Equipe Raisonnement Induction Statistique

Citations/Quotes

"The field of statistics continues to flourish despite, and partly because of, its foundational controversies." (Efron, 1978)

Citations sur l'inférence statistique
Quotes about statistical inference

Mise à jour 01 mars 2005 / Updating 01 march 2005

Bruno Lecoutre & Jacques Poitevineau

Auteurs / Authors

Bruno LECOUTRE

[bruno.lecoutre@univ-rouen.fr]
Directeur de recherche C.N.R.S.
LMRS Laboratoire de Mathématiques Raphaël Salem, UMR 6085
C.N.R.S. et Université de Rouen
Mathématiques Site Colbert
F 76821 Mont Saint Aignan Cedex (France)

Jacques POITEVINEAU

[poitevin@ccr.jussieu.fr]
Ingénieur d'études C.N.R.S.
LAM / LCPE UMR 7604
CNRS - Université Paris 6 - Ministère de le Culture
11 rue de Lourmel
F 75015 Paris (France)

Generalities

"What is done in practice is to use the confidence procedure on a series of different problems - not use the confidence procedure for a series of repetitions of the same problem with different data (which would typically make no sense in practice)." (Bayarri & Berger, 2003, page 5)

"ANOVA may be the most commonly used statistical procedure. It is assuredly the most commonly misused statistical procedure!" (Berry, 1996, page 395)

"There were far too many studies to plan and too much data to analyze to worry seriously about what the p-values and confidence coefficients produced by the package actually meant." (Breslow, 1990, page 269)

"Statistical techniques must be chosen and used to aid, but not to replace, relevant thought." (Bryan-Jones & Finney, 1983)

"My recommendation is to give always a look at the data, since the eye of the expert is in most simple (i.e. low-dimensional) cases better that automatic tests." (D'Agostini, 2000)

"[...] statisticians believe that statistics exists as a discipline in its own right, even if they can't agree on its exact nature." (Efron, 1978)

"[...] the familiar optimality criteria of statistics are in fact in conflict with scientific principles [...]" (Fraser & Reid, 1990)

"We need statistical thinking, not rituals." (Gigerenzer, 1998)

"Statistics has unfortunately achieved almost the status of a superstition in some quarters in psychology, and I hope, in all humility, that this text sets a slightly more liberal and rational one." (Hays, 1963, pages vi-vii)

"But if there is ever a conflict between the use of a statistical technique and common sense, then common sense comes first." (Hays, 1973, page 386)

"For many years, statistics textbooks have followed this 'canonical' procedure: (1) the reader is warned not to use the discredited methods of Bayes and Laplace, (2) an orthodox method is extolled as superior and applied to a few simple problems, (3) the corresponding Bayesian solution are *not* worked out or described in any way. The net result is that no evidence whatsoever is offered to substantiate the claim of superiority of the orthodox method. [...] The orthodox results are satisfactory only when they agree closely (or exactly) with the Bayesian results. No contrary example has yet been produced. [...] We conclude that orthodox claims of superiority are totally unjustified; today the original statistical methods of Bayes and Laplace stand in apposition of proven superiority in actual performance, that places them beyond the reach of mere ideological or philosophical attacks. It is the continued teaching and use of orthodox methods that is in need of justification and defense." (Jaynes, 1976, page 175)

"[...] I should think that orthodox teachers would be very troubled by the following situation. Who have made the important advances in statistical practice in this Century? Others will judge differently, but my own list is: 'Student', Jeffreys, Fisher, Wiener, von Neumann, Shannon, Wald, Zellner, Burg, Skilling. Here we find a chemist, a physicist, a eugenicist, two mathematicians, an economist, an astromer, tow engineers – and only one professional statistician! Whatever list one makes, I think he will find that most of the important advances have come from outside the profession, and had to make their way against the opposition of most statisticians." (Jaynes, 1985, page 46)

"The statistician can provide guidance as to what the statistics mean; but the individual consumer of the statistics remains the ultimate judge of whether the evidence of any experiment is convincing. Statistics cannot substitute for good judgement, nor can it transform a flawed experiment into a valid one. Where an experiment cannot distinguish between two equally capable explanations, no amount of statistical analysis will change that situation. Where data are at the margins of detectablility, the solution is to design a better experiment, not more statistics." (Jefferys, 1992)

"The difficulty is that the solution to this problem [finding the most powerful test] has no relevance *per se* to the problems of applied statistics..." (Kempthorne, 1977)

"I know of no field where the foundations are of such practical importance as in statistics." (Lindley, 1972)

"At any rate what I feel quite sure at the moment to be needed is simple illustration of the new [Bayesian] notions on real, everyday statistical problems." (E.S. Pearson, 1962)

"This points to the difference between statistics as an effort to learn, to get at the truth, and decision theory — a difference that was emphasized by Fisher in some of his disputes with Neyman." (Lehmann, 1998)

"But we must question the value of statistical research 'stimulated by its mathematical, rather than practical, aspects' [McDermott & Wang, in Perlman & Wu, 1999, page 375] when such work produces impractical procedures that are them promoted (fortunately unsuccessfully) to the applied community." (Perlman & Wu, 1999, page 378)

"We hope that we have alerted statisticians to the dangers inherent in uncritical application of the NP [Neyman & Pearson] criterion, and, more generally, convinced them to join Fisher, Cox and many others in carefully weighing the scientific relevance and logical consistency of any mathematical criterion proposed for statistical theory." (Perlman & Wu, 1999, page 381)

"Neyman and Pearson contributed vitally to our understanding by their *formulation* of statistical problems, but they have never claimed their *methods* were more than ad hoc procedures with some pleasant properties. Their methods, while extremely ingenious and useful, are not completely satisfactory, let alone uniquely objective and scientific." (Pratt, 1962)

"Statistical 'recipes' are followed blindly, and ritual has taken over from scientific thinking." (Preece)

"I cannot see how anyone could now agree with this [Fisher's 1935 quote about experiments and null hypotheses]." (Preece)

"[Neyman-Pearson theory] does not address the problem of representing and interpreting statistical evidence, and the decision rules derived from NP theory are not appropriate tools for interpreting data as evidence." (Royall, 1997, page 58)

[A need for] "[...] development of diagnostic tools with a greater emphasis on assessing the usefulness of an assumed model for specic purposes at hand rather than on whether the model is true." (Tiao & Xu, 1993)

"It is far better to arrive at an appropriate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise." (Tukey, 1962)

Null Hypothesis Significance Testing [NHST]

"Somehow there has developed a widespread belief that statistical analysis is legitimate only if it includes significance testing. This belief leads to, and is fostered by, numerous introductory statistics texts that are little more than catalogues of techniques for performing significance tests." (Altman, 1985)

"The test is like a gauge on the dashboard" (Anonymous psychology researcher, *in* Lecoutre M.-P., 2000, page 77)

"Tests of the null hypothesis that there is no difference between certain treatments are often made in the analysis of agricultural or industrial experiments in which alternative methods or processes are compared. Such tests are [...] totally irrelevant. What are needed are estimates of magnitudes of effects, with standard errors." (Anscombe, 1956)

"The common practice of reporting only the significance level of the test and not the data on which it was calculated (often justified on grounds of space) ensures that no conflict with other research can be detected." (Atkins & Jarrett, 1981, page 101)

"It is hardly surprising that empirical views of science, and the structure of careers and institutions in social science, provide fertile ground for the use of procedures [significance tests] which tend to disguise the inadequacy of measurements, and the lack of developed theoretical explanations – as well as discouraging debate about alternative procedures." (Atkins & Jarrett, 1981, page 105)

"The test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; [...] a great deal of mischief has been associated with its use." (Bakan, 1966, page 423).

"We need to get on with the business of generating psychological hypotheses and proceed to do investigations and make inferences which bear on them, instead of, as so much of our literature would attest, testing the statistical null hypothesis in any number of contexts in which we have every reason to suppose that it is false in the first place." (Bakan, 1967, *in* Morrison & Henkel, 1970, page 251)

"When we reach a point where our statistical procedures are substitutes instead of aids to thought, and we are led to absurdities, then we must return to common sense." (Bakan, 1967, *in* Morrison & Henkel, 1970)

"Statistics looks very bad when it recommends a conclusion that clearly contradicts common sense." (Berger & Wolpert, 1988, page 141)

"In a world in which only significant results are published, this makes researchers into gamblers, whose careers depend on the outcome of the chance events they are attempting to control for." (Blaich, 1998, page 194)

"It is our belief that the great reliance placed by many sociologists on tests of significance is chiefly an attempt to provide scientific legitimacy to empirical research without adequate theoretical significance." (Camilleri, 1962)

[NHST] "is not only useless, it is also harmful because it is interpreted to mean something it is not." (Carver, 1978, page 392)

[NHST] a "corrupt form of the scientific method" (Carver, 1993, page 288)

"The best research articles are those that include *no* tests of statistical significance." (Carver, 1993, page 289)

"[...] a Bayesian is someone who doesn't understand what a frequentist is, and a frequentist is someone who doesn't

understand what a Bayesian is" [from charles@clef.demon.co.uk, http://www.lns.cornell.edu/spr/2002-03/msg0040564.html, June 2003]

"Many of tests reported in the *Journal* [*of Wildlife Management*] and the [*Wildlife Society*] *Bulletin* are unnecessary." (Cherry, 1998, page 947)

"[NHST] does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (Cohen 1994, page 997).

"[NHST] has not only failed to support the advance of psychology as a science but also has seriously impeded it." (Cohen, 1994, page 997).

"When passing null hypothesis tests becomes the criterion for successful predictions, as well as for journal publications, there is no pressure on the psychology researcher to build a solid, accurate theory; all he or she is required to do, it seems, is produce 'statistically significant' results." (Dar, 1987,page 149)

"[NHST] An automatic routine" (Falk & Greenbaum, 1995)

"[NHST] fail[s] to give us the information we need [...] induce[s] the illusion that we have it." (Falk & Greenbaum, 1995, page 94)

"Rigid dependence upon significance tests in single experiments is to be deplored." (Finney)

"The statistical examination of a body of data is thus logically similar to the general alternation of inductive and deductive methods throughout the sciences. A hypothesis is conceived and defined with all necessary exactitude; its logical consequences are ascertained by a deductive argument; these consequences are compared with the available observations; if these are completely in accord with the deductions, the hypothesis is justified at least until fresh and more stringent observations are available." (Fisher, 1990/1925, page 8)

"[...] for the tests of significance are used as an aid to judgement, and should not be confused with automatic acceptance tests, or 'decision functions'.'" (Fisher, 1990/1925, page 28).

"Though recognizable as a psychological condition of reluctance, or resistance to the acceptance of a proposition, the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to, and verifiable by, other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief it engenders. It is more primitive, or elemental than, and does not justify, any exact probability statement about the proposition." (Fisher, 1990/1956, page 46)

"Whereas, the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination through the hypotheses which he has decided to test, or usually indeed of some specific aspects of these hypotheses." (Fisher, 1990/1956, page 81)

"[The] statistician who advertises the [scientifically unacceptable] procedure is guilty of professional misconduct." (Fraser & Reid, *in* Brown, 1990, page 507)

"A way of thinking that has survived decades of ferocious attacks is likely to have some value." [An anonymous reviewer, *in* Frick, 1996, page 379]

"Thus null hypothesis testing is an optimal method for demonstrating sufficient evidence for an ordinal claim." (Frick, 1996, page 379)

[NHST] "A mechanical behavior" (Gigerenzer, 1991)

"In psychology, it [NHST] has been practiced like ritualistics handwashing and sustained by wishful thinking about its utility." Gigerenzer, 1998)

"This ritual [NHST] discourages theory development by providing researchers with no incentive to specify hypotheses." Gigerenzer, 1998)

"It is misleading to tell a Student he must decide on his significance test in advance, although it is correct according to the Fisherian technique." (Good, 1976, page 54)

The "star worshippers" (Guttman, 1983)

"Despite their wide use in scientific journals such as The Journal of Wildlife Management, statistical hypothesis tests add very little value to the products of research. Indeed, they frequently confuse the interpretation of data." (Johnson, 1999, page 63)

"There is a rising feeling among statisticians that hypothesis tests [...] are not the most meaningful analyses." (Jones, 1984)

"At its worst, the results of statistical hypothesis testing can be seriously misleading, and at its best, it offers no informational advantage over its alternatives." (Jones & Matloff, 1986)

"[...] in fact, focusing on *p* values and rejecting null hypotheses actually distracts us from reaching our goals: deciding whether data support our scientific hypothesis and are practically significant or useful." (Kirk, 1996, page 755)

"Seventy-five years of null hypothesis testing has taught us the folly of blindly adhering to a ritualized procedure." (Kirk, 2001, page 217)

"I believe that clear rationales for hypothesis testing (unified or not) should replace murky decision-theoretic metaphors, and that this replacement will facilitate improvements in both teaching and practice." (Krantz, 1999, page 1380)

"Because of the relative simplicity of its structure, significance testing has been overemphasized in some presentations of statistics, and as a result some students come mistakenly to feel that statistics is little else than significance testing." (Kruskal)

"Il est hélas tentant, lorsque le problème est complexe et possède trop de degrés de liberté, de les [les tests statistiques] utiliser mécaniquement et de s'en remettre à leur 'froid jugement'. C'est une erreur qui a été maintes fois relevée dans la littérature statistique [...] et qui relève peut-être de ces '*restes de magie qui subsistent au cœur de chacun*' dont parlait Alfred Sauvy." (Ladiray, 2002, pages 6-7)

"However the use of NHST is such an integral part of scientists' behavior that its misuses and abuses should not be discontinued by flinging it out of the window." (Lecoutre, Lecoutre & Poitevineau, 2001, page 413)

"Few concepts in the social sciences have wielded more discriminatory power over the status of knowledge claims than that of statistical significance." (Litle, 2001, page 363)

"I believe part of the difficulty with the current use of NHST is the exaggerated practical implications that have come to be attached to its results." (Locascio, 1999)

"Despite the stranglehold that hypothesis testing has on experimental psychology, I find it difficult to imagine a less insightful means of transiting from data to conclusions." (Loftus, 1991, page 103)

"Null Hypothesis Statistical Testing, as typically utilized, is barren as a means of transiting from data to conclusions." (Loftus, 1996)

"Problems stemming from the fact that hypothesis tests do not address questions of scientific interest." (Matloff, 1991)

"I suggest to you that Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology." (Meehl, 1978, page 817)

"Some hesitation about the unthinking use of significance tests is a sign of statistical maturity." (Moore & McCabe, 1993)

"[...] thus, any difference in the groups on a particular variable in a given assignment will have some calculable probability of being due to errors in the assignment procedure..." (Morrison & Henkel, 1969, *in* Morrison & Henkel, 1970, pages 195-196)

"The test provides neither the necessary nor the sufficient scope or type of knowledge that basic scientific social research requires." (Morrison & Henkel, 1969, *in* Morrison & Henkel, 1970, page 198)

"In addition to important technical errors, fundamental errors in the philosophy of science are frequently involved in this indiscriminate use of the tests [of significance]." (Morrison & Henkel, 1969, *in* Morrison & Henkel, 1970)

"The question many researchers (especially those interested in the application of science to solve practical problems) want to ask is whether the effects are large enough to make a real difference. The statistical tests most frequently encountered in the social and behavioral sciences do not directly address this question." (Murphy & Myors, 1999, page 234)

"The grotesque emphasis on significance tests in statistics courses of all kinds [...] is taught to people, who if they come

away with no other notion, will remember that statistics is about tests for significant differences. [...] The apparatus on which their statistics course has been constructed is often worse than irrelevant, it is misleading about what is important in examining data and making inferences." (Nelder)


"I contend that the general acceptance of statistical hypothesis testing is one of the most unfortunate aspects of 20th century applied science." (Nester, 1996)


"[...] if the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data." (Nunnally, 1960)


"Probably few methodological issues have generated as much controversy among sociobehavioral scientists as the use of [Null Hypothesis Significance] tests." (Pedhazur & Schmelkin, 1991, page 198)


"This reinforces the well-documented but oft-neglected fact that the Neyman-Pearson theory desideratum of a more (or most) powerful size alpha test may be scientifically inappropriate; the same is true for the criteria of unbiasedness and alpha-admissibility." (Perlman & Wu, 1999, page 355)


"[...] Berger and Hsu (1996, page 192) make the following statement: "We believe that notions of size, power, and unbiasedness are more fundamental than 'intuition'..." In our opinion, such a statement places the credibility of statistical science at serious risk within the scientific community. If we are indeed teaching our students to disregard intuition in scientific inquiry, then a fundamental reassessment of the mission of mathematical statistics is urgently needed." (Perlman & Wu, 1999, page 366) [Berger's reply: "If we are indeed teaching our students to disregard intuition in scientific inquiry, then a fundamental reassessment of the mission of mathematical statistics is urgently needed." (page 373)]


"[...] Le chercheur qui présente un résultat significatif, tel le vainqueur d'une épreuve sportive, fait souvent l'objet de suspicion et doit satisfaire à un contrôle avant que son résultat soit homologué (publié). C'est le rôle des éditeurs et rapporteurs des revues aux réserves desquels l'expérimentateur est souvent confronté. Malheureusement, la norme est si bien établie que ces réserves portent presque exclusivement sur la validité des tests (A-t-on utilisé le bon test? Les conditions d'application sont-elles satisfaisantes? Etc.) et non sur leur pertinence (Le test répond-il vraiment à la question posée?)." (Poitevineau, 1998, page 11)


"Tests [of hypotheses] provide a poor model of most real problems, usually so poor that their objectivity is tangential and often too poor to be useful." (Pratt, 1976)


"This reduces the role of tests essentially to convention. Convention is useful in daily life, law, religion, and politics, but it impedes philosophy." (Pratt, 1976)


"Over-emphasis on significance-testing continues." (Preece)


"Given the many attacks on it, null-hypothesis testing should be dead." (Rindskopf, 1997, page 319)


"[...] there is the current prestige of *exact tests* in statistics. The magic of 'exactness' must be qualified of course. Student's *t* -test was (and still is) an exact test to!" (Rouanet & Bert, 2000, page 121)


"The stranglehold that conventional null hypothesis significance testing has clamped on publication standards must be broken." (Rozeboom, 1960, *in* Morrison & Henkel, 1970, page 230)


"The traditional null hypothesis significance-test method, more appropriately called 'null hypothesis decision [NHD] procedure',of statistical analysis is here vigorously excoriated for its inappropriateness as a method of *inference*." (Rozeboom, 1960, *in* Morrison & Henkel, 1970, page 230)


"[NHST] Surely the most bone-headedly misguided procedure ever institutionalised in the rote training of science students." (Rozeboom, 1997, page 335)


"The 'religion of statistics' with its rites such as the use of the profoundly mysterious symbols of the religion *NS*, *, **, and *mirabile dictu* ***." (Salsburg, 1985)


"One of the dangers of small samples is the discarding of valid results simply because of the relatively high probability that they *might* have occurred by chance." (Selvin, 1957, *in* Morrison & Henkel, 1970, page 110)


[NHST] "such tests do not provide the information that many researchers assume they do" (Shaver, 1993, page 294)


[NHST] "diverts attention and energy from more appropriate strategies such as replication and consideration of the practical or theoretical significance of results" (Shaver, 1993, page 294)

"One of the chief drawbacks on the F-test in ANOVA is that by itself, *F*(dfb,dfw) tells us hardly anything useful about what effects our experiment has had." (Smithson, 2000, page 238)

"Fisher [...] appears to have placed an undue emphasis on the significance test." (Street, 1990)

"Superficial understanding of significance testing has led to serious distortions, such as researchers interpreting significant results involving large effect sizes." (Thompson, 1989, page 2)

"Tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and they are tired." (Thompson, 1993a, page 363).

"Never use the unfortunate expression 'accept the null hypothesis'. " (Wilkinson and Task Force on Statistical Inference, 1999)

"We believe that although unreasonable claims are sometimes made for the test of significance and that although many have sinned in implicitly treating statistical significance as proof of a favored explanation, still the social scientists is better off for using the significance test that for ignoring it. More precisely, it is our judgment that although the test of significance is irrelevant to the interpretation of the cause of a difference, still it does provide a relevant and useful way of assessing the relative likelihood that a real difference exists and is worthy of interpretive attention, as opposed to the hypothesis that the set of data could be a haphazard arrangement." (Winch & Campbell, 1969, *in* Morrison & Henkel, 1970, page 199)

"The present writers think that the indiscriminate cataloguing of trivial effects is, in fact, a major problem in psychology today...". (Wilson *et al.*, 1967)

"We reason that it is very important to have a formal and nonsubjective way of deciding whether a given set of data shows haphazard or systematic variation. [...] And we believe it is important not to leave the determination of what is systematic or haphazard arrangement of data to the intuition of the investigator." (Winch & Campbell, 1969, *in* Morrison & Henkel, 1970, page 206)

"The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective. Results are significant or not significant and this is the end of it." (Yates, 1951)

Bayesian inference

"I have never had a student suggest that the Bayesian approach does not make sense to them. But I have had students in frequentist courses (with no prodding from me!) make statements such as: 'Why would anyone want to calculate probabilities assuming that the null hypothesis is true (or false)? It seems to me that want to know is the probability that the null hypothesis is true'." (Berry, 1997)

"But since the Bayesian approach fits neatly with the scientific perspective, it forces the statistician to take a broad view rather than one limited to the results of a particular experiment." (Berry, 1997)

"[...] scientists think and reason like Bayesians, whether or not they know Bayes' theorem." (Berry, 1997)

"We should indeed argue that noninformative prior Bayesian analysis is the single most powerful method of statistical analysis." (Berger, 1985, page 90)

"At the very least, use of noninformative priors should be recognized as being at least as objective as any other statistical techniques." (Berger, 1985, page 110).

"Its flexibility makes the Bayesian approach ideal for analysing clinical trials." (Berry, 1993, page 1377)

"In fact, I find it easier teaching Bayesian statistics than frequentist statistics. There is a single, pivotal notion - Bayes' rule - that describes the process of learning. Bayes' rule is especially easy to teach, and it is easy for students to use." (Berry, 1995)

"Bayesian statistics is difficult in the sense that thinking is difficult." (Berry, 1997)

"Bayes procedures [...] can even define the class of optimal *frequentist* procedures, thus 'beating the frequentist at his own game'." (Carlin & Louis, 2000, page 10))

"It [statistical significance testing] still gives us an estimate of $p(D'|H_0)$, when what we want is $p(H_0|D')$, $p(R|D')$ and $p(H_1|D')$." (Carver, 1978, page 392).

"But Bayesianism is far more than a bag of tricks for helping other specialists out with their tricky problems - It is a totally original way of thinking about the world we live in." (Dawid, 2000, page 1)

"An objective scientific report is a report of the whole prior-to-posterior mapping of a relevant range of prior probability distributions, keyed to meaningful uncertainty interpretations." (Dickey, 1986, page 135)

"Subjective Bayesianism must face the challenge of scientific objectivity. This is the ultimate stronghold of the frequentist viewpoint. If the 21st century is Bayesian, my guessing is that it will be some combination of subjective, objective and empirical Bayesian, not significantly less complicated and contradictory that the present situation." (Efron, 1978)

"A widely accepted objective Bayes theory, which fiducial inference was intended to be, would be of immense theoretical and practical importance. A successful objective Bayes theory would have to provide good frequentist properties in familiar situations, for instance, reasonable coverage probabilities for whatever replaces confidence intervals." (Efron, 1998)

"Bayesian inference might, in principle, fill the void created by abandoning significance-testing [...] Implementation of Bayesian analysis, however, requires subjective assessments of prior distributions, and often involves technical problems." (Falk & Greenbaum, 1995)

"It is still wonder they [Bayesians] are still treated as a kind of lunatic fringe preaching a doctrine so pure and untainted by the real world as to make it useful for little other than academics furthering their research careers'." (Freeman, 1993, page 1450)

"Bayesian posterior probabilities are exactly what scientists want." (Goodman & Berlin, 1994, page 203)

"Confidence intervals should play an important role when setting sample size, and power should play no role once the data have been collected. [...] In this commentary, we present the reasons why the calculation of power after a study is over is inappropriate and how confidence intervals can be used during both study design and study interpretation." (Goodman & Berlin, 1994, page 200).

"For interpretation of observed results, the concept of power has no place, and confidence intervals, likelihood, or Bayesian methods should be used instead." (Goodman & Berlin, 1994, page 205).

"In fact, one can argue that the objections to subjective probability and Bayes are a nineteenth century aberration which has been redressed in the last thirty years - except by psychologists." (Gregson, 1998, page 202).

"[Bayesian analysis provides] direct probability statements - which are what most people wrongly assume they are getting from conventional statistics." (Grunkemeier & Payne, 2002, page 1901)

"It could be argued that since most physicians use statement A [the probability the true mean value is in the interval is 95%] to describe 'confidence' intervals, what they really want are 'probability' intervals. Since to get them they must use Bayesian methods, then they are really Bayesians at heart!" (Grunkemeier & Payne, 2002, page 1904)

"As long as we are uncertain about values of parameters, we will fall into the Bayesian camp." (Iversen, 2000, page 10)

"Bayesian statistics, because of its straightforward interpretation, and because the assumptions are out in the open, offers a way to clarify and sharpen our thinking about experiments, and by giving us new insight about why parapsychological experiments are not having their intended effect of convincing a skeptical scientific world, they can point out research directions that might be more fruitful." (Jefferys, 1992)

"Again it is not clear why such a set [the confidence interval] should be of interest unless one makes the natural error of thinking of the parameter as random and the confidence set as containing the parameter with a specified probability. Again, this is a statement only a Bayesian can make, although confidence intervals are often so misinterpreted. I find the classical quantities useless for making decisions and believe that they are widely misinterpreted as Bayesian because the Bayesian quantities are more natural." (Kadane, 1995, page 316)

"The utility of the Bayesian approach is increasingly being recognized by the scientific establishment." (Krueger & Funder,

2001)

"In our view, the way in which the Bayesian approach is used in an area of research reflects the maturity of this field." (Krueger & Funder, 2001)

"Depuis 1973, l'Analyse des Comparaisons a intégré les techniques bayésiennes, classiques et contemporaines (Jeffreys, Lindley, etc.), mais en les utilisant avec une motivation fiduciaire (Fisher). Ces techniques nous paraissent en effet les mieux adaptées pour pallier les insuffisances [...] des tests de signification traditionnels." (Lecoutre, Rouanet & Denhière, 1988, page 384)

"We [statisticians] will all be Bayesians in 2020, and then we can be a united profession." (Lindley, *in* Smith, 1995, page 317)

"This is the likelihood principle according to which values of x, other than that observed, play no role in inference." (Lindley, 2000)

"A non-Bayesian states that there is a 95% chance that the [obtained] confidence interval contains the true value of the population mean. A Bayesian would say there is a 95% chance that the population mean falls between the obtained limits. One is a probability statement about the interval, the other about the population parameter." (Phillips, 1973, page 335)

"Null-hypothesis tests are not completely stupid, but Bayesian statistics are better." (Rindskopf, 1998)

"The motivation for using this methodology [a Bayesian approach] is practical rather than ideological." (Spiegelhalter, Freedman & Parmar, 1994, page 357)

"This state of affairs [the reluctance of scientists to use Bayesian inferential procedures in practice] appears to be due to a combination of factors including philosophical conviction, tradition, statistical training, lack of 'availability', computational difficulties, reporting difficulties, and perceived resistance by journal editors." (Winkler, 1974, page 129).

0.05/Choice of the level of significance

"Another problem associated with the test of significance. The particular level of significance chosen for an investigation is not a logical consequence of the theory of statistical inference." (Camilleri, 1962, *in* Morrison & Henkel, 1970)

"It is convenient to draw the line at about the level at which we can say: 'Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials.' [...] If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or once in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance." (Fisher, 1926, page 504)

"[...] the sanctified (and sanctifying) magic .05 level. This basis for decision has played a remarkable role in the social sciences and in the lives of social scientists. In governing decisions about the status of null hypotheses, it came to determine decisions about the acceptance of doctoral dissertations and the granting of research funding, and about publication, promotion, and whether to have a baby just now. Its arbitrary unreasonable tyranny has led to data fudging of varying degrees of subtlety from grossly altering data to dropping cases where there 'must have been errors'." (Cohen, 1990, page 1307)

"Do people, scientists and nonscientists, generally feel that an event which occurs 5% of the time or less is a rare event? If the answer [...]. is 'Yes,' [...] then the adoption of the level as a criterion for judging outcomes is justifiable." (Cowles and Davis, 1982, page 557)

"Fisher "advocated 5% as the standard level." (Lehmann, 1993) [But see Fisher, 1990c/1956, p. 45: [...] for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of evidence and his ideas."].

"[...] the degree of conviction is not even approximately the same in two situations with equal significance levels. 5% in to-day's small sample does not mean the same as 5% in to-morrow's large one." (Lindley, 1957, page 189)

"The current obsession with .05 [...] has the consequence of differentiating significant research findings and those best forgotten, published studies from unpublished ones, and renewal of grants from termination. It would not be difficult to document the joy experienced by a social scientist when his *F* ratio or *t* value yields significance at .05, nor his horror when the table reads 'only' .10 or .06. One comes to internalize the difference between .05 and .06 as 'right' *vs.* 'wrong', 'creditable' *vs.* 'embarrassing', 'success' *vs.* 'failure'." (Skipper, Guenther & Nass, 1967)

"Blind adherence to the .05 level denies any consideration of alternative strategies, and it is a serious impediment to the interpretation of data." (Skipper, Guenther & Nass, 1967)

"Surely, God loves the .06 nearly as much as the .05." (Rosnow & Rosenthal, 1989, page 1277)

"It may not be an exaggeration to say that for many Ph.D. students, form whom the .05 has acquired almost an ontological mystique, it can mean joy, a doctoral degree, and a tenure-track position at a major university if their dissertation *p* is less than .05. However, if the *p* is greater than .05, it can mean ruin, despair, and their advisor's thinking of a new control condition that should be run." (Rosnow & Rosenthal, 1989, page 1277)

[alpha=.05] "A deliberate attempt to offer a standardized, public method for objectifying an individual scientist's willingness to make an inference." (Wilson, Miller & Loweret, 1967, page 191)

Conception of probability

"Identifying probability with frequency is like confusing a table with the English world 'table'." (D'Agostini, 2000)

"The subject of a probability statement if we know what we are talking about, is singular and unique; we have some degree of uncertainty about its value, and it so happens that we can specify the exact nature and extent of our uncertainty by means of the concept of Mathematical Probability as developed by the great mathematicians of the 17th century Fermat, Pascal, Leibnitz, Bernoulli and their immediate followers." [...] "The probability statements refer to the particular throw or to the particular result of shuffling the cards, on which the gambler lays his stake. The state of rational uncertainty in which he is placed may be *equated* to that of the different situation which can be imagined in which his throw is chosen at random out of an aggregate of throws, or of shufflings, which might equally well have occurred, though such aggregates exist only in imagination." (Fisher, 1959, page 22)

"For Fisher, probability appears as a measure of uncertainty applicable in certain cases but, regretfully, not in all cases. For me, it is solely the answer to the question 'How frequently this or that happens.'" (Neyman, 1952, page 187)

"[...] Isn't this equivalent to discussing the probabilities of hypotheses themselves, which would be useless? E.g., it would be useless to discuss the probability of Student's hypothesis because this would be the same as the probability of *mu*= 0. As *mu* is an unknown constant, the probability of *mu* being equal to zero must be either $P\{mu = 0\} = 0$ or $P\{mu = 0\} = 1$ and, without obtaining precise information as to whether *mu* is equal to zero or not, it would be impossible to decide what is the value of $P\{mu = 0\}$. Undoubtedly, *mu* is an unknown constant and, as far as we deal with the theory of probability as described in my first two lectures, it is useless to consider $P\{mu = 0\}$." (Neyman, 1952, page 56)

Confidence intervals/Interval estimates

"[Confidence intervals] in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended." (APA *Publication Manual*, 2001, page 22).

"Significance testing in general has been a greatly overworked procedure, and in many cases where significance statements have been made it would have been better to provide an interval within which the value of the parameter would be expected to lie." (Box, Hunter & Hunter, 1978)

"Confidence intervals avoid the problems of classic significance tests. They do not require a-priori hypotheses, nor do they test trivial hypotheses. Confidence intervals comprise the information of a significance test and are considerably easier to understand, which results in their didactic superiority." (Brandstaetter, 1999, page 43)

"The question as to whether significance tests should replace confidence intervals or not can be answered with a guarded "yes". Confidence intervals contain the information of a significance test, therefore there is no loss of information and no risk involved when confidence intervals replace significance tests. Taken together, confidence intervals in addition to replications, graphic illustrations and meta-analyses seem to represent a methodically superior alternative to significance tests. Hence, in the long run, confidence intervals appear to promise a more fruitful avenue for scientific research." (Brandstaetter, 1999, page 43)

"Not all statistically significant differences are clinically significant. Fortunately, confidence intervals can address both clinical and statistical significance." (Braitman, 1991, page 515)

"In a large majority of problems (especially location problems) hypothesis testing is inappropriate: Set up the confidence interval and be done with it!" (Casella & Berger, 1987)

"Scientists often finish their analysis by quoting a P-value, but this is not the right place to stop. One still wants to know how large the effect is, and a confidence interval should be given where possible." (Chatfield, 1988, page 51)

"It is far more in-formative to provide a confidence interval." (Cohen, 1990, page 1310)

"Objection has sometimes been made that the method of calculating Confidence Limits by setting an assigned value such as 1% on the frequency of observing 3 or less (or at the other end of observing 3 or more) is unrealistic in treating the values less than 3, which have not been observed, in exactly the same manner as the value 3, which is the one that has been observed. *This feature is indeed not very defensible save as anapproximation* ". (Fisher 1956, page 66, italics added)

"[...] a confidence interval can function to indicate which values could not be rejected by a two-tailed test with alpha at .05. In this function, the confidence interval could replace the report of null hypothesis for just one value, instead of communicating the outcome of the tests of all values as null hypotheses." (Frick, 1996, page 383)

"Confidence intervals should play an important role when setting sample size, and power should play no role once the data have been collected. [...] In this commentary, we present the reasons why the calculation of power after a study is over is inappropriate and how confidence intervals can be used during both study design and study interpretation." (Goodman & Berlin, 1994, page 200).

"For interpretation of observed results, the concept of power has no place, and confidence intervals, likelihood, or Bayesian methods should be used instead." (Goodman & Berlin, 1994, page 205).

"When making inferences about parameters [...] hypothesis tests should seldom be used if confidence intervals are available [...] the confidence intervals could lead to opposite practical conclusions when a test suggests rejection of H_0 [...] even though H_0 is not rejected, the confidence interval gives more useful information." (Graybill, 1976)

"We cannot escape the logic of NHST by turning to point estimates and confidence intervals". (Hagen, 1997, page 22)

"For problems where the usual null hypothesis defines a special value for a parameter, surely it would be more informative to give a confidence range for that parameter." (Hinkley, 1987)

"How about 'alpha and beta risks' and 'testing the null hypothesis'? [...] The very beginning language employed by the statistician describes phenomena in which engineers/physical scientists have little practical interest! They want to know how many, how much, and how well [...] Required are interval estimates. We offer instead hypothesis tests and power curves." (Hunter, 1990)

"Point estimates and their associated CIs are much easier for students and researchers to understand and, as a result, are much less frequently misinterpreted. Any teacher of statistics knows that it is much easier for students to understand point estimates and CIs than significance testing with its strangely inverted logic. This is another plus for point estimates and CIs." (Hunter & Schmidt, 1997, page 56).

"Reporting of results in terms of confidence intervals instead of hypothesis tests should be strongly encouraged." (Jones, 1984)

"We recommend that authors display the estimate of the difference and the confidence limit for this difference." (Jones & Matloff, 1986)

"Prefer confidence intervals when they are available." (Jones & Tukey, 2000)

"Again it is not clear why such a set [the confidence interval] should be of interest unless one makes the natural error of thinking of the parameter as random and the confidence set as containing the parameter with a specified probability. Again, this is a statement only a Bayesian can make, although confidence intervals are often so misinterpreted. I find the classical quantities useless for making decisions and believe that they are widely misinterpreted as Bayesian because the Bayesian quantities are more natural." (Kadane, 1995, page 316)

"The preference of many mathematical statisticians for confidence interval procedures over significance tests is understandable since both procedures involve the same assumptions, but confidence interval procedures provide an experimenter with more information" (Kirk, 1982)

"I believe that science is best served when researchers focus on the size of effects and their practical significance. Questions regarding the size of effects are addressed with descriptive statistics and confidence intervals. It is hard to understand why researchers have been so reluctant to embrace confidence intervals." (Kirk, 2001, page 214

"It is easy to [...] throw out an interesting baby with the nonsignificant bath water. Lack of statistical significance at a conventional level does not mean that no real effect is present; it means only that no real effect is clearly seen from the data. That is why it is of the highest importance to look at power and to compute confidence intervals." (Kruskal)

"Estimation procedures provide more information [than significance tests]: they tell one about reasonable alternatives and not just about the reasonableness of one value." (Lindley, 1986)

"It is usually wise to give a confidence interval for the parameter in which you are interested." (Moore and McCabe)

"The researcher armed with a confidence interval, but deprived of the respectability of statistical significance must work harder to convince himself and others of the importance of his findings. This can only be good." (Oakes, 1986, page 66)

"Although the underlying logic is essentially similar they [confidence intervals] are not couched in the pseudo scientific hypotheses testing language of significance tests. They do not carry with them decision-making implications, but, by giving a plausible range for the unknown parameter, they provide a basis for a rational decision should one be necessary. Should sample size be inadequate this is signaled by the sheer width of the interval." (Oakes, 1986, pages 66-67)

"Above all, interval estimates *are* estimates of effect size. It is incomparably more useful to have a plausible range for the value of a parameter than to know, what whatever degree of certitude, what single value is untenable." (Oakes, 1986, page 67)

"A confidence interval certainly gives more information than the result of a significance test alone [...] I [...] recommend its use [standard error of each mean]." (Perry, 1986)

"A non-Bayesian states that there is a 95% chance that the [obtained] confidence interval contains the true value of the population mean. A Bayesian would say there is a 95% chance that the population mean falls between the obtained limits. One is a probability statement about the interval, the other about the population parameter." (Phillips, 1973, page 335)

"Frequentist reasoning allows that investigators may use the word *confidence* for the specific numerical interval, but they are explicitely forbidden to use the term *probability* when making inferences for the same interval. It is perhaps not suprising that students often have difficulty with this distinction." (Prusek, 1997 pages 288-289)

"Unfortunately, knowing that 95% of an infinite number of 95% confidence intervals would contain the population mean is not the inference that a researcher ordinarily desires. What usually is desired is not an inference about psi [the parameter of interest] based on an infinite number of confidence intervals but an inference about psi based on the results of the specific confidence intervals that is obtained in practice" (Reichardt & Gollob, 1997, page 263).

"It would not be scientifically sound to justify a procedure [confidence intervals] by frequentist arguments and to interpret it in Bayesian terms." (Rouanet, 1998, page 54).

"Whenever possible, the basic statistical report should be in the form of a confidence interval." (Rozeboom, 1960, *in* Morrison & Henkel, 1970, page 227)

"Prior to the appearance of Fisher's 1932 and 1935 texts, data analysis in individual studies was typically conducted using point estimates and confidence intervals." (Schmidt, 1996, page 121).

"If we mindlessly interpret a confidence interval with reference to whether the interval subsumes zero, we are doing little more than nil hypothesis statistical testing" (Thompson, 1998, page 800)

"An improved quantitative science would emphasize the use of confidence intervals (CIs), and especially CIs for effect sizes." (Thompson, 2002, page 25)

"Probably the greatest ultimate importance among all types of statistical procedures we now know, belongs to confidence procedures." (Tukey, 1960)

"Interval estimates should be given for any effect sizes involving principal outcomes. Provide intervals for correlations and other coefficients of association or variation whenever possible." (Wilkinson and Task Force on Statistical Inference, 1999)

Effect sizes/Magnitude of effects

"The general principle to be followed [...] is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship." (APA Publication Manual, 2001 page 26)

"Tests of the null hypothesis that there is no difference between certain treatments are often made in the analysis of agricultural or industrial experiments in which alternative methods or processes are compared. Such tests are [...] totally irrelevant. What are needed are estimates of magnitudes of effects, with standard errors." (Anscombe, 1956)

"If the test of significance is really of such limited appropriateness [...]. At the very least it would appear that we would be much better if we were to attempt to estimate the magnitude of the parameters in the populations; and recognize that we then need to make other inferences concerning the psychological phenomena which may be manifesting themselves in these magnitudes." (Bakan, 1967, *in* Morrison & Henkel, 1970, page 250)

"Nothing is more important in educational and psychological research than making sure that the effect size of results is evaluated when tests of statistical significance are used." (Carver, 1993, page 289)

"In many experiments it seems obvious that the different treatments must have produced some difference, however small, in effect. Thus the hypothesis that there is no difference is unrealistic: the real problem is to obtain estimates of the sizes of the differences." (Cochran & Cox, 1957)

"Estimates and measures of variability are more valuable than hypothesis tests." (Cormack, 1985)

"Statistical significance is quite different from scientific significance and [...] therefore estimation, at least roughly, of the magnitude of effects is in general essential regardless of whether statistically significant departure from the null hypothesis is achieved." (Cox, 1977, page 61)

"The primary purpose of analysis of variance is to produce estimates of one or more error mean squares, and not (as is often believed) to provide significance tests." (Finney)

"I conclude that effect sizes are the single best index of the relationship between theoretical predictions and the obtained data." (Harris, 1991)

"The commonest agricultural experiments [...] are fertilizer and variety trials. In neither of these is there any question of the population treatment means being identical [...] the objective is to measure how big the differences are." (Healy)

"Preoccupation with testing 'is there an interaction' in factorial experiments, [...] emphasis should be on 'how strong is the interaction?'" (Jones, 1984)

"A null hypothesis rejection means that the researcher is pretty sure of the direction of the difference. Is this any way to develop psychological theory? I think not. How far would physics have progressed if their researchers had focused on discovering ordinal relationships? What we want to know is the size of the difference between A and B and the error associated with our estimate; knowing that A is greater than B is not enough." (Kirk, 1996, page 754)

"The tests of null hypotheses of zero differences, of no relationships, are frequently weak, perhaps trivial statements of the researcher's aims [...] in many cases, instead of the tests of significance it would be more to the point to measure the magnitudes of the relationships, attaching proper statements of their sampling variation. The magnitudes of relationships cannot be measured

in terms of levels of significance." (Kish, *in* Morrison & Henkel, 1970)

"It is unfortunate that the reporting of effect sizes has been framed as a controversy. Reporting of effect sizes is, instead, simply good scientific practice." (Hyde, 2001, page 228)

"The experimental aim should not be to establish whether changes have occurred, but rather to estimate whether changes have occurred in excess of some stipulated magnitude and importance. When a 'significant difference' has been established, investigators must then measure the size of the effect and consider whether it is of any biological or medical importance." (Lutz & Nimmo, 1977, page 77)

"The question many researchers (especially those interested in the application of science to solve practical problems) want to ask is whether the effects are large enough to make a real difference. The statistical tests most frequently encountered in the social and behavioral sciences do not directly address this question." (Murphy & Myors, 1999, page 234)

"Unfortunately, many researchers do not report [...] [effect sizes] along with their *F*-test results. This is a pity. (Smithson, 2000 pages 245)

"The most commonly occurring weakness [...] is [...] undue emphasis on tests of significance, and failure to recognise that in many types of experimental work estimates of treatment effects, together with estimates of the errors to which they are subject, are the quantities of primary interest." (Yates)

Fiducial inference

"Maybe Fisher's biggest blunder [the fiducial inference] will become a big hit in the 21st century." (Efron, 1998, page 107)

"It is sometimes asserted that the fiducial method generally leads to the same results as the method of Confidence Intervals. It is difficult to understand how this can be so, since it has been firmly laid down that the method of confidence intervals does not lead to probability statements about parameters." (Fisher, 1959, page 26)

"When knowledge *a priori* in the form of mathematically exact probability statements is available, the fiducial argument is not used, but that of Bayes. Usually exact knowledge is absent, and, when the experiment can be so designed that estimation can be exhaustive, similar probability statements *a posteriori* may be inferred by the fiducial argument." (Fisher, 1990/1935 page 198)

"[...] for there is no other method [the fiducial method] ordinarily available for making correct statements of probability about the real world." (Fisher, 1990/1935, pages 198-199)

"Depuis 1973, l'Analyse des Comparaisons a intégré les techniques bayésiennes, classiques et contemporaines (Jeffreys, Lindley, etc.), mais en les utilisant avec une motivation fiduciaire (Fisher). Ces techniques nous paraissent en effet les mieux adaptées pour pallier les insuffisances [...] des tests de signification traditionnels." (Lecoutre, Rouanet & Denhière, 1988, page 384)

"It seems reasonable to postulate that the no-knowledge *a priori* distribution in classical inverse probability theory should be that distribution which, when experimental data capable of yielding a fiducial argument are now given, results in an *a posteriori* distribution identical with the corresponding fiducial distribution." (Rozeboom, 1960, page 229)

"The fiducial philosophy of inference is an alternative to, and compensates for, the deficiencies of the other two procedures of inference [Bayesian inference, frequentist inference]. It is unfortunate that its importance has been unduly overlooked." (Wang, 2000, page 105)

Frequentist (orthodox, classical) inference

"The classical design of clinical trials is dictated by the eventual analysis. If the design varies from that planned then classical analysis is impossible." (Berry, 1987, page 181).

"The steamroller of frequentism is not slowed by words." (Berry, 1993).

"In contrast to the logical development and intuitive interpretations of the Bayesian approach, frequentist methods are nearly impossible to understand, even for the best students." (Berry, 1997)

"Students in frequentist courses may learn very well how to calculate confidence intervals and P values, but they cannot give them correct interpretations. I stopped teaching frequentist methods when I decided that they could not be learned." (Berry, 1997)

"In attempts to teach the 'correct' interpretation of frequentist procedures, we are fighting a losing battle." (Freeman, 1993, page 1446)

"Maybe we should banish our use of the world *probability* and substitute *how often*, instead, if we stay with the frequentist approach. Then, perhaps we can stay frequentists and still be honest with ourselves." (Iversen, 2000, page 9)

"Some say that Bayesianism has feet of clay (the need to specify a prior); but at least its feet are out in the open for everyone to see and criticise. By contrast, frequentist statistics has no clothes, for it calculates an irrelevant number and pretends that this tells us something important about the hypotheses we are interested in." (Jefferys, 1995, page 122)

"Classical statistics was invented to make statistical inference 'objective.' In fact, classical statistics is no more objective than Bayesian statistics, but by hiding its subjectivity it gives the illusion of objectivity." Jefferys, 1992)

"Interestingly, the sampling distribution that orthodox theory does allow us to use is noting more than a way of describing our prior knowledge bout the 'noise' (measurement errors). This, orthodox thinking is in the curious position of holding it decent to use prior information about nosie, but indecent to use prior information about the 'signal' of interest." (Jaynes, 1985, page 30)

Misinterpretations of NHST and confidence intervals

"[the confidence level] a measure of the confidence we have that the interval does indeed contain the parameter of interest" Aczel (1995, page 205)

"[a significant result] indicates that the chances of the finding being random is only 5 percent or less." (Azar, 1999)

"The psychological literature is filled with misinterpretations of the nature of the tests of significance." (Bakan, 1967, *in* Morrison & Henkel, 1970, page 239)

"In addition to the overall interpretative bias there was a very strong interaction between the training and the transfer problems [$chi^2(1)=14.71$, $p<0.001$]." (Bassock *et al.*, 1995)

"Subject's performance was not affected by differences in the size of the assigned and the receiving sets [$chi^2(1)=0.08$, n.s.], so we combined the results of subjects [...]" (Bassock *etal.*, 1995)

"An alternative approach to estimation is to extend the concept of error bound to produce an interval of values that is likely to contain the true value of the parameter. " (Bhattacharyya & Johnson, 1977, page 243)

"Many instructors err in describing confidence intervals and even some texts err. But whether texts or instructors err in explaining them, students do not understand them. And they carry this misunderstanding with them into later life. Calculating a confidence interval is easy. But everyone except the cognoscenti believes that when one calculates 95% confidence limits of 2.6 and 7.9 say, the probability is 95% that the parameter in question lies in the interval from 2.6 to 7.9." (Berry, 1997)

"P values are nearly as obscure as confidence intervals." (Berry, 19997)

"Students in frequentist courses may learn very well how to calculate confidence intervals and P values, but they cannot give them correct interpretations. I stopped teaching frequentist methods when I decided that they could not be learned." (Berry, 1997)

"Inevitably, students (and essentially everyone else) give an inverse or Bayesian twist to frequentist measures such as confidence intervals and P values." (Berry, 1997)

"[...] when a statistician rejects the null hypothesis at a certain level of confidence, say .05, he may be fairly well assured (*p* = .95) that the alternative statistical hypothesis is correct." (Bolles, 1962, page 639)

"Ask your colleagues how they perceive the statement '95% confidence level lower bound of 77.5 GeV/c2 is obtained for the mass of the Standard Model Higgs boson'. I conducted an extensive poll in July 1998, personally and by electronic mail. The result is that for the large majority of people the above statement means that 'assuming the Higgs boson exists, we are 95% confident that the Higgs mass is above that limit, i.e. the Higgs boson has 95% chance (or probability) of being on the upper side, and 5% chance of being on the lower side', which is not what the operational definition of that limit implies." [D'Agostini, 2000]

"[...] we assert that the population mean probably falls within the interval that we have established. "(Elifson, Runyon & Haber, 1990, page 367)

"Comme nous l'avons dit, on a avantage à rechercher si une transformation de l'échelle des *x* peut conduire à un schéma linéaire, c'est-à-dire à un F_2 non significatif." (Faverge, 1975, tome 2, page 268)

"Further, two additional 2×2 chi-square tests found class status (graduate vs. undergraduate) to be independent of whether students appear to hold misconceptions (chi2 = 3.5, *df*= l, *p*>.05) and whether students passed the test (chi2 = 3.02, *df*=l, *p*>.05)." (Hirsch &O'Donnell, 2001, page 10)

"I see these answers [about confidence intervals: '95% of the intervals would fall between the two values of the parameter', '95% chance that the actual value will be contained within the confidence interval' [...] as cries in the wilderness about how the world view we try to construct for our customers is not a world view our customers are comfortable with." (Iversen, 200, page 8)

"Dobyns, Jahn, and others […] publish the *observed p*- value, calling this 'the probability of obtaining this result by chance.' Such a use of *p*-values is illegitimate and not condoned by standard statistical theory." (Jefferys, 1995, page 595)

"A random sample can be used to specify a segment or interval on the number line such that the parameter has a high probability of lying on the segment. The segment is called a confidence interval." Kirk (1982, page 42)

"We can be 95% confident that the population mean is between 114.06 and 119.94." (Kirk, 1982, page 43)

"La consultation des tables permet simplement de dire que l'on ne peut pas refuser l'hypothèse posée au début. Il est vrai que, dans la pratique, beaucoup diront, et cela par un abus de langage strict, que les 3 groupes ne présentent pas de différence significative entre eux, qu'ils appartiennent à la même population. L'interprétation correcte est bien : 'on ne peut pas refuser l'hypothèse posée au départ'." (Mialaret, 1996, page 127) "La valeur 0 étant comprise dans l'intervalle de confiance on ne peut pas refuser l'hypothèse nulle selon laquelle les deux séries de valeurs ont la même moyenne. On dira, en d'autres termes, que l'ensemencement n'a pas eu d'effet sur la prise des pêcheurs." (Mialaret, 1996, page 112)

"In these conditions [a *p*-value of 1/15], the odds of 14 to 1 that this loss was caused by seeding [of clouds] do not appear negligible to us." (Neyman *et al.*, 1969)

"[...] 95 [chances] out of 100 that the observed difference will hold up in future investigations." (Nunnally, 1975, page 195; quoted by Carver, 1978)

"[95% CI] an interval such that the probability is 0.95 that the interval contains the population value." (Pagano, 1990, page 288)

"En fonction de la valeur du Khi-deux et du nombre de degrés de liberté, le logiciel calcule la probabilité exacte. Si l'on se donne un seuil de 5% de risques, une probabilité inférieure à ce seuil signifie que l'erreur d'échantillonnage est faible, on suppose qu'il existe une dépendance entre les 2 variables ligne et colonne. Le hasard intervient seulement dans moins de 5 chances sur 100, dans la répartition observée des effectifs dans le tableau. Le hasard, l'erreur d'échantillonnage sont considérés comme négligeables. L'hypothèse d'indépendance est rejetée." (*QUESTION*, Grimmer logiciels, 1993).

"Par exemple, si dans un sondage de taille 1000, on trouve [fréquence] = 0,613, la proportion *pi* à estimer a une probabilité 0,95 de se trouver dans la fourchette: [...] [0,58; 0,64]." (Robert Cl., 1995, pages 221-222)

"La majorité des chercheurs en psychologie ont recours à une épreuve de *signification statistique* pour décider si les résultats obtenus confirment ou infirment leur hypothèse. Cette épreuve permet d'établir quelle est la probabilité d'obtenir de tels résultats plutôt que ceux correspondant à l'hypothèse nulle, soit un postulat statistique attribuant les variations comportementales à des erreurs d'échantillonnage et de mesure, ainsi qu'au hasard." (Robert M., 1995, page 66)

"In summary, the probability [of the effect] was established for several samples of psychologists. For one *N* of 20, *p*=.88; for one *N* of 19, *p*=.996; for a smaller *N* of 2, *p*='1.00' and for another *N* of 2, *p*='0.00'." (Rosenthal & Gaito, 1964)

"From the table the probability is 0.9985 [1-p] or the odds are 666 to 1 that [soporific] 2 is the better soporific." (Student, 1908, page 21).

"The fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial..." (Tryon, 1998, page 796).

"Most psychologists and other users of statistics believe that this minimum significance level is the 'probability that the results are due to chance' and many applied statistics texts support this belief." (Wilson, 1961, page 230)

Multiple comparison test procedures

"One of the attractions of the Bayesian approach is that there is no need to introduce a penalty term for performing thousands of simultaneous tests; Bayesian testing has a built-in penalty or Ockham’s razor effect'." (Scott & Berger, 2003).

"I have failed to find a single instance in which the Duncan test was helpful, and I doubt whether any of the alternative tests [multiple range significance tests] would please me better." (Finney)

"The blind need frequent warnings and help in avoiding the multiple comparison test procedures that some editors demand but that to me appear completely devoid of practical utility." (Finney)

"Multiple comparison methods have no place at all in the interpretation of data." (Nelder)

"The ritualistic use of multiple-range tests-often when the null hypothesis is a priori untenable [...] is a disease." (Preece)

Null hypothesis

"There is really no good reason to expect the null hypothesis to be true in any population [...] Why should any correlation coefficient be exactly .00 in the population? [...] why should different drugs have exactly the same effect on any population parameter." (Bakan, 1967, *in* Morrison & Henkel, 1970)

"[this] is patently absurd and not in fact what scientists do. They do not test the same hypothesis over and over again." (Camilleri, 1962)

"Statistical significance testing sets up a straw man, the null hypothesis, and tries to knock him down." (Carver, 1978, page 381)

"The research worker has been oversold on hypothesis testing. Just as no two peas in a pod are identical, no two treatment means will be exactly equal. [...] It seems ridiculous [...] to test a hypothesis that we a priori know is almost certain to be false." (Chew, 1976)

"In many experiments it seems obvious that the different treatments must have produced some difference, however small, in effect. Thus the hypothesis that there is no difference is unrealistic: the real problem is to obtain estimates of the sizes of the differences." (Cochran & Cox, 1957)

"Exact truth of a null hypothesis is very unlikely except in a genuine uniformity trial." (Cox)

"In typical applications, one of the hypotheses-the null hypothesis-is known by all concerned to be false from the outset." (Edwards, Lindman & Savage, 1963)

"A null hypothesis that yields under two different treatments have identical expectations is scarcely very plausible, and its rejection by a significance test is more dependent upon the size of an experiment than upon its untruth." (Finney)

"Is it ever worth basing analysis and interpretation of an experiment on the inherently implausible null hypothesis that two (or more) recognizably distinct cultivars have identical yield capacities?" (Finney)

"[...] it would therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as they are contradicted by the data: but that they are never capable of establishing them as certainly true." (Fisher, 1929, page 192)

"[...] and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist in order to give the facts a chance of disproving the null hypothesis." (Fisher, 1990/1935, page 16)

"It is evident that the null hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the 'problem of distribution', of which the test of significance is the solution." (Fisher, 1990/1935, page 16)

"A large enough sample will usually lead to the rejection of almost any null hypothesis [...] Why bother to carry out a statistical experiment to test a null hypothesis if it is known in advance that the hypothesis cannot be exactly true." (Good, 1983)

"The commonest agricultural experiments [...] are fertilizer and variety trials. In neither of these is there any question of the population treatment means being identical [...] the objective is to measure how big the differences are." (Healy)

"When we formulate the hypothesis that the sex ratio is the same in two populations, we do not really believe that it could be exactly the same." (Hodges & Lehmann, 1954)

"All populations are different, a priori." (Jones & Matloff, 1986)

"Because point hypotheses, while mathematically convenient, are never fulfilled in practice, convert them to limiting cases of interval hypotheses." (Jones & Tukey, 2000)

"Welcome to the *Journal of Articles in Support of the Null Hypothesis*. In the past other journals and reviewers have exhibited a bias against articles that did not reject the null hypothesis. We seek to change that by offering an outlet for experiments that do not reach the traditional significance levels ($p < .05$). Thus, reducing the file drawer problem, and reducing the bias in psychological literature. Without such a resource researchers could be wasting their time examining empirical questions that have already been examined. We collect these articles and provide them to the scientific community free of cost." (Journal of Articles in Support of the Null Hypothesis, 2002...)

"No one, I think, really believes in the possibility of sharp null hypotheses that two means are absolutely equal in noisy sciences." (Kempthorne,)

"It is ironic that a ritualistic adherence to null hypothesis significance testing has led researchers to focus on controlling the Type I error that cannot occur because all null hypotheses are false." (Kirk, 1996, page 747)

"Another criticism of standard significance tests is that in most applications it is known beforehand that the null hypothesis cannot be exactly true." (Kruskal)

"Unless one of the variables is wholly unreliable so that the values obtained are strictly random, it would be foolish to suppose that the correlation between any two variables is identically equal to 0.0000 [...] (or that the effect of some treatment or the difference between two groups is exactly zero)." (Lykken, 1968, *in* Morrison & Henkel, 1970)

"The test is asking whether a certain condition holds exactly, and this exactness is almost never of scientific interest." (Matloff, 1991)

"With regard to a goodness-of-fit test to answer whether certain ratios have given exact values, 'we know a priori this is not true; no model can completely capture all possible genetical mechanisms'." (Matloff, 1991)

"The number of stars by itself is relevant only to the question of whether H_0 is exactly true-a question which is almost always not of interest to us, especially because we usually know a priori that H_0 cannot be exactly true." (Matloff, 1991)

"We usually know in advance of testing that the null hypothesis is false." (Morrison & Henkel, 1969, *in* Morrison & Henkel, 1970)

"The null-hypothesis models [...] share a crippling flaw: in the real world the null hypothesis is almost never true, and it is usually nonsensical to perform an experiment with the sole aim of rejecting the null hypothesis." (Nunnally, 1960)

"If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data." (Nunnally, 1960)

"The mere rejection of a null hypothesis provides only meager information." (Nunnally, 1960)

"And when, as so often, the test is of a hypothesis known to be false [...] the relevance of the conventional testing approach remains to be explicated." (Pratt, 1976)

"Null hypotheses of no difference are usually known to be false before the data are collected [...] when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science." (Savage, 1957)

"One feature [...] which requires much more justification than is usually given, is the setting up of unplausible null hypotheses. For example, a statistician may set out a test to see whether two drugs have exactly the same effect, or whether a regression line is exactly straight. These hypotheses can scarcely be taken literally." (Smith, 1960)

"Most researchers mindlessly test only nulls of no difference or of no relationship because most statistical packages only test such hypotheses." (Thompson, 1998, page 799)

"It is foolish to ask 'Are the effects of A and B different?' They are always different – for some decimal place." (Tukey, 1991, page 100)

"The worst, i.e., most dangerous, feature of 'accepting the null hypothesis' is the giving up of explicit uncertainty […] Mathematics can sometimes be put in such black-and-white terms, but our knowledge or belief about the external world never can." (Tukey, 1991)

"Never use the unfortunate expression 'accept the null hypothesis'." (Wilkinson and Task Force on Statistical Inference, 1999)

"In many experiments [...] it is known that the null hypothesis customarily tested, i.e. that the treatments produce no effects, is certainly untrue [...]." (Yates, 1964, page 320)

"The occasions [...] in which quantitative data are collected solely with the object of proving or disproving a given hypothesis are relatively rare." (Yates)

One-sided *vs* two-sided tests

"I regard the one-sided vs. two-sided *p* value debate to be silly." (Berry, 1991, page 86)

"While the popularity of one-tailed tests is undoubtedly attributable in part to the overwillingness of psychologists as a group to make use of the statistical recommendations they have most recently read, [...]" (Burke, 1953)

Power

"In either case it is inappropriate for authors to claim that there is no effect if the results of a test yield a nonsignificant *P*-value and low power." (Cherry, 1998, page 949

"[...] the question of the probability that his investigation would lead to statistically significant results, i.e., its power?" (Cohen, 1969, page vii)

"A salutary effect of power analysis is that it draws one forcibly to consider the magnitude of effects. In psychology, and especially in soft psychology, under the sway of the Fisherian scheme, there has been little consciousness of how big things are. [...] Because science is inevitably about magnitudes, it is not surprising how frequently *p* values are treated as surrogates for effect sizes. [...] In retrospect, it seems to me simultaneously quite understandable yet also ridiculous to try to develop theories about human behavior with *p* values from Fisherian hypothesis testing and no more than a primitive sense of effect size. And I wish I were talking about the long, long ago." (Cohen, 1990, page 1309).

"Confidence intervals should play an important role when setting sample size, and power should play no role once the data have been collected. [...] In this commentary, we present the reasons why the calculation of power after a study is over is inappropriate and how confidence intervals can be used during both study design and study interpretation." (Goodman & Berlin, 1994, page 200).

"For interpretation of observed results, the concept of power has no place, and confidence intervals, likelihood, or Bayesian methods should be used instead." (Goodman & Berlin, 1994, page 205).

Predictive probabilities

"An essential aspect of the process of evaluating design strategies is the ability to calculate predictive probabilities of potential results." (Berry, 1991, page 81)

p-values

"A Neyman-Pearson error probability, alpha, has the actual frequentist interpretation that a long series of alpha level tests will reject no more than 100alpha% of true *H* 0, but the data-dependent P-values have no such interpretation. P-values do not even fitt easily into any of the conditional frequentist paradigms." (Berger & Delampady, 1987)

"*P*-values calculated assuming fixed sample sizes may be reasonable as measures of extremity, to answer the question 'How unusual are the data if H_0 is true?', but one should not take them too seriously." (Berry, 1985, page 525).

"P values are nearly as obscure as confidence intervals." (Berry, 1997)

"It is very bad practice to summarise an important investigation solely by a value of P." (Cox)

"p-values can be used to spot a possible problem, but certainly not to draw scientific conclusions or to take decisions." (D'Agostini, 2000, page 18)

"Although *p*-values are not the most direct index of this information [about the strength of evidence], they provide a reasonable surrogate within the constraints posed by the mechanics of traditional hypothesis testing." (Dixon, 1998, page 391)

"The actual value of *p* [...] indicates the strength of the evidence against the hypothesis." (Fisher, 1990/1925, page 80)

"The current widespread practice of using *p*-values as the main means of assessing and reporting the results of clinical trials cannot be defended." (Freeman, 1993, page 1443)

"Even statisticians seem to have very little idea of how the interpretation of *p*-values should depend on sample size." (Freeman, 1993)

"[...] *whatever assumptions one makes,the observed p-value is not a valid estimate of the probability that the nullhypothesis is true*, and in fact, it always underestimates this probability by a large factor." (Jefferys, 1995, page 597)

"P-values, as provided by orthodox statistical methods, can be and often are misunderstood even by those who use them every day. Data-dependent P-values contain subtle traps that makes their interpretation hazardous." (Jefferys, 1992)

"If P is small, that means that there have been unexpectedly large departures from prediction. But why should these be stated in terms of P? The latter gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value is nearly always negligible. *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure." (Jeffreys, 1961, page 385, italics added)

"It is a travesty to describe a *p* value [...] as 'simple, objective and easily interpreted' [...] To use it as a measure of closeness between model and data is to invite confusion." (Healy)

"Editors must be bold enough to take responsibility for deciding which studies are good and which are not, without resorting to letting the *p* value of the significance tests determine this decision." (Lykken, 1968)

"[...] *p* values may be preferred to confidence intervals precisely because *p* values, in some ways, present less information than confidence intervals." (Reichardt & Gollob, 1997, page 275).

"There is no statistical sense to significance levels." (Rubin, 1969)

"Need we – should we – stick to *p*=0.05 if what we seek is a relatively pure list of appearances? No matter where our cutoff comes, we will not be sure of all appearances. Might it not be better to adjust the critical *p* moderately – say to 0.03 or 0.07 – whenever such a less standard value seems to offer a greater fraction of presumably real appearances among those significant at the critical *p*? We would then use different modifications for different sets of data. No one, to my knowledge, has set himself the twin problems of how to do this and how well doing this in a specific way performs." (Tukey, 1969, page 85)

Replication of experiments

"[...] smaller samples produce statistics more frequently which deviate widely from parameter than do large samples. Thus the large differences in a small sample must always be replicated in large samples to assess substantive importance." (Gold, 1958, *in* Morrison & Henkel, 1970, page 108)

"The essence of science is replication: a scientist should always be concerned about what would happen if he or another scientist were to repeat his experiment." (Guttman, 1983)

Scientific/experimental research

"We find that null hypothesis testing is uninformative when no estimates of means or effect size and their precision are given. Contrary to common dogma, tests of statistical null hypotheses have relatively little utility in science and are not a

fundamental aspect of the scientific method." (Anderson, Burnham & Thompson, 2000, page 912)

"Nor do you find experimentalists typically engaged in disproving things. They are looking for appropriate evidence for affirmative conclusions. Even if the mediate purpose is the disestablishment of some current idea, the immediate objective of a working scientist is likely to be gain affirmative evidence in favor of something that will refute the allegation which is under attack." (Berkson, 1942)

"However, statistics is not merely a set of methods for analyzing data. It is also a way for integrating data into the scientific process." (Berry, 1995, preface)

"There is no doubt that most scientists would disavow knowledge of Bayesian methods. But these same scientists think and reason like Bayesians, whether or not they know Bayes' theorem. Namely, they update what they think on the basis of the results of experiments." (Berry, 1997)

"The resultant magnification of the importance of formal hypothesis tests has inadvertently led to underestimation by scientists of the area in which statistical methods can be of value and to a wide misunderstanding of their purpose." (Box, 1976)

"In the past, the need for probabilities expressing prior belief has often been thought of, not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief […] I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters." (Box, 1980).

"In problem of scientific inference we would usually, were it possible, like the data to 'speak by themselves'." (Box & Tiao, 1973, page 2)

"But in psychology, like it or not, NHSTP [Null Hypothesis Significance Test Procedure] is the principal tool for testing substantive hypotheses in theory-corroborating studies. and in that capacity it is not only inadequate, but may be destructive to psychology as a scientific discipline." (Dar, 1998, page 196)

"The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to supposed frequencies of possible statements, based on them, being right or wrong, thus seems to miss the essential nature of such tests. [...] However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." (Fisher, 1990/1956, pages 44-45)

"I shall emphasize some of the various ways in which this operation [acceptance procedures] differs from that by which improved theoretical knowledge is sought in experimental research. This emphasis is primilary necessary because the needs and purposes of workers in the experimental sciences have been so badly misunderstood and misrepresented." (Fisher 1990/1956, pages 81-82)

"None of the meaningful questions in drawing conclusions from research results – such as how probable are the hypotheses? how reliable are the results? what is the size and impact of the effect that was found? – is answered by the test." (Falk & Greenbaum, 1995, page 94)

"The single most important problem with null hypothesis testing provides researcher with no incentive to specify either their own research hypotheses [...] Testing an unspecified hypothesis against chance may be all we can do in situations where we know very little. But when used as a general ritual, this method ironically ensures that we continue to know very little." (Gigerenzer, 1998, page 200)

"The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science." Good (1973)

"However, conventions about significant result should not be turned into canons of good scientific practice. Even more emphatically, a convention must not be made a superstition. [...] It is a grave error to evaluate the 'goodness' of an experiment only in terms of the significance level of its results." (Hays, 1973, page 385)

"Based on my own experience, most good research psychologists consult only occasionally with statistical experts. Thus [...] experts [...] more often see poor practice [...]. Such situations [...] offer the statistician little insight concerning the effective roles of statistical methods in good scientific work." (Krantz, 1999, page 1375)

"Psychology will be a much better science when we change the way we analyze data." (Loftus, 1996)

"Problems stemming from the fact that hypothesis tests do not address questions of scientific interest." (Matloff, 1991)

"The test provides neither the necessary nor the sufficient scope or type of knowledge that basic scientific social research requires." (Morrison & Henkel, 1969, *in* Morrison &Henkel, 1970, page 198)

"The use of significance tests involves the researcher in the process of making firm 'reject' or 'accept' decisions on each test of each null hypothesis on the basis of a formal, firm, and frequently arbitrary criterion, the significance level. This *decision making process* is antithetical to the *information accumulation process* of scientific inference." (Morrison & Henkel, 1970, page 309)

"Why do editors think that *P*-value dominated analysis constitutes a scientific procedure?" (Nelder, 1996, quoted by Cherry, 1998, page 947)

"Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong." (Neyman & Pearson, 1933a, page 291)

"We have suggested that a statistical test may be regarded as a rule of behaviour to be applied repeatedly in our experience when faced with the same set of alternative hypotheses." (Neyman & Pearson, 1933b, page 509)

"[...] Berger and Hsu (1996, page 192) make the following statement: "We believe that notions of size, power, and unbiasedness are more fundamental than 'intuition'..." In our opinion, such a statement places the credibility of statistical science at serious risk within the scientific community. If we are indeed teaching our students to disregard intuition in scientific inquiry, then a fundamental reassessment of the mission of mathematical statistics is urgently needed." (Perlman & Wu, 1999, page 366) [Berger's reply: "If we are indeed teaching our students to disregard intuition in scientific inquiry, then a fundamental reassessment of the mission of mathematical statistics is urgently needed." (page 373)]

"We hope that we have alerted statisticians to the dangers inherent in uncritical application of the NP [Neyman & Pearson] criterion, and, more generally, convinced them to join Fisher, Cox and many others in carefully weighing the scientific relevance and logical consistency of any mathematical criterion proposed for statistical theory." (Perlman & Wu, 1999, page 381)

"While several serious objections to the method [the null hypothesis significance-test method] are raised, its most basic error lies in mistaking the aim of a scientific investigation to be a *decision*, rather than a *cognitive* evaluation of propositions." (Rozeboom, 1960, *in* Morrison & Henkel, 1970, page 230)

"The null-hypothesis significance test treats 'acceptance' or 'rejection' of a hypothesis as though these were decisions one makes. *But the primary aim of a scientific experiment is notto precipitate decisions, but to make an appropriate adjustment in the degreeto which one accepts, or believes, the hypothesis or hypotheses being tested*." (Rozeboom, 1960, *in* Morrison &Henkel, 1970, page 221)

"It is the claim of this paper that, rather than being the 'only correct way to analyse a clinical trial', this paradigm [the 'intent to treat' paradigm] is a warning that we should heed Fisher's original observation that the N-P [Neyman-Pearson] formulation is irrelevant to scientific research." (Salsburg, 1994, page 334)

"Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution. After decades of unsuccessful efforts, it now appears possible that reform of data analysis procedures will finally succeed. If so, a major impediment to the advance of scientific knowledge will have been removed." (Schmidt & Hunter, in Harlow, Mulaik & Steiger, 1997, chap. 3)

[Statistical significance tests] "provide a façade of scientism in research. For many in educational research, being quantitative is equated with being scientific…despite the fact that some scientists and many psychologists [...] have managed very well without inferential statistics. (Shaver, 1992, page 2)

"A common problem for statistical inference is to determine, in terms of probability, whether observed differences between two samples signify that the populations sampled are themselves really different." (Siegel, 1956, page 2)

"Establishing that a correlation exists between two variables may be the ultimate aim of a research, [...]." "It is, of course, of some interest to be able to state the degree of association between two sets of scores from a given group of subjects. But it is perhaps of greater interest to be able to say whether or not some observed association in a sample of scores indicates that the variables under study are most probably associated in the population from which the sample was drawn." (Siegel, 1956, page 195)

"Science becomes an automated, blind research for mindless tabular asterisks using thoughtless hypotheses." (Thompson, 1998, page 799)

"The tyranny of the N-P [Neyman-Pearson] theory in many branches of empirical science is detrimental, not advantageous, to the course of science." (Wang, 1993)

"In many experiments [...] it is known that the null hypothesis customarily tested, i.e. that the treatments produce no effects, is certainly untrue; such experiments are in fact undertaken with the different purpose of assessing the magnitude of the effects. Fisher himself was of course well aware of this, as is evinced in many of his own analyses of experimental data, but he did not, I think, sufficiently emphasise the point in his expository writings. Scientists were thus encouraged to expect final and definite answers from their experiments in situations in which only slow and careful accumulation of information could be hoped for. And some of them, indeed, came to regard the achievement of a significant result as an end in itself." (Yates, 1964, page 320)

"The emphasis given to formal tests of significance [...] has resulted in [...] an undue concentration of effort by mathematical statisticians on investigations of tests of significance applicable to problems which are of little or no practical importance [...] and [...] it has caused scientific research workers to pay undue attention to the results of the tests of significance [...] and too little to the estimates of the magnitude of the effects they are investigating." (Yates)

"The unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective." (Yates)

Statistical significance/nonsignificance

"Typically, mere difference from zero is totally uninteresting." (Abelson, 1997, page 121)

"Rather than ask if these differences are statistically significant, it seems more important to ask if they are of educational importance." (Chatfield, 1985)

"1. A significant effect is not necessarily the same thing as an interesting effect; 2. A non-significant effect is not necessarily the same thing as no difference." (Chatfield, 1988, page 51)

"Researchers and journal editors as a whole tend to (over)rely on 'significant differences' as the definition of meaningful research." (Craig, Eison & Metze, 1976, page 282)

"However, conventions about significant result should not be turned into canons of good scientific practice. Even more emphatically, a convention must not be made a superstition. [...] It is a grave error to evaluate the 'goodness' of an experiment only in terms of the significance level of its results." (Hays, 1973, page 385)

"Acceptability of a statistically significant result [...] promotes a high output of publication. Hence the argument that the techniques work has a tempting appeal to young biologists, if harassed by their seniors to produce results, or if admonished by editors to conform to a prescribed ritual of analysis before publication. [...] the plea for justification by works [...] is therefore likely to fall on deaf ears, unless we reinstate reflective thinking in the university curriculum." (Hogben, 1957)

"I believe we can already detect signs of such deterioration in the growing volume of published papers – especially in the domain of animal behaviour – recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration." (Hogben, 1957, *in* Morrison & Henkel, 1970, page 21)

"We can already detect signs of such deterioration in the growing volume of published papers [...] recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration." (Hogben, 1957)

"To use statistics adequately, one must understand the principles involved and be able to judge whether obtained results are statistically significant *and* whether they are meaningful in the particular research context." (Kerlinger, 1989, pages 318-319)

"Significance should stand for meaning and refer to substantive matter. [...] I would recommend that statisticians discard the phrase 'test of significance'." (Kish, 1957, *in* Morrison & Henkel, 1970)

"Statistical significance of a sample bears no necessary relationship to possible subject-matter significance." (Kruskal)

"It is easy to [...] throw out an interesting baby with the nonsignificant bath water. Lack of statistical significance at a conventional level does not mean that no real effect is present; it means only that no real effect is clearly seen from the data. That is why it is of the highest importance to look at power and to compute confidence intervals." (Kruskal)

"We are also concerned about the use of statistical significance - P values - to measure importance; this is like the old confusion of substantive with statistical significance." (Kruskal & Majors, 1989)

"Statistical significance (alpha and *p* values) and practical significance (effect sizes) are not *competing* concepts. They are *complementary* ones." (Levin, 1993, page 379)

"It is important to ask whether we really want to test the existence or nonexistence of a relation. Suppose a relation is extremely weak: Is such a relation of interest? Probably not, in most cases; yet a large enough sample would find such a relation to be significantly different from chance. On the other hand, an extremely strong relationship would be found not significantly different from chance if the sample were very small. (Lipset, Trow & Coleman, 1956, *in* Morrison & Henkel, 1970, page 85)

"The idea that one should proceed no further with an analysis, once a non-significant *F*-value for treatments is found, has led many experimenters to overlook important information in the interpretation of their data." (Little, 1981)

"The moral of this story is that the finding of statistical significance is perhaps the least important attribute of a good experiment: it is *never* a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence – or that an experimental report ought to be published." (Lykken, 1968)

"The finding of statistical significance is perhaps the least important attribute of a good experiment." (Lykken, *in* Morrison & Henkel, 1970)

"Editors must be bold enough to take responsibility for deciding which studies are good and which are not, without resorting to letting the *p* value of the significance tests determine this decision." (Lykken, 1968)

"Scientists care about whether a result is statistically significant, but they should care much more about whether it is meaningful." (McCloskey, 1995)

"Too many users of the analysis of variance seem to regard the reaching of a mediocre level of significance as more important than any descriptive specification of the underlying averages." (McNemar, 1960)

"So much of what should be regarded as preliminary gets published, then quoted as the last word, which it usually is because the investigator is too willing to rest on the laurels that come from finding a significant difference. Why should he worry about the degree of relationship or its possible lack of linearity." (McNemar, 1960)

"In psychological and sociological investigations involving very large numbers of subjects, it is regularly found that almost all correlations or differences between means are statistically significant." (Meehl, 1967, *in* Morrison & Henkel, 1970)

[significant] "cancerous" "misleading" (Meehl, 1997, page 421)

"[...] I am cautioning that we must not get caught up in the misguided belief that having statistically significant things makes our research significant." (Moore, 1992)

"[...] there is ample evidence that it is impossible to use the term 'significant' in a statistical context and avoid the erroneous connotations of that term (for writers *and* readers)." (Morrison & Henkel, 1969, *in* Morrison & Henkel, 1970, page 198)

"To have the latter [scientific inference] we will have to have much more than the façade that claims of significance provide." (Morrison & Henkel, 1969, *in* Morrison & Henkel, 1970, page 198)

"We should not feel proud when we see the psychologist smile and say 'the correlation is significant beyond the .01 level.' Perhaps that is the most that he can say, but he has no reason to smile." (Nunnally, 1960)

"To make measurements and then ignore their magnitude would ordinarily be pointless. Exclusive reliance on tests of significance obscures the fact that statistical significance does not imply substantive significance." (Savage, 1957)

"I hope most researchers understand that *significant* (statistically) and *important* are two different things. Surely the term *significant* was ill chosen" (Schafer, 1993, page 387)

"For many years, medical research has overrated the importance of *p*-values and thereby statistical and clinical significance have been mixed up and misinterpreted." (Schmidt, 1995, page 483)

"Is there anybody who would believe that the two values are exactly the same? The problem is to get a reliable estimates for the difference. You want not statistical significance but practical significance." [Schmitt, 1969, page 255]

"Nonsignificance was generally interpreted [in the *Journal of Abnormal Psychology*, 1984] as confirmation of the null hypothesis (if this was the research hypothesis), although the median power was as low as .25 in these cases." (Seldmeier & Gigerenzer, 1989)

"Many users of tests confuse statistical significance with substantive importance or with size of association." (Selvin, 1957, *in* Morrison & Henkel, 1970, page 106)

"Moreover, the tendency to dichotomy resulting from judging some results 'significant' and other 'nonsignificant' can be misleading both to professional and lay audiences." (Skipper, Guenther & Nass, 1967)

"There is no guarantee, form SS [Statistical Significance] that the mean difference is greater than infinitesimal." (Sohn, 1998, page 299)

"In many experiments it is well known [...] that there are differences among the treatments. The point of the experiment is to estimate [...] and provide [...] standard errors. One of the consequences of this emphasis on significance tests is that some scientists [...] have come to see a significant result as an end in itself." (Street, 1990)

"The emphasis on significance levels tends to obscure a fundamental distinction between the size of an effect and its statistical significance." (Tversky & Kahneman, 1971)

"The interpretations which have commonly been drawn from recent studies indicate clearly that we are prone to conceive of statistical significance as equivalent to social significance. These two terms are essentially different and ought not to be confused. [...] Differences which are statistically significant are not always socially important. The corollary is also true: differences which are not shown to be statistically significant may nevertheless be socially significant." (Tyler, 1931, pages 115-117)

"The experimenter must keep in mind that significance at the 5% level will only coincide with practical significance by chance!" (Upton, 1992)

"But is vital that a *statistically significant* difference should not necessarily be assumed to be an *important* difference. [...] It is extremely important that doctors give thought to these matters and that they are not persuaded by advertisers or others to accept statistically significant differences in the performance of drugs as necessarily indicating a difference of practical importance of value." (Wade & Waterhouse, 1977, page 412).

"The word 'significant' could be abolished [...] Based on a dictionary definition, one might expect that results that are declared significant would be important, meaningful, or consequential. Being 'significant at an arbitrary probability level' [...] ensures none of these." (Warren, 1986)

"Results are significant or not significant and this is the end of it." (Yates, 1951)

Editorial policies/Guidelines/Propositions

"We estimated that 47% (SE=3.9%) of the P-values in the Journal of Wildlife Management lacked estimates of means or effect sizes or even the sign of the difference in means or other parameters. We find that null hypothesis testing is uninformative when no estimates of means or effect size and their precision are given. Contrary to common dogma, tests of statistical null hypotheses have relatively little utility in science and are not a fundamental aspect of the scientific method. We recommend their use be reduced in favor of more informative approaches." (Anderson, Burnham & Thompson, 2000, page 912)

"If pressed, we would probably argue that Bayesian statistics (with emphasis on objective Bayesian methodology) should be the type of statistics that is taught to the masses, with frequentist statistics being taught primarily to advanced statisticians" (Bayarri & Berger, 2003, page 3)

"In presenting the main results of a study it is good practice to provide confidence intervals rather than to restrict the analysis to significance tests. Only by doing so can authors give readers sufficient information for a proper conclusion to be

done." (Berry, 1986, *The Medical Journal of Australia*, Editorial)

"Therefore, intending authors are urged to express their main conclusions in confidence interval form (possibly with the addition of a significance test, although strictly that would provide no extra information)." (Berry, 1986, *TheMedical Journal of Australia*, Editorial).

"Significance tests are intended solely to address the viability of the null hypothesis that a treatment has no effect, and not to estimate the magnitude of the treatment effect. Researchers are advised to move away from significance tests and to present instead an estimate of effect size bounded by confidence intervals." (Borenstein, 1997, *Annals of Allergy, Asthma, & Immunology*, page 5)

"The statistical descriptors known as confidence intervals can increase the ability of readers to evaluate conclusions drawn form small trials." (Braitman, 1988, *Annals of Internal Medicine*, Editorial)

"[...] the point estimate both summarizes the sample and infers the true value; it should always be reported. Confidence intervals should be used to assess the clinical significance as well as the statistical significance of the main study results. When space permits, presenting all the raw data for important results (for example, in a graph) is best; this is practical only for relatively small studies. In reporting results of statistical tests, exact *P* values are preferable to verbal statements of 'statistical significance' (or $P<0.05$) or of nonsignificance ($P>0.05$) because they contain more information." (Braitman, 1991, *Annals ofInternal Medicine*, Editorial)

"In a large majority of problems (especially location problems) hypothesis testing is inappropriate: Set up the confidence interval and be done with it!" (Casella & Berger, 1987)

"My main recommendation is for wildlife researchers to stop taking statistical testing so seriously." (Cherry, 1998, page 951)

"Since power is a direct monotonic function of sample size, it is recommended that investigators use larger sample sizes than they customarily do. It is further recommended that research *plans* be routinely subjected to power analysis, using as conventions the criteria of population effect size employed in this survey." (Cohen, 1962, page 153)

"Rather, we are proposing that indices of association are another part of a composite picture a researcher is building when he reports data suggesting one or more variables are important in understanding a particular behavior." (Craig, Eison & Metze, 1976, page 282)

"I think that much clarity will be achieved if we remove from scientific parlance the misleading expressions 'confidence intervals' and 'confidence levels'." (D'Agostini, 2000)

"Together with many recent critics of NHT, we also urge reporting of important hypothesis tests in enough descriptive detail to permit secondary uses such as meta-analysis." (Greenwald, Gonzalez, Harris & Guthrie, 1996, page 175)

"As statisticians, we owe it to researchers using statistics in their research to make clear the impact statistics has on their work and enable them to choose Bayesian methods. We should train researchers well enough to make it possible for them to understand the role Bayesian statistics can play in their work." (Iversen, 2000, page 10)

"Authors are required to report and interpret magnitude-of-effect measures in conjunction with every *p* value that is reported." (Heldref Foundation, 1997, *Journal of ExperimentalEducation*, Guidelines)

"As a teacher, I therefore feel that to continue the time honoured practice – still in effect in many schools – of teaching pure orthodox statistics to students, with only a passing sneer at Bayes and Laplace, is to perpetuate a tragic error which has already wasted thousands of man-years of our fines mathematical talent in pursuit of false goals. If this talent had been directed toward understanding Laplace's contributions and learning how to user them properly, statistical practice would be far more advanced than it is." (Jaynes, 1976, page 256)

"Reporting of results in terms of confidence intervals instead of hypothesis tests should be strongly encouraged." (Jones, 1984)

"We recommend that authors display the estimate of the difference and the confidence limit for this difference." (Jones & Matloff, 1986)

"The only remedy [...] is for journal editors to be keenly aware of the problems associated with hypothesis tests, and to be sympathetic, if not strongly encouraging, toward individuals who are taking the initial lead in phasing them out." (Jones & Matloff, 1986)

"The reader will find that no traditional significance tests have been reported in connection with the statistical results in this volume. This is intentional policy rather than accidental oversight." (Kendall, 1957, *in* Morrison & Henkel, 1970, page 87)

"Evaluations of the outcomes of psychological treatments are favourably enhanced when the published report includes not only statistical significance and the required effect size but also a consideration of clinical significance. That is, [...] it is also important for the evaluator to consider the degree to which the outcomes are clinically significant (e.g., normative comparisons)." (Kendall, 1997, *Journal of Consulting and Clinical Psychology*, Editorial)

"As a remedial step, I would recommend that statisticians discard the phrase 'test of significance', perhaps in favor of the somewhat longer but proper phrase 'test against the null hypothesis or the abbreviation 'TANH'. (Kish, 1959, *in* Morrison & Henkel, 1970, page 139)

"We suggest that the sole effective therapy for curing its [NHST] 'ills' is a *smooth transition* towards the Bayesian paradigm." (Lecoutre, Lecoutre & Poitevineau, 2001, page 413)

"In this book, no statistical tests of significance have been used." (Lipset, Trow & Coleman, 1956, *in* Morrison & Henkel, 1970, page 81)

"In particular, I offer the following guidelines. 1. By default, data should be conveyed as a figure depicting sample means *with associated standard errors and/or, where appropriate, standarddeviations*. 2. More often than not, inspection of such a figure will immediately obviate the necessity of any hypothesis-testing procedures. In such situations, presentation of the usual hypothesis information (*F* values, *p* values, etc.) will be discouraged." (Loftus, 1993, *Memory & Cognition*, Editorial comment)

"When a 'significant difference' has been established, investigators must then measure the size of the effect and consider whether it is of any biological or medical importance." (Lutz & Nimmo, 1977, *European Journal of Clinical Investigation*, Editorial)

"It is usually wise to give a confidence interval for the parameter in which you are interested." (Moore & McCabe)

"If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit." (Murphy, 1997, *Journal of Applied Psychology*, Editorial)

"In reporting results, authors should still provide measures of variability and address the issue of the generalizability and reliability of their empirical findings across people and materials. There are a number of acceptable ways to do this, including reporting MSEs and confidence intervals and, in case of within-subject or within-items designs, the number of people or items that show the effect in the reported direction." (Neeley, 1995, *: Learning, Memory, and Cognition*, *21*, page 261)

"A confidence interval certainly gives more information than the result of a significance test alone [...] I [...] recommend its use [standard error of each mean]." (Perry, 1986)

"The norm should be that only a standard error is quoted for comparing means from an experiment." (Preece)

"[...] confidence intervals are unlikely to be widely reported in the literature unless their use is encouraged, or at least not penalized, by the publication criteria of journals." (Reichardt & Gollob, 1997, page 282)

"Bayesian hypothesis testing is reasonably well developed […] and well worth inclusion in the arsenal of any data analyst." (Robinson & Wainer, 2002, page 270)

"In the past, journals have encouraged the routine use of tests of statistical significance; I believe the time has now come for journals to encourage routine use of confidence intervals instead." (Rothman, 1978, *The NewEngland Journal of medicine*, Editorial)

"Whenever possible, the basic statistical report should be in the form of a confidence interval." (Rozeboom, 1960, *in* Morrison & Henkel, 1970)

"Accepting the proposition that significance testing should be discontinued and replaced by point estimates and confidence intervals entails the difficult effort of changing the beliefs and practices of a lifetime. Naturally such a prospect provokes resistance. Researchers would like to believe there is a legitimate rationale for refusing to make such a change." (Schmidt & Hunter, 1997, page 49)

"It just seemed high time that someone stirred the Bayesian pot on an elementary level so that practitioners, rather than

theorists, could start discussions and supply feedback to one another." (Schmitt, 1969, preface)


"It is recommended that, when inferential statistical analysis is performed, CIs [confidence intervals] should accompany point estimates and conventional hypothesis tests wherever possible." (Sim & Reid, 1999, page 186)


"We will go further [than mere encouragement]. Authors reporting statistical significance will be *required* to both report and interpret effect sizes. However, these effect sizes may be of various forms, including standardized differences, or uncorrected (e.g., r^2, R^2, eta^2) or corrected (e.g., adjusted R^2, $omega^2$) variance-accounted-for statistics." (Thompson, 1994, *Educational and Psychological Measurement*, Guidelines)


"It is proposed to judge the clinical relevance and importance by means of four values, fixed in discussions with the clinician before commencement of the study, and to proceed by testing non-zero nullhypotheses (shifted nullhypotheses) where the 'clinically relevant difference' is the shift parameter." (Victor, 1987, page 109)


"It would seem to us to be easier for those who design clinical trials to continue to use the usual form of tests of significance based on the null hypothesis. But is vital that a *statisticallysignificant* difference should not necessarily be assumed to be an *important* difference. [...] It is extremely important that doctors [...] are not persuaded by advertisers or others to accept statistically significant differences in the performance of drugs as necessarily indicating a difference of practical importance of value." (Wade & Waterhouse, 1977, *British Journal of Clinical Pharmacology*, Editorial).


"It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual *p* value or, better still, a confidence interval. Never use the unfortunate expression 'accept the null hypothesis.' Always provide some effect-size estimate when reporting a *p* value." (Wilkinson and Task Force on Statistical Inference, 1999)


"Interval estimates should be given for any effect sizes involving principal outcomes. Provide intervals for correlations and other coefficients of association or variation whenever possible." (Wilkinson and Task Force on Statistical Inference, 1999)


"Always present effect sizes for primary outcomes. If the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day), then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure." (Wilkinson and Task Force on Statistical Inference, 1999)


"Provide information on sample size and the process that led to sample size decisions. Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations. Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size." (Wilkinson and Task Force on Statistical Inference, 1999)


"We encourage researchers to use CIs to present their research findings, rather than relying on p-values alone." (Wolfe & Cumming, 2004, page 138).