Ciencia Ergo Sum
ISSN: 1405-0269
ciencia.ergosum@yahoo.com.mx
Universidad Autónoma del Estado de México
México

Effect-Size Reporting in Mexican Psychology Journals: What it Says about the Quality of Research within the Field

Vera García, Fernando
Effect-Size Reporting in Mexican Psychology Journals: What it Says about the Quality of Research within the Field
Ciencia Ergo Sum, vol. 24, no. 3, 2017
Universidad Autónoma del Estado de México, México
Available in: https://www.redalyc.org/articulo.oa?id=10452159006

Ciencias Humanas y de la Conducta

Effect-Size Reporting in Mexican Psychology Journals: What it Says about the Quality of Research within the Field

Reporte del tamaño del efecto en revistas de psicología en México: lo que dice de la Calidad de la Investigación en la Disciplina

Fernando Vera García
University of Edinburgh, Escocia
fvera87@hotmail.com

Redalyc: https://www.redalyc.org/articulo.oa?id=10452159006

Received: 04 October 2016
Accepted: 01 February 2017

Abstract:

The purpose of this study was to determine the extent and context of effect-size (ES) reporting in Mexican psychology journals. 1554 statistical tests, distributed across 70 articles published in Mexican psychology journals in 2015, were analyzed with a checklist for es reporting practices. Results suggest that es reporting varies according to different types of tests, with journals that require it show slightly higher percentages; however, percentages of justification, definition and discussion on its importance are low, across all types of tests. Qualitative results about the context of es reporting suggest a broader problem in psychologists understanding and usage of null hypothesis significance testing. Recommendations are made for journal editors, which aim at a better usage and understanding of these statistical methods.

Keywords: effect-size, null hypothesis significance test, confirmation bias.

Resumen:

Se determina el grado y contexto del reporte de tamaño del efecto (TE) en revistas mexicanas de psicología. Se analizaron con una lista de cotejo sobre reporte de te 1554 pruebas estadísticas distribuidas a través de 70 artículos publicados en 2015. Los resultados sugieren que el reporte de te varía según la prueba estadística, y que las revistas que lo requieren muestran resultados ligeramente superiores. Los porcentajes de justificación, definición o explicación de la importancia del te fueron bajos para todas las pruebas. Los resultados cualitativos sugieren un problema mayor en cuanto a la comprensión y uso del método de prueba de hipótesis nula. Se hacen recomendaciones a editores de revistas, cuyo objetivo es mejorar el uso y comprensión de estos métodos estadísticos.

Palabras clave: tamaño del efecto, prueba de significancia de hipótesis nula, sesgo de confirmación.

Introduction

In the context of null hypothesis significance testing (NHST), effect size (ES) is a parameter that estimates the degree of departure from the null hypothesis (Cohen, 1988). ES is but one parameter used in statistical inference, the others being the significance criterion (viz. the probability of obtaining a sample result when the null hypothesis is true), sample size, and statistical power (viz. the probability that a test will lead to a rejection of the null hypothesis) (Cohen, 1988). In most cases, three out of these four parameters can be used to determine the remaining, meaning that ES can be thought of as a function of power, significance and sample size.

The importance of reporting ES can be divided into two broad categories. The first one is *a priori*, in that it can be used (in conjunction with two more parameters) to determine the required sample size given certain α and β values (measures of significance and power). In his now classic work on statistical power, Cohen

Author notes

fvera87@hotmail.com

(1988) actually provided power tables for such an analysis. Hence, if a researcher is interested in finding an ES of *x*, given a certain statistical power and significance criterion, he or she can learn that the sample required for such finding is of *n*. The second category of potential use is *a posteriori*, in that it is used to analyze results, whether to determine their practical significance, or in the context of meta-analysis, where they help with the task of replicability.

In determining practical significance, ES interpretation has been widely influenced by a set of cut-off points on the numerical values of ES indices proposed by Cohen (1988). He suggested that ES indices be interpreted as "small", "medium" or "large" for practical purposes. Ironically, despite his warning that these conventions should be used with caution, avoiding them if possible (1988, p. 532), they have become an unofficial standard for interpreting ES.[1] Other measures of practical significance, such as *odds ratio* and *relative risk* are more readily interpretable in probabilistic terms.

Despite their clear importance, statistical power and effect size have been consistently neglected by psychology researchers throughout the twentieth century. For instance, Huberty (2002) traced criticisms to the excessive attention placed by researchers on significance values, at the detriment of magnitude of effects, as far back as 1951. However, the subject persisted as one attended to solely by statisticians and methodologists until the 80´s and 90´s. Increased awareness of the limitations of both, the NHST method itself, and how it is (miss)used by researchers increased during this time (see Cohen, 1988; Hunter, 1997; Meehl, 1990). This prompted the APA to appoint a task force on the practice of NHST with specific recommendations which included the reporting of ES in NHST research reports, as an important complement to *p* values (American Psychological Association, 1996). Going further, several journals have started requiring that researchers report ES. However, even after such concrete actions have taken place, progress has been slow. For example, Alhija and Levy (2009) found that for tests other than correlation and regression,[2] ES was reported in 69% of articles where the journal required it, and on 57% of articles from journals not requiring it. A conceptual replication by Sun, Pan and Wang (2010) found that from a sample of 1,243 articles, 60% reported ES. Results such as these suggest that the problem is far from reaching a satisfactory outcome, but that at least some progress is being made.

In Mexico, a literature review suggests that the issue has not been addressed at all. A search was conducted by the author on the Mexican open access database Redalyc, which hosts nine major Mexican psychology journals, on August 2016. By searching, either in the title or as keyword, for the terms "*tamaño del efecto*", "*volumen del efecto*", "*potencia estadística*", "*effect size*", and "*statistical power*", and restricting the range of search to psychology as discipline and Mexico as country, resulted in no articles. On the other hand, by conducting the same search, but with no country restriction, eight articles directly relevant to the issue and published in Spanish speaking countries were found. Going even further, these eight articles plus two additional ones (Alhija & Levy, 2009; Sun, Pan & Wang, 2010) were used for citation analysis on both Web of Science and Scopus. Both searches revealed that only Alhija and Levy´s (2009) had been cited in an article with a Mexican-affiliated author. Yet, this sole citation took place in an article published in *Psychology in Russia: State of the Art*, which by its very title would make it difficult for someone not explicitly looking for the topic to come across it.

Against both, the background of ES reporting in general, and the apparent lacunae within Mexican psychology, the purpose of the present study is to determine the extent and context of ES reporting in Mexican psychology journals.

1. Method

1. 1. Sample

Nine major Mexican psychology journals were considered for the present study. Table 1 presents the profile of each journal, in terms of editorial institution and whether or not their current guidelines for authors require that researchers report effect sizes. As it can be seen from the table, only one journal explicitly requires authors to report ES.

TABLE 1.
Journals Selected for the Study.

Journal	Professional Association (PA) or Higher Education Institution (HEI)	Requires ES reporting
Acta Comportamentalia	PA	-
Acta de Investigación Psicológica	HEI	Yes
Enseñanza e Investigación en Psicología	PA	-
Journal of Behavior, Health and Social Issues	PA	-
Psicología Iberoamericana	HEI	-
Revista Intercontinental de Psicología e Investigación	HEI	-
Revista Latinoamericana de Psicología Conductual	PA	-
Revista Mexicana de Análisis de la Conducta	PA	-
Revista Mexicana de Psicología	PA	-

Source: author elaborated.

Note:. Acta Comportamentalia stipulates that authors adhere to the latest edition of the APA Manual, which includes reporting of ES. However, since it´s not explicitly stated, it was not marked as such (Universidad Nacional Autónoma de México, n/d)..

As exclusion criteria, articles or tests where NHST was used for assessing psychometric properties were not considered (except where these were accompanied by empirical research results). The rationale was that the interest of the present study lies in confirmation bias within empirical results (in the sense that they make claims about the world), as opposed to results regarding the structure of a given test. After applying these exclusion criteria, a total of 70 articles published in 2015, were analyzed.

1. 2. Instruments

A modified version of Alhija and Levy´s (2007) checklist was used to analyze articles (Annexed 1). As in the original study, since a single article can report more than one statistical test, one checklist was applied for each statistical test reported. However, when a test consisted of more than one significance analysis, only the primary one was registered (for instance, if an ANOVA test was followed by a *post hoc* analysis, only the former was registered; likewise, where logistic regression with more than one *odds ratio* or *relative risk* calculations was conducted, they were considered as one). The rationale for this approach was that derived (significance) analyses are but a part of one test.

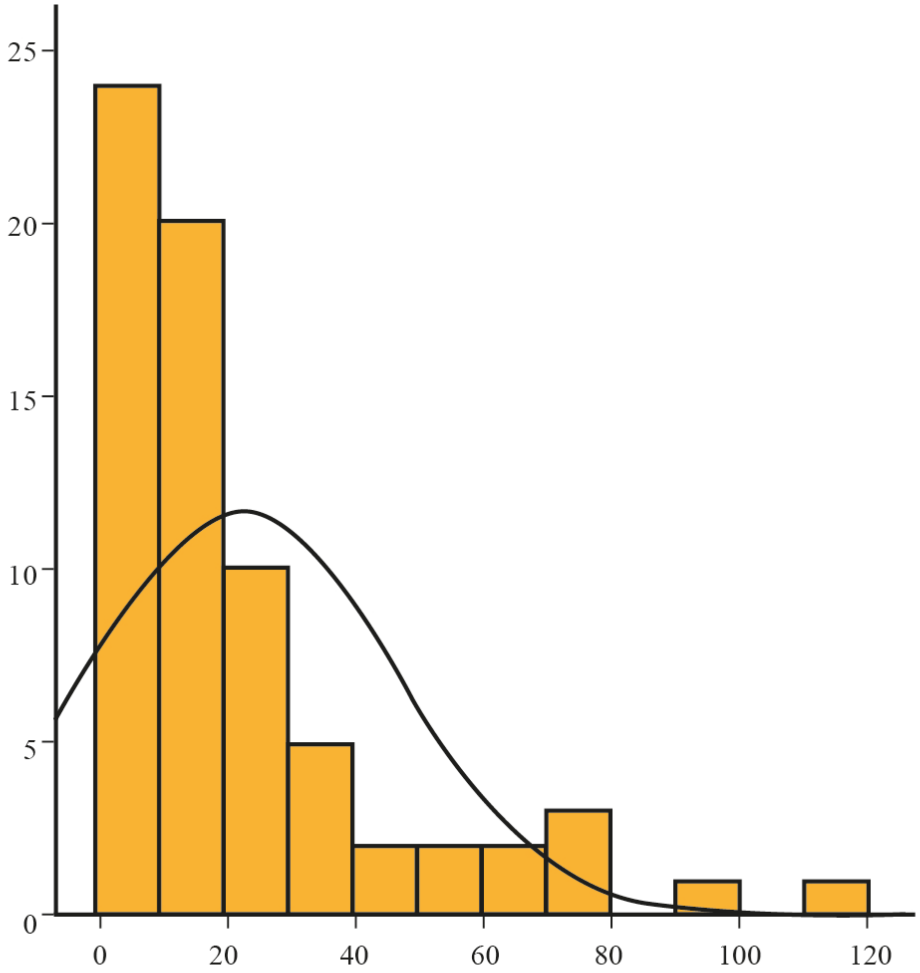
1. 3. Analysis of Information

Results are reported at both the test and article levels. In terms of type of analysis, the checklist was used for quantitative breakdown, but it was complemented with qualitative observations about the context in which NHST and ES analyses were conducted. Qualitative results presented below are discussed within the broader context of confirmation bias in psychology, and its associated phenomena. Since this was an exploratory analysis, only descriptive statistics were employed.

2. Results

2. 1. Quantitative Results

At the article level, a total of 70 articles accounted for 1554 statistical tests: [3] the mean amount of tests per article was of 22.2. Results showed a high dispersion, with a standard deviation of 23.91 and a range of 1-117 tests per article (Graph 1). Out of the 70 articles of the study, 43 (61%) reported es for at least one of their statistical tests. However, this figure is unlikely to be indicative of broad understanding of the concept and relevance of ES, since it is affected by correlation analyses, which are ES measures themselves, and thus inflate ES reporting rate drastically. This had already been reported by Alhija and Levy (2007), and is supported by further findings on correlation coefficients bellow. Once Pearson´s *r* and Spearman´s *rho* were removed from the analysis (along with articles reporting these tests exclusively), the amount of articles reporting at least one measure of es dropped down to 28 (42% of 67 remaining articles).



GRAPH 1.
Frequency distribution of tests per article.
Source: author elaborated.

At the test level, table 2 shows the total amount of tests, as well as reporting of ES measures. Results are presented categorizing tests by their complexity (as done by Sun, Pan & Wang, 2010), as *simple, general linear models*, and complex models. Likewise, they are separated by whether the publishing journal requires authors to report ES or not.

TABLE 2.
Tests conducted and reporting of ES, by journal type.

	Journal requires ES reported		Journal does not require ES reported	
	Tests	ES reported F (% of test total)	Tests	ES reported F (%)
Simple tests	147	32 (22%)	236	32 (14%)
t Test	72	0	167	22 (13%)
Z Score	2	2 (100%)	0	
Mann-Whitney U	18	18 (100%)	21	5 (24%)
X^2	19	0	38	5 (13%)
Kruskal-Wallis	10	0	1	0
Wilcoxon signed-rank test	12	12 (100%)	9	0
Unspecified bivariate analysis	14	0	0	
General Linear Models	295	244 (83%)	831	672 (81%)
ANOVA	52	13 (25%)	142	43 (31%)
MANOVA	0		2	0
Correlation	216	211 (98%)	634	601 (95%)
Regression	27	20 (74%)	51	28 (55%)
Pearson´s Contingency	0		2	0
Complex Models	12	12 (100%)	5	4 (80%)
Structural Equation Modelling (SEM)	12	12 (100%)	4	4 (100%)
Multilevel Modelling	0		1	0
Undefined Statistical Test	1	0	27	0

Source: author elaborated.

Notes: . Correlation includes both Pearson´s r and Spearman´s rho, as well as 108 unspecified correlation analyses. Likewise, regression covers all variants (i.e. linear, multiple, stepwise, logistic, etc.)..

While in the aggregate results do not seem as different between journal types, there is a tendency towards more reporting of es when the journal requires it. The type of test does show an increase in ES reporting as complexity increases. Structural equation modelling and correlation showed the highest percentages of ES reports. [4]

Table 3 shows the various indices used for each nhst test.

TABLE 3.
NHST Methods by ES Indices used.

NHST Method ES Index	Frequency	NHST Method ES Index	Frequency
Student´s T		Correlation	
Cohen´s *d*	22	*r*	812
Z Test		Regression	
Cohen´s *d*	2	*r*	2
Mann-Whitney U		r^2	1
Cohen´s *d*	5	*R*	2
r	18	R^2	15
X^2		ΔR^2	1
Cramer´s *V*	5	Adjusted R^2	22
Wilcoxon Signed Rank Test		Nagelkerke´s R^2	9
r	12	Odds ratio	11
ANOVA		Relative risk	1
Cohen´s *d*	5	SEM	
Eta Squared	20	Goodness of Fit Indices	16
Eta	7	Variance Explained	8
Partial Eta Squared	24	Latent mean differences	2

Source: author elaborated.

Note: . discrepancies between total ES indices (1022) and total times ES was reported (996) is due to cases where more than one ES index was reported for an NHST method. This was the case for regression analyses and SEM. See qualitative results for further treatment of ES in SEM..

Table 4 presents the various practices for tests that reported ES. As it can be seen, with the exception of SEM methods, ES was almost never justified (with only three other tests doing it, and always in percentages bellow 50%). Likewise, a definition of ES, whether as formula or text, was only provided for 6 uses of SEM (38%); the same was the case for discussions on the importance of ES. Interpretation of ES measures obtained the highest scores.[5] Finally, only in seven instances across three tests were discrepancies between significance and ES results reported.

TABLE 4.
ES reporting practices.

	ES Reported	ES Justified	Definition of ES	Importance of ES discussed	ES Interpreted	Discrepancy between sig and ES reported
	F	*f* (%)	*f* (%)	*f* (%)	*f* (%)	*f* (%)
Simple tests						
t Test	22	7 (32%)	0	0	15 (68%)	3 (14%)
Z Score	2	0	0	0	1 (50%)	1 (50%)
Mann-Whitney U	23	5 (22%)	0	0	5 (22%)	0
X^2	5	0	0	0	5 (100%)	3 (60%)
Wilcoxon signed-rank test	12	0	0	0	0	0
General Linear Models						
ANOVA	56	0	0	0	2 (4%)	0
Correlation	812	57 (7%)	0	0	299 (37%)	0
Regression	48	0	1 (2%)	0	44 (92%)	0
Complex Models						
SEM	16	10 (63%)	6 (38%)	6 (38%)	14 (88%)	0

Source: author elaborated.

Finally, it's important to analyze the reporting practices of ES, conditional on statistical significance. Out of the 1554 statistical tests recorded, 1040 were significant (67%). For the remaining 514 non-significant tests, ES was reported in 277 instances (54%). Of these 277, 231 (83%) cases occurred in correlation analysis (usually in the context of a correlation matrix), with only 30 explicitly reporting the significance value. A narrative reconstruction of these results would be: the majority of tests reported are statistically significant; out of the non-significant 43%, ES is reported in only 54% of the cases, primarily in correlations that lack significance values to accompany them.

2. 2. Qualitative Results

a) The Use of Effect Size Indices

As noted previously, even when authors do report ES values, they seldom justify or interpret them. In the case of correlation coefficients, this was already hinted at by the fact that only 36% of the times was the correlation coefficient interpreted (thus supporting the thesis that it not even understood as a measure of ES, but rather as just another type of NHST test). In light of their variability (and sometimes controversy), it´s important for authors to clearly justify their choice for ES index. Examples of problematic cases beyond that of correlation are provided below.

In the case of ANOVA tests, the choice between η^2 and partial η^2 is dependent on the number of factors. As Cohen (1973) pointed out, partial η^2 *partials out* other sources of variance. While in a one fixed factor ANOVA design, both formulas are interchangeable (since there´s nothing to partial out), with more than one factor, the first formula yields smaller results (see figure 1). In the present context, one article reporting 11 uses of η^2 had more than one factor, making the choice of ES questionable (see also Pierce, Block & Aguinis, 2004).

$\eta^2 = \frac{SS_A}{SS_T}$	Where SS_A is the sum of squared of factor A, and SS_T it the total sum of squares.
partial η^2 $(YA \bullet BC \ldots J) = \frac{df_A F_A}{df_A F_A + df_E} = \frac{SS_A}{SS_A + SS_E}$	Where: At the right: df_A and df_E are the degrees of freedom for factor A and *error* respectively; and F_A is the value of the F statistic for factor A. At the left: Y is the dependent variable, A is the factor in question, and B, $C \ldots J$ are all other sources of variance.

FIGURE 1.

Relationship between η^2 and partial η^2.

Note: in a one factor ANOVA design, the total sum of squares (SS_T) will be equal to $SS_A + SS_E$, thus making it interchangeable with partial η^2. With more than one factor, use of regular η^2 will inevitably lead to smaller values. Source: Cohen (1973: 107-108).

As for regression analysis, while all indices used were variations of the coefficient of determination *r*, the variability of potential es indices makes their justification especially important. For example, one study conducting stepwise linear regression reported ES as both percentages of variance explained and ΔR^2 for each step; however the subtraction of $\Delta R^2(\text{Step}_{n-1})$ from $\Delta R^2(\text{Step}_n)$ (which is the real ΔR^2) did not match the reported ΔR^2 value. [6]

Varying R^2 measures are particularly problematic for logistic regression. Because the dependent variable is measured at the ordinal level, authors must rely on seudo-R^2 measures. And even though out of the two articles reporting logistic regression results, one did use Nagelkerke´s (1991) R^2, the other one reported a regular r^2 index, which is problematic.[7]

The case of structural equation modelling (SEM) is one where no clear consensus exists. At several stages of SEM analysis, effect size measures are interpreted. The first and most agreed on measure is that of overall model data fit. In sem one aims at confirming the null hypothesis that the reproduced covariance matrix is equal to the covariance matrix of the population (Cui, 2012; Bowen and Guo, 2011). However, it should be noted that the rationale for goodness of fit measures to be considered as ES measures (e.g. Kelley & Preacher, 2012), is not an agreed one. A widely adopted framework for power analysis developed by MacCallum, Browne and Sugawara (1996), defined effect size as a difference between a rmsea value reflecting the null hypothesis, and an inconsistent one ($\delta = e_0 - e_a$). Yet since this index is not standardized, it´s not without critics (see also Cui, 2012; Bowen & Guo, 2011). A second measure, is that of R^2 for factors or latent variables. In this regard, while measuring specific variable combinations, the coefficient of determination does not provide a measure of the degree of departure from SEM´s null hypothesis. In the present study, all instances of sem reported fit indices (including RMSEA). In addition, eight instances across four articles reported variance explained among the model´s variables, while another article reported latent mean differences between variables of two models, with ES measures being Cohen´s *d*. Thus, more stringent criteria such as MacCallum and colleagues would not count raw RMSEA values as ES indices, while mere focus on the R^2 of a model´s composing variables would fail to account for overall model data fit.

Finally, the case of non-parametric tests is equally problematic, since they do not conform to traditional ES measures (see Grissom & Kim, 2001). The reported ES measures for Mann-Whitney U and Wilcoxon signed-rank tests presented in table three are sensitive to parametric assumptions.

b) The 0.05 rule of thumb

Critics of the way NHST has been used historically have pointed out the damaging 0.05 consensus for rejecting or not the null hypothesis (Cohen, 1994; Sun, Pan & Wang, 2010; Anderson, Burnham & Thompson, 2000). Not only is this rule of thumb arbitrary, but it is also inappropriate to the extent that it treats categorically a probability measure that is a continuum. In the context of ES, marginally significant results may be accompanied by very small ES values; but conversely, α levels slightly above the 0.05 criterion may be associated with large ES values.

Several findings suggest that rule of thumb (whether 0.05 or 0.01) is being a-critically summoned. First, out of the 514 documented non-statistically significant results, test statistics (and *p* values) were only provided for 176 of them (34%). Second, by authors reporting p values as ranges (viz. <0.05 and >0.05); and there were even articles that reported precise *p* values for significant results but only ranges for non-significant (i.e. >0.05). Third, specifically in the context of es reporting and interpreting, within articles that did report ES, it was interpreted in terms of research question on 10% less of the times where the significance result was negative (31% vs 41% of the total times es was reported). However, it must be noted that the majority of reported es measures related to non-significant results were small (by Cohen´s 1988 conventions); yet even then, there were medium effect sizes reported for eight non-significant results, of which only one was interpreted.

c) Distinguishing Relevant from Spurious Findings

The aforementioned range of tests (1-117) with a mean of 22.2 tests per article may be indicative of a broader problem regarding the practice of NHST. It was not uncommon to find studies were statistical tests were conducted on all possible combinations of variables, without clear hypotheses. Some examples of this are: a correlation between perceived stress and PTSD; depression negatively correlated to satisfaction with being alive, healthy, among others (out of 65 correlation analyses conducted); an inverse correlation between marital satisfaction and having marital problems; or a positive correlation between use of condom and acceptance of its use.

These types of results share three characteristics: they are the result of running statistical tests on several possible combinations, in lack of a clear hypothesis; they are expectable, in the sense that one could infer them out of either common sense (i.e. people having marital problems being dissatisfied with their marriage) or analyticity (i.e. perceived stress in people suffering from a stress disorder); and that they are symptomatic of the misuse of the NHST method that leads to the null hypothesis being almost always false. This last point has been argued on two grounds: first, that by taking the null hypothesis to be of the form $H_0 = 0$[8] (as in zero differences between means, or zero correlation), one is almost always bound to find at least slight differences. Secondly, because of what has sometimes been referred to as the crud factor, or ambient correlation noise (Meehl, 1990; Cohen, 1994): the fact that in nature (especially in biological and psychological sciences), everything is to a larger or lesser extent correlated. A proper application of NHST (a context referred to by Hunter in 1997 as the *debunking* context) is one where the probability of H_0 is actually larger, one where $H_0 = n$, where n is a specific value such that $H_0 \neq 0$.

d) *Misinterpretation of statistical significance*

2Several authors have criticized the misinterpretation of significance values as part of a broader criticism of the excessive reliance in such parameter (see Verdam, Oort & Sprangers, 2014). This was the case in various tests: one study reported the results of anova and Mann-Whitney U tests as “Statistically significant correlations”; another study reported a statistically significant difference in rates of variable-x between groups as a significantly lower rate of variable-x between groups; a third study reported a statistically significant difference in mean depression rates using the use of internet as independent variable, as depression being the result of the use of internet; a fourth study reported a lack of statistically significant differences between groups as there being “no differences”. What these interpretations share is a misunderstanding of what NHST provides. Contrary to their claims, a p value of 0.05 or lower is not synonymous with meaningful differences, nor does it prove causality. Conversely, a p value higher than 0.05 is not indicative of there being no differences, nor does it prove that there is no effect of x on y. Rather, all such p values provided is a probability of getting such results with the null hypothesis being true. Once again, the probability of $H_0 = 0$ being true is always quite low. .

How this ties to the concept of ES is easy to grasp: differences between groups were not necessarily “significant”, nor rates in *variable-x* significantly lower. A proper estimation of es (as well as confidence intervals) would provide appropriate estimates of just how significant such results actually were. The difference between practical and statistical significance is precisely what ES allows.

e) *The File Drawer Problem*

Described by Rosenthal (1979), the “file drawer” problem consists of the fact that published results are by no means the total amount of results obtained, but rather the positive ones. Since that time, this problem has been documented time and again in psychology as one that contributes to confirmation bias. In several cases, this practice was evidenced by authors′ language and explicit references to other tests. Eight articles were found, where authors made claims of the following type: “Correlation coefficients were calculated between all variables, but only x correlation was found”, or “out of all tests conducted, only n were significant”. Thus, the 22.2 tests per article (along with associated frequency statistics) are clearly an underestimation of the total amount of tests being conducted for a given published article.

Clearly authors making such claims were unaware of why selective reporting is problematic, which shows how pervasive the bias for positive results is within psychology. This has already been documented extensively. For instance, Fanelli (2010) showed that between 80% and 90% of papers from behavioral and social sciences obtained results supporting their hypothesis, scoring highest among sampled fields. In a separate paper, the same author (Fanelli, 2012) found that psychology and other social sciences have seen the largest decrease in published negative results across time.

3. Discussion

The results of the present study suggest that reporting of ES in Mexican psychology journals is considerably low, and even where such reporting takes place, proper understanding of the concept and justification for its specific use, is limited. As the qualitative results have shown, this is related to a broader misuse of the NHST method. A narrative reconstruction of these results would be that psychologists (in general) apply a whole set of significance tests, but rarely report them appropriately. That they take these methods to be ways of proving hypotheses, and interpret their findings in precisely that way. The incompleteness, selectiveness, and a-critical way in which these methods are employed is cause for concern.

Several limitations of the present study must be discussed. First, it has focused only on psychology journals, excluding inter-disciplinary ones. While there seems to be no reason to suppose that interdisciplinary journals will show a different tendency, this is something that has to be determined empirically. Second, it has focused only on articles published in 2015, given that its interest lies in the *current* state of affairs; it is possible that reporting of ES show changes year after year (which would likely indicate changes in the quality of research teaching or development of the field). Third, it has avoided to include psychometric analyses that rely on NHST. The rationale was that the main interest (and interpretative framework) lied in confirmation bias. Once again, there is no reason to believe that authors who fail to report ES for a t test between groups in an experimental context will do so for a pre-test/post-test reliability analysis; or that they will fail to report ES on a factorial analysis of variables affecting schizophrenia onset age, but will do so for a factorial analysis of construct validity for a test. It is suggested that future research focuses on both other social science disciplines (i.e. education), as well as expanding its scope through time.

The prospective recommendations that stem from this study are both clear and straightforward, aimed at journal editors and reviewers:

a) To change journal´s guidelines for authors, so as to explicitly require them to report complete test statistics for all nhst analyses employed (whether they are statistically significant or not). This requirement should not just include ES, but also confidence intervals, and statistical power.
b) They should also emphasize and require an integral interpretation of results (there is no point in reporting es if all that gets interpreted is α level) beyond the 0.05/0.01 rule of thumb. Especially emphasizing practical significance of results.
c) Finally for editors and reviewers, a more critical assessment of the tests employed, and the specific hypotheses being assessed, is required. This recommendation can take the form of a grading checklist, including items such as "is the statistical test required?", or "is the interpretation of NHST methods accurate?"

The rationale for aiming recommendations at journal editors and reviewers is the following: to the extent that academic journals are the primary outlet of scientific results, it is to be expected that these requirements will be subsequently cascaded down to other instances, especially graduate and postgraduate classrooms. Ideally, professional associations like the *Sociedad Mexicana de Psicología*, or the *Consejo Nacional para la Enseñanza e Investigación en Psicología* (which publish their own journals) would have the necessary infrastructure for communication of this renewed focus on the quality of nhst. With today´s technology it is quite possible to publish online manuals and calculators for ES, sample size, among other tools; it is equally possible to spread the word about freeware downloadable software like G-Power (Faul, Erdfelder, Lang & Buchner, 2007), which allows for estimations based on the four NHST parameters discussed.

Unlike other authors (i.e. Hunter, 1997), the present study does not necessarily support the thesis that NHST should be discarded altogether. Rather, it argues for not reducing experimental findings to *p* values, and yes/no decisions. Campbell´s (1982: 698) words may be just as applicable for today´s Mexican psychology as they were more than 20 years ago:

> Books [...] have been written to dissuade people from the notion that smaller *p* values mean more important results or that statistical significance has anything to do with substantive significance. It is almost impossible to drag authors away from their *p* values, and the more zeros after the decimal point, the harder people cling to them.

References

Alhija, F.N.A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement*, *69*(2), 245-265.

American Psychological Association. (1996). *Task force on statistical inference* (Initial report). Retrieved September 21, 2016 from http://www.apa.org/science/leadership/bsa/statistical/tfsi-initial-report.pdf .

Anderson, D., Burnham, K., & Thompson, W. (2000). Null hypothesis testing: problems, Prevalence, and an Alternative. *The Journal of Wildlife Management*, *64*(4), 912-923.

Bowen, N. & Guo, S. (2011). *Structural equation modelling.* Oxford: Oxford University Press.

Campbell, J. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology, 67*(6), 691-700.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement, 6*(2), 135-147.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* 2nd Edition. New York: Taylor & Francis Group.

Cohen, J. (1994). The earth is round (p < 0.05). *American Psychologist, 49*(12), 997-1003.

Cui, M. (2012).*Effect-size index for evaluation of model-data fit in structural equation modeling* (Unpublished Master of Science dissertation). The Florida State University, Florida.

Ellis, P. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results.* Cambridge: Cambridge University Press.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*(3), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891-904.

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Gómez-Benito, J., Hidalgo, MD. & Zumbo, B. (2013). Effectiveness of combining statistical tests and effect sizes when using logistic discriminant function regression to detect differential item functioning for polytomous ITEMS . *Educational and Psychological Measurement, 73*(5): 875-897.

Grissom, R., & Kim, J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6*(2), 135-146.

Huberty, C. (2002). A history of effect size indices. *Educational and Psychological Research, 62*(2), 227-240.

Hunter, J. (1997). Needed: A ban on the significance test. *Psychological Science, 8*(1), 3-7.

Kelly, K., & Preacher, K. (2012). On effect size. *Psychological methods, 17*(2), 137-152.

Le, H. & Marcus, J. (2012). The overall odds ratio as an intuitive effect size index for multiple logistic regression: examination of further refinements. *Educational and Psychological Measurement, 72*(6): 1001-1014.

MacCallum, R., Browne, M., & Sugawara, H. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149.

Meehl, P. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(Monograph Suppl. 1-V66), 195-244.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691-692.

Pierce, C., Block, R., & Aguinis, H. (2004). Cautionary note on reporting Eta-squared values from multifactor anova designs. *Educational and Psychological Measurement, 64*(6), 916-924.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Rosnow, R. & Rosenthal, R. (2003). Effect sizes for experimental psychologists. *Canadian Journal of Experimental Psychology, 57*(3): 221-237.

Sun, S., Pan, W., & Wang, L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102*(4), 989-1004.

Universidad Nacional Autónoma de México (n.d.) *Acta Comportamentalia: Revista latina de análisis del comportamiento. Normas para autores.* Retrieved on October 10, 2016 [http://www.revistas.unam.mx/index.php/acom/about/submissions#authorGuidelines].

Verdam, M., Oort, F., & Sprangers, M. (2014). Significance, truth and proof of p values: reminders about common misconceptions regarding null hypothesis significance testing. *Quality of Life Research, 23*(1), 5-7.

Annexed

ANNEXED 1.
Checklist for Analyzing Reporting of Effect-size.

No	Item	Note/Check
1	Title of the article	
2	Year	
3	Journal name	
4	Type of journal (Requires ES reported)	Yes/No
5	Research topic (First keyword)	
6	Purpose of statistical test	
7	Participants	
8	Research design (Experimental)	Yes/No
9	Sample size	
10	Statistical procedure (NHST Method)	
11	Statistically significant	Yes/No
12	Significance value	
13	ES reported	Yes/No
14	ES Index used	
15	ES value	
16	ES reported for non-significant	Yes/No/Non-applicable
17	ES justified	Yes/No
18	ES justification	
19	Definition of ES reported	Yes/No
20	Importance of ES discussed	Yes/No
21	ES interpreted	Yes/No
22	Discrepancy of results between statistical and practical significance discussed	Yes/No

Source: Adapted from Alhija and Levy (2009); and Sun, Pan and Wang (2010).

Notes: . ES was considered to be interpreted when a practical and verbal explanation of the index was given (i.e. Cohen´s conventions of small, medium, large; or phrases such as “close to perfect”, etc.). ES was considered to be defined when either a formula, or a verbal description of its calculation was provided..

Notes

1. A more thorough discussion of the problem of interpreting es indices was provided by Ellis (2010), who gave examples of cases where context and prior knowledge determine interpretations conflicting with Cohen´s conventions (see also Rosnow & Rosenthal, 2003). This is a similar problem to the one encountered by the 0.05/0.01 significance criterion, where an acritical rule replaces in-depth interpretation.
2. Correlation and regression are excluded from this report because they are es measures themselves. How much the use of correlation and regression reflect researchers´ understanding of ES will be addressed below.
3. This figure accounts for statistical tests reported or described, regardless of whether or not test statistics were reported. For example, if an article reported that no statistically significant differences were found between groups in a given variable, without reporting test statistics, it was nonetheless quantified.
4. Alhija and Levy (2007) found that correlation analysis had a 100% rate of es reporting. Actually, the 98 and 95% reporting rates found in the present study are problematic insofar as they represent cases where a correlation analysis was conducted but no test statistics were provided. This issue will be addressed in full detail in the qualitative results section.
5. See Note on the Appendix.
6. One can only guess as to the reason for this discrepancy, but one possibility is that since SPSS´s output provides both R^2 and adjusted R^2, the authors may have used one for variance explained reporting, and another for changes in R^2.
7. Alternative es measures for logistic regression are *odds ratio* and *relative risk* measurements. In the present study, it was found that out of three articles reporting logistic regression results, two accompanied R^2 measures with *odds ratio*, while one reported only *odds ratio*. The choice between one and the other is not without controversy (see Gómez-Benito, Hidalgo & Zumbo, 2013; Le & Marcus, 2012).
8. Cohen (1994) called this form of the null hypothesis the “Nil hypothesis” in the sense that the null hypothesis has a value of nil, zero.