

Interpretation of the p value: A national survey study in academic psychologists from Spain

Laura Badenes-Ribera¹, Dolores Frías-Navarro¹, Héctor Monterde-i-Bort¹ and Marcos Pascual-Soler²

¹ Universitat de València and ² ESIC Business & Marketing School (Valencia)

Abstract

Background: Incorrect interpretations of p values affect professionals' decisions and jeopardize the quality of psychological interventions and the accumulation of valid scientific knowledge. This study analyzes the errors made by Spanish academic psychologists in interpreting p values. **Method:** We surveyed academic psychologists about their interpretations of p values. The sample is composed of 418 academic psychologists from Spanish public universities. The mean number of years working as a university professor is 14.16 (SD = 9.39). **Results:** Our findings suggest that many academic psychologists do not know how to correctly interpret p values. The inverse probability fallacy presents the greatest comprehension problems. Methodology instructors also interpret the significance of the p value erroneously. **Conclusions:** These results highlight the importance of the statistical re-education of professors.

Keywords: Interpretation, p value, significance testing, statistical inference.

Resumen

Interpretación del valor p : una encuesta nacional entre profesores españoles de Psicología. **Antecedentes:** las interpretaciones incorrectas de los valores p afectan a las decisiones de los profesionales y ponen en peligro la calidad de las intervenciones psicológicas y la acumulación de conocimiento científico válido. Este estudio analiza los errores de interpretación del valor p y su interpretación correcta entre el profesorado de las universidades de Psicología de España. **Método:** se encuestó a profesores universitarios sobre sus interpretaciones de los valores p . La muestra está compuesta por 418 profesores de Psicología de las universidades públicas españolas. La media de años como profesor universitario es 14,16 (DT = 9,39). **Resultados:** nuestros hallazgos sugieren que muchos profesores universitarios no saben interpretar correctamente los valores de p . La falacia de la probabilidad inversa presenta los mayores problemas de comprensión. Los profesores de Metodología también cometen errores de interpretación del valor p . **Conclusiones:** estos resultados resaltan la importancia de la re-educación estadística de los profesores.

Palabras clave: interpretación, valor p , prueba de significación, inferencia estadística.

Knowing how to interpret p values of probability is a core competence of the professional in Psychology and any discipline where statistical inference is applied.

The p value linked to the results of a statistical test is the probability of the observed result or a more extreme value if the null hypothesis is true (Fisher, 1925; Gill, 1999; Hubbard & Lindsay, 2008; Johnson, 1999). The definition is clear and precise, but researchers' incorrect interpretations of this p value continue to be numerous and repetitive (Cumming, 2012; Verdam, Oort, & Sprangers, 2014; Wagenmakers, 2007).

The most common incorrect interpretations of the p value are the "inverse probability" fallacy, the "replication" fallacy, the "effect size" fallacy, and the "clinical or practical significance" fallacy (Carver, 1978; Cohen, 1994; Falk & Greenbaum, 1995; Fidler, 2005; Gill, 1999; Goodman, 2008; Johnson, 1999; Kirk,

1996; Kline, 2004; Levine, Weber, Hullett, Sun, & Massi, 2008; 2013; Nickerson, 2000; Thompson, 1996).

The "inverse probability" fallacy is the false belief that the p value indicates the probability that the null hypothesis (H_0) is true, given certain data ($\Pr(H_0|Data)$).

The "inverse probability fallacy" is related to confusing the probability of the result, assuming that the null hypothesis is true, with the probability of the null hypothesis, given certain data. The null hypothesis statistical testing (NHST) procedure does not provide information about the conditional probability of the null hypothesis based on the data obtained in the study (Kirk, 1996; Sharver, 1993).

The "replication fallacy" states that the p value is the degree of replicability of the result. That is, a replication of the study has a $1-p$ probability of obtaining a statistically significant result (Carver, 1978). Based on this false belief, given a value of $p < .05$, a researcher could erroneously infer that the probability of replication is $> .95$ (Fidler, 2005; Kline, 2013).

The "effect size" fallacy represents one of the strongest criticisms of the NHST procedure. This fallacy involves the belief that the p value provides direct information about the effect size (Gliner, Vaske, & Morgan, 2001). However, the effect size can only

Received: December 16, 2014 • Accepted: April 28, 2015

Corresponding author: Dolores Frías-Navarro

Facultad de Psicología

Universitat de València

46010 Valencia (Spain)

e-mail: M.Dolores.Frias@uv.es

be determined by directly estimating its value with the appropriate statistic and its confidence interval (Cumming, 2012; Cumming, Fidler, Kalinowski, & Lai, 2012; Kline, 2004, 2013). This fallacy could underlie deficiencies in scientific reports published in high-impact journals when reporting effect size statistics (Fidler, 2005; Kirk, 2001).

The “clinical or practical significance” fallacy links statistical significance to the importance of the effect size. However, a statistically significant result does not indicate that the result is important, in the same way that a non-statistically significant result might still be important (Kirk, 1996). The importance of the finding is a subjective decision made by the expert; it is not a question of statistics or effect size. Therefore, the clinically important value of the effect size depends on the construct being studied and certain contextual variables (Frías-Navarro, 2011).

Previous studies found that 97% of academic psychologists (Oakes, 1986), 80% of methodology instructors, and 89.7% of scientists who do not teach methodology (Haller & Kraus 2002) made some type of misinterpretation of the *p* value.

Vacha-Haase and Ness (1999) analyzed the articles published in the journal *Professional Psychology: Research and Practice* between 1990 and 1997. They found that 77% of the reports used statistical significance tests, and less than 20% of the articles correctly used the term statistical significance.

Finch, Cumming and Thomason (2001) reviewed the articles published in the *Journal of Applied Psychology* during the past sixty years. They found that in 38% of them, the statistically non-significant results were interpreted to mean that the null hypothesis was considered true.

Moreover, Hoekstra, Finch, Kiers, and Johnson (2006), in a sample of articles published in journals indexed in the *Journal Citation Report*, found that more than half of the articles incorrectly interpreted the lack of statistical significance of the results as proof of the absence of an effect, and 20% of them interpreted the presence of statistical significance as proof of an effect.

These incorrect interpretations are made by the researcher, based on certain beliefs and attributions about the significance of the results.

The purpose of the present study is to detect the statistical reasoning errors that academic psychologists make when presented with the results of a statistical inference test. Thus, their vision and interpretation of the findings is a quality filter that should not be susceptible to erroneous beliefs and interpretations of the statistical procedure that is the basic tool for obtaining scientific knowledge.

Consequently, this article analyzes two questions. The first is the extension of the fallacies among academic psychologists, and the second is the extent to which *p* values are correctly interpreted. To address these questions, we surveyed academic psychologists.

Method

Participants

The sample is composed of 418 academic psychologists. The mean number of years of the professors in the University is 14.16 years (SD = 9.39). Men represent 48.6% and women 51.4%. Regarding university departments, 23.4% of the university

professors (n = 98) belong to the area of Personality, Evaluation and Psychological Treatments, 16% to the area of Behavioral Sciences Methodology (n = 67), 13.4% to the area of Basic Psychology (n = 56), 17.7% to the area of Social Psychology (n = 74), 6.9% to the area of Psychobiology (n = 29) and 22.5% to the area of Developmental and Educational Psychology (n = 94).

Instruments

A structured questionnaire was elaborated. First, the questionnaire includes items related to information about socio-demographic variables: sex, knowledge area, and years of experience as an academic psychologist.

Second, the instrument includes a set of 10 questions that analyze the interpretations of the *p* value (Table 1).

Finally, the instrument evaluates other questions, such as statistical practice or knowledge about the statistical reform, which are not analyzed in this paper.

Procedure

The e-mail addresses of academic psychologist were recorded after consulting publicly accessed sources, obtaining a sample framework consisting of 4,066 academics. The data collection was carried out during the 2013 and 2014 school years, by means of a CAWI (Computer Assisted Web Interviewing) system.

This study is framed within the line of research on statistical thinking and education that our research team has been developing for several years.

Data analysis

The data analyses were performed using the statistical program IBM SPSS v. 20 for Windows.

<i>Table 1</i> Questionnaire on interpretation of the <i>p</i> -value
<i>Let's suppose that a research article indicates a value of $p = 0.001$ in the results section ($\alpha = 0.05$). Mark which of the following statements are true (T) or false (F).</i>
Inverse probability fallacy
1. The null hypothesis has been shown to be true
2. The null hypothesis has been shown to be false
3. The probability of the null hypothesis has been determined ($p = 0.001$)
4. The probability of the experimental hypothesis has been deduced ($p = 0.001$)
5. The probability that the null hypothesis is true, given the data obtained, is 0.01
Replication fallacy
6. A later replication would have a probability of 0.999 (1-0.001) of being significant
Effect size fallacy
7. The value $p = 0.001$ directly confirms that the effect size was large
Clinical or practical significance fallacy
8. Obtaining a statistically significant result indirectly implies that the effect detected is important
Correct interpretation and decision made
9. The probability of the result of the statistical test is known, assuming that the null hypothesis is true
10. Given that $p = 0.001$, the result obtained makes it possible to conclude that the differences are not due to chance

Results

Of the 4,066 academic psychologists who were sent an e-mail with the link to access the survey, 418 filled it out (10.26%). Therefore, the results must be qualified by the low response rate. However, it is possible that the participants who responded to the survey felt more confident about their statistical knowledge than those who did not respond. In this case, these results could underestimate the extension of the fallacies about the *p* value among Spanish academic psychologists in public universities.

Inverse probability fallacy

Table 2 shows the percentage of responses by participants who endorsed the false statements about the *p* value related to the inverse probability fallacy, according to the Psychology knowledge areas. The table shows that the majority of the academic psychologists perceived some of the false statements about the *p* value to be true.

The participants in the area of Methodology made fewer incorrect interpretations of the *p* value than the rest of the participants.

The false statements that received the most support were “the null hypothesis has been shown to be false” and “the probability of

the null hypothesis has been determined ($p = .001$)”. The percentage of those who rated the 5 statements correctly ranged from 0% for the participants from the area of Psychobiology to 19.4% for the participants from the area of Methodology.

Replication fallacy

Table 3 shows the participants’ responses that endorsed the false statements of the *p* value as a degree of replicability of the result, by knowledge area. It shows that the majority of the participants correctly evaluated the false statements.

Effect size fallacy and clinical significance fallacy

Table 4 shows the percentage of participant responses that endorsed the false statements about the *p* value as an effect size and as having clinical or practical significance, by knowledge area.

The false statement that received the most support was the one related to the clinical or practical significance of the findings.

The percentage of participants who rated both statements correctly ranged from 48.9% in the area of Developmental and Educational Psychology to 76.1% in the area of Methodology.

Table 2
Fallacy of the inverse probability by Psychology knowledge area (%)

Ítem	1 n = 98	2 n = 67	3 n = 56	4 n = 74	5 n = 29	6 n = 94	Total n = 418
1. The null hypothesis has been shown to be true	8.2	1.5	7.1	5.4	6.9	12.8	7.4
2. The null hypothesis has been shown to be false	65.3	35.8	60.7	66.2	55.2	61.7	58.6
3. The probability of the null hypothesis has been determined ($p = 0.001$)	51	58.2	67.9	62.2	62.1	56.4	58.4
4. The probability of the experimental hypothesis has been deduced ($p = 0.001$)	40.8	13.4	23.2	36.5	37.9	43.6	33.7
5. The probability that the null hypothesis is true, given the data obtained, is 0.01	32.7	19.4	25	31.1	41.4	36.2	30.6
% Participants who correctly rate the 5 statements as false	4.1	19.4	5.4	2.7	0	4.3	6.2

Note: 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

Table 3
Replication fallacy by knowledge area (%)

Ítem	1 n = 98	2 n = 67	3 n = 56	4 n = 74	5 n = 29	6 n = 94	Total n = 418
6. A later replication would have a probability of 0.999 (1-0.001) of being significant	34.7	16.4	35.7	39.2	27.6	45.7	34.7
% Participants who correctly rate the item as false	65.3	83.6	64.3	60.8	72.4	54.3	65.3

Note: 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

Table 4
Effect size fallacy and clinical/practical significance fallacy by knowledge area (%)

Ítem	1 n = 98	2 n = 67	3 n = 56	4 n = 74	5 n = 29	6 n = 94	Total n = 418
7. The value $p = 0.001$ directly confirms that the effect size was large	12.2	3	8.9	16.2	24.1	18.1	13.2
8. Obtaining a statistically significant result indirectly implies that the effect detected is important	39.8	22.4	28.6	35.1	27.6	45.7	35.2
% Participants who correctly rate the 2 statements as false	55.1	76.1	67.9	59.5	58.6	48.9	59.8

Note: 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

Correct interpretation of the *p* value and the decision made

Figure 1 shows the proportion of the different groups of participants endorsing each of the two statements. The majority of the participants in the different knowledge areas had problems with the probabilistic interpretation of the *p* value.

The interpretation improved when performed in terms of the statistical conclusion, compared to the probabilistic interpretation of the *p* value, except in the academic psychologists from the area of Methodology, where this improvement was not observed. In this case, the professors presented greater problems with the statistical interpretation of the *p* value than with the probabilistic interpretation.

Discussion

The results indicate that the comprehension and correct application of many statistical concepts continue to be problematic among Spanish academic psychologists.

The “inverse probability fallacy” is the most frequently observed misinterpretation. This means that participants confuse the probability of obtaining a result or a more extreme result if the null hypothesis is true ($\Pr(\text{Data}|\text{H}_0)$) with the probability that the null hypothesis is true given some data ($\Pr(\text{H}_0|\text{Data})$). As Kirk (1996) points out, statistical inference tests do not respond to what researchers want to know. When researchers perform a statistical inference test, they want to find out the probability that the Null Hypothesis (H_0) is true, given certain data (D), that is, $\Pr(\text{H}_0|\text{Data})$. However, the statistical inference tests indicate the probability of obtaining a result or a more extreme result if the null hypothesis is true (p. 747). Not rejecting the null hypothesis does not imply the truth of the null hypothesis. For this reason, it should never be stated that the null hypothesis is “accepted” when the *p* value is less than alpha; the null hypothesis is rejected or not rejected (Palmer & Sesé, 2013).

Furthermore, the *p* value does not provide information about the degree of replicability of the result, or about the effect or

importance of the result, in spite of researchers’ erroneous interpretations (Grant, 1962; Sharver, 1993). The presentation of a lot of asterisks along with the *p* value of probability or very small *p* values only highlights that, in this design, the null hypothesis is not very plausible, but from this, it cannot be inferred that the effect found is important (Gliner et al., 2001). Thus, to distinguish between the importance or practical significance of the findings and their statistical significance, the term “statistically significant” should be used to describe the results linked to a value of $p < \alpha$ (Cumming, 2012; Frías-Navarro, Pascual-Llobel, & García-Pérez, 2000; Gliner et al., 2001; Kline, 2013; Monterde-i-Bort, Frías-Navarro, & Pascual-Llobel, 2010; Thompson, 1996).

The results also indicate that academic psychologists from the area of Methodology are not immune to erroneous interpretations. However, they show fewer problems than their colleagues from other areas. These data are consistent with previous studies (Haller & Kraus 2002; Lecoutre, Poitevineau, & Lecoutre, 2003).

The differences between the psychologists from the area of Methodology and those from the rest of the areas in the correct valuation of the *p* value can be due to the fact that the probabilistic interpretation requires thinking about the significance of the *p* value as a conditional probability and a random variable, whereas the statistical interpretation is only based on the valuation of the *p* value compared to the alpha value. The results of the statistical programs include the *p* value and only require the researcher to routinely apply the $p < \alpha$ rule. In contrast, the probabilistic interpretation involves statistical reasoning; that is, it means reflecting on the statistical processes involved in the behavior of the *p* value when the null hypothesis is not rejected (Ben-Zvi & Garfield, 2004).

Problems in understanding the *p* value influence the conclusions that professionals draw from their data (Hoekstra, Morey, Rouder, & Wagenmakers, 2014), jeopardizing the quality of the results of psychological research (Frías-Navarro, 2011). The value of the evidence depends on the quality of the statistical analyses and their interpretation (Faulkner, Fidler, & Cumming, 2008). To avoid these erroneous interpretations, Wilkinson and the Task Force on Statistical Inference (1999) and the American Psychological Association manuals (APA, 2001, 2010) recommend accompanying *p* values with estimators of the effect size and its confidence intervals (CIs). However, following or adhering to these recommendations continues to be an uncommon practice. Caperos and Pardo (2013) analyzed the studies published in 2011 in Spanish Psychology journals indexed in the JCR. Their results point out that only 24.3% of the *p* values reported were accompanied by an effect size statistic (similarly, see Badenes-Ribera, Frías-Navarro, Monterde-i-Bort, & Pascual-Soler, 2013).

Reporting the effect size and its CIs would enhance the body of scientific knowledge and lead to a better understanding of the *p* value (Cumming, 2014, 2013; Cumming et al., 2012; Frías-Navarro et al., 2000; García-García, Ortega-Campos, & De la Fuente, 2011; Gliner et al., 2001; Monterde-i-Bort et al., 2010; Palmer & Sesé, 2013; Pascual-Llobel, Frías-Navarro, & Monterde-i-Bort, 2004; Thompson, 2002; Vacha-Haase, 2001).

Regarding the CIs, studies reveal that they have advantages over the NHST procedures (Cumming, 2012; Cumming & Finch, 2005; Fidler, 2005; Gardner & Altman, 1986). Empirical studies have provided evidence that CIs avoid some of the incorrect interpretations of *p* values (Fidler & Loftus, 2009; Hoekstra, Johnson, & Hal, 2012). However, they are not immune to incorrect

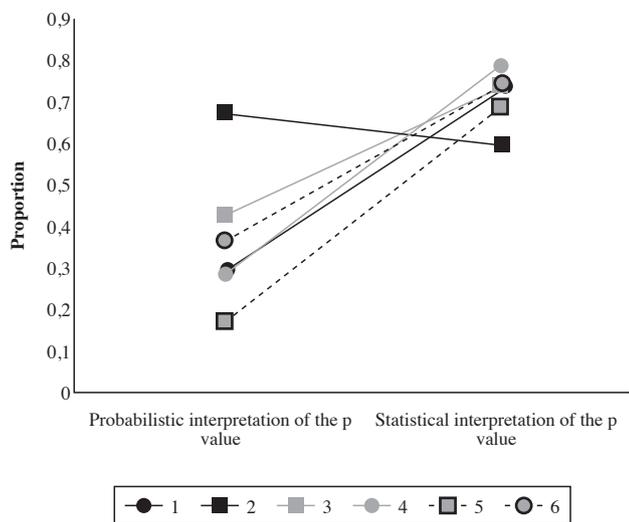


Figure 1. Proportion of correct interpretation and statistical decision adopted by knowledge area

Note: 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology

interpretations either (Belia, Fidler, Williams, & Cumming, 2005; Hoekstra et al., 2014).

Finally, the fact that academic psychologists from the area of Methodology erroneously interpret the p value hinders the students' statistical training and facilitates the transmission of these false beliefs, as well as their perpetuation (Haller & Kraus, 2002; Kirk, 2001; Kline, 2013). It is, therefore, necessary to improve the statistical education or training of academic psychologists and the content of statistics textbooks in order to guarantee high quality training of future professionals (Cumming, 2012; Gliner, Leech, & Morgan, 2002; Kline, 2013; Haller & Kraus, 2002).

Thus, as Kirk (2001) points out, to promote good statistical practices it is necessary to use a multi-faceted approach (as in Vacha-Haase, 2001), an approach that involves textbook authors, professors who teach in degree and postgraduate programs, authors of statistical software packages, journal editors and publication manuals.

The literature developed on statistical thinking and instruction has a whole line of research open on this issue (Beyth-Marón, Fidler, & Cumming, 2008; Garfield, Ben-Zvi, Chance, Medina,

Roseth, & Zieffler, 2008; Garfield, Zieffler, Kaplan, Cobb, Chance, & Holcomb, 2011).

The low response rate (10.26%) affects the representativity of the sample and, therefore, the generalization of the results. Furthermore, it should be kept in mind that this study is descriptive.

Nonetheless, the results of the present study agree with the findings of previous studies (e.g., Gordon, 2001; Haller & Kraus, 2002; Lecoutre et al., 2003; Mittag & Thompson, 2000; Oakes, 1986) in indicating the need to adequately train Psychology professionals in order to produce valid scientific knowledge and improve professional practice.

Acknowledgments

We thank the participants who responded to our questions. Study financed by the Spanish Ministry of Economy and Competitiveness (I+D+i) (Spain) (EDU2011-22862) and the VALi+d program for pre-doctoral Training of Research Personnel (ACIF/2013/167). Conselleria d'Educació, Cultura i Esport, Generalitat Valenciana (Spain).

References

- American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th Ed.). Washington, DC: American Psychological Association.
- American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2013). Informar e interpretar el tamaño del efecto en Psicología y Educación [Reporting and interpreting the effect size in Psychology and Education]. XIV Congreso Virtual de Psiquiatría.com. Interpsiquis, 2013: 1-28 February.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389-396.
- Ben-Zvi, D., & Garfield, J. (Eds.). *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17-45). Dordrecht: Kluwer Academic Publishers.
- Beyth-Marón, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal, 7*, 20-39.
- Caperos, J.M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*, 408-414.
- Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.
- Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist, 49*, 997-1003.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- Cumming, G. (2013). The new statistics: A how to guide. *Australian Psychologist, 48*, 161-170.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29.
- Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64*, 138-146.
- Cumming, G., & Finch, S. (2005). Inference by eye. Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.
- Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*, 75-98.
- Faulkner, C., Fidler, F., & Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behavior Research and Therapy, 46*, 270-281.
- 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.
- Fidler, F., & Loftus, G.R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology, 217*, 27-37.
- Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181-210.
- Frías-Navarro, D. (2011). *Técnica estadística y diseño de investigación [Statistical technique and research design]*. Valencia (Spain): Palmero Ediciones.
- Frías-Navarro, D., Pascual-Llobel, J., & García-Pérez, F. (2000). Tamaño del efecto del tratamiento y significación estadística [Effect size of the treatment and statistical significance]. *Psicothema, 12*, 236-240.
- Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.
- García-García, J., Ortega-Campos, E., & De la Fuente-Sánchez, L. (2011). The use of the effect size in JCR Spanish journals of psychology: From theory to fact. *The Spanish Journal of Psychology, 14*, 1050-1055.
- Gardner, M.J., & Altman, D.G. (1986). Confidence intervals rather than p-values: Estimation rather than hypothesis testing. *British Medical Journal, 292*, 746-750.
- Garfield, J.B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). *Developing students' statistical reasoning. Connecting research and teaching practice*. New York, NY: Springer Publishers
- Garfield, J., Zieffler, A., Kaplan, D., Cobb, G., Chance, B., & Holcomb, J.P. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician, 65*, 1-10.
- Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly, 52*, 647-674.
- Gliner, J.A., Vaske, J.J., & Morgan, G.A. (2001). Null hypothesis significance testing: Effect size matters. *Human Dimensions of Wildlife, 6*, 291-301.
- 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.

- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology, 45*, 135-140.
- Gordon, H.R.D. (2001). American vocational education research association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Educational Research, 26*, 1-18.
- Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics on investigating theoretical models. *Psychological Reviews, 69*, 54-61.
- 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. Retrieved July 30, 2014, from <http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf>.
- Hoekstra, R., Finch, S., Kiers, H.A.L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review, 13*, 1033-1037.
- Hoekstra, R., Johnson, A., & Hal, K. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement, 72*, 1039-1052.
- Hoekstra, R., Morey, R.D., Rouder, J.N., & Wagenmakers, E. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*, 1157-1164.
- 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.
- 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. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
- Kline, R.B. (2013). *Beyond significance testing: Statistic reform in the behavioral sciences*. Washington, DC: American Psychological Association.
- 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.
- Levine, T.R., Weber, R., Hullett, C., Sun, H. & Massi, L.L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34*, 171-187.
- 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.
- Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chicester: John Wiley & Sons.
- Palmer, A., & Sesé, A. (2013). Recommendations for the use of statistics in clinical and health psychology. *Clinica y Salud, 24*, 47-54.
- Pascual-Llobel, J., Frías-Navarro, D., & Monterde-i-Bort, H. (2004). Tratamientos psicológicos con apoyo empírico y práctica clínica basada en la evidencia [Psychological treatments with empirical support and evidence-based clinical practice]. *Papeles del Psicólogo, 87*, 1-8.
- 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.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect size. *Educational Researcher, 31*, 25-32.
- Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61*, 219-224.
- 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.
- 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.
- Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779-804.
- Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist, 54*, 594-604.