Methodological and Didactical Controversies around Statistical Inference

Carmen Batanero and Carmen Díaz

Methodological and Didactical Controversies around Statistical Inference
Conference invite, 38émes Journées de Statistique, Societé Française de Statistique,
Paris, Clamart, 2006

Abstract

The incorrect interpretations and researchers' excessive confidence in statistical inference have lead to intensive criticism by different professional organisations that have had scarce impact on the didactical practices at University level. In this paper we analyse this situation and conclude with some reflections about the teaching of statistical inference to undergraduates.

Empirical sciences rely heavily on establishing the existence of effects using the statistical analysis of data. Statistical inference dates back almost 300 years. However, since the logic of statistical inference is difficult to grasp, its use and interpretation are not always adequate and have been criticized for nearly 50 years (for example, in Yates, 1951). Extensive review of criticisms against significance testing can be found in Morrison and Henkel (1970) and Harlow, Mulaik, and Steiger (1997).

This controversy has increased in the past ten years within professional organizations (Menon, 1993; Ellerton, 1996; Levin, 1998; Levin & Robinson, 1999; Robinson & Levin, 1997; Ares, 1999; Glaser, 1999; Wilkinson, 1999; Batanero, 2001; Fidler, 2002; López, 2003; Kline, 2004) which are suggesting important shifts in their editorial policies regarding the use of statistical significance testing.

Despite the arguments that statistical tests are not adequate to justify scientific knowledge, researchers persist in relying on statistical significance (Hager, 2000; Borges, San Luis, Sánchez & Cañadas, 2001; Finch, Cumming, y Thomason, 2001). Some explanations for this persistence include inertia, conceptual confusion, lack of better alternative tools, and psychological mechanisms such as invalid generalization from deductive logic to inference under uncertainty (Falk & Greenbaum, 1995). In this paper we summarise some of the problems that were analysed in Batanero (2000) and Díaz and de la Fuente, (2004). and finally suggest possible ways in which statistical education might contribute to the better understanding and application of statistical inference.

Common Errors in Interpreting Statistical Tests

Misconceptions related to statistical tests refer mainly to the level of significance, α, which is defined as the probability of rejecting a null hypothesis, given that it is true. The most common misinterpretation of this concept consists of switching the two terms in the conditional probability, that is, interpreting the level of significance as the probability that the null hypothesis is true, once the decision to reject it has been taken. For example, Birnbaum (1982) reported that his students found the following definition reasonable: "*A level of significance of 5% means that, on average, 5 out of every 100 times we reject the null hypothesis, we will be wrong*". Falk (1986) found that most of her students believed that α was the probability of being wrong when rejecting the null hypothesis at a significance level α. Similar results were described in Pollard and Richardson (1987), Lecoutre, Lecoutre and Poitevineau (2001) and Haller and Krauss (2002) in their study using researchers.

Another common error is the belief in the conservation of the significance level value when successive tests are carried out on the same data set, which produces the problem of multiple comparisons. If we carry out 100 comparisons on the same data set using in all of them a level of significance .05 it is expected that about 5 out of the 100 tests will be significant just by chance, even when the null hypothesis is true. This makes it difficult to interpret the results (Moses, 1992).

Some people believe that the p-value is the probability that the result is due to chance. That this is a misconception can be deduced from the fact that, even when the null hypothesis is true a significant result might be due to other factors. For example, in testing the differences between two groups of students, the students in the experimental group might work harder than their counterparts to prepare for the test. Here we can see the relevance of experimental control to try to ensure that all the conditions are held constant in the two groups. The p-value is the probability of obtaining the particular result or one more extreme when the null hypothesis is true *and* there are no other possible factors influencing the result. What is rejected in a statistical test is the null hypothesis, and therefore we cannot infer the existence of a particular cause in an experiment from a significant result.

Another erroneous belief is that the .05 and .01 levels of significance are justified by mathematical theory. In his book “Design of Experiments”, Fisher (1935) suggested selecting a significance level of 5% as a convention to recognize significant results in experiments. In later writings, however, Fisher considered that every researcher should select the significance level according to the circumstances, stating that "in fact, no scientific worker has a fixed level of significance at which from year to year and in all circumstances, he rejects hypotheses" (Fisher, 1956, p. 42). Instead, Fisher suggested publishing the exact p-value obtained in each particular experiment which, in fact, implies establishing the significance level after the experiment. In spite of these recommendations, research literature shows that the common arbitrary levels of .05, .01, .001 are almost universally selected for all types of research problems and are sometimes used as criteria for publication.

Misinterpretations of the significance level are linked to misinterpreting significant results, about which there was another disagreement between Fisher and Neyman and Pearson. A significant result, for Fisher, implied that the data provided evidence against the null hypothesis, while for Neyman and Pearson it just stated the relative frequency of times that we would reject a true null hypothesis (the Type I error) in the long run. On the other hand, we should distinguish between statistical and practical significance, since we might have obtained a higher level of significance with a smaller experimental effect and a larger sample size. Practical significance involves statistical significance plus a sufficiently large experimental effect.

Philosophical and Psychological Issues

In fact there are several reasons that explain the difficulties in understanding statistical tests. On one hand, statistical tests involve a series of concepts such as null and alternative hypotheses, Type I and Type II errors, probability of errors, significant and non significant results, population and sample, parameter and statistics, sampling distribution. Some of these concepts are misunderstood or confused by students and experimental researchers.

Moreover, the formal structure of statistical tests is superficially similar to that of proof by contradiction. However, there are fundamental differences between these two types of reasoning that are not always well understood. In proof by contradiction we reason in the following way: If A implies B cannot happen, then, if B happens, we deduce A is false. In statistical testing, it is tempting to apply similar reasoning as follows: If A implies B is very unlikely to happen. However, this does not imply that if B happens, A is very unlikely and herein lies the confusion.

The controversy surrounding statistical inference involves the philosophy of inference and the logical relations between theories and facts. We expect from statistical testing more than it can provide us, and underlying this expectation is the philosophical problem of finding scientific criteria to justify inductive reasoning, as stated by Hume. The contribution made by statistical inference in this direction is important but it does not give a complete solution to this problem (Hacking, 1975; Seidenfeld, 1979, Cabria, 1994).

On the other hand, there are two different views about statistical tests that sometimes are confused or mixed. Fisher saw the aims of significance testing as confronting a null hypothesis with observations and for him a p-value indicated the strength of the evidence against the hypothesis (Fisher, 1958). However, Fisher did not believe that statistical tests provided inductive inferences from samples to population, but, rather, a deductive inference from the population of possible samples to the particular sample obtained in each case.

For Neyman (1950), the problem of testing a statistical hypothesis occurs when circumstances force us to make a choice between two courses of action. To accept a hypothesis means only to decide to take one action rather than another. This does not mean that one necessarily believes that the hypothesis is true. For Neyman and Pearson, a statistical test is a rule of inductive behaviour; a criterion for decision-making, which allows us to accept or reject a hypothesis by assuming some risks.

The dispute between these authors has been hidden in applications of statistical inference in psychology and other experimental sciences, where it has been assumed that there is only one statistical solution to inference (Gingerenzer et al, 1989), Today, many researchers apply the statistical tools, methods, and concepts of the Neyman-Pearson theory with a different aim, namely, to measure the evidence in favour of a given hypothesis (Royal, 1997). Therefore, the current practice of statistical tests contains elements from Neyman-Pearson (it is a decision procedure) and from Fisher (it is an inferential procedure, whereby data are used to provide evidence in favor of the hypothesis), which apply at different stages of the process. We should also add that some researchers often give a Bayesian interpretation to the result of (classical) hypothesis tests, in spite of the fact that the view from Bayesian statistics is very different from the theories of either Fisher or Neyman and Pearson.

The above practice of statistical tests can explain the belief that statistical inference provides an algorithmic solution to the problem of inductive inference, and the consequent mechanistic behavior that is frequently displayed in relation to statistical tests (Gingerenzer, 1993). Moreover, biases in inferential reasoning can be seen simply as examples of adults' poor reasoning in probabilistic problems (Nisbett & Ross, 1980; Kahneman, Slovic, & Tversky, 1982). In the specific case of misinterpreting statistical inference results, Falk and Greenbaum (1995) describe *the illusion of probabilistic proof by contradiction*, which consists of the erroneous belief that one has rendered the null hypothesis improbable by obtaining a significant result. This illusion of probabilistic proof by contradiction is, however, apparently difficult to eradicate, in spite of clarification in many statistics textbooks. In other cases, this misconception is implicit in textbooks.

Misconceptions around the significance level are also related to difficulties in discriminating between the two directions of conditional probabilities, otherwise known as *the fallacy of the transposed conditional* (Diaconis and Friedman, 1981), which have been long recognized as pervasive among students and even professionals. Although α is a well defined conditional probability, the expression "Type I error" is not conditionally phrased, and does not spell out to which combination of the two events it refers. This leads us to interpret the significance level as the conjunction of the two events "the null hypothesis is true" and "the null hypothesis is rejected" (Menon, 1993).

The Statistical Tests Controversy

For many years, criticisms have been raised against statistical testing, and many suggestions have been made to eliminate this procedure from academic research. However, significant results continue to be published in research journals, and errors around statistical tests continue to be spread throughout statistics courses and books, as well as in published research. An additional problem is that other statistical procedures suggested to replace or complement statistical tests (such as confidence intervals, measuring the magnitude of experimental effects, power analysis, and Bayesian inference) do not solve the philosophical and psychological problems we have described (see Fidler, 2002; Cumming, Williams, & Fidler, 2004). Below we revisit some frequent criticisms that either are not justified or refer to researchers' use of statistical tests more than to the procedure itself.

Criticism 1. *The null hypothesis is never true and therefore statistical tests are invalid, as they are based on a false premise (that the null hypothesis is true).*

This criticism is not pertinent. because what is asserted in a test is that a significant result is improbable, given that the null hypothesis is true. This is a mathematical property of the sampling distribution that has nothing to do with the truth or falsity of the null hypothesis.

Criticism 2. *Statistical significance is not informative about the practical significance of the data, since the alternative hypothesis says nothing about the exact magnitude of the effect.*

In significance testing (Fisher's approach) the aim of experimental research is directed towards theory confirmation in providing support for a substantive hypothesis and the magnitude of effect is not so important. In the context of taking a decision (Neyman-Pearson), however, the magnitude of the effect could be relevant to the decision. In these cases, the criticism applies and statistical tests should be complemented with power analysis and/ or estimates of the magnitude of the effects (Levin, 1998; Frías, Pascual & García, 2000; Vacha-Haase, 2001).

Criticism 3. The choice of the level of significance is arbitrary; therefore some data could be significant at a given level and not significant at another different level.

It is true that the researcher chooses the level of significance. This arbitrariness does not, however, mean that the procedure is invalid. Moreover, it is also possible, following the approach of Fisher, to use the exact p-value to reject the null hypothesis at different levels, though in the current practice of statistical testing it is advisable to chose the significance level before taking the data to give more objectivity to the decision.

Criticism 4. Statistical significance is not informative as to the probability of the hypothesis being true. Nor is statistical significance informative of the true value of the parameter.

Even when tests are not informative of the probability of the hypothesis being true or the probability of replication confidence intervals are not informative of this probability either (Sohn, 1998), the posterior probability of the null hypothesis, given a significant result, depends on the prior probability of the null hypothesis, as well as on the probabilities of having a significant result given the null and the alternative hypotheses. These probabilities cannot be determined in classical inference. It is only within Bayesian inference that posterior probability of the hypotheses can be computed, although these are subjective probabilities. What we can do at best, and using Bayesian inference, is to revise our personal degree of belief in the hypothesis, in view of the result (Cabria, 1994; Lecoutre, 1999; 2006).

Criticism 4. Type I error and Type II errors are inversely related. Researchers seem to ignore Type II errors while paying undue attention to Type I error.

Though the probabilities of the two types of errors are inversely related, there is a fundamental difference between them. While the probability of Type I error α is a constant that can be chosen before the experiment is done, the probability of Type II error is a function of the true value of the parameter which is unknown. To solve this problem, power analysis assumes different possible values for the parameter and computes the probability of Type II error for these different values.

Teaching and Learning Inference Concepts

The above discussion suggests an educational problem, since most professionals had taken at least a Statistics course along his/her undergraduate education. To add complexity to the situation, statistics is taught at University level by lecturers with a variety of backgrounds, the majority of whom are statisticians, but that also includes economists, health care professionals, engineers, psychologists or educators, and very rarely mathematicians or mathematics educators. Some of these lecturers may transmit incorrect conceptions to their students. Moreover, education is for these lecturers only a secondary research field, what explains the fact that, research in advanced stochastic teaching and learning is still scarce so that learning difficulties in advanced statistics are still not well known.

Statistical educators are not indifferent to these problems, since in fact our conceptions about statistical inference and how it should be used in applied research also affect the way we teach statistics. Below we describe some challenges that we have to face in order to improve the teaching and practice of statistics.

Setting the teaching of statistics in a wider context

At University level, statistics is studied mainly as a tool to solve problems in other fields such as education, geography or medicine. These service courses, however, often emphasize the teaching of formulas for calculating statistics (e.g. correlation coefficients or confidence intervals) without much concern towards the data context or interpretative activities. In other cases, the courses are over-mathematised for students who often meet concepts of advanced stochastic without any prior experience of advanced algebra or calculus. As a consequence, many students after these courses are able to manipulate definitions and algorithms with apparent competence, but they do not know what statistical procedure to apply when they face a real problem of data analysis (Pimenta, 2006).

Statistical inference is just a part of the more general process of scientific inference. However, we frequently teach statistics in isolation without connecting it with a more general framework of research methodology and experimental design. From our point of view, it is necessary to discuss the role of statistics within experimental research with the students and make them conscious of the possibilities and limitations of statistics in experimental work. Since statistics is not a way of doing, but a way of thinking that helps us to solve problems in science and everyday life, teaching statistics should begin with real problems through which students develop their ideas, working through the different stages of solving a real problem (planning a solution, collecting and analyzing data, checking initial hypotheses, and taking appropriate decisions).

Different orientation for different students

The controversy around inference also influences changes in the role given to probability within the curriculum, from being the central core to trying to teach statistics without resort to probability (focus on exploratory data analysis only) or favouring classical, Bayesian, computer-intensive (resampling methods) or mathematical-abstract approaches to inference. Each of these approaches might be better suited for a particular type of student and we need to find the best ways to introduce a given approach.

We should also concentrate on clarifying why current teaching of statistics does not improve stochastic intuition and think of some alternative ways, for example including some elements of psychology or philosophy in this teaching. *"In any case statistics should be taught in conjunction with material on intuitive strategies and inferential errors"... "It seems to us that this would have the advantages both of clarifying the underlying principles of statistics and probability and of facilitating an appreciation of their applications to concrete judgmental tasks"* (Nisbett & Ross, 1980, p.281).

Technology

Fortunately, increasingly easy access to powerful computing facilities has saved time previously devoted to laborious calculations and encouraged less formal, more intuitive approaches to statistics (Biehler, 2003; Chauchat, 2003; Oriol & Régnier, 2003). In particular, computer simulations could contribute to improving students' understandings of many stochastic ideas. However, del Mas et al. (1999) warn that the use of technology and activities based on research results did not always produce effective understanding. The new activities and the learning of the software might be too demanding for some students, and the amount of new information about the sofware might interfere with the students' learning.

Advances in technology and increasing student enrollment numbers have also led many universities to offer on-line courses although few studies have compared online and traditional methods of teaching of advanced statistics and results are inconclusive. At the IASE Satellite Conference on Statistics Education and the Internet, most presentations focused on analysing Internet resources for teaching statistics or presenting examples of these resources. This research is becoming more important with new directions of teaching statistics at secondary school level that include some elements of inference (Parzysz, 2003; Chaput, & Henry, 2005).

Final remarks

The number of useful statistical methods and the quick pace of change and development in statistics mean that the statistical knowledge acquired in statistics courses at University is insufficient for future professionals to be independent. A realistic vision of training University students should recognise the complex character of statistical knowledge, even when increasing the teaching time and improving the didactical resources. It is difficult for students who are not specialising in statistics to acquire a complete mastering of statistical concepts and methods, beyond the most basic content, or that which becomes familiar due to its frequent use.

Consequently, it is unrealistic to expect these professionals to be their own statisticians and solve all their data analysis problems by themselves. It is therefore necessary to increase the appreciation of statistical work on the part of these students. An important goal of education is making students realize that they need the collaboration of professional statisticians in their future work.

Given the relevance of a correct understanding and application of statistical inference to improve empirical research (including research in mathematics education) we also need more didactical research on teaching statistics at University level, which is still very scarce.

Note: This research has been supported by the grant SEJ2004-00789 M.E.C. (Madrid).

References

Ares, V.M. (1999) La prueba de significación de la «hipótesis cero» en las investigaciones por encuesta, *Metodología de Encuestas*, 1, 47-68.

Biehler, R. (2003) Interrelated learning and working environments for supporting the use of computer tools in introductory courses. CD-ROM, *Proceedings IASE Satellite conference on Teaching Statistics and the Internet,* Berlin: IASE.

Batanero, C. (2000) Controversies around significance tests, *Mathematical Thinking and Learning, 2(1-2),* 75 – 98.

Batanero, C. (2001) *Training researchers in the use of statistics,* Granada: International Statistical Institute.

Birnbaum, I. (1982) Interpreting statistical significance, *Teaching Statistics,* 4, 24–27.

Borges, A., San Luis, C., Sánchez, J. A. & Cañadas, I. (2001) El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa, *Psicothema,* 13(1), 174-178.

Cabriá, S. (1994) *Filosofía de la estadística*, Valencia: Servicio de Publicaciones de la Universidad.

Chauchat, J. H. (2003). Teaching statistical inference using many samples from a real large dataset Invited paper in the *International Statistical Institute 53 Session*. Berlin, 2003.

Chaput, B. & Henry, M. (2005). (Coord.). *Statistique au lycée. Volume 1. Les outils de la statistique*. Commisaion Inter.-IREM Statistique et Probabilités.

Cumming, G.; Williams, J. & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics,* 3, 299-311.

DelMas, R. C., Garfield, J. B., & Chance, B. (1999) *Exploring the role of computer simulations in developing understanding of sampling distributions*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Díaz, C. & de la Fuente, I. (2004) Controversias en el uso de la inferencia en la investigación experimental, *Metodología de las Ciencias del Comportamiento,* Volumen especial 2004, 161-167.

Diaconis, P., & Freedman, D. (1981) The persistence of cognitive illusions, *Behavioral and Brain Sciences,* 4, 378-399.

Ellerton, N. (1996) Statistical significance testing and this journal, *Mathematics Education Research Journal,* 8(2), 97–100.

Falk, R. (1986) Misconceptions of statistical significance, *Journal of Structural Learning,* 9, 83–96.

Falk, R., & Greenbaum, C. W. (1995) Significance tests die hard: The amazing persistence of a probabilistic misconception, *Theory and Psychology,* 5 (1), 75-98.

Fidler, F. (2002). The fifth edition of the APA publication manual: Why its statistics recommendations are so controversial. *Educational And Psychological Measurement*, 62 (5), 749-770.

Finch, S., Cumming, G., y Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform, *Educational and Psychological Measurement,* 61, 181-210.

Fisher, R. A. (1935) *The design of experiments*, Edimburgh: Oliver & Boyd.

Fisher, R. A. (1956) *Statistical methods and scientific inference,* Edinburgh: Oliver & Boyd.

Fisher, R. A. (1958) *Statistical methods for research workers* (Thirteenth edition), New York: Hafner.

Frías, M. D, Pascual, J., García, J. F. (2000) Tamaño del efecto del tratamiento y significación estadística, *Psicothema,* 12, 2 supl, 236-240

Glaser, D. N. (1999) The controversy of significance testing: Misconceptions and alternatives, *American Journal of Critical Care,* 5. 291-296.

Gingerenzer, G. (1993) The superego, the ego and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.)*, A handbook for data analysis in the behavioral sciences: Methdological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Hacking, I. (1975) *The logic of statistical inference*, Cambridge: Cambridge University Press.

Harlow, L. L. (1997) Significance testing: Introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.)*, What if there were no significance tests?* (pp. 1-20). Mahwah, NJ: Lawrence Erlbaum Associates.

Hager, W. (2000) About some misconceptions and the discontent with statistical tests in psychology. *Methods on Psychological Research, 5(1).* On line. Disponible en http://www.mpr-online.de

Haller, H., & Krauss, S. (2002) Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research,* 7(1). On line: http://www.mpronline.de/issue16/art1/haller.pdf.

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

Kahneman, D., Slovic, P., & Tversky, A. (1982) *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kline, R. B. (2004) *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC:American Psychological Association.

Lecoutre, B. (1999) Beyond the significance test controversy: Prime time for Bayes? *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute* (Tome 58, Book 2) (pp. 205-208). Helsinki, Finland: International Statistical Institute.

Lecoutre B. (2006) Training students and researchers in Bayesian methods for experimental data analysis. *Journal of Data Science*, 4, (in press).

Lecoutre B.; Lecoutre M.P.; y Poitevineau J. (2001) Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical Review*, 69, 399-418.

Levin, J. R. (1998) To test or not to test H_0? *Educational and Psychological Measurement,* 58, 313-333.

Levin, J. R., & Robinson, D. H. (1999) Further reflections on hypothesis testing and editorial policy for primary research journals, *Educational Psychological Review,* 11, 143-155.

López, E. (2003) Las pruebas de significación: una polémica abierta, *Bordón,* 55, 241-252.

Menon, R. (1993) Statistical significance testing should be discontinued in mathematics education research, *Mathematics Education Research Journal, 5*(1), 4–18.

Moses, L. E. (1992) The reasoning of statistical inference. In D. C. Hoaglin & D. S. Moore (Eds.), *Perspectives on contemporary statistics* (pp. 107-122). Washington, DC: Mathematical Association of America.

Morrison, D. E., & Henkel, R. E. (1970*) The significance tests controversy. A reader*, Chicago: Aldine.

Neyman, J. (1950) *First course in probability and statistics*, New York: Henry Holt.

Nisbett, R., & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgments*, Englewood Cliffs, NJ: Prentice Hall.

Oriol, J.C & Régnier J.C. (2003) Fonctionnement didactique de la simulation en statistique :exemple de l'enseignement du concept d'intervalle de confiance. *Actes des Journées de Statistique SFDS* Lyon, 2003.

Parzysz, B. (2003). From frequency to probability. some questions posed by the new french senior high school curricula. Invited paper in the *International Statistical Institute 53 Session.* Berlin, 2003.

Pimenta, R. (2006) *Assessing statistical reasoning through project work.* Paper to be presented at the Seventh International Conference on Teaching Statistics. Salvador de

Bahia, Brazil.

Pollard, P., & Richardson, J. T. E. (1987) On the probability of making Type I errors, *Psychological Bulletin,* 10, 159-163.

Robinson, D. H., & Levin, J. T. (1997) Reflections on statistical and substantive significance, with a slice of replication, *Educational Researcher*, 26 (5), 21-26.

Royal, R. (1997) *Statistical evidence. A likelihood paradigm*, London: Chapman & Hall.

Seidenfeld, T. (1979) *Philosophical problems of staistical inference: Learning from R. A. Fisher*. Dordrecht, The Netherlands: Reidel.

Sohn, D. (1998) Statistical significance and replicability: Why the former does not presage the latter, *Theory & Psychology*, 8(3), 291-311.

Vacha-Haase, T. (2001) Statistical significance should not be considered one of life's guarantees: Effect sizes are needed, *Educational and Psychological Measurement*, 61, 219-224.

Wilkinson, L. (1999) Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist,* 54, 594-604.

Yates, F. (1951) The influence of "Statistical methods for research workers" on the development of the science of statistics, *Journal of the American Statistical Association,* 46, 19-34.