Psicothema

Psicothema 2016, Vol. 28, No. 4, 448-456
doi: 10.7334/psicothema2016.24

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2016 Psicothema
www.psicothema.com

Knowledge level of effect size statistics, confidence intervals and meta-analysis in Spanish academic psychologists

Laura Badenes-Ribera[1], Dolores Frias-Navarro[1], Marcos Pascual-Soler[2] and Héctor Monterde-i-Bort[1]
[1] University of Valencia and [2] ESIC Business & Marketing School (Valencia)

Abstract

Background: The statistical reform movement and the American Psychological Association (APA) defend the use of estimators of the effect size and its confidence intervals, as well as the interpretation of the clinical significance of the findings. **Method:** A survey was conducted in which academic psychologists were asked about their behavior in designing and carrying out their studies. The sample was composed of 472 participants (45.8% men). The mean number of years as a university professor was 13.56 years (SD= 9.27). **Results:** The use of effect-size estimators is becoming generalized, as well as the consideration of meta-analytic studies. However, several inadequate practices still persist. A traditional model of methodological behavior based on statistical significance tests is maintained, based on the predominance of Cohen's d and the unadjusted R^2/η^2, which are not immune to outliers or departure from normality and the violations of statistical assumptions, and the under-reporting of confidence intervals of effect-size statistics. **Conclusion:** The paper concludes with recommendations for improving statistical practice.

Keywords: Statistical reform, effect size, confidence interval, meta-analysis, descriptive survey study.

Resumen

Grado de conocimiento sobre los tamaños del efecto, intervalos de confianza y meta-análisis en psicólogos académicos españoles. **Antecedentes:** el movimiento de la reforma estadística y la Asociación Americana de Psicología (APA) defienden el uso de estimadores del tamaño del efecto y sus intervalos de confianza, así como la interpretación de la significación clínica de los hallazgos. **Método:** se realizó una encuesta a psicólogos académicos sobre su conducta en el diseño y realización de estudios. La muestra estuvo compuesta de 472 participantes (45,8% hombres). La media en años como académico fue 13,56 (DT= 9,27). **Resultados:** el uso de estadísticos del tamaño del efecto se está generalizando, también la consideración de los estudios meta-analíticos. Sin embargo, persisten prácticas estadísticas inadecuadas. Se mantiene un modelo tradicional de comportamiento metodológico basado en las pruebas de significación estadística, predominio de la d de Cohen, y del R^2/η^2 no ajustado que no son inmunes a la existencia de outliers y violaciones de las asunciones y un escaso uso de los intervalos de confianza de los estadísticos del tamaño del efecto. **Conclusiones:** se concluye con recomendaciones para la mejora de la práctica estadística.

Palabras clave: reforma estadística, tamaño del efecto, intervalo de confianza, meta-análisis, estudio descriptivo de encuesta.

The "statistical reform" movement recommends an important change in researchers' behavior: to change their perspective from "how probable or improbable the sample result is" (application of the traditional statistical significance tests and dichotomous statistical decisions based on the comparison of the *p*-value and the alpha-value) to new analytic strategies that estimate the effect size (ES) and its confidence intervals (CIs) and favor the replication of the findings, as well as their practical/clinical significance (Frias-Navarro, 2011; Kline, 2013; Wilkinson & the Task Force on Statistical Inference, 1999). This way of proceeding facilitates the development of "meta-analytic" thinking among researchers (Cumming, 2014; Henson, 2006; Peng, Chen, Chiang, & Chiang, 2013; Thompson, 2002), redirecting the design, analysis and interpretation of the results toward the ES value and, in addition, contextualizing its size within a specific area of investigation. This method emphasizes the importance of the ES obtained compared to the previously estimated effect within a specific research context, and it avoids routine interpretations of the small, medium or large ES based on Cohen's (1988) three classic values of 0.2, 0.5 and 0.8, respectively.

These recommendations were incorporated into the revised fifth edition of the Publication Manual of the American Psychological Association (APA, 2001), and they were again included in the sixth edition (APA, 2010), in an attempt to minimize researchers' excessive confidence in statistical significance tests and the dichotomous decisions supported by the *p* values of probability.

Nevertheless, several studies have pointed out that these recommendations were hardly being followed by academic psychologists. For example, García, Ortega, and De la Fuente (2008) carried out a review of articles published in the year 2007 in all the Psychology journals included in the Redalyc database, dedicated to publications in the Ibero-American setting. They found that almost a decade after the report by the APA and the

Received: January 21, 2016 • Accepted: July 5, 2016
Corresponding author: Dolores Frias-Navarro
Facultad de Psicología
Universidad de Valencia
46010 Valencia (Spain)
e-mail: M.Dolores.Frias@uv.es

statistical inference group (APA & TFSI, 1999) and the fifth APA Manual (2001), where the estimation of the ES and its CI had already been explicitly recommended, these recommendations were hardly being followed. Only 12.26% of the articles studied (n= 452) estimated the ES. The authors did not provide information about the estimation of the CIs.

Recently, Caperos, and Pardo (2013) analyzed the articles published in 2011 in four multidisciplinary Spanish Psychology journals indexed in the JCR database. Again, they found that the recommendations of APA (2010) were hardly being followed. Only 24.3% of the statistical inference tests were accompanied by an ES statistic (along the same lines, Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, & Pascual-Soler, 2013). Moreover, Badenes-Ribera, et al. (2013) found that only 9.5% of the ES statistics were accompanied by their CIs (along the same lines, Frias-Navarro, Monterde-i-Bort, Pascual-Soler, Pascual-Llobell, & Badenes-Ribera, 2012).

Finally, Sesé and Palmer (2012) analyzed the use of statistics in the articles published in 2010 in eight Clinical and Health Psychology journals with impact indexes in the Journal Citation Reports (JCR) database (e.g., International Journal of Clinical and Health Psychology from Spain). Their findings showed that 52.78% of the studies published reported the ES, but, only 18.87% of them reported CIs.

On the other hand, there are dozens of ES measures available (Henson, 2006). Overall, they can be classified into two broad groups: measures of mean differences and measures of strength of relations (Henson, 2006; Kline, 2013; Rosnow & Rosenthal, 2009). The former is based on the standardized group mean difference (e.g., Cohen's *d*, Glass' *g*, Hedges' *g*, Cohen's *f*); the latter is based on the proportion of variance accounted for or correlation between two variables (e.g., R^2/r^2, η^2, w^2).

Prior studies have pointed out that the most frequently reported ES measures are the unadjusted R^2, Cohen's *d*, and η^2 (Badenes-Ribera et al., 2013; Lakens, 2013; Peng et al., 2013; Sesé & Palmer, 2012; Sun, Pan, & Wang, 2010). These statistics have been criticized for bias (i.e., they tend to be positively biased), lack of robustness to outliers, and instability under violations of statistical assumptions (Fritz, Morris, & Richler, 2012; Grissom & Kim, 2012; Kline, 2013; Thompson, 2002; Wang & Thompson, 2007).

Within this context of change and methodological advances, systematic, meta-analytic reviews of studies have gained considerable relevance and prevalence in the most prestigious journals (Borenstein, Hedges, Higgins, & Rothstein, 2009). Meta-analytic studies offer several advantages over narrative reviews: meta-analysis involves a scientifically-based research process that depends on the rigor and transparency of each of the decisions made during its elaboration, and it can provide a definitive answer about the nature of an effect when there are contradictory results (Botella & Sánchez-Meca, 2015; Ellis, 2010). Meta-analyses facilitate more precise ES estimations, they make it possible to rate the stability of the effects, and they help researchers to contextualize the ES values obtained in their study (Cumming, Fidler, Kalinowski, & Lai, 2012). Moreover, the results of a meta-analytic study help to plan the size of the sample by providing the value of the estimated ES in a specific research context.

The main purpose of our study is to analyze what academic psychologists know about ESs, their CIs, and meta-analyses, given that this is one of the main recommendations proposed by the APA (2010) to improve statistical practice and favor the accumulation of knowledge and the replication of findings. For this purpose, the participants were asked about their statistical knowledge and statistical analyses performed.

Method

Participants

A non-probabilistic (convenience) sample consisted of 472 academic psychologists. The professors' mean number of years teaching in the University is 13.56 years (SD= 9.27). Men represent 45.8% (n= 216) and women 54.2% (n= 256).

Regarding university departments, 23.9% of the university professors (n= 113) belong to the area of Personality, Evaluation and Psychological Treatments, 14.8% to the area of Behavioral Sciences Methodology (n= 70), 16.1% to the area of Basic Psychology (n= 76), 16.7% to the area of Social Psychology (n= 77), 6.8% to the area of Psychobiology (n= 32) and 22% to the area of Developmental and Educational Psychology (n= 104). Regarding kind of university, 87.9% belong to public university (12.1% private university). Finally, 64.9% of the participants have been a reviewer of scientific journals in the last year.

Instrument

The survey consisted of two sections. The first one included items related to information about sex and years of experience as an academic psychologist, Psychology knowledge areas, kind of university (public/private).

The second section included the items related to statistical knowledge and statistical practice of the researcher. They are the following:

1. Knowledge and use of statistical terms, evaluated with 4 questions.
   A. "*What terms from the following list do you know sufficiently: standard deviation, sedimentation graph, forest plot, ANOVA, funnel plot, correlation, meta-analysis, regression analysis, ES*". On this item, more than one response can be chosen.
   B. "*Can you give the name of an ES statistic?*".
   C. "*If your answer is Yes, please specify its name*" (open-ended question).
   D. "*In your reports, what type of statistics do you use more often*?" Likert-type response scale with 5 response ratings that range from 0=*not at all*, to 4=*used often*.
2. Opinions about meta-analysis, evaluated with 1 question.
   A. "*What type of review do you think has the most credibility and objectivity?*" (select only one response):
      a) The narrative review carried out by experts (such as those performed in the "Annual Review").
      b) The quantitative review or meta-analysis.
      c) The qualitative review.
3. Use of meta-analytic study, evaluated with 1 question: "*Have you read or used a meta-analytic study?*"
   a) I have never read or used one.
   c) Yes: I have read or used 1 -2 meta-analytic studies.
   d) Yes, I have read or used more than 2 meta-analytic studies.

4. Researcher's behavior, evaluated with 11 questions related to research design (e.g., estimate a priori sample size, strategies used for it, so on), reporting on p-value, and interpretation of *p*-value (see Table 5).

Procedure

The e-mail addresses of academic psychologists were found by consulting the webs of the Spanish universities, resulting in 4,463 potential participants. The data collection was performed during the 2013-2014 school year. Potential participants were invited to complete a survey through the use of a CAWI (Computer Assisted Web Interviewing) system. A follow-up message was sent two weeks later to non-respondents. The response rate was 10.58%.

This study is framed within the line of research on cognition and statistics education that our research group has been developing for many years.

Data analysis

Analysis included statistics descriptive of the variable evaluated. These analyses were performed with the statistical program IBM SPSS v. 20 for Windows.

To assess a possible social desirability, we analyzed the relationship between the number of years as academic and the use of ES statistic in report research and the use of confidence intervals in report research. Finally, we assessed whether there was a difference in the number of years as academic psychologists according to knowing or not knowing the name of ES statistics.

To ensure normality, Kline (2011) suggested the cut off of absolute values of 3.0 and 10 for skewness and kurtosis, respectively. The absolute values of skewness and kurtosis for the scores on the three variables: use of ES statistics, use of confidence intervals and the number years as academic were within the acceptable range of the normal distribution (univariate skewness ranged from -0.16 to -0.69, and univariate kurtosis ranged from -0.39 to -1.16). Therefore, no adjustments were made to the scores on the variables measured in our study. Pearson's correlation was calculated. Cohen (1988) established a conventional interpretation of ESs in which $r = 0.10$ is considered a small effect, $r = 0.30$ is a medium-sized effect, and $r = 0.50$ is a large effect. These guidelines were used throughout this article for interpreting results.

Results

Table 1 shows the participants' responses by psychology knowledge areas to the item that rates their knowledge about the statistical terms. It can be noted that more than 90% of the participants state they had adequately known about standard deviation, correlation, analysis of variance and regression analysis.

In addition, more of 80% of them adequately know the statistical terms of ES and meta-analysis. However, it is noteworthy that the graphics that usually accompany meta-analytic studies (forest plot and funnel plot) were rated as sufficiently known by a very low percentage of the participants, especially the forest plot graphic, where the mean ES and its CI are presented along with the ESs and CIs of the primary studies.

Regarding their knowledge about ES statistics, 72.3% of the participants (n= 341) stated they know the ES statistic. However, only 68.4% of them indicated the name of an ES statistic (n= 323). By Psychology knowledge areas, the percentage of them who stated knowing ES statistics were 78.8% in the area of Personality, Evaluation and Psychological Treatments, 97.1% in Methodology, 65.8% in Basic Psychology, 70.1% in of Social Psychology, 40.6% in Psychobiology and 64.4% in Developmental and Educational Psychology. Moreover, the percentage of them who gave the name of an ES statistic were 75.2% in the area of Personality, Evaluation and Psychological Treatments, 91.4% in Methodology, 61.8% in Basic Psychology, 64.9% in of Social Psychology, 37.5% in Psychobiology and 62.5% in Developmental and Educational Psychology. Consequently, there is greater knowledge about term of the ES than about ES statistics.

The statistics most familiar to the participants were those that evaluate the differences between the means of the groups analyzed (differences in standardized means), followed by the proportion of explained variance (η^2) and the correlation coefficients (Table 2).

There were no statistically significant differences between academics who could name an ES statistic (*M* = 13.87, *SD* = 9.21) and academics who could not (*M* = 12.73, *SD* = 9.21) for the variable number of years as academic psychologist (F(1, 470) = 1.817, *p* = .178, *d* = 0.13, 95% *CI* [-0.06, 0.33], small effect) .

Concerning the use ES statistics in research reports (Table 3), 40.7% of the participants stated that they use the ES *a lot* in

Table 1
Statistical terms the participants know sufficiently (%)

Statistical terms	1 (n= 113)	2 (n= 70)	3 (n= 76)	4 (n= 77)	5 (n= 32)	6 (n= 104)	Total (n= 472)
Standard deviation	100	100	100	98.7	93.8	98.1	98.9
Correlation	98.2	100	100	98.7	96.9	97.1	98.5
ANOVA	97.3	100	98.7	94.8	100	96.2	97.5
Regression analysis	96.5	98.6	93.4	96.1	90.6	90.4	94.5
Effect size	94.7	92.9	89.5	84.4	78.1	77.9	87.1
Meta-analysis	92.9	91.4	81.6	85.7	71.9	86.5	86.9
Sedimentation graphic	57.5	67.1	27.6	45.5	15.6	38.5	45.1
Forest plot	12.4	27.1	9.2	6.5	6.3	4.8	11
Funnel plot	6.2	22.9	2.6	6.5	6.3	1	7

Note: More than one response could be selected. 1= Personality, Evaluation and Psychological Treatments; 2= Behavioral Sciences Methodology; 3= Basic Psychology; 4= Social Psychology; 5= Psychobiology; 6= Developmental and Educational Psychology their studies, but only 24.4% of them estimated the CI around the ES. Approximately 36.8% of the participants said they use the ES *little* or *not at all* in their statistical reports. Furthermore, most of participants (57.8%) recognized that they use ESs and their CIs *very little* or *not at all* (*not utilized*, *scarcely utilized* and *somewhat utilized*). Finally, the correlation and analysis of variance (ANOVA) were the most widely used statistics by the participants (more than 50%) in their studies.

On another hand, the study of the relationship between the number of years as academic psychologist and use of ES statistics in research reports showed that there is no statistically significant link between the two variables (r =.08, p = .08, 95% CI [-.01, .17] small effect). Nevertheless, there was a small statistically significant relationship between the number of years as academic psychologist and the use of confidence intervals in research reports (r =.01, p = .04, 95% *CI* [.01, .18], small effect).

Table 2
Known effect size statistics (responses of 323 participants)

Effect size statistics	n	%
Cohen's d	228	70.6
η^2	142	44
Correlation coefficient (Pearson, Spearman, biserial, phi, Cramer's V)	80	24.8
Hedges' g	35	10.8
R^2	32	9.9
Omega/Omega^2	26	8.1
Odds Ratio	19	5.9
Cohen's f/Cohen's f^2)	9	2.8
Relative Risk	8	2.5
Glass' delta	6	1.9
Beta	3	0.9
Number Needed to Treat (NNT)	3	0.9
Wilk's Lambda	2	0.6
Epsilon/Epsilon^2	2	0.6
Cliff's delta	1	0.3
Common Language (CL)	1	0.3

Note: The majority of the participants reported knowing more than one effect size statistic

Table 3
Use of the statistics (%) (N= 472)

Statistics	Quite utilized	Fairly utilized	Somewhat utilized	Scarcely utilized	Not utilized
ANOVA	65.3	25.4	6.1	2.6	0.6
Correlation	55.7	26.1	12.9	4.7	0.6
T tests	44.7	29.2	17.6	7.9	0.6
Regression	44.5	27.5	18.4	8.3	1.3
Effect size	40.7	22.5	14.8	14.2	7.8
Confidence intervals	26.1	22	22.5	23	6.4
Exploratory factorial analysis	24.8	23.9	21.2	22.5	7.6
Effect Size and Confidence interval	24.4	17.8	18	15.7	24.1
MANOVA	21.6	22.5	20.7	28.2	7
Confirmatory factor analysis	19.9	22.2	18.1	28.4	11.4
Structural equations	13.3	18.1	15	31.8	21.8
Discriminant analysis	5.1	10.2	26	40.3	18.4

Table 4 shows that most of the participants (57.4%) pointed out that meta-analytic studies are the type of review with the most credibility and objectivity. Nevertheless, 42.6% said they grant more importance to narrative reviews carried out by experts and/or qualitative reviews. By Psychology knowledge areas, the percentage of participants who granted more importance to narrative reviews carried out by experts and/or qualitative reviews ranged from 24.3% in the area of Methodology to 59.4% in the area of Psychobiology.

In addition, the majority of the participants said they have used or read a meta-analytic study for their research. By knowledge areas, the percentage of participants who stated utilizing meta-analytic studies ranged from 75% in the area of Psychobiology to 92.2% in the area of Social Psychology.

Finally, we analyzed the profile of researchers based on whether or not they could indicate the name of an ES statistic. The results indicated that the behavior of academics who could name an ES statistic is closer to good statistical practices and research design (see Table 5).

In this way, academics who could name an ES statistic, compared to participants who could not, had higher proportion of participants who had read or used meta-analysis studies, had been reviewers for scientific journals, had published an article in journals with impact factor JCR (Journal Citation Reports of WoS) and thought that meta-analysis studies are the type of review with the most credibility.

Regarding their behavior when they plan or prepare a study, academics who could name an ES statistic perform better methodological practices than the rest of the participants, as a larger proportion of them estimate a priori sample size (both groups have a high proportion), plan the number of participants, and use statistical criteria so the sample will represent the characteristics of the population. It is noteworthy that academics who named an ES statistic confuse to a lesser extent planning the statistical power a priori with a strategy to adjust the significance level or alpha value and, also make to a lesser extent the clinical or practical size fallacy where the statistical significance of the effect is related to its importance, although in both groups of academics, more than 30% of the subjects believe in that association. However, a statistically significant effect can be found, but will not have any clinical importance, and vice versa. The clinical or practical importance of the findings should be described by an expert in the field, and not placed in the statistics alone.

In addition, they follow the APA recommendations to avoid expressions of p-value as p < alpha or p > alpha and use its exact value to a higher degree than the rest of the participants.

Finally, a high proportion of both groups of academics said that they do not know any checklist to assess the design quality of a study (91.9% of academics who could not name a statistical ES and 78% of academics who could) and that they did not know that there is currently some kind of open debate on statistical issues or research design (79.9% the group of academics who could not name a statistical ES and 53.9% in the group of academics who could).

Discussion

The low response rate could affect the representativeness of the sample and, therefore, the generalizability of the results. Furthermore, it should be kept in mind that this study is descriptive. However, it is possible that the participants who responded to the interview felt more confident about their statistical knowledge than those who did not respond. In that case, the results might overestimate the extension of the impact of the statistical reform in Spanish academic psychologists.

Taking into account these limitations, the results are novel because, until now, there were no self-report data about the following of the statistical reform and the APA Manual recommendations among Spanish researchers, even though these recommendations must be followed in almost all of the psychological journals.

The results of our study indicate that the emphasis the statistical reform places on the use of the ES and its CI has also had an impact on participants, especially the estimation of the ES. The majority of the interviewees state that they use ES statistics (63.2%) a fair amount or a lot. Moreover, 42.2% of them also say that they use ESs and their CIs a fair amount or a lot. Therefore, our results point to an increase in the use of the ES and CIs compared to previous studies (Badenes-Ribera et al., 2013; Caperos & Pardo, 2013; Frias-Navarro et al., 2012; García, Ortega, & De la Fuente, 2008; Sesé & Palmer, 2012). For example, Caperos and Pardo (2013) found that only 24.3% of the statistical inference tests were accompanied by an ES statistic. It could be a sign of the change in the analytic behavior of the researcher.

However, CIs were reported not nearly as frequently as ES point estimate (along the same lines, Badenes-Ribera et al., 2013; Fritz et al., 2012; Peng et al., 2013; Sesé & Palmer, 2012). This result goes against the APA recommendation, such as, "*Whenever possible, provide a confidence interval for each ES reported to indicate the precision of estimation of the ES"* (APA, 2010, p. 34). This will probably improve in future studies, given that the change in statistical practices takes time.

Nevertheless, the change in statistical practice is slow, if we take into account that the recommendations about using the ES and its CI appeared in the 1999 report by the statistical inference workgroup of the American Psychological Association (Wilkinson & TFSI, 1999). The elaboration of this report was the APA's response to a broad set of criticisms of the null hypothesis statistical technique (NHST), proposing an improvement in statistical practices (Balluerka, Vergara, & Arnau, 2009; Nickerson, 2000; Monterde-i-Bort, Frias-Navarro, & Pascual-Llobell, 2010).

Regarding the type of ES statistic they know, the participants mention more frequently the ES statistics from the family of standardized differences in means and η_2 (parametric ES statistics). These findings are in line with previous research that analyzes the use of ES statistics in journals. For example, Peng et al. (2013) found that the most frequently reported ES measures were R^2 and Cohen's *d*. Nevertheless, standardized differences in means (e.g., Cohen's *d*, Glass' δ, Hedges' *g*,) and from the family of correlation (Pearson correlation, R^2, η^2, omega^2, and so on) have been criticized for lack of robustness against outliers or departure from normality, and instability under violations of statistical assumptions (Algina, Keselman, & Penfield, 2005; Grissom & Kim, 2012; Kline, 2013; Peng & Chen, 2014; Wang & Thompson, 2007).

There are theoretical reasons and empirical evidence that outliers and violations of assumptions are common in practice (Erceg-Hurn & Mirosevich, 2008; Grissom & Kim, 2001). Consequently, researchers should consider using ES statistics that are more resistant to outliers and violations of statistical assumptions (Erceg-Hurn & Mirosevich, 2008; Grisom & Kim, 2012; Keselman, Algina, Lix, Wilcox, & Deerin, 2008; Kline, 2013). Moreover, CIs are not immune to outliers or departure from normality and the violations of statistical assumptions.

There are alternatives for parametric ES statistics: on the one hand, non-parametric ES statistics, such as, the Spearman correlation, Cliff's δ, so on, and, on the other hand, the modern robust ES statistics , such as the robust standardized mean difference based on robust estimators (trimmed means and winsorized variances); the probability of superiority (PS), which is defined as the probability that a randomly sampled score from one population is larger than a randomly sampled score from a second population; the number needed to treat (NNT), an ES index appropriate for conveying information in psychotherapy outcome studies or other behavioral research that involves comparisons between treatments or treatment and control or placebo conditions (e.g., Arnau, Bendayan, Blanca, & Bono, 2013; Erceg-Hurn & Mirosevich, 2008; Grisom & Kim, 2012; Keselman et al., 2008; Wilcox & Keselman, 2003).

Our results point out that the modern robust statistical methods are not known by some participants, or at least, the participants did not give the name of robust ES statistics. In fact, only 0.9% of the

Table 4
Opinions about the review with the most credibility and objectivity and use of meta-analytic studies (%)

	1	**2**	**3**	**4**	**5**	**6**	**Total**
	(n= 113)	**(n= 70)**	**(n= 76)**	**(n= 77)**	**(n= 32)**	**(n= 104)**	**(n= 472)**
Opinions about the review with the most credibility and objectivity							
The quantitative review or meta-analysis	64.6	75.7	52.6	51.9	46.9	48.1	57.4
The narrative review carried out by experts	27.4	20	42.1	39	40.6	40.4	34.3
The qualitative review	8	4.3	5.3	9.1	12.5	11.5	8.3
Reading or use of meta-analytic studies							
I have never read or used one	9.8	12.9	13.2	7.8	25	23.1	14.4
I have read or used 1-2 meta-analytic studies	21.2	27.1	34.2	31.2	34.4	36.5	30.1
I have read or used more than 2 meta-analytic studies	69	60	52.6	61	40.6	40.4	55.5

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 5
Researcher profile according to knowing or no knowing the name of effect size statistics (%)

Item	**Not know (n = 149)**	**Know (n = 323)**
Researcher's behavior		
1. Have you read or used a meta-analytic study?		
-I have never read or used one	28.9	7.8
- Yes: I have read or used 1 -2 meta-analytic studies	**36.9**	26.9
- Yes, I have read or used more than 2 meta-analytic studies	34.2	**65.3**
2. Have you been reviewer for scientific journals in the last year?		
-No	**48.3**	29.1
-Yes: 1-2 reviewed articles	38.3	33.4
-Yes: more than 2 reviewed articles	13.4	**37.5**
3. Have you published an article in a journal indexed in the WoS with JCR impact factor in the last year?		
-No	**39.6**	21.7
-Yes: 1-2 published articles	**39.6**	**43**
-Yes: more than 2 published articles	20.8	35.3
4. Do you know checklist for assessing research design of a study?		
-No	**91.9**	**78**
-Sí	8.1	22
5. What type of review do you think has the most credibility and objectivity?		
- The narrative review carried out by experts	**40.9**	31.3
- The quantitative review or meta-analysis	**41.6**	**64.7**
- The qualitative review	17.4	4
6. In your opinion, what statistical questions or issues related to the study design are currently being debated?		
-I don't know	**79.9**	**53.9**
- I don't think there are any debates open	2	2.2
- There is some debate	18.1	44
Researcher's methodological behavior		
7. When you plan a study, do you estimate a priori the sample size you will need?		
-No	21.5	14.6
-Yes	**78.5**	**85.4**
8. What kind of strategy do you use when you want to plan the sample size of a study?		
- You try to achieve the greatest number of participants possible	33.1	25.17
- You use software or tables to estimate the sample size according to the statistical criteria	25.2	34.7
- You try to make the sample represent the characteristics of the population	33.8	37.41
- You do not use any strategy because it isn't part of your research interests.	7.9	2.7
9. In your opinion, what is the purpose of calculating the statistical power a priori?		
- To adjust the significance level or alpha value	47	33.3
- To explore the reliability of the scales	13.6	4.8
- To estimate the sample size	**39.4**	**61.9**
10. In your opinion, obtaining a statistically significant result implies indirectly that the detected effect is important		
-No	45.6	**69.7**
-Yes	**54.4**	30.3
11. When you perform a statistical test, do you consider it a priority to always report the statistical significance obtained?		
-No	5.4	3.7
- Yes, and using expressions like p<0.05, p>0.05	**59.7**	41.8
- Yes, and using expressions with the p value of exact probability	34.9	**54.5**

participant (n= 3) gave the name of a robust ES statistic (NNT). As Erceg-Hurn and Mirosevich (2008) pointed out, this might be due to lack of exposure to these methods. Thus, "the psychology statistics curriculum, journal articles, popular textbooks, and software are dominated by statistics developed before the 1960s" (Erceg-Hurn & Mirosevich, p. 593).

Regarding the knowledge of meta-analytic studies, the majority of the participants give more credibility and objectivity to systematic reviews and meta-analytic studies than to other types of literature reviews. Also, they have an adequate knowledge of meta-analyses. However, they have a poor knowledge of graphical displays for meta-analyses (i.e., forest plots and funnel plots), which can in a become misinterpretation of results. The graphical presentation of results is an important part of a meta-analysis and it has become the primary tool for presenting the results of multiple studies on the same research question (Anzures-Cabrera & Higgins 2010; Borenstein et al., 2009; Ellis, 2010; Sánchez-Meca & Marín-Martínez, 2010).

The analysis of researchers' behavior associated with their methodological practices reveals that academics who know some ES statistics present a profile closer to good statistical practices and research design, participate more actively in the process of peer review and, publish in journals with impact.

However, three issues warn about the knowledge that both groups of academics have about the ES and the validity of statistical conclusions in general: they wrongly associate the ES with the importance of a finding (clinical or practical significance fallacy), they continue to use to a large degree *p*-value expressions that revolve around the oracle of the value of alpha, and they do not know the purpose of planning a priori statistical power.

Finally, two events such as the open debate on the uses and abuses of statistical significance tests (which started almost at the same time it began to be used) and the development of check tools such checklists (CONSORT, STROBE, PRISMA...), which have allowed science to debate on statistical procedures, progress towards a statistical reform and present greater transparency and quality of studies, continue to be unknown by a high proportion of Spanish academic psychologists.

Nevertheless, there is currently an open scientific and social debate that can change the course of statistical practices among researchers. For example, for the last three years, criticism of the classical statistical inference procedure based on the probability value *p* and the dichotomous decision to keep or reject the null hypothesis has gained strength (Allison et al., 2016; Nuzzo, 2014). In addition, the low proportion of replication studies, publication bias that leads to an overestimation of the magnitude of effects, questionable statistical practices, and fraud also are current issues under discussion (Earp & Trafimow, 2015; Ioannidis, 2005; Kepes, Banks, & Oh, 2014).

Within this context of change and methodological advances (Spellman, 2015), the purpose of our study has been especially to emphasize the need for statistical re-education among Spanish academic psychologists, to disseminate the use of checklists, as tools for assessing the methodological quality of studies and, to motivate the development of manuals that conceptually describe the statistical tests and point out the consequences of poor statistical practice on the accumulation of scientific knowledge. Our purpose has also been to note the need for incorporating the modern robust ES statistics in statistical programs, such as SPSS.

We acknowledge some limitations of this study that need to be mentioned. Firstly, the low response rate could affect the representativeness of the sample and, therefore, the generalizability of the results. Moreover, it is possible that the participants who responded to the survey had higher levels of statistical knowledge than those who did not respond. Should this be the case, the results might overestimate the extension of the impact of the statistical reform in Spanish academic psychologists. Furthermore, it must also be acknowledged that some participants do not use quantitative methods at all. These individuals may also have been less likely to respond. Nevertheless, our findings are in line with previous research that analyzes the use of ES statistics in journals. For example, Sesé and Palmer, (2012) found that the most frequently reported ES measures were R^2 and Cohen's *d*. In addition, Peng et al. (2013) pointed out that robust ES statistics were reported less than non robust statistics, such as standardized mean differences.

In addition, it is possible that there has been an effect of social desirability, as is usual when data are collected using self-report questionnaires. For example, the percentage of participants who stated that they could give the name of an ES statistic was higher than the percentage that actually did so. A way to control this bias in future research would be to formulate the questions (e.g., what is the correct interpretation of a specific forest plot, funnel plot, ES or regression analysis?) with a three- or four-response format, or with an open question. These response formats would permit us to assess the level of knowledge of the statistical terms, thus, they would have been far more informative. In addition, we did not find a statistically significant relationship between the number of years as an academic and the use of ES reports, and neither were there statistically significant differences between academics who could name an ES statistic and academics who could not as a function of the variable number of years as academic psychologists. The presence of social desirability would be linked to participants with more years as academics reporting greater use of statistical ESs and greater knowledge of their names. Therefore, it is difficult that the degree of social desirability existed, unless we assume that is a stable and constant feature in all sub-samples of social extraction and independent of content, whether it is sensitive or purely formal, and with the voluntary nature of the survey.

"Evidence-based Practice" requires professionals to critically evaluate the results of psychological research studies in order to decide whether or not they are appropriate (Frias-Navarro, 2011; Beyth-Maron, Fidler, & Cumming, 2008; Sánchez-Meca & Botella, 2010). The information provided by the studies depends on the statistical analyses performed; therefore, their value largely depends on the quality of the statistical analyses and the interpretation of the results (Cumming, 2012; Cumming et al., 2012; Kline, 2013; Palmer & Sesé, 2013; Wilkinson & the TFI, 1999).

Estimating ESs means contextualizing their value within a research area, and not only deciding whether an effect is statistically significant or not. The interpretation of the magnitude of the effect implies making a judgment about its magnitude within a specific research context, indicating whether it is a small, medium or large effect. To make this judgment, the researcher must pose questions of practical and/or clinical significance, abandoning the emphasis on whether the result was or was not statistically significant (Cumming et al., 2012). Estimating effects and evaluating their magnitude within a specific context means that researchers' statistical practices must be complemented by their experience and judgment, fomenting Evidence-based Practice. This type of performance facilitates better comprehension of the study results, and it aids professionals (practitioners) in the true interpretation of the findings and their possible use in their clinical practice.

The proposed change is not easy because it requires joining forces in different areas (Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, & Pascual-Soler, 2015; Balluerka, Gómez, & Hidalgo, 2005; Cumming et al., 2007; Erceg-Hurn & Mirosevich, 2008; Kirk, 2001; Vacha-Haase, 2001).

The change must first arise from university teaching, and new programs and manuals of statistics that include alternatives to traditional statistical are needed (e.g., Campitelli, 2015). These should consider statistics that are more resistant to outliers and robust to violations of the assumptions of population normality and homogeneity of variance than means and variances (e.g., modern robust statistical methods, such as trimmed means and winsorized variance). Statistical software programs should also be updated. There are several websites that offer computing routines/

programs for general or specific ESs estimators and CIs of various ESs (Erceg-Hurn & Mirosevich, 2008; Fritz et al., 2012; Peng et al., 2013). Furthermore, the change requires the editorial policies of the psychology journals to clearly and decisively ask authors to follow the recommendations of the APA Manual (APA, 2010). As Peng et al. (2013) note, the ES reporting rate is higher for journals requiring ES reporting than for journals that do not require it.

Acknowledgments

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

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10*, 317-328.

Allison, D. B., Brown, A. W., George, B. J., & Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. *Nature*, *530*, 27-29.

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.

Anzures-Cabrera, J., & Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods, 1*, 66-80.

Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*, 873-879.

Badenes-Ribera, L., Frias-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value. A national survey study in academic psychologists from Spain. *Psicothema, 27,* 292-295.

Badenes-Ribera, L., Frias-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2013, February). *Informar e interpretar el tamaño del efecto en Psicología y Educación* [Reporting and interpreting the effect size in Psychology and Education]. Paper presentet at the XIV Congreso Virtual de Psiquiatria.com. Interpsiquis, Palma de Malloca, Spain.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology, 1,* 55-70.

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

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to Meta-analysis*. Chichester: Wiley.

Botella, J., & Sánchez-Meca, J. (2015). *Meta-análisis en ciencias sociales y de la salud* [Meta-analysis in social and health sciences]. Madrid, Spain: Sintesis.

Campitelli, G. (2015). Answering research questions without calculating the mean. *Frontiers in Psychology, 6,* 1379-1381.

Caperos, J. M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*, 408-414.

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

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29.

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., Fidler, F., Leonard, M., Kalinowski, P, Christiansen, A., Kleining, A., Lo, L, McMenamin & Wilson, S. (2007). Statistical reform in Psychology: Is anything changing? *Psychological Science, 18*, 230-232.

Earp, B. D., & Trafimow, D. (2015). Replication, falsifications, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621

Ellis, P. D. (2010). *The essential guide to effect size*. *Statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591-601.

Frias-Navarro, D. (2011). Reforma estadística. Tamaño del efecto. En D. Frias-Navarro (Ed.), *Técnica estadística y diseño de investigación* [*Statistical technique and research design*]. Valencia, Spain: Palmero Ediciones.

Frias-Navarro, D., Monterde-i-Bort, H., Pascual-Soler, M., Pascual-Llobell, J., & Badenes-Ribera, L. (2012, July). *Improving statistical practice in clinical production: A case Psicothema*. Poster presented at the V European Congress of Methodology, Santiago de Compostela, Spain.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology, General*, *141,* 2-18.

García, J., Ortega, E., & De la Fuente, L. (2008). Tamaño del efecto en las revistas de Psicología indizadas en Redalyc [Effect size in psychology journals indexed in redalyc]. *Informes Psicológicos*, *10,* 173-188.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6*, 135-146.

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

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *Counseling Psychologist, 34,* 601-629.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8): e124.

Kepes, S., Banks, G. C., & Oh, I.-S. (2014). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology*, *29*, 183-203.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deerin, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13,* 110-129.

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

Kline, R. E. (2011). *Principles and practice of structural equations modelling* (3rd ed.). Guilford Press, London.

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4,* 1-11.

Monterde-i-Bort, H., Frias-Navarro, D., & Pascual-Llobell, J. (2010). Uses and abuses of statistical significance tests and other statistical resources: A comparative study. *European Journal of Psychology of Education, 25*, 429-447.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.

Nuzzo, R. (2014). Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature*, *130*, 150-152.

Palmer, A., & Sesé, A. (2013). Recommendations for the use of statistics in clinical and health psychology. *Clínica y Salud, 24*, 47-54.

Peng, C.-Y J., Chen, L.-T, Chiang, H., & Chiang, Y. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review, 25*, 157-209.

Peng, C.-Y. J., & Chen, L.-T. (2014). Beyond Cohen's d: Alternative effect size measures for between subject designs. *The Journal of Experimental Education, 82,* 22-50.

Rosnow, R. L., & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Journal of Psychology, 217*, 6-14.

Sánchez-Meca, J., & Botella, J. (2010). Revisiones sistemáticas y meta-análisis: herramientas para la práctica profesional [Systematic reviews and meta-analyzes: Tools for professional practice]. *Papeles del Psicólogo, 31,* 7-17.

Sánchez-Meca, J., & Marín-Martínez, F. (2010). Meta-analysis in psychological research. *International Journal of Psychological Research, 3,* 150-162.

Sesé, A., & Palmer, A. (2012). El uso de la estadística en psicología clínica y de la salud a revisión [The current use of statistics in clinical and health psychology under review]. *Clínica y Salud*, *23*, 97-108.

Spellman, B. A. (2015). A Short (Personal) Future History of Revolution 2.0. *Perspectives on Psychological Science*, *10*, 886-899.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 10,* 989-1004.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect size. *Educational Researcher, 31*, 25-32.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432.

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.

Wang, Z., & Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? *Journal of Experimental Education, 75*, 109-125.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8*, 254-274.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist*, *54*, 594-604.