Psicothema
2000. Vol. 12, Supl. nº 2, pp. 236-240

ISSN 0214 - 9915 CODEN PSOTEG

Tamaño del efecto del tratamiento y significación estadística

María Dolores Frías Navarro, Juan Pascual Llobell y José Fernando García Pérez
Universidad de Valencia

En la actualidad la atención por la sensibilidad y validez de conclusión estadística del diseño de la investigación ha aumentado, especialmente en el tratamiento que reciben en las ediciones actuales de los manuales de diseños experimentales aunque quizás en el ámbito aplicado (donde la estimación del tamaño del efecto cobra su mayor importancia) no se ha desarrollado todo lo que sería de desear como lo demuestran los estudios de potencia de los trabajos publicados. El principal propósito de este trabajo es analizar la repercusión o impacto que tienen las indicaciones de los consejos editoriales sobre los trabajos de investigación publicados respecto al cálculo e interpretación conjunta de las medidas de magnitud del efecto junto con los valores de significación estadística.

Effect size and statistical significance. Currently, there is a growing interest in the study of the sensitive and validity of the statistical conclusions of experimental design. Although most of books on experimental design stress these issues, many students on applied psychology still do not take advantage of these advances, as can be deduced by low statistical power. The goal of this article is to examine the impact of the guidelines of the editorial Board of peer reviewed respect to the computation and interpretation of the measures of effect size as well as the values of statistical significance.

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 los fenómenos. Ya sea para comprobar teorías o para estimar efectos de un tratamiento, los investigadores tienen que realizar un proceso de comprobación de hipótesis traduciendo la hipótesis científica a hipótesis estadística. Por ello, la técnica estadística del contraste de hipótesis y el diseño de la investigación se han necesitado mutuamente durante décadas. Sin embargo, históricamente el contraste y comprobación de hipótesis estadísticas ha sido causa de confusión, crítica y controversia entre los investigadores (Bakan, 1966, 1967; Cohen, 1990, 1994; Falk y Greenbaum, 1995; Hagen, 1997; Thompson, 1988, 1996), provocando interpretaciones erróneas de los resultados (Carver, 1978; Craig, Eison y Metze, 1976; Manzano, 1997; Thompson, 1989) que nada han favorecido la acumulación científica del conocimiento. El problema ha continuado durante décadas, reavivándose en ocasiones, de tal modo que actualmente estamos viviendo un momento de polémica con posturas enfrentadas, en algunos casos de forma extrema, entre los defensores (ej. Abelson, 1997a, 1997b; Cortina y Dunlap, 1997; Fritz, 1995, 1996; Greenwald, Gonzalez, Harris y Guthrie, 1996; Hagen, 1997, Levin, 1993) y detractores (ej. Chow, 1988; Cohen, 1994; Cowles, 1989; Meehl, 1978; Morrison y Henkel, 1970; Murphy, 1997; Schimdt, 1996) de las pruebas de significación estadística como instrumento válido para el progreso científico.

Durante la década de los noventa la polémica sobre el uso e interpretación de las pruebas de significación estadística se ha reavivado de nuevo (Shea, 1996), produciéndose momentos de destacadas reflexiones teóricas que han originado debates en foros como la revista *Journal of Experimental Education* (volumen 61 de 1993) o la revista *American Psychologist* (volumen 49 de 1994) que en el número de Julio de 1998 retoma de nuevo el tema con un conjunto de trabajos que critican y valoran la defensa que Hagen (1997) realizó en esta misma publicación sobre las pruebas de la hipótesis nula. También se han publicado libros (Chow, 1996; Gigerenzer, Swijtink, Porter, Daston, Beatty y Krüger, 1989; Harlow, Mulaik y Steiger, 1997; Henkel, 1976; Morrison y Henkel, 1970) especializados en recopilar y analizar el proceso de decisión estadística, revisando los principios de las pruebas de comprobación de hipótesis estadísticas.

Las reuniones científicas también se hacen eco de la polémica en torno al uso e interpretación de las pruebas de significación estadística y dedican sesiones al debate de la controversia, planteándose incluso su posible abandono (Carver, 1978, 1993; Schmidt, 1996). Como McClure y Suen (1994) anotaron: «*A misguided reliance on statistical significance would pose a serious threat to the archives of scientific knowledge*» (pág. 89). Por ejemplo en las reuniones anuales de la *American Psychological Association* (A.P.A.) y de la *American Psychological Society* (A.P.S.) celebradas en 1996 se planteó la siguiente cuestión: «*should significance tests be banned*». En 1997 esta misma pregunta ha sido recogida por Hunter (1997). Y el mismo Jacob Cohen y Bruce Thompson

Correspondencia: María Dolores Frías Navarro
Facultat de Psicologia
Universitat de València
46010 Valencia (Spain)
E-mail: M.Dolores.Frias@uv.es

fueron invitados a participar ese mismo año en el Congreso que se realizó en Chicago promovido por la *American Psychological Association* con dos trabajos cuyos títulos fueron «*Much ado about nothing*» (Cohen, 1997) y «*If statistical significance tests are broken/misused, what practices should supplement or replace them*» (Thompson, 1997, 1999a). El interrogante planteado es uno de los temas de mayor actualidad, proponiéndose distintas alternativas de análisis (Valera y Sánchez, 1997). El debate y la popularidad del tema sigue vigente, casi podríamos decidir que con una aparición cíclica como algunos fenómenos importantes de la naturaleza.

Ciertamente las peculiaridades de los objetivos de la investigación psicológica han favorecido que hoy en día se plantee la necesidad de ir más allá de la significación estadística tradicional obtenida con las pruebas de contraste estadístico. Esta búsqueda de la utilización de otros recursos para la investigación científica está directamente motivada por la *significación practica* que el área de la psicología aplicada demanda con insistencia (Aiken, West, Sechrest y Reno, 1990; Kirk, 1996), donde la significación estadística usual no da respuestas satisfactorias a las preguntas relacionadas con la magnitud de los efectos detectados. Los investigadores, especialmente aquellos que están interesados en la aplicación de la ciencia para solucionar problemas prácticos, no desean conocer si el tratamiento tuvo *algún* efecto sino que desean conocer si el tratamiento tiene el efecto que ellos plantean (Fowler, 1985) o también puede suceder que el cambio estadísticamente significativo no indique el verdadero valor terapéutico (Howlin, 1997). De ahí que algunos autores como Schmidt (1996) sugieran que el contraste estadístico es innecesario, recomendando centrarse en la estimación del tamaño del efecto. El *tamaño del efecto* es un índice en una métrica común que indica la magnitud de una relación o efecto (Cohen, 1988), por ejemplo se puede expresar en términos de diferencias estandarizadas como la media del grupo experimental menos la media del grupo control dividido por la desviación estándar común (véase para el cálculo por ejemplo Kirk, 1996, Friedman, 1982 y Snyder y Lawson, 1993)

Muy brevemente, las pruebas de significación estadística facilitan al investigador un test o prueba que informa de la probabilidad de conseguir la diferencia obtenida, o mayor que la observada, si la hipótesis nula es cierta. La prueba estadística asume que la hipótesis nula es cierta en la población y calcula la probabilidad del resultado de la muestra. Si el valor de probabilidad o $p_{\text{CALCULADO}}$ es igual o menor que 0.05 se concluye que la probabilidad de que sea el azar o la variabilidad muestral la explicación del resultado obtenido es muy baja y por lo tanto se rechaza la hipótesis de nulidad de no diferencias entre las medias. El resultado es *estadísticamente significativo*. Y lo que es muy importante, por lo común, el procedimiento implica la comprobación de la hipótesis de que el tratamiento no tiene ningún efecto o que la correlación entre dos variables es igual a cero; hipótesis conocidas como «*nil hypothesis*» en términos de Cohen (1994), diferenciándolas de la categoría general de prueba de la hipótesis nula donde el investigador puede contrastar la hipótesis de que la diferencia entre dos tratamientos es igual a cualquier valor, incluyendo pero no limitándolo a cero.

Las asociaciones científicas y la misma política editorial de las revistas apuestan por detallar en los informes de investigación la estimación del tamaño del efecto junto con la significación estadística. Con estas recomendaciones se pretende que las pruebas de «*nil hypothesis*» permitan al investigador evaluar la probabilidad que tiene un efecto (o mayor que el encontrado en una muestra dada) de ser obtenido a partir de una población en la que no existe efecto, ($d = 0$), facilitando un instrumento que permita conocer la credibilidad de la evidencia producida por un estudio (Fritz, 1996).

Por ejemplo, en la cuarta edición del manual publicado en el verano de 1994 por la *American Psychological Association* (A.P.A.) se realizan ciertas recomendaciones sobre el estilo de los informes de investigación y se enfatiza que los valores *p* no son índices aceptables de la magnitud del efecto —dependen del tamaño de la muestra—, estimulando a los investigadores a proporcionar información sobre el tamaño del efecto junto con los valores de probabilidad aportados por las pruebas de significación estadística, promoviendo la interpretación sustantiva de los resultados obtenidos en la investigación y destacando la falta de conexión entre resultado improbable (resultado con un valor *p* pequeño) y resultado interesante o importante (véanse ejemplos en Shaver, 1985; Thomspon, 1993) o significación estadística y replicación del resultado (Cohen, 1994; Thompson, 1989, 1996, 1999b). Afortunadamente nada que ver con las recomendaciones que en 1962 realizaba Arthur Melton como editor del *Journal of Experimental Psychology* donde señalaba que los manuscritos que no rechazaran la hipótesis nula nunca serían publicados, los resultados estadísticamente significativos al nivel 0.05 apenas serían aceptados mientras que los estadísticamente significativos al 0.01 merecerían un lugar en la revista, añadiendo que los resultados negativos son sinónimos de «no rechazar la hipótesis nula» y los resultados positivos de «rechazarla».

También, cada vez más los consejos editoriales de las revistas recomiendan que los autores informen e interpreten medidas de la magnitud del efecto junto con los valores de probabilidad de significación estadística. Por ejemplo, han adoptado dicho criterio la revista *Memory and Cognition* (Loftus, 1993), la revista *Educational and Psychological Measurement* (Thompson, 1994), la revista *Measurement and Evaluation in Counseling and Development* (Hansen, 1995) y más recientemente el *Journal of Experimental Education* (Heldref Publicactions, 1997) y el *Journal of Applied Psychology* (Murphy, 1997).

Conviene tener en cuenta que el tamaño del efecto y el valor de *p*, se encuentran inversamente relacionados, de tal manera que cuanto mayor es el primero, menor es el segundo y a la inversa. En el caso de que se cumpla con los supuestos estadísticos, la prueba de la hipótesis nula permite conocer la probabilidad de obtener por azar un tamaño del efecto, medido con un estadístico, igual o mayor que el encontrado. De nuevo nos encontramos con un procedimiento estadístico basado en «*nil hypothesis*».

Significación estadística y tamaño del efecto

Pero conocidos los problemas y limitaciones de las pruebas significación estadística ¿ha cambiado el comportamiento del científico? ¿Continúan las pruebas de significación estadística de «*nil hypothesis*» dominando la interpretación de los datos cuantitativos? Realmente ¿qué impacto o repercusión tienen las indicaciones de los consejos editoriales sobre los trabajos de investigación publicados respecto al cálculo e interpretación de las medidas de la magnitud del efecto junto con los valores de probabilidad de significación estadística? Quizás nos encontremos de nuevo ante una situación semejante a la del cálculo de la potencia de la prueba estadística: todos conocemos su importancia pero pocos planifican su presencia. Así, pese a los esfuerzos, encabezados por Cohen (1962, 1969, 1990, 1994), para popularizar el estudio de la po-

tencia y el control del *error de Tipo II*, el trabajo de Sedlmeier y Gigerenzer (1989), y más recientemente el de Clark-Carter (1997), indica que los estudios de la potencia han tenido poca trascendencia en la conducta de los investigadores, no variando sus hábitos de investigación (con una potencia media de 0.50 y 0.59 respectivamente para detectar un tamaño del efecto medio).

• En primer lugar, los resultados de los estudios empíricos confirman el comportamiento tradicional del científico en el uso de la significación estadística, de manera que el procedimiento de contraste estadístico de la hipótesis nula de efecto cero como medio de análisis e interpretación de los fenómenos de la realidad sigue arraigado dentro del proceso del diseño de la investigación casi como único (Murphy y Myors, 1999; Vacha-Haase y Ness (1999).

• En segundo lugar, los estudios de Kirk (1996), Snyder y Thompson (1998), Thompson (1999c, 1999d) Thompson y Snyder (1997, 1998), Vacha-Haase y Nilsson (1998), y Vacha-Haase y Ness (1999) confirman la escasa repercusión que las recomendaciones de la *American Psychological Association* han tenido sobre los informes de investigación, destacando también el uso e interpretación inapropiado que aún realizan algunos investigadores de la prueba de significación estadística.

En el estudio de Vacha-Haase y Ness (1999), donde se revisaron 256 artículos publicados entre 1990 y 1997 en la revista *Professional Psychology: Research and Practice*, el 77% de los informes utilizaron pruebas de significación estadística y menos del 20% usaron correctamente el término significación estadística. El 81.9% de los autores de los artículos sí informaron siguiendo el estilo de la A.P.A., incluyendo los grados de libertad, el nivel de alfa y el valor de los estadísticos pero la mayoría de los artículos no mencionan el tamaño del efecto. Mención también escasa en *Exceptional Children* (Thompson, 1999d).

En ocasiones, y quizás forzada por la política editorial, sí se indica el tamaño del efecto junto con los resultados de significación estadística pero sin llegar a englobar la interpretación dentro del contexto de nivel alfa, tamaño de la muestra y tamaño del efecto, tal y como concluyen Vacha-Haase y Nilsson (1998) al revisar desde 1990 a 1996 la revista *Measurement and Evaluation in Counseling and Development*, publicada por la *Association for Assessment in Counseling* de la *American Counseling Association*, cuyos editores recomiendan desde 1988 que se analice la significación estadística junto con el tamaño de la muestra y el tamaño del efecto. Únicamente el 7.3% de los trabajos contextualizaron el resultado de la significación estadística con el del tamaño de la muestra, el 35.3% informó del tamaño del efecto y sólo una minoría menciona el alfa seleccionado (13.2%).

Reflexiones metodológicas

• Conviene tener claro desde el principio que el valor de la estimación del tamaño del efecto debe ser interpretado en el contexto de un estudio y área concreta de investigación ya que un pequeño tamaño del efecto puede ser de gran importancia práctica en un contexto concreto por ejemplo de intervención clínica.

• El investigador debe analizar posibles violaciones de la validez de conclusión estadística de la investigación, comprobando los supuestos estadísticos y conociendo el comportamiento de los estimadores ya que por ejemplo, los índices del tamaño del efecto están afectados por el tamaño de la muestra. Así el cómputo de eta cuadrado con muestras pequeñas tiende a sobrestimar los efectos, recomendándose otros índices como omega cuadrado (Young, 1993).

• La aleatorización (muestreo o asignación) es una de las piezas claves del procedimiento de la significación estadística de la hipótesis nula ya que sin ella dicho contraste estadístico es irrelevante dado que la hipótesis nula será falsa a priori.

• La interpretación de la significación estadística deja de tener sentido cuando el tamaño de la muestra es tan grande que cualquier diferencia detectada, por pequeña que sea, permitirá rechazar la hipótesis de nulidad de diferencias. Del mismo modo cuando se plantean hipótesis triviales desde el punto de vista teórico donde la hipótesis nula es razonablemente falsa de tal modo que rechazarla es cuestión de potencia estadística, realizar el contraste estadístico también resulta absurdo.

• Facilitar la comprensión de la relación entre potencia, tamaño del efecto, nivel de alfa y significación estadística favorecerá interpretaciones correctas y contextualizadas de los datos y el diseño de la investigación. Únicamente la planificación cuidadosa del diseño de investigación validará los resultados obtenidos.

• Cuando las hipótesis intentan determinar la probabilidad de diferencias de grupos o efectos de intervención, hipótesis ordinales o cualitativas en términos de Fritz (1996), la aplicación de las pruebas de significación es correcta ya que no se especifica un tamaño del efecto concreto sino únicamente algún efecto. Estos resultados nos permitirán plantear hipótesis teóricas más elaboradas que planteen efectos de tratamiento concretos (hipótesis cuantitativas en términos de Fritz) donde dichas pruebas no tienen cabida ya que no fueron elaboradas con dicho fin. Por supuesto, cuando un área de conocimiento determinada ha alcanzado el consenso de que la hipótesis nula es falsa entonces las pruebas de significación estadística son totalmente innecesarias.

• Quizá, la explicación del uso intensivo que se hace en Psicología de la prueba de significación estadística de la hipótesis nula puede estar en la naturaleza ordinal de la mayor parte de las leyes y teorías de nuestra disciplina.

• Quizá, poder contrastar hipótesis nulas con efecto distinto a cero (hipótesis «non-nil nulls» en términos de Cohen) enriquecería nuestras teorías psicológicas, avanzado el conocimiento y eliminando ciertas polémicas sobre la trivialidad de testar hipótesis con efecto cero al mismo tiempo que evitaríamos la interpretación de resultados estadísticamente significativos sin importancia práctica. Los trabajos de Serlin y Lapsley (1985, 1993) acerca de «good-enough hypothesis» y Rouanet (1996) con métodos bayesianos profundizan en esta perspectiva.

• Recientemente Murphy y Myors (1999) ofrecen un método sencillo para el cálculo de hipótesis de efectos mínimos, que implica elaborar las tablas de la distribución no central *F* cuya construcción está determinada por los grados de libertad de la hipótesis (L1), los grados de libertad del error (L2) y el parámetro de no centralidad λ (cuando se contrastan nil hypothesis λ es igual a cero) que puede estimarse con:

$$\frac{\nu_2 \times PV}{1 - PV}$$

donde PV es el porcentaje de varianza en la variable dependiente que está explicada por la variable o variables independientes del diseño. Cuanto mayor el valor de λ (y por lo tanto mayor PV) mayor será el valor empírico de *F* que se necesitara para rechazar la hipótesis nula. La definición de los efectos mínimos daría sentido a la formulación sustantiva de las hipótesis cuyos efectos dependerán del área psicológica concreta en la que se formulen. Otra

ventaja que los autores añaden (Murphy y Myors, 1998) es que evitaría que un resultado no estadísticamente significativo simplemente lo fuera al aumentar el tamaño de la muestra ya que con el método de los efectos mínimos si los efectos reales del tratamiento son triviales, la probabilidad de rechazar la hipótesis de un efecto mínimo no se incrementa a medida que el tamaño de la muestra aumenta sino que decrece.

• En conclusión, la responsabilidad de la construcción teórica de los enunciados psicológicos no corresponde al método de investigación seleccionado, y por extensión a las técnicas matemáticas de cálculo sino que los criterios deben ser de orden teórico puesto que en resumidas cuentas, la inferencia estadística únicamente proporciona, si se hace correctamente, la precisión, o incertidumbre, de un enunciado científico. En definitiva, las pruebas de significación de la hipótesis nula serán adecuadas cuando se ajusten a los objetivos teóricos planteados por el investigador pero querer ir más allá o no ajustarse a sus supuestos implícitos es querer obtener algo que ella no nos puede dar.

Referencias

Abelson, R. P. (1997a). 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?* Mahwah, NJ: Lawrence Erlbaum Associates.

Abelson,R. P. (1997b). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15.

Aiken, L. S., West, S. G., Sechrest, L. y Reno, R. R (1990). Graduate training in statistics, methodology and measurement in psychology. *American Psychologist*, *45*, 721-734.

American Psychological Association (A.P.A.)(1994) *Publications manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Bakan, D. (1966). The effect of significance testing in psychological research. *Psychological Bulletin*, *66*, 423-437.

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation*. San Francisco: Jossey-Bass.

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

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287-292.

Chow, S. L.(1988). Significance test or effect size? *Psychological Bulletin*, *103*, 15-110.

Chow, S. L.(1996). *Statistical significance. Rationale, validity and utility*. London, UK: Sage Publications.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the *British Journal of Psychology*. *British Journal of Psychology*, *88*, 71-83.

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

Cohen, J. (1969): Statistical power analysis for the behavioral sciences. New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

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

Cohen, J. (1997, August). *Much ado about nothing*. Lecture presented at the annual meeting of the American Psychological Association, Chicago.

Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

Cowles, M.(1989). *Statistics in psychology: An historical perspective*. Hillsdale, NY: Lawrence Erlbaum Associates.

Craig, J. R.,Eison,C.L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and omega-squared. *Bulletin of the Psychonomic Society*, *7*, 280-282.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory and Psychology*, *2*, 75-98.

Fowler, R. L.(1985). Testing for substantive significance in applied research by specifying nonzero null hypotheses. *Journal of Applied Psychology*, *70*, 215-218.

Friedman, H. (1982). Simplified determinations of statistical power, magnitude of efect and research sample size. *Educational and Psychological Measurement*, *42*, 521-526.

Fritz, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*, 132-138.

Fritz,R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379-390.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect size and *p*-values: What should be reported and what should be replicated? *Psychophysiology*, *33*, 175-183.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15-24.

Hansen, J. C.(1995). Revised APA style manual recommended to authors. *Measurement and Evaluation in Counseling and Development*, *28*, 67-68.

Harlow, L.L.,Mulaik, S. A., & Steiger, J. H.(Eds.) (1997). *What if there were no significances tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Heldref Publications (1997). Guidelines for contributors. *Journal of Experimental Education*, *65*, 95-96.

Henkel, R. E. (1976). *Tests of significance*. London, UK: Sage Publications. Quantitative Applications in the Social Sciences series, Vol. 4.

Howlin, P. (1997). When is a significant change not significant?. *Journal of Autism and Developmental Disorders*, *27*, 347-348.

Hunter, J. E.(1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3-7.

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

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition*, *21*, 1-3.

Manzano, V. (1997). Usos y abusos del error de Tipo I. *Psicológica. Revista de Metodología*. *18*, 153-169.

McClure, J., & Suen, H. K.(1994). Interpretation of statistical significance testing: A matter of perspective. *Topics in Early Childhood Special Education*., *14*, 88-100.

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

Morrison, D. E., & Henkel, R. E. (Eds.) (1970). *The significance test controversy: a reader*. Chicago: Aldine.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology*, *82*, 3-5.

Murphy, K. R. & Myors, B. (1998). Statistical power analysis: A simple and general model for traditional and modern hypothesis tests. Mahwah, NJ: Lawrence Erlbaum Associates.

Murphy, K. R. & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimun-effect tests in the general linear model. *Journal of Applied Psychology*, *84*, 234-2484.

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin*, *119*, 149-158

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.

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

Serlin,R.C., & Lapsley, D. K.(1985). Rationality in psychological research: the good-enough principle. *American Psychologist*, *40*, 73-83.

Serlin, R.C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good enough principle. En G. Keren y C. Lewis, (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues*. Hillsdale, NJ: Lawrence Erlbaum.

Shaver, J. (1985). Chance and nonsense. *Phi Delta Happan*, *67*, 57-60.

Shea, C. (1996). Psychologists debate accuracy of «significance test». *Chronicle of Higher Education*, *42*, A12, A16.

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

Snyder, P. A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practice and suggested alternatives. *School Psychology Quarterly*, *13*, 335-348.

Thompson, B. & Snyder, P. A.,(1998). Statistical significance testing y reliability analyses in recent JCD research articles. *Journal of Counseling and Development*, *76*, 436-441.

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

Thompson, B. (1988). A note about significance testing. *Measurement and Evaluation in Counseling and Development*, *20*, 146-148.

Thompson, B. (1989). Asking «what if» questions about significance tests. *Measurement and Evaluation in Counseling and Development*, *22*, 66-68.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361-377.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

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

Thompson, B. (1997, August). *If statistical significance tests are broken/misused, what practices should supplement or replace them?*. Paper presented at the annual meeting of the American Psychological Association, Chicago.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory and Psychology*, *9*, 165-181.

Thompson, B. (1999b). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory and psychology*, *9*, 191-196.

Thompson, B. (1999c). Why «encouraging» effect size reporting is not working: The etiology of researcher resistance to changing practices. *Journal of Psychology*, *133*, 133-140.

Thompson, B. (1999d). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children*, *65*, 329-337.

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

Vacha-Haase, T., & Nilsson, J. E.(1998). Statistical significance reporting: Current trends and usages within MECD. *Measurement and Evaluation in Counseling and Development*, *31*, 46-57.

Valera, A. & Sánchez, J. (1997). Pruebas de significación y magnitud del efecto: reflexiones y propuestas. *Anales de Psicología*, *13*, 85-90.

Young, M. A. (1993). Supplementing tests of statistical significance: Variation accounted for. *Journal of Speech and Hearing Research*, *36*, 644-656.