804