Clínica y Salud 24 (2013) 47-54

Clínica y Salud

www.elsevier.es/clysa

Recommendations for the use of statistics in Clinical and Health Psychology

Alfonso Palmer and Albert Sesé*

Universitat de les Illes Balears, Spain

INFORMACIÓN ARTÍCULO

Historia artículo:
Recibido: 25/09/2012
Aceptado: 15/11/2012

Keywords:
Clinical and Health Psychology
Methodological recommendations
Statistical use
Publication Rules

ABSTRACT

The generation of scientific knowledge in Psychology has made significant headway over the last decades, as the number of articles published in high impact journals has risen substantially. Breakthroughs in our understanding of the phenomena under study demand a better theoretical elaboration of work hypotheses, efficient application of research designs, and special rigour concerning the use of statistical methodology. Anyway, a rise in productivity does not always mean the achievement of high scientific standards. On the whole, statistical use may entail a source of negative effects on the quality of research, both due to (1) the degree of difficulty inherent to some methods to be understood and applied and (2) the commission of a series of errors and mainly the omission of key information needed to assess the adequacy of the analyses carried out. Despite the existence of noteworthy studies in the literature aimed at criticising these misuses (published specifically as improvement guides), the occurrence of statistical malpractice has to be overcome. Given the growing complexity of theories put forward in Psychology in general and in Clinical and Health Psychology in particular, the likelihood of these errors has increased. Therefore, the primary aim of this work is to provide a set of key statistical recommendations for authors to apply appropriate standards of methodological rigour, and for reviewers to be firm when it comes to demanding a series of sine qua non conditions for the publication of papers.

Recomendaciones para el uso de estadísticos en Psicología Clínica y de la Salud

Palabras clave:
Normas de publicación
Psicología Clínica y de la Salud
Recomendaciones metodológicas
Usos estadísticos

RESUMEN

La generación de conocimiento científico en Psicología ha experimentado una importante progresión durante las últimas décadas, ya que el número de artículos publicados en revistas con factor de impacto ha incrementado sensiblemente. Los avances en la comprensión de los fenómenos objeto de estudio exigen una mejor elaboración teórica de las hipótesis de trabajo, una aplicación eficiente de los diseños de investigación y un gran rigor en la utilización de la metodología estadística. Por esta razón, sin embargo, no siempre un incremento en la productividad supone alcanzar un alto nivel de calidad científica. Los usos estadísticos pueden ser, en general, una fuente de efectos negativos sobre la calidad de la investigación, tanto por el grado de dificultad que la comprensión y aplicación de algunos métodos requiere, como por la comisión de un conjunto de errores como, sobre todo, por la omisión de información fundamental para evaluar la adecuación de los análisis realizados. A pesar de que haya notables trabajos dedicados a la crítica de estos malos usos, publicados específicamente como guías de mejora, la incidencia de mala praxis estadística todavía permanece en niveles mejorables. Dada la creciente complejidad de las teorías elaboradas en la psicología en general y en la psicología clínica y de la salud en particular, la probabilidad de ocurrencia de tales errores se ha incrementado. Por este motivo, el objetivo fundamental de este trabajo es presentar un conjunto de recomendaciones estadísticas fundamentales para que los autores consigan aplicar un nivel de rigor metodológico adecuado, así como para que los revisores se muestren firmes a la hora de exigir una serie de condiciones sine qua non para la publicación de trabajos.

*Correspondence on this article should be sent to Albert Sesé.
E-mail: albert.sese@uib.es

1130-5274/$ - see front matter

DOI: http://dx.doi.org/10.5093/cl2013a6

In the words of Loftus (1996), "Psychology will be a much better science when we change the way we analyse data". Empirical data in science are used to contrast hypotheses and to obtain evidence that will improve the content of the theories formulated. However it is essential to establish control procedures that will ensure a significant degree of isomorphism between theory and data as a result of the representation in the form of models of the reality under study.

Over the last decades, both the theory and the hypothesis testing statistics of social, behavioural and health sciences, have grown in complexity (Treat and Weersing, 2005). Hence, the degree of sophistication in quantitative research in the area of Psychology in general and in the area of Clinical and Health Psychology in particular, is increasing in such a way that the novice researcher is faced with such a variety of options that he/she can feel mixed up sometimes. Anyway, the use of statistical methodology in research has significant shortcomings (Sesé and Palmer, 2012).

This problem has also consequences for the editorial management and policies of scientific journals in Psychology. For example, Fiona, Cummings, Burgman, and Thomason (2004) say that the lack of improvement in the use of statistics in Psychology may result, on the one hand, from the inconsistency of editors of Psychology journals in following the guidelines on the use of statistics established by the American Psychological Association and the journals' recommendation and, on the other hand from the possible delays of researchers in reading statistical handbooks.

Whatever the cause, the fact is that the empirical evidence found by Sesé and Palmer (2012) regarding the use of statistical techniques in the field of Clinical and Health Psychology seems to indicate a widespread use of conventional statistical methods except a few exceptions. Yet, even when working with conventional statistics significant omissions are made that compromise the quality of the analyses carried out, such as basing the hypothesis test only on the levels of significance of the tests applied (Null Hypothesis Significance Testing, henceforth NHST), or not analysing the fulfilment of the statistical assumptions inherent to each method. Hill and Thomson (2004) listed 23 journals of Psychology and Education in which their editorial policy clearly promoted alternatives to, or at least warned of the risks of, NHST. Few years later, the situation does not seem to be better. This lack of control of the quality of statistical inference does not mean that it is incorrect or wrong but that it puts it into question.

Apart from these apparent shortcomings, there seems to be is a feeling of inertia in the application of techniques as if they were a simple statistical cookbook –there is a tendency to keep doing what has always been done. This inertia can turn inappropriate practices into habits ending up in being accepted for the only sake of research corporatism.

Therefore, the important thing is not to suggest the use of complex or less known statistical methods "per se" but rather to value the potential of these techniques for generating key knowledge. This may generate important changes in the way researchers reflect on what are the best ways of optimizing the research-statistical methodology binomial. Besides, improving statistical performance is not merely a desperate attempt to overcome the constraints or methodological suggestions issued by the reviewers and publishers of journals. Paper authors do not usually value the implementation of methodological suggestions because of its contribution to the improvement of research as such, but rather because it will ease the ultimate publication of the paper.

Consequently, this work gives a set of non-exhaustive recommendations on the appropriate use of statistical methods, particularly in the field of Clinical and Health Psychology. We try to provide a useful tool for the appropriate dissemination of research results through statistical procedures.

Statistical Recommendations

In line with the style guides of the main scientific journals, the structure of the sections of a paper is: 1. Method; 2. Measurement; 3. Analysis and Results; and 4. Discussion. Authors will include accordingly the statistical information related to his/her research.

1. Method

1.1 Designs

It is necessary to provide the type of research to be conducted, which will enable the reader to quickly figure out the methodological framework of the paper. Studies cover a lot of aims and there is a need to establish a hierarchy to prioritise them or establish the thread that leads from one to the other. As long as the outline of the aims is well designed, both the operationalization, the order of presenting the results, and the analysis of the conclusions will be much clearer.

Sesé and Palmer (2012) in their bibliometric study found that the use of different types of research was described in this descending order of use: Survey (31.8%), Quasi-experimental (28.4%), Experimental (19.7%), Theoretical (7.5%), Instrumental (3.2%), Qualitative (2.6%), Meta-analysis (1.4%), among others. It is worth noting that some studies do not establish the type of design, but use inappropriate or even incorrect nomenclature. In order to facilitate the description of the methodological framework of the study, the guide drawn up by Montero and León (2007) may be followed.

1.2 Population and Samples

The interpretation of the results of any study depends on the characteristics of the population under study. It is essential to clearly define the population of reference and the sample or samples used (participants, stimuli, or studies). If comparison or control groups have been defined in the design, the presentation of their defining criteria cannot be left out. The sampling method used must be described in detail, stressing inclusion or exclusion criteria, if there are any. The size of the sample in each subgroup must be recorded. Do not forget to clearly explain the randomization procedure (if any) and the analysis of representativeness of samples. Concerning representativeness, by way of analogy, let us imagine a high definition digital photograph of a familiar face made up of a large set of pixels. The minimum representative sample will be the one that while significantly reducing the number of pixels in the photograph, still allows the face to be recognised. For a deeper understanding, you may consult the classic work on sampling techniques by Cochran (1986), or the more recent work by Thompson (2012).

Whenever possible, make a prior assessment of a large enough size to be able to achieve the power required in your hypothesis test. There are statistical programmes that enable you to carry out these tasks in a simple way, such as G*Power (Erdfelder, Faul, & Buchner, 1996), or the R programme (R Development Core Team, 2012), which are free and can be downloaded directly from the Internet.

1.3 Assignment

Random assignment. For a research which aims at generating causal inferences, the random extraction of the sample is just as important as the assignment of the sample units to the different levels of the potentially causal variable. Random selection guarantees the representativeness of the sample, whereas random assignment makes it possible to achieve better internal validity and thereby greater control of the quality of causal inferences, which are more free from the possible effects of confounding variables.

Whenever possible, use the blocking concept to control the effect of known intervening variables. For instance, the R programme, in its *agricolae* library, enables us to obtain random assignation schematics of the following types of designs: Completely randomized, Randomized blocks, Latin squares, Graeco-Latin squares, Balanced incomplete blocks, Cyclic, Lattice and Split-plot.

For some research questions, random assignment is not possible. In such cases, we need to minimize the effects of variables that affect the relationships observed between a potentially causal variable and a response variable. These variables are usually called confusion variables or co-variables. The researcher needs to try to determine the relevant co-variables, measure them appropriately, and adjust their effects either by design or by analysis. If the effects of a co-variable are adjusted by analysis, the strong assumptions must be explicitly established and, as far as possible, tested and justified. Describe the methods used to mitigate sources of bias, including plans to minimize dropout, non-compliance and missing values.

2. Measurement

2.1 Variables

Explicitly define the variables of the study, show how they are related to the aims and explain in what way they are measured. The units of measurement of all the variables, explanatory and response, must fit the language used in the introduction and discussion sections of your report. Consider that the goodness of fit of the statistical models to be implemented depends on the nature and level of measurement of the variables in your study. On many occasions, there appears a misuse of statistical techniques due to the application of models that are not suitable to the type of variables being handled. The paper by Ato and Vallejo (2011) explains the different roles a third variable can play in a causal relationship.

2.2 Instruments

The use of psychometric tools in the field of Clinical and Health Psychology has a very significant incidence and, therefore, neither the development nor the choice of measurements is a trivial task. Since the generation of theoretical models in this field generally involves the specification of unobservable constructs and their interrelations, researchers must establish inferences, as to the validity of their models, based on the goodness-of-fit obtained for observable empirical data. Hence, the quality of the inferences depends drastically on the consistency of the measurements used, and on the isomorphism achieved by the models in relation to the reality modelled. In short, we have three models: (1) the theoretical one, which defines the constructs and expresses interrelationships between them; (2) the psychometric one, which operationalizes the constructs in the form of a measuring instrument, whose scores aim to quantify the unobservable constructs; and (3) the analytical model, which includes all the different statistical tests that enable you to establish the goodness-of-fit inferences in regards to the theoretical models hypothesized.

The theory of psychological measurement is particularly useful in order to understand the properties of the distributions of the scores obtained by the psychometric measurements used, with their defined measurement model and how they interact with the population under study. Psychometric models such as the Classical Test Theory (CTT), Item Response Theory (IRT), or the Generalizability Theory (GT) constitute powerful tools that enable researchers to develop and understand the behaviour of the measurements generated under certain assumptions. This information is fundamental, as the statistical properties of a measurement depend, on the whole, on the population from which you aim to obtain data. The knowledge of the type of scale defined for a set of items (nominal, ordinal, interval) is particularly useful in order to understand the probability distribution underlying these variables. If we focus on the development of tests, the measurement theory enables us to construct tests with specific characteristics, which allow a better fulfilment of the statistical assumptions of the tests that will subsequently make use of the psychometric measurements.

For the purpose of generating articles, in the "Instruments" subsection, if a psychometric questionnaire is used to measure variables it is essential to present the psychometric properties of their scores (not of the test) while scrupulously respecting the aims designed by the constructors of the test in accordance with their field of measurement and the potential reference populations, in addition to the justification of the choice of each test. You should also justify the correspondence between the variables defined in the theoretical model and the psychometric measurements (when there are any) that aim to make them operational. The psychometric properties to be described include, at the very least, the number of items the test contains according to its latent structure (measurement model) and the response scale they have, the validity and reliability indicators, both estimated via prior sample tests and on the values of the study, providing the sample size is large enough. It is compulsory to include the authorship of the instruments, including the corresponding bibliographic reference.

The articles that present the psychometric development of a new questionnaire must follow the quality standards for its use, and protocols such as the one developed by Prieto and Muñiz (2000) may be followed. Lastly, it is essential to express the unsuitability of the use of the same sample to develop a test and at the same time carry out a psychological assessment. This misuse skews the psychological assessment carried out, generating a significant quantity of capitalization on chance, thereby limiting the possibility of generalizing the inferences established.

For further insight, both into the fundamentals of the main psychometric models and into reporting the main psychometric indicators, we recommend reading the International Test Commission (ITC) Guidelines for Test Use (2000) and the works by Downing and Haladyna (2006), Embretson and Hershberger (1999), Embretson and Reise (2000), Kline (2005), Martínez-Arias (2005), Muñiz (1997, 2002, 2010), Olea, Ponsoda, and Prieto (1998), Prieto and Delgado (2010), and Rust and Golombok (2008). All these references have an instructional level easily understood by researchers and professionals.

In the field of Clinical and Health Psychology, the presence of theoretical models that relate unobservable constructs to variables of a physiological nature is really important. Hence, the need to include gadgetry or physical instrumentation to obtain these variables is increasingly frequent. In these situations researchers must provide enough information concerning the instruments, such as the make, model, design specifications, unit of measurement, as well as the description of the procedure whereby the measurements were obtained, in order to allow replication of the measuring process. It is important to justify the use of the instruments chosen, which must be in agreement with the definition of the variables under study.

2.3 Procedure

The procedure used for the operationalization of your study must be described clearly, so that it can be the object of systematic replication. Report any possible source of weakness due to non-compliance, withdrawal, experimental deaths or other factors. Indicate how such weaknesses may affect the generalizability of the results. Clearly describe the conditions under which the measurements were made (for instance, format, time, place, personnel who collected the data, etc.). Describe the specific methods used to deal with possible bias on the part of the researcher, especially if you are collecting the data yourself. Some publications

require the inclusion in the text of a flow chart to show the procedure used. This option may be useful if the procedure is rather complex.

2.4 Power and sample size

Provide the information regarding the sample size and the process that led you to your decisions concerning the size of the sample, as set out in section 1.2 of this article. Document the effect sizes, sampling and measurement assumptions, as well as the analytical procedures used for calculating the power. As the calculation of the power is more understandable prior to data compilation and analysis, it is important to show how the estimation of the effect size was derived from prior research and theories in order to dispel the suspicion that they may have been taken from data obtained by the study or, still worse, they may even have been defined to justify a particular sample size.

In the study by Sesé and Palmer (2012), the analysis of the power (prior and/or observed) was only referred to in 7.8% of the studies reviewed. Meanwhile, the results were presented in the form of confidence interval in 94 of the 498 studies, that is, in 18.9%. From these data, it follows that it is necessary to continue to insist on researchers using these statistical resources, as overlooking them means generating reasonable doubt as to the empirical value of the results.

2.5 Checking Statistical Assumptions

A statistical assumption can be considered a prerequisite that must be fulfilled so that a certain statistical test can function efficiently. Nearly every statistical test poses underlying assumptions so that, if they are fulfilled, these tests can contribute to generating relevant knowledge. But if there is a certain degree of non-fulfilment, the results may lead to distorted or misleading conclusions. It is important to point out that this is not a binary question of fulfilment/non-fulfilment, but rather a question of degree of fulfilment/non-fulfilment.

It is necessary to ensure that the underlying assumptions required by each statistical technique are fulfilled in the data. For instance, Wilkinson (1999) establishes that it is necessary to carry out a good analysis of the results of the statistical model applied. However, an analysis of the literature enables us to see that this analysis is hardly ever carried out.

When the size of the sample increases, and hence the power, sometimes the fulfilment of assumptions is ruled out when actually the degree of non-fulfilment does not have significant effects on the result of the subsequent contrast test (e.g. normality tests). Therefore, whenever possible it is more advisable to plot the analysis of the assumptions on a graph.

It is worth noting that attention must be paid to the underlying assumptions of the statistical method chosen, while simultaneously considering a series of specifications that are crucial to the study, such as the definition of the population, the sampling procedures, the choice or development of measuring instruments, the estimation of power and the determination of sample size or the control of extraneous variables, to name but a few.

Clearly an appropriate analysis of the assumptions of a statistical test will not improve the implementation of a poor methodological design, although it is also evident that no matter how appropriate a design is, better results will not be obtained if the statistical assumptions are not fulfilled (Yang and Huck, 2010). Hence, the study requires an analysis of the fulfilment of the corresponding statistical assumptions, since otherwise the quality of the results may be really jeopardised.

Due to the great importance of checking statistical assumptions as regards the quality of subsequent inferences, take into account the analysis of their fulfilment, even before beginning to collect data. The verification of the assumptions is thereby less likely to be overlooked or treated as an addition with a reactive nature –and not proactive as it should be (Wells and Hintze, 2007). This proactive nature of a prior planning of assumptions will probably serve to prevent possible subsequent weaknesses in the study, as far as decision-making regarding the statistical models to be applied is concerned.

Lastly, it is interesting to point out that some statistical tests are robust in the case of non-fulfilments of some assumptions, in which the distribution of reference will continue to have a behaviour that will enable a reasonable performance of the statistical test, even though there is no perfect fulfilment. Nevertheless, this does not mean it should not be studied. If the degree of non-fulfilment endangers the validity of the estimations, fall back on alternative procedures such as non-parametric tests, robust tests or even exact tests (for instance using bootstrap). This type of tests applied in experimental research, can be consulted in Palmer (2011a, b).

For a review of the underlying assumptions in each statistical test consult specific literature. We recommend that you read the papers by McPherson (1990), Wells and Hintze (2007) and Wilcox (1988). Data collected in the study by Sesé and Palmer (2012) regarding articles published in the field of Clinical and Health Psychology indicate that assessment of assumptions was carried out in 17.3% the cases, a figure far from what should be desirable.

3. Analysis and Results

3.1 Previous (missing, protocol violations, etc.)

Before presenting the results, comment on any complications, non-fulfilment of protocol, and any other unexpected events that may have occurred during the data collection. This includes missing values, withdrawals, or non-responses. Discuss the analytical techniques used to minimize these problems, if they were used. Describe statistical non-representation, informing of the patterns and distributions of missing values and possible contaminations. Document how the analyses carried out differ from the analyses that were proposed before the appearance of complications. Use techniques to ensure that the results obtained are not produced by anomalies in the data (for instance, outliers, influencing points, non-random missing values, selection biases, withdrawal problems, etc.), as a standard behaviour in every analysis. All these variations can undermine the validity of the study and, therefore, it is essential to refer to them in the text so that the reader can assess the degree of influence on the inferences established.

3.2 Selection of Statistical Techniques

The determination of a suitable statistical test for a specific research context is an arduous task, which involves the consideration of several factors:

1) Questions/hypotheses of the research study
2) Typology of independent or explanatory variables and dependent or response variables.
3) Research design

These factors condition decision-making regarding the identification of a set of possible appropriate statistical techniques. The huge variety of modern quantitative methods places researchers in the nontrivial situation of fitting the techniques and the design to the research questions. Although complex designs and novel methods are sometimes necessary, in order to efficiently direct studies simpler classical approaches may offer sufficient, elegant answers to important issues.

When it comes to creating a study, it is not a question of choosing a statistical method in order to impress readers or, perhaps, to divert

possible criticism as to the fundamental issues under study. If the assumptions and the power of a simpler method are reasonable for handling the data and the research issue, you should not hesitate to use it. The principle of parsimony (Occam's razor) should not only be applied to the formulation of theories, but also to the application of statistical methodology.

When it comes to describing a data distribution, do not use the mean and variance by default for any situation. These are non-resistant indices and are not valid in non-symmetrical distributions or with the presence of outliers. In these cases use a resistant index (e.g., an M-estimator). You can consult Palmer (1993), Palmer (1999), Palmer, Beltrán, and Cortiñas (2006), and Cajal, Gervilla, and Palmer (2011).

In the study by Sesé and Palmer (2012) it was found that the most used statistical procedure was Pearson's linear correlation coefficient. Therefore, we will make some reflections concerning this coefficient.

3.3 Difference between statistical interpretation and practical interpretation

Obtaining a significant correlation is not the same as saying that the existing relationship between variables is important at a practical or clinical level. This is so, among other reasons, because the significance of the correlation coefficient depends on the size of the sample used in such a way that with large sample sizes, low correlation coefficients become significant, as shown in the following table (Palmer, 2011a) which relates these elements.

	n-2	8	10	20	30	50	100	200	500	1000	2000
Significant to .05	r	.71	.58	.42	.35	.27	.20	.14	.09	.06	.04
	r^2	.50	.33	.18	.12	.07	.04	.02	.01	.004	.002

From the above table it can be observed that if, for instance, there is a sample of 202 observations, a correlation coefficient of .14 is significant at 5%, despite the fact that the proportion of variance explained is only 2%, which may turn out to be of very little practical relevance.

Thus, we must not confuse statistical significance with practical significance or relevance. Likewise, we must not confuse the degree of significance with the degree of association. On the other hand, this example does allow us to understand that a very large sample size enables us to obtain statistical significances with very low values, both in terms of relationship and association.

Lastly, it is very important to point out that a linear correlation coefficient equal to 0 does not imply there is no relationship. It is often frequent, on obtaining a non-significant correlation coefficient, to conclude that there is no relationship between the two variables analysed. At the risk of abusing language, it goes without saying that there is no *linear* relationship between the variables, which does not mean that these two variables cannot be related to each other, as their relationship could be *non-linear* (e.g. quadratic, exponential, etc.). In this sense, it is always recommended, prior to the estimation of models, to analyse the scatterplot of the variables involved.

3.4 Statistical software

There are many very good programmes for analysing data. However, verifying the results, understanding what they mean, and how they were calculated is more important than choosing a certain statistical package. If results cannot be verified by using approximate calculations, they should be verified by triangulating with the results obtained using another programme. Do not fail to report the statistical results with greater accuracy than that arising from your data simply because this is the way the programme offers them. An example of this misuse happens when, on obtaining a small probability for a contrast test it is referred to in the text as p = .000 (generally, exactly the way it appears in the outputs of the programme), when actually it should be recorded as p < .001, or still better, provide the exact value in scientific notation, which can be achieved through the statistical package used itself, or through statistical tables available on the Internet.

Using a computer is an opportunity to control your methodological design and your data analysis. If a programme does not implement the analysis needed, use another programme so that you can meet your analytical needs, but do not apply an inappropriate model just because your programme does not have it. Never assume that by using a highly recommendable, sound programme you are acquitted of the responsibility of judging whether its results are plausible. Two obvious things concerning this: if a certain statistical programme does not implement a certain calculation, it does not mean that this calculation does not exist; and remember that you are the one doing the statistical analysis, not the statistical programme.

When you document the use of a technique, do not only include the reference of the programme handbook, but the relevant statistical literature related to the model you are using. As opposed to the commercial software normally used (SPSS, SAS, Stata, Splus, Minitab, Statistica, etc.) we recommend free software (R, Weka, etc.) and specifically the use of the R programme. Paraphrasing the saying, "What is not in the Internet, it does not exist", we could say, "What cannot be done with R, cannot be done". It is necessary for you to specify the programme, or programmes, that you have used for the analysis of your data. In the work by Sesé and Palmer (2012), SPSS was undoubtedly the most used statistical package, whereas only 0.6% of the studies reviewed used R. Nowadays, there is a large quantity of books based on R which can serve as a reference, such as Cohen and Cohen (2008), Crawley (2007), Ugarte, Militino and Arnholt (2008) and Verzani (2005).

3.5 Hypothesis tests

It is about time we started to banish from research the main errors associated with the limitations of the NSHT. On the whole, we can speak of two fundamental errors:

1) The lower the probability value p, the stronger the proven relationship or difference, and
2) Statistical significance implies a theoretical or substantive relevance.

This inappropriate use remains more widespread than expected in current psychological research (Gliner, Leech & Morgan, 2002), despite the efforts some authors have devoted to minimizing it (Cohen, 1994; Mulaik, Raju & Harshman, 1997). Kirk (1996) explains that NHST is a trivial exercise as the null hypothesis is always false, and rejecting it clearly depends on having sufficient statistical power. Therefore, with a large enough sample size, practically any pair of variables will show a significant relationship (remember the example explained above regarding linear correlation) or differ significantly.

The purpose of scientific inference is to estimate the likelihood that the null hypothesis (H_0) is true, provided a set of data (n) has been obtained, that is, it is a question of conditional probability $p(H_0|D)$. Nevertheless, what the NHST procedure really offers us is the likelihood of obtaining these or more extreme data if the null hypothesis is true, that is, the opposite conditional probability $p(D|H_0)$. For this reason, "acceptance" of the null hypothesis should never be expressed, thus it is either rejected or not.

In order to avoid the effects of this confusion between statistical significance and practical relevance, it is recommended that if the measurement of the variables used in the statistical tests is understandable confidence intervals are used. If, on the other hand, the units of measurement used are not easily interpretable, measurements regarding the effect size should be included. In a

combined way, it is possible to provide the confidence intervals regarding the effect sizes (Steiger & Fouladi, 1997).

The most important thing is to be clear on the fact that when applying a statistical test a decision to "reject" the null hypothesis, by itself, is not indicative of a significant finding (Huck, 2000, p. 199). Since this malpractice has even been condemned by the Task Force on Statistical Inference (TFSI) of the American Psychological Association (APA) (Wilkinson, 1999), it is absolutely essential that researchers do not succumb to it, and reviewers do not issue favourable reports of acceptance for works that include it. For a more in-depth view of the issue, you can look, among others, at the works of Chow (1996), Cohen (2010), Mittag and Thompson (2000), and Nickerson (2000) and, in our context, those by Balluerka, Vergara, and Arnau (2009); Borges, San Luis, Sánchez-Bruno, and Cañadas (2001), and Monterde, Pascual, and Frías (2006).

3.6 Effect Sizes Estimation

One of the main ways to counter NHST limitations is that you must always offer effect sizes for the fundamental results of a study. If the units of measurements are significant at a practical level (for instance, number of cigarettes smoked in a day), then a non-standardised measurement is preferable (regression coefficient or difference between means) to a standardized one (f^2 o d). It is extremely important to report effect sizes in the context of the extant literature. This context analysis enables researchers to assess the stability of the results through samples, designs and analysis. If you include the effect sizes in your articles, they can be used in the future for meta-analytical studies. To go further into the analysis of effect sizes, you can consult Rosenthal and Rubin (1982), Cohen (1988), Cohen (1994), or Rosenthal, Rosnow, and Rubin, (2000).

According to the bibliometric study of Sesé and Palmer (2012), in 304 of the 498 possible studies a measurement of the effect size was provided, that is, in 61% of cases. The most used effect size, in all the journals analysed, was the R square determination coefficient (33.9%), followed by the partial square and by some Cohen effect size measurements. Even though these results do not pose a negative scenario, they clearly leave room for improvement, such that reporting the effect size becomes a habit, which is happening as statistical programmes include it as a possible result.

3.7 Interval estimates

A confidence interval (CI) is given by a couple of values, between which it is estimated that a certain unknown value will be found with a certain likelihood of accuracy. In a formal way, it is calculated from the data of a sample concerning an unknown population parameter following a certain theoretical distribution. The likelihood of success in the estimation is represented as 1-alpha and is called confidence level. The width of the interval depends fundamentally on the inverse sample size, that is, a narrower CI will be obtained and therefore a more accurate estimate (lower error), the larger the sample size.

CIs should be included for any effect size belonging to the fundamental results of your study. It is even necessary to include the CI for correlations, as well as for other coefficients of association or variance whenever possible. Normally the estimation of the CI is available in most of the statistical programmes in use. It is also important to highlight the CI of previous research, in order to be able to compare results in such a way that it is possible to establish a more profound analysis of the situation of the parameters. For a more in-depth view, read for instance Schmidt (1996).

3.8 Number of comparisons

The analysis of the hypotheses generated in any design (inter, block, intra, mixed, etc.) requires the use of contrasts. It is essential to distinguish the contrasts "a priori" or "a posteriori" and in each case use the most powerful test. There are manifold comparison tests (Dunn with the Bonferroni or Sidak correction, Holm, Tukey, Dunnett, Scheffé, etc.) and it is necessary in each case to use the one with the maximum potential to be able to discover the difference, if it exists. Likewise, bear in mind the fulfilment or not of the assumption of homogeneity of variance when it comes to choosing the appropriate test. You will find extensive information on this issue in Palmer (2011a).

In the research by Sesé and Palmer (2012) it was found that only in 11% of the studies, in which some type of design was used, contrast analyses were carried out. The use of contrasts to assess hypotheses is fundamental in an experimental study, and this analysis in a study with multiple contrasts requires special handling, as otherwise the Type 1 error rate can rise significantly, i.e., the likelihood of rejecting the null hypothesis when it is true increases in the population. Thus, it is the responsibility of the researcher to define, use, and justify the methods used. The texts of Palmer (2011b, c, d) widely address this issue.

Hence for instance, when all the existing correlations between a set of variables are obtained it is possible to obtain significant correlations simply at random (Type I error), whereby, on these occasions, it is essential to carry out a subsequent analysis in order to check that the significances obtained are correct. If the sample is large enough, the best thing is to use a cross-validation through the creation of two groups, obtaining the correlations in each group and verifying that the significant correlations are the same in both groups (Palmer, 2011a).

3.9 Causality

Inferring causality from non-randomised designs can be a risky enterprise. Researchers who use non-randomised designs incur an extra obligation to explain the logic the inclusion of co-variables follows in their designs and to alert the reader to possible alternative hypotheses that may explain their results. Even in randomized experiments, attributing causal effects to each of the conditions of the treatment requires the support of additional experimentation. Statistical technique never guarantees causality, but rather it is the design and operationalization that enables a certain degree of internal validity to be established.

In a non-experimental context, as is the case of selective methodology, and related with structural equation models (SEM), people make the basic mistake of believing that the very estimation of an SEM model is a "per se" empowerment for inferring causality. This has been helped by the fact that, in the literature, these models have been labelled "causal" models. However, the possibility of inferring causality from a model of structural equations continues to lie in the design methodology used. We would like to reiterate that it is not the technique that confers causality, but rather the conditions established by the research design to obtain the data. For a more in-depth look, you can consult the works of Cheng (1997) and Griffiths and Tenenbaum (2005).

3.10 Tables and Figures

Although tables are used to present the exact results of the statistical models estimated, well-designed figures should not be exempt from preciseness. Figures attract the readers' eye and help transmit the overall results. Since as subjects we have different ways of processing complex information, the inclusion of tables and figures often helps. This works better when the figures are small enough to leave enough room for both formats. Complex figures should be avoided when simple ones can represent relevant information adequately. Neither should a scientific graph be converted into a commercial diagram. Avoid three dimensions when the information being transmitted is two-dimensional. Remember to

include the confidence intervals in the figures, wherever possible. For a good development of tables and figures the texts of Everett (2000), Tufte (2001), and Good and Hardin (2003) are interesting.

4. Discussion

4.1 Interpretation

When effects are interpreted, try to analyse their credibility, their generalizability, and their robustness or resilience, and ask yourself, are these effects credible, given the results of previous studies and theories? Do the data analysed in the study, in accordance with the quality of the sample, similarity of design with other previous ones and similarity of effects to prior ones, suggest they are generalizable? Are the designs and analytical methods robust enough to generate powerful conclusions? The appropriate answer to these questions, well fitted to reality, means you have achieved a good interpretation of the empirical results obtained.

Avoid making biased interpretations such as, for instance when faced with a probability value associated to a contrast of hypothesis concerning the comparison of two means whose value was .052 (statistically non-significant) you interpret it as if it were by resorting to the fact that "despite not being significant there is a tendency towards difference". At any rate, it is possible to resort to saying that in your sample no significance was obtained but this does not mean that the hypothesis of the difference being significantly different to zero in the population may not be sufficiently plausible from a study in other samples. Think that the validity of your conclusions must be grounded on the validity of the statistical interpretation you carry out.

4.2 Conclusions

Do not conclude anything that does not derive directly and appropriately from the empirical results obtained. The quality of your conclusions will be directly related to the quality obtained from the data analysis carried out. You can use speculation, but it should be used sparsely and explicitly, clearly differentiating it from the conclusions of your study. If the results have partially satisfied your hypotheses, do not conclude part of it as if it were the whole. Do not try to maximize the effect of your contribution in a superficial way either. You must help the reader to value your contribution, but by being honest with the results obtained. Do not interpret the results of an isolated study as if they were very relevant, independently from the effects contributed by the literature. The results of one study may generate a significant change in the literature, but the results of an isolated study are important, primarily, as a contribution to a mosaic of effects contained in many studies.

It also helps in this task to point out the limitations of your study, but remember that recognising the limitations only serves to qualify the results and to avoid errors in future research. This sort of confession should not seek to dismantle possible critiques of your work. Recommendations for future studies should be very well drawn up and well founded in the present and on previous results. Gratuitous suggestions of the sort, "further research needs to be done…" only take up space and are more than obvious. Therefore, refrain from including them.

Finally, we would like to highlight that currently there is an abundant arsenal of statistical procedures, working from different perspectives (parametric, non-parametric, robust, exact, etc.). On each occasion, choose the most powerful procedure. Do not allow a lack of power to stop you from discovering the existence of differences or of a relationship, in the same way as you would not allow the non-fulfilment of assumptions, an inadequate sample size, or an inappropriate statistical procedure to stop you from obtaining valid, reliable results. Meanwhile, do not direct your steps directly towards the application of an inferential procedure without first having carried out a comprehensive descriptive analysis through the use of exploratory data analysis. You can consult, to this end, the text by Palmer (1999).

By way of summary

The basic aim of this article is that if you set out to conduct a study you should not overlook, whenever feasible, the set of elements that have been described above and which are summarised in the following seven-point table:

1. Choose the most appropriate design methodology according to the nature of the object under study.
2. Develop a procedure of data collection that will optimise the quality of the measurements and the representativeness of the empirical model to be compared.
3. Obtain a large enough sample size so as to reach an appropriate level of power.
4. Choose the most powerful statistical tests, according to the nature of the variables that will make up the analysis and the intended aim.
5. Analyse the fulfilment of the statistical assumptions underlying the statistical models to be estimated.
6. Use the most appropriate statistical software to analyse your aims and the one that will enable you to obtain the maximum amount of information.
7. Provide the effect size and the confidence intervals obtained.

To finish, we echo on the one hand the opinions Hotelling, Bartky, Deming, Friedman, and Hoel (1948) expressed in their work *The teaching statistics*, in part still true 60 years later: "Unfortunately, too many people like to do their statistical work as they say their prayers – merely substitute a formula found in a highly respected book written a long time ago" (p. 103); and, on the other hand, we also echo what was expressed by McCloskey (1996) regarding investigative judgement: "Focusing on the calculation, if it causes us to forget this obvious human job of making human judgements, it is going to make us forget what we are doing".

Conflicts of Interest

The auhors declare that they have no conflicts of interest.

References

Ato, M., & Vallejo, G. (2011). Los efectos de terceras variables en la investigación psicológica. *Anales de Psicologia*, *27*, 550-561.

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.

Borges, A., San Luis, C., Sánchez-Bruno, A., & Cañadas, I. (2001). El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa. *Psicothema, 13*, 173-178.

Cajal, B., Gervilla, E., & Palmer, A. (2011). When the mean fails, use an M-estimator. *Anales de Psicologia*, *28*, 281-288.

Cheng, P.W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review, 104*, 367-405.

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

Cochran, W.G. (1986). *Técnicas de muestreo (6ª Ed.)*. Mexico: Ed. Continental.

Cohen, B.H. (2010). *Null Hypothesis Significance Testing*. Corsini Encyclopedia of Psychology. 1–2.

Cohen, J. (1988, 2ªed). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Cohen, Y., & Cohen, J.Y. (2008). *Statistics and data with R*. New Jersey: John Wiley and Sons.

Crawley, M. J. (2007). *The R book*. New Jersey: John Wiley and Sons.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S.E., & Hershberger, S.L. (1999). *The new rules of measurement: What every psychologist and educator should know*. New York: Taylor Francis.

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum Publishers.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1-11.
Everett, G. (2000). *How to lie with charts*. Lincoln: Authors Choice Press.
Fiona, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics, 33*, 615-630.
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.
Good, P., & Hardin, J. (2003). *Common errors in statistics (and how to avoid them)*. New Jersey: John Wiley and Sons.
Griffiths, T.L., & Tenenbaum, J.B. (2005). Strength and structure in causal induction. *Cognitive Psychology, 51*, 334-384.
Hill, C.R., & Thomson, B. (2004). Computing and interpreting effects sizes. In J.C. Smart, J.C. (Ed.), *Higher Education: Handbook of Theory and Research (pp.175-196)*. Kluwer: New-York.
Hotelling, H., Bartky, W., Deming, W.E., Friedman, M., & Hoel, P. (1948). The teaching of statistics. *Annals of Mathematical Statistics, 19*, 95-115.
Huck, S.W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.
International Test Commission (ITC) (2000). *International Guidelines for Test Use*. Downloaded electronically on 15/06/2011 from www.intestcom.org/itc_projects.htm
Kline, T. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation*. London: Sage.
Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current directions in psychological science, 5*, 161-171.
Martínez-Arias, R. (2005). *Psicometría: Teoría de los tests psicológicos y educativos.* Madrid: Síntesis.
McPherson, G. (1990). *Statistics in Scientific Investigation: Its Basis, Application and Interpretation*. New York: Springer-Verlag.
Mittag, K.C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*, 14-20.
Monterde, H., Pascual, J., & Frías, D. (2006). Errores de interpretación de los métodos estadísticos: importancia y recomendaciones. *Psicothema, 18*, 848-856.
Montero y León (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology, 7*, 847-862.
Mulaik, S.A., Raju, N.S., & Harshman, R.A. (1997). There is a time and place for significance testing. In L.L. Harlow, S.A. Mulaik and J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–116). Mahwah, NJ: Erlbaum.
Muñiz, J. (1997). *Introducción a la Teoría de la Respuesta a los Ítems*. Madrid: Pirámide.
Muñiz, J. (2002). *Teoría Clásica de los Tests (2nd ed.)*. Madrid: Pirámide.
Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo, 31*, 57-66.
Nickerson, R.S. (2000). Null Hypothesis Significance Tests: A Review of an Old and Continuing Controversy. *Psychological Methods, 5*, 241–301.
Olea, J., Ponsoda, V., & Prieto, G. (Eds.). (2008). *Tests informatizados: Fundamentos y aplicaciones*. Madrid: Pirámide.
Palmer, A. (1993). M-estimadores de localización como descriptores de las variables de consumo. *Adicciones, 5(2)*, 171-184.
Palmer, A. (1999). *Análisis de datos. Etapa exploratoria*. Madrid: Ed. Pirámide.
Palmer, A. (2011a). *Procedimientos estadísticos con SPSS y R para la comparación de dos medias.* Palma de Mallorca: Edicions UIB.
Palmer, A. (2011b). *Análisis unifactorial de la variancia con SPSS y R.* Palma de Mallorca: Edicions UIB.
Palmer, A. (2011c). *Análisis de datos en el diseño unifactorial de medidas repetidas*. Madrid: Ed. La Muralla. Colección Cuadernos de Estadística, 39.
Palmer, A. (2011d). *Análisis de datos en diseños experimentales.* Palma de Mallorca: Edicions UIB.
Palmer, A., Beltrán, M., & Cortiñas, P. (2006). Robust estimators and bootstrap confidence intervals applied to tourism spending. Tourism Management 27(1), 42-50.
Prieto, G., &Delgado, A. (2010). Fiabilidad y Validez. *Papeles del Psicólogo, 31*, 67-74.
Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77*, 65-71.
R Development Core Team (2012). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
Rosenthal, R., & Rubin, D.B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.
Rosenthal, R., Rosnow, R.L., & Rubin, D.B. (2000). *Contrasts and effect sizes in behavioural research: A correlational approach.* New York: Cambridge University Press.
Rust, J., & Golombok, S. (2008). *Modern Psychometrics: The Science of Psychological Assessment*.
Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*, 115-129.
Sesé, A., & Palmer, A. (2012). The current use of statistics in clinical and health psychology under review. *Clínica y Salud 23(1)*, 97-108.
Steiger, J.H., & Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A., Mulaik, and J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–258). Mahwah, NJ: Erlbaum.
Thompson, S.K. (2012). *Sampling (3rd Ed.)*. New York: Wiley.
Treat, T.A., & Weersing, V.R. (2005). Clinical Psychology. In B.S. Everitt and D.C. Howell, *Encyclopedia of Statistics in Behavioral Science*. New York John Wiley and sons.
Tufte, E.R. (2001). *The visual display of quantitative information*. (2nd ed.). Cheshire: Graphics Press.
Ugarte, M. D., Militino, A. F., & Arnholt, A. T. (2008). *Probability and Statistics with R*. New York: CRC Press.
Verzani, J. (2005). *Using R for introductory statistics*. New York: Chapman and Hall/CRC Press.
Wells, C.S., & Hintze, J.M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools, 44*, 495-502.
Wilcox, R.R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*, 300-314.
Wilkinson, L. (1999). Statistical methods in Psychology Journals: Guidelines and Explanations. *American Psychologist, 54*, 594-604.
Yang, H., & Huck, S.W. (2010). The importance of attending to underlying statistical assumptions. *Newborn & Infant Nursing Reviews, 10*, 44-49.