Equipe Raisonnement Induction Statistique

Bibliographie / Bibliography

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Bibliographie sur la pratique de l'inférence statistique

Bibliography about the practice of statistical inference

Mise à jour 01 mars 2005 / Updating 01 march 2005

Des centaines de références

sur la pratique de l'inférence statististique dans le traitement des données expérimentales (essentiellement en psychologie, et aussi dans les essais cliniques):

- Références théoriques et méthodologiques sur l'usage des tests
- Critiques des tests ("la controverse des tests de signification"
- Exemples d'abus d'utilisation
- Fondements de l'inférence statistique
- Solutions de rechange (en particulier méthodes bayésiennes) et exemples d'applications

La sélection d'une référence dans cette bibliographie n'est pas un jugement de valeur.

Several hundred references

about the practice of statistical inference in the analysis of experimental data (especially in psychology, also in clinical trials):

- Theoretical and methodological references about the use of significance tests
- Criticisms of significance tests ("the significance test controversy")

- Exemples of abuses and misuses
- Fundations of statistical inference
- Alternative solutions (especially Bayesian methods) and examples of applications

The selection of a reference in this bibliography is not a judgment of value.

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)

A

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Abdi, H. (1987). *Introduction au Traitement Statistique des Données Expérimentales*. Grenoble: Presses Universitaires de

Grenoble.

Abelson, R.P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 129-133.

Abelson, R.P. (1995). *Statistics as Principled Argument.* Hillsdale, NJ: Erlbaum.

Abelson, R.P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15

Abelson, R.P. (1997). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). *In* L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What If There Were No Significance Tests?* Hillsdale, NJ: Erlbaum, 117-141

Acklin, M.W., McDowell, C.J., & Orndoff, S. (1992). Statistical power and the Rorschach: 1975-1991. *Journal of Personality Assessment*, *59*, 366-379.

Acree, M.C. (1978). *Theories of Statistical Inference in Psychological Research: A Historico-critical Study*. Dissertation Abstracts International, 39, 5073B (University Microfils N° H790 H7000).

Aczel, A.D. (1995). *Statistics: Concepts and applications*. Chicago: Irwin.

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of Ph.D. programs in North America. *American Psychologist*, *45*, 721-734.

Albert, J. (1995). Teaching Inference about Proportions Using Bayes and Discrete Models. *Journal of Statistics Education*, *3*(3).

Albert, J. (1997). Teaching Bayes' rule: A data-oriented approach. *The American Statistician*, *51*, 247-253.

Algina J., Moulder B.C. (2001). Sample sizes for confidence intervals on the increase in the squared multiple correlation coefficient. *Educational and Psychological Measurement*, *61*, 633-649.

Altham, P.M.E. (1969). Exact Bayesian analysis of a 2x2 contingency table and Fisher's "exact" significance test. *Journal of the Royal Statistical Society*, *Series B*, *31*, 261-269.

Altman, D.G. (1982). Statistics in medical journals. *Statistics in Medicine*, *1*, 59-71.

Altman, D. G. (1985). Discussion of Dr. Chatfield's paper. *Journal of the Royal Statistical Society*, *Series A*, *148*, 242.

Altman, D. G. (1992). Confidence intervals in research evaluation. *ACP J Club.*, *Suppl 2*, A8-9.

Altman, D.G., & Bland J. (1991). Improving doctors' understanding of statistics. *Journal of the Royal Statistical Society*, *Series A*, *154*, 223-267.

Altman, D. G., Gore, S. M., Gardner, M. J., & Pocock, S. J. (1983). Statistical guidelines for contributors to medical journals. *British Medical Journal*, *286*, 1489-1493.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th edition). Washington, DC: Author.

American Psychological Association, Board of Scientific Affairs (1996). Task force on statistical inference initial report (draft).

Amery, W.K., Hoing, M., Debroye, M., & Dom, F. (1987). Some comments on the use of statistics in the evaluation of drug trials in migraine. *Neuroepidemiology*, *6*, 220-227.

Amorim, M.A. (1999). A neurocognitive approach to human navigation. *In* R.G. Golledge (Ed.), *Wayfinding Behavior*, Baltimore, MA: The Johns Hopkins University Press, 152-167.

Amorim, M.A., Glasauer, S., Corpinot, K., & Berthoz, A. (1997). Updating an object's orientation and location during nonvisual navigation: A comparison between two processing modes. *Perception and Psychophysics*, *59*, 404-418.

Amorim, M.A., Loomis, J.M., & Fukusima, S.S. (1998). Reproduction of object shape is more accurate without the continued availability of visual information. *Perception*, *27*, 69-86.

Amorim, M.-A., & Stucchi, N. (1997). Viewer- and object-centered mental explorations of an imagined environment are not equivalent. *Cognitive Brain Research, 5*, 229-239.

Amorim, M.-A., Trumbore, B., & Chogyen, P.L. (2000). Cognitive repositioning inside a desktop VE: The constraints introduced by first-versus third-person imagery and mental representation richness. *Presence: Teleoperators and Virtual Environments, 9*, 165-186.

Anderson, D.R., Burnham, K.P., & Thompson, W.L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*, 912-923.

Anderson, D.R., Link, W.A., Johnson, D.H., & Burnham, K.P. (2001). Suggestions for presenting the results of data analyses. *Journal of Wildlife Management, 65*, 373-378.

Anderson, & Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics, Theory and Methods, 12*, 2663-2692.

Anderson, W. T. (1992). Trouble in paradigms: robobuyer versus the blob - part 2. *Marketing and Research Today, 20*, 87-94.

Anscombe, F.J. (1956). Discussion of paper by F.N. David and N.L. Johnson. *Journal of the Royal Statistical Society, Series B, 18*, 24-27.

Anscombe, F.J. (1963). Sequential clinical trials. *Journal of the American Statistical Association, 58*, 365-383.

Anscombe, F.J. (1990). The summarizing of clinical experiments by significance levels. *Statistics in Medicine, 9*, 703-708.

Arabie, P., & Hubert, L.J. (1996). An overview of Combinatorial Data Analysis. *In* P. Arabie, L.J. Hubert & G. de Soete (Eds.), *Clustering and Classification*, World Scientific.

Arbuthnott, J. (1710). An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society, 27*, 186-190 [Reprinted *in* M. Kendall &R.L. Plackett (Eds.), *Studies in the History of Statistics and Probability, Vol. II*, London: Griffin].

Argimon, J.M. (2002). El intervalo de confianza: algo más que un valor de significación estadística [Confidence intervals: something more than a statistical significance test]. *Medicina Clinica, 118*, 382-384.

Aron, A., & Aron, E. N. (1999). *Statistics for Psychology* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Arvey, R.D., Cole, D.S., Hazucha, J.F., & Hartanto, F.M. (1985). Statistical power of training evaluation designs. *Personal Psychology, 38*, 493-507.

Atkins, L., & Jarrett, D. (1981). The significance of significance tests *In* J. Irvine, I. Miles & J. Evans (Eds), *Demystifying Social Statistics*, London: Pluto Press, 87-109.

Atkinson, D.R., Furlang, M.J., & Wampold, B.E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology, 29*, 189-194.

Azar, B. (1997). APA task force urges a harder look at data. *APA Monitor Online, 28*, 26.

Azar, B. (1999). APA statistics task force prepares to release recommendations for public comment. *APA Monitor Online, 30*.

B

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Bacher, F. (1998). L'utilisation des modèles dans l'analyse des structures de covariance. *L'Année Psychologique*, sous presse.

Bacon, F.T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 241-252.

Bailar, J.C., & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations. *Annals of Internal Medicine*, *108*, 266-273 [Reprinted *in* J. C. Bailar & F. Mosteller (Eds.), *Medical uses of statistics* (2nd edition), Boston, Mass.: New England Journal of Medicine Books, 313-331].

Bailar, J.C., & Mosteller, F. (1992). Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations. *In* J.C. Bailar & F. Mosteller (Eds.), *Medical uses of statistics* (2nd edition), Boston, Mass.: New England Journal of Medicine Books, 313-331.

Baird, D. (1984). Tests of significance violate the rule of implication. *Proceedings of the 1984 Biennal Meeting of the Philosophy of Science Association*, *1*, 81-92.

Bakan, D. (1967/1966). The test of significance in psychological research. *On Method*, San Francisco: Jossey-Bass, 1-29 [Reprinted in Morrison & Henkel, 1970, 231-251; early version in *Psychological Bulletin*, 1966, *66*, 423-437].

Bakan, D. (1967). *On Method: Toward a Reconstruction of Psychological Investigation*. San Francisco, CA: Jossey-Bass, Inc.

Bandt, C. L., & Boen, J. R. (1972). A prevalent misconception about sample size, statistical significance, and clinical importance. *Journal of Periodontics*, *43*, 181-183.

Baril, G.L., & Cannon, J.T. (1995). What is the probability that null hypothesis testing is meaningless [Comment]. *American Psychologist*, *50*, 1098-1099. Barlow, D.H. (1981). On the relation of clinical research to clinical practice: Current issues, new directions. *Journal of Consulting and Clinical Psychology*, *49*, 147–155>

Barnard, G.A. (1947). The meaning of a significance level. *Biometrika*, *34*, 179-182.

Barnard, G.A. (1989). On alleged gains in power from lower *p* values. *Statistics in Medicine*, *8*, 1469-1477.

Barnard, G.A. (1990). Must clinical trials be large? The interpretation of *p*-values and the combination of tests results. *Statistics in Medicine*, *9*, 601-614.

Barnard, G. A. (1992). Statistics and OR - some needed interactions. *Journal of the Operational Research Society*, *43*, 787-795.

Barnard, G. (1998). Letter. *New Scientist*, *157*, 47.

Barndorff-Nielsen, O. (1977). Discussion of D. R. Cox's paper. *Scandinavian Journal of Statistics*; *4*, 67-69.

Barnett, V. (1982). *Comparative Statistical Inference* (2nd edition). New York: Wiley.

Barnett, M.L., & Mathlsen, A. (1997). Tyranny of the p-value: The conflict between statistical significance and common sense [Editorial]. *Journal Dent. Research*, *76*, 534-536.

Bartko, J.J. (1991). Proving the null hypothesis [Comment]. *American Psychologist*, *46*, 1089.

Bassok, M., Wu, L.L., & Olseth, K.L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition*, *23*, 354-367.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Reasoning*, *2*, 75-98.

Bayarri, M. J. & Berger, J. O. (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, *19*, 58–80.

Bayes, T. (1763). Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370-418. [Reprinted in *Biometrika*, 1958, *45*, 293-315]

Beale, D.K. (1972). What's so significant about .05? *American Psychologist*, *27*, 1079-1080.

Beauchamp, K.L., & May, R.B. (1964). Replication report: Interpretation of levels of significance by psychological researchers. *Psychological Reports*, *14*, 272.

Beaven, E.S. (1935). Discussion on Dr. Neyman's Paper. *Journal of the Royal Statistical Society*, *Supplement 2*, 159-161.

Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, *25*, 16-39.

Beck-Bornholdt, H.P., & Dubben, H.H. (1994). Potential pitfalls in the use of *p*-values in the interpretation of significance levels. *Radiotherapy and Oncology*, *33*, 177-178.

Becker, G. (1991). Alternative methods of reporting research results. *American Psychologist*, *46*, 654-655.

Begg, I., Armour, V., & Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioral Science*, *17*, 199-214.

Bellhouse, D.R. (1993). Invited commentary: *p* values, hypothesis tests, and likelihood. *American Journal of Epidemiology*, 137, 497-499.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*, *Series B*, 289-300.

Bennett, J.H. (1990). *Statistical Inference and Analysis (Selected Correspondence of R.A. Fisher*). Oxford, U.K.: Clarendon Press.

Berg, A.O. (1979). Some non-random views of statistical significance. *Journal of Family Practice*, *8*, 1011-1014.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer Verlag.

Berger, J. O. (1986). Are P-values reasonable measures of accuracy? *In* I. S. Francis, B. F. J. Manly, & F. C. Lam (Eds.), *Pacific Statistical Congress*, New York: Elsevier, 21-27.

Berger, J.O., & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity. *The American Scientist*, *76*, 159-165.

Berger, J.O., & Berry, D.A. (1988). The Relevance of Stopping Rules in Statistical Inference, *Statistical Decision Theory and Related Topics IV*, *1*, New York: Springer-Verlag, 29-72.

Berger, J.O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317-352.

Berger, J.O., & Sellke, T. (1987). Testing a point null hypothesis: the irreconciability of *P* values and evidence [With discussion]. *Journal of the American Statistical Association*, *82*, 112-139.

Berger J.O., & Wolpert R.L. (1988). *The Likelihood Principle* (2nd edition), Hayward, CA: Institute of Mathematical Statistics.

Berger, R.L., & Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, *11*, 283-319.

Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics In Medicine*, *19*, 1319-1328.

Bergin, A.E., & Strupp, H.H. (1972). *Changing Frontiers in the Science of Psychotherapy*. Chicago: Aldine-Atherton.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526-542.

Berkson, J. (1941). Comments on Dr. Madow's "Note on tests of departure from normality" with some remarks concerning tests of significance. *Journal of the American Statistical Association*, *46*, 539-541.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325-335.

Berkson, J. (1943). Experience with tests of significance: A reply to Professor R.A. Fisher. *Journal of the American Statistical Association*, *38*, 242-246.

Bernard, J.-M. (1986). Méthodes d'inférence bayésienne sur des fréquences. *Informatique et Sciences Humaines*, *68-69*, 89-133.

Bernard, J.-M. (1996). Bayesian interpretation of frequentist procedures for a Bernoulli process. *The American Statistician*, *50*, 7-13.

Bernard, J.-M. (2000). Bayesian inference for categorized data. *In* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference* (2nd edition), Bern: Peter Lang, 159-226.

Bernard, J.-M., Blancheteau, M., & Rouanet, H. (1985). Le comportement prédateur chez un forficule, Eurobellia Moesta (Géné). *Biology of Behaviour*, *10*, 1-22.

Bernardo J.M., & Smith, A.F.M. (1994). *Bayesian Theory*. Chichester: John Wiley & Sons.

Berry, D.A. (1980). Statistical inference and the design of clinical trials. *Biomedicine*, *32*, 4-7.

Berry, D.A. (1985). Interim analysis in clinical trials: Classical vs. Bayesian approach. *Statistics in Medicine*, *4*, 521-526.

Berry, D.A. (1987). Statistical inference, designing clinical trials, and pharmaceutical company decisions. *The Statistician*, *36*, 181-189.

Berry, D.A. (1987). Interim analysis in clinical trials: The role of the likelihood principle, *The American Statistician*, *41*, 117-122.

Berry, D.A. (1988). Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective [with discussion]. *In* J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith (Eds.), *Bayesian Statistics 3*, Oxford, England: Oxford University Press, 79-94.

Berry, D.A. (1989). Monitoring accumulating data in a clinical trial. *Biometrics*, *45*, 1197-1211.

Berry, D.A. (1991). Experimental design for drug development: A Bayesian approach. *Journal of Biopharmaceutical Statistics*, *1*, 81-101.

Berry, D.A. (1991). Bayesian methodology in phase III trials. *Drug Information Association Journal*, *25*, 345-368.

Berry, D.A. (1993). A case for Bayesianism in clinical trials. *Statistics in Medicine*, *12*, 1377-1393 [With discussion, pages 1395-1404].

Berry, D.A. (1994). *Basic Statistics: A Bayesian Approach*. Belmont: Wadsworth.

Berry, D.A. (1995). Decision analysis and Bayesian methods in clinical trials. *Cancer Treat. Research*, *75*, 125-154.

Berry, D.A. (1995). Decision Analysis and Bayesian Methods in Clinical Trials. *In* P. Thall (Ed.), *Recent Advances in Clinical Trial Design and Analysis*, New York: Kluwer Press, 125-154.

Berry, D.A. (1996). *Statistics: A Bayesian Perspective*. Belmont, CA: Duxbury Press.

Berry, D.A. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician*, *51*, 241-246.

Berry, G. & Armitage, P. (1995). Mid-*P* confidence intervals: a brief review. *The Statistician*, *44*, 417-423.

Berry, D.A., & Hochberg Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 215-227.

Berry, D.A., & Lindgren, B.W. (1996). *Statistics: Theory and Methods* (2nd edition). Duxbury Press.

Berry, D.A., & Stangl D.K. (1996). *Bayesian Biostatistics*. New York: Marcel Dekker:

Berry, D.A., Stangl D.K. (1996). Bayesian methods in health-related research. *In* D.A. Berry and D.K. Stangl (Eds.), *Bayesian Biostatistics*, New York: Marcel Dekker, 1-66.

Berry, G. (1986). Statistical significance and confidence intervals [Editorial]. *The Medical Journal of Australia*, *144*, 618-619.

Beshers J (1958). On "A critique of tests of significance in survey research". *American Sociological Review*, *23*, 199 [Reprinted in Morrison & Henkel, 1970, 111-112].

Bezeau, S.; Graves, R. (2001). Statistical power and effect sizes of clinical Neuropsychology research. *Neuropsychology Development and Cognition, Section A Journal of Clinical and Experimental Neuropsychology*, *23*, 399-406.

Bhattacharyya & Johnson (1997). *Statistical Concepts and Methods*. New York: Wiley.

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *70*, 107-115.

Bird, K.D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement*,. *62*,

197-226.

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association*, *56*, 246-249.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association* [With discussion], *57*, 269-306.

Birnbaum, A. (1977). The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. *Synthese*, *36*, 19-49.

Birnbaum, I. (1982). Interpreting Statistical Significance. *Teaching Statistic*, *4*, 24-27.

Blackwelder, W.C. (1982). Proving the null hypothesis in clinical trials. *Controlled Clinical Trials*, *3*, 345-353.

Blackwelder, W.C., & Chang, M.A. (1984). Simple size graphs for proving the null hypothesis. *Controlled Clinical Trials*, *5*, 97-105.

Blaich, C.F. (1998). The null-hypothesis significance-test procedure: Can't live with it, Can't live without it. *Behavioral and Brain Sciences*, *21*, 194-195.

Blalock, H. M., Jr. (1972). *Social statistics* (2nd edition). New York: McGraw-Hill.

Boardman, T. J. (1994). The statistician who changed the world: W. Edwards Deming, 1900-1993. *The American Statistician*, *48*, 179-187.

Bofinger, E. (1985). Expanded confidence intervals. *Communications in Statistics A*, *14*, 1849-1864.

Bofinger, E. (1992). Expanded confidence intervals, one-sided tests and equivalence testing. *Journal of Biopharmaceutical Statistics*, *2*, 181-188.

Boik, R.J. (1993). The analysis of two-factor interactions in fixed effects linear models. *Journal of Educational Statistics*, *18*, 1-40.

Bolles, R. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, *11*, 639-645.

Bolles, R., & Messick, S. (1958). Statistical utility in experimental inference. *Psychological Reports*, *4*, 223-227.

Bondy, W.A. (1969). A test of an experimental hypothesis of negligible difference between means. *The American Statistician*, *23*, 28-30.

Boos, D.D. & Hughes-Oliver, J.M. (2000). How large does *n* have to be for *Z* and *t* intervals? *The American Statistician*, *54*, 121-128.

Borak, J., & Veilleux, S. (1982). Errors of intuitive logic among physicians. *Soc. Sci. Medicine*, *16*, 1939-1947

Borenstein, M. (1994). A note on the use of confidence intervals in psychiatric research. *Psychopharmacology Bulletin*, *30*, 235-238.

Borenstein, M. (1994). The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials*, *15*, 411-428.

Borenstein, M. (1997). Hypothesis testing and effect size estimation in clinical trials. *Annals of Allergy, Asthma, & Immunology*, *78*, 5-16.

Boring, E.G. (1919). Mathematical versus scientific significance. *Psychological Bulletin*, *16*, 335-338.

Bourke, S. (1993). Babies, bathwater and straw person: A response to Menon. *Mathematics Education Research Journal*, *5*, 19-22.

Box, G. E. P. 1976. Science and statistics. *Journal of the American Statistical Association*, *71*, 791-799.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A*, *143*, 383-430.

Box, G. E. P. (1983). An apology for ecumenism in statistics. *In* G. E. P. Box, T. Leonard and C. F. Wu (Eds.), *Scientific Inference, Data Analysis, and Robustness*, San Diego, CA: Academic Press, Inc., 51-84.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: John Wiley & Sons.

Box, G.E.P., & Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison Wesley.

Bozdogan, H. (1994). Editor's general preface. *In* H. Bozdogan (Ed.), *Engineering and Scientific Applications, Vol. 3. Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, Dordrecht, Netherlands: Kluwer, ix-xii.

Braithwaite, R. B. (1953). *Scientific Explanation: A Study of the Function of Theory, Probability and Law in Science*. Cambridge, England: Cambridge University Press.

Braitman, L.E. (1988). Confidence intervals extract clinically useful information from data [Editorial]. *Annals of Internal Medicine*, *108*, 296-298.

Braitman, L.E. (1991). Confidence intervals assess both clinical significance and statistical significance [Editorial]. *Annals of Internal Medicine*, *114*, 515-517.

Braitman, L.E. (1993). Statistical estimates and clinical trials. *Journal of Biopharmaceutical Statistics*, *3*, 249-256.

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online*, *4*(2), 33-46.

Bredenkamp, J. (1972). *Der Signifikanztest in der Psychologischen Forschung*. Frankfurt am Main: Akademische Verlagsgesellschaft.

Brennan, P., & Croft, P. (1994). Interpreting the results of observational research. *British Medical Journal*, *309*, 727-730.

Breslow, N. (1990). Biostatistics and Bayes [With comments]. *Statistical Science*, *5*, 269-298.

Brewer, J.K. (1972). On the power of statistical tests in the American Educational Research Journal. *American Educational Research Journal*, *9*, 391-401.

Brewer, J.K. (1985). Behavioral statistics textbooks: sources of myths and misconceptions? *Journal of Educational Statistics*, *10*, 252-268.

Brewer, J.K., & Owen, P.W. (1973). A note on the power of statistical tests in the Journal of Educational Measurement. *Journal of Educational Measurement*, *10*, 71-74.

Bristol, D.R. (1995). Delta: The true clinically significant difference to be detected. *Drug Information Journal*, *29*, 33-36.

Brophy, J.M., & Joseph, L. (1995). Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *Journal of the American Medical Association*, *273*, 871-875.

Bross, I.D. (1990). How to eradicate fraudulent statistical methods: Statisticians must do science. *Biometrics*, *46*, 1213-1225.

Brown, F.L. (1973). Introduction to statistical methods in psychology. *In* G.A. Miller & R. Buckhout (Eds.), *Psychology: The Science of Mental Life*, New York: Harper & Row.

Brown, J., & Hale, M.S. (1992). The power of statistical studies in consultation-liaison psychiatry. *Psychosomatics*, *33*, 437-443.

Brown, L.D. (1990). An ancillarity paradox which appears in multiple linear regression [with discussion]. *The Annals of Satistics*, *18*, 471-538.

Brown, L.D., Hwang, J.T.G., & Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics*, *25*, 2345-2367.

Browne, R.H. (1979). On visual assessment of the significance of a mean difference. *Biometrics*, *35*, 657-665.

Browne, R.H. (1995). Bayesian analysis and the GUSTO trial. Global Utilization of Streptokinase and Tissue Plasminogen Activator in Occluded Coronary Arteries [Letter]. *Journal of the American Medical Association*, 1995, *274*, 873.

Browner, W.S., & Newman, T.B. (1987). Are all significant *P* values created equal? The analogy between diagnostic tests and

clinical research. *Journal of the American Medical Association*, *257*, 2459-2463.

Bryan-Jones, J., & Finney, D.J. (1983). On an error in "Instructions to Authors". *HortScience*, *18*, 279-282.

Bryk, A.S., & Raudenbush, S.W. (1988). Heterogeneity of variance in experimental studies: a challenge to conventional interpretations. *Psychological Bulletin*, *104*, 396-404.

Buchanan-Wollaston, H.J. (1935). The philosophic basis of statistical analysis. *Journal of the International Council for the Exploration of the Sea*, *10*, 249-263.

Bulmer, M.G. (1957). Confirming statistical hypotheses. *Journal of the Royal Statistical Society, Series B*, *19*, 125-132.

Bulpitt, C.J. (1987). Confidence intervals. *Lancet*, *1*, 494-497.

Burke, C.J. (1954). Further remarks on one-tailed tests. *Psychological Bulletin*, *51*, 587-590.

C

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Camilleri, S.F. (1962). Theory, probability, and induction in social research. *American Sociological Review*, *27*, 170-178 [Reprinted in Morrison & Henkel, 1970, 142-154].

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

Campbell, M. (1992). Letter. *Royal Statistical Society News & Notes*, *18*, 5.

Campillo, A. C. (1996). Erroneous interpretation of *p* values [Spanish]. *Atencion Primaria*, *17*, 221-224.

Capone, C.A., Jr., & Seaman, S.L. (1989). Uses and misuses of hypothesis testing. *Journal of Business Forecasting Methods and Systems*, *8*, 18-27.

Capraro R.M., Capraro M.M. (2002). Treatments of effect sizes and statistical significance tests in textbooks. *Educational and Psychological Measurement*, *62*, 771-782.

Carlin, C, & Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (2nd edition). London: Chapman and Hall.

Carlson, R. (1976). The logic of tests of significance. *Philosophy of Science*, *43*, 116-128.

Carpenter, JA. (2001). Deliberations of the Task Force on Statistical Inference, *Journal of Studies on Alcohol*, *62*, 405-408.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Carver, R.P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287-292.

Casella, G., & Berger, L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem [With discussion]. *Journal of the American Statistical Association*, *82*, 106-135.

Chaloner, K. (1996). Elicitation of prior distributions. *In* D. Berry D., & D. Stangl D. [Eds], *Bayesian Biostatistics*. New York: Marcel Dekker.

Chaloner, K., Church, T., Louis, T., & Matts, J. (1993). Graphical elicitation of a prior distribution for a clinical trial. *The Statistician*, 1993, *42*, 341-353.

Charron, C. (2002). Conceptualization of fractions and categorization of problems for adolescent. *European Journal of Psychology of Education*, *17*, 115-128.

Chase, L.J., & Baran, S.J. (1976). An assessment of quantitative research in mass communication. *Journalism Quarterly*, *53*, 308-311.

Chase, L.J., & Chase, R.B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61*, 234-237.

Chase, L.J., Chase, R.B., & Tucker, R.K. (1978). Statistical power in physical anthropology: A technical report. *American Journal of Physical Anthropology, 49*, 133-138.

Chase, L.J., & Tucker, R.K. (1975). A power-analytic examination of contemporary communication research. *Speech Monographs, 42*, 29-41.

Chase, L.J., & Tucker, R.K. (1976). Statistical power: Derivation, development; and data-analytic implications. *The Psychological Record, 42*, 29-41.

Chatfield, C. (1985). The initial examination of data [With discussion]. *Journal of the Royal Statistical Society, Series A, 148*, 214-253.

Chatfield, C. (1988). *Problem Solving: A Statistician's Guide*. London: Chapman & Hall.

Chatfield, C. (1989). Comments on the paper by McPherson. *Journal of the Royal Statistical Society, Series A, 152*, 234-238.

Chatfield, C. (1991). Avoiding statistical pitfalls. *Statistical Science, 6*, 240-268.

Chernoff , H. (1986). Comment. *The American Statistician, 40*, 5-6.

Cherry, S. (1998). Statistical tests in publications of The Wildlife Society. *Wildlife Society Bulletin, 26*, 947-953.

Chew, V. (1976). Comparing treatment means: a compendium. *HortScience, 11*, 348-357.

Chew, V. (1977). Statistical hypothesis testing: an academic exercise in futility. *Proceedings of the Florida State Horticultural Society, 90*, 214-215.

Chew, V. (1980). Testing differences among means: correct interpretation and some alternatives. *HortScience, 15*, 467-470.

Chia, K.S. (1997). "Significant-itis" - an obsession with the P-value. *Scand. Journal Work Environ. Health, 23*, 152-154.

Choi, S.C., & Pepple, P.A. (1989). Monitoring clinical trials based on predictive probability of significance. *Biometrics, 45*, 317-323.

Chow, S.L. (1988). Significance tests or effect size? *Psychological Bulletin, 103*, 105-110.

Chow, S.L. (1989). Significance tests and deduction: Reply to Folger (1989). *Psychological Bulletin, 106*, 161-165.

Chow, S.L. (1991). Conceptual rigor versus practical impact. *Theory and Psychology, 1*, 337-360.

Chow, S.L. (1991). Rigor and logic: A response to comments on "conceptual rigor". *Theory and Psychology, 1*, 389-400.

Chow, S.L. (1991). Some reservation about power analysis [Comment]. *American Psychologist, 46*, 1088-1089.

Chow, S.L. (1996). *Statistical Significance: Rationale, Validity and Utility*. London: Sage.

Chow, S. L. (1998). What Statistical Significance Means. *Theory & Psychology, 8*, 323-330.

Chow, S.L. (1998). Open Peer Commentary and author's response / Statistical Significance: Rationale, Validity and Utility. *Behavioral and Brain Sciences, 21*, 194-239.

Chow, S. L. (2002). Issues in statistical inference. *History and Philosophy of Psychology Bulletin, 14*, 30-41.

Christensen, J.E., & Christensen, C.E. (1977). Statistical power analysis of health, physical education, and recreation research. *Research Quarterly, 48*, 204-208.

Ciancia, F., Maitte, M., Honoré, J., Lecoutre, B., & Coquery, J.-M. (1988). Orientation of attention and sensory gatting: An evoked potential and RT study in cat. *Experimental Neurology, 100*, 274-287.

Clark, C.A. (1963). Hypothesis testing in relation to statistical methodology. *Review of Educational Research, 33*, 455-473.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the *British Journal of Psychology*. *British Journal of Psychology*, *88*, 71-83.

Clément, E., & Richard, J.-F. (1997). Knowledge of domain effects in problem representation: the case of Tower of Hanoi isomorphs. *Thinking and Reasoning*, *3*, 133-157.

Clements, M.A. (1993). Statistical significance testing: Providing historical perspective for Menon's paper. *Mathematics Education Research Journal*, *5*, 23-27.

Clopper, C.J., & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404-413.

Coats, W. (1970). A case against the normal use of inferential statistical models in educational research. *Educational Researcher*, *June*, 6-7.

Cochran, W.G., & Cox, G.M. (1957). *Experimental designs* (2nd edition). New York: John Wiley & Sons, Inc.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153.

Cohen, J. (1965). Some statistical issues in psychological research. *In* B.B. Wolman (Ed.), *Handbook of Clinical Psychology*, New York: McGraw-Hill, 95-121.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*, 426-443.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edition, 1st edition: 1969). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*, 98-105.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist*, *49*, 997-1003. With comment and replies replies *in American Psychologist*, 1995, *50*, 1098-1103.

Cohen, L.H. (1979). Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, *47*, 421-423.

Connolly, R.A. (1991). A posterior odds analysis of the weekend effect. *Journal of Econometrics*, *49*, 51-104.

Cooke, R.W., & Weindling, A.M. (1993). Clinical trials and *P* values. *Pediatrics*, *92*, 188-189.

Cooper & Topher (1994). Anomalous propagation. *Journal of Scientific Exploration*, *8*, 401-402.

Cooper, H.M. (1989). *Integrating Research: A Guide for Literature Reviews* (2nd edition). Beverly Hills, CA: Sage.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*, 447-452.

Cooper, H., & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin*, *8*, 168-173.

Cooper, H.M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, *87*, 442-449.

Cormack, R. M. (1985). Discussion of Dr. Chatfield's paper. *Journal of the Royal Statistical Society, Series A*, *148*, 231-233.

Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician*, *20*, 18-23.

Cornfield, J. (1966). A Bayesian test of some classical hypotheses - with applications to sequential clinical trials. *Journal of the American Statistical Association*, *61*, 577-594.

Cornfield, J. (1969). The Bayesian outlook and its application. *Biometrics*, *25*, 617-657.

Corroyer, D., Devouche, E., Bernard, J.-M., Bonnet, P., & Savina, Y. (2003). Comparaison de six logiciels pour l'analyse de la variance d'un plan S<A2*B2> déséquilibré. *L'Année Psychologique*, *103*, 277-312.

Corroyer, D., Rouanet, H. (1994). Sur l'importance des effets et des indicateurs dans l'analyse statistique des données. *L'Année Psychologique*, *94*, 607-624.

Cortina, J.M., & Dunlap, W.P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage.

Coursol, A., & Wagner, E.E. (1986). Effect of positive findings on submission and acceptance rates. *Professional Psychology: Research and Practice*, *17*, 136-137.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*, 553-558.

Cowles, M. (1989). *Statistics in psychology: An historical perspective*. Hillsdale, NJ: Erlbaum.

Cox, D.R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, *29*, 357-372.

Cox, D.R. (1977). The role of significance tests [With discussion]. *Scandinavian Journal of Statistics*, *4*, 49-70.

Cox, D.R. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology*, *14*, 325-331.

Cox, D.R. (1986). Some general aspects of the theory of statistics. *International Statistical Review*, *54*, 117-126.

Cox, D.R. (2001). Another comment on the role of statistical methods. *British Mediacl Journal*, *322*, 231.

Cox, D.R., & Snell, E.J. (1981). *Applied statistics: Principles and examples*. London: Chapman and Hall.

Craig, J.R., Eison, C.L., & Metze, L.P. (1976). Significance tests and their interpretation: An example utilizing published research and *omega-square*. *Bulletin of the Psychonomic Society*, *7*, 280-282.

Cronbach, L.J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116-127.

Cronbach, L.J., & Snow, R.E. (1977). *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York: Irvington.

Crow, E.L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology" [Comment]. *American Psychologist*, *46*, 1083.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-575.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist*, *60*, to appear.

Cumming, G., Thomason, N., Howard, A., Les, J., & Zangari, M. (1995). The StatPlay software for statistical understanding: Confidence intervals and hypothesis testing. *In* J. Pearce, A. Ellis, C. McNaught, & G. Hart (Eds.), *Learning with technology. Proceedings of the 12th annual Australian Society for Computers in Learning in Tertiary Education '95 conference*, University of Melbourne, 4-6 December, 104-112.

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

Cutler, S., Greenhouse, S., Cornfield, J., *et al.* (1966). The role of hypothesis testing in clinical trials. *Journal of Chronic Diseases*, *19*, 857-882.

D

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

D'Agostini, G. (1999). Teaching statistics in the physics curriculum. Unifying and clarifying role of subjective probability. *American Journal of Physics*, *67*, 1260-1268.

D'Agostini, G. (1999). Bayesian reasoning versus conventional statistics in high energy physics. Proceedings XVIII International Workshop on Maximum Entropy and Bayesian Methods, Garching, Germany. Kluwer Academic, 157-170.

D'Agostini, G. (2000). Role and meaning of subjective probability: some comments on common misconceptions. *AIP Conference Proceedings*, Melville, *568*, 23-30.

D'Agostini, G. (2000). Teaching Bayesian statistics in the scientific curricula, *The ISBA Newletter*, *7*, 1, 18.

D'Agostini, G. (2000). Confidence limits: what is the problem? Is there *the* solution? *Workshop on Confidence* at CERN, January 17-18.

D'Agostini, G. (2003). Bayesian inference in processing experimental data: principles and basic applications. Invited paper for *Reports on Progress in Physics*.

D'Agostini, G. (2003). *Bayesian reasoning in data analysis - A critical introduction*. River Edge, NJ: World Scientific Publishing.

D'Andrade, R., & Dart, J. (1990). The interpretation of *r* versus *r*2 or why percent of variance accounted for is a poor measure of size of effect. *Journal of Quantitative Anthropology*, *2*, 47-59.

Dahl, H. (1999). Teaching hypothesis testing. Can it still be useful? *Bulletin of the International Statistical Institute*, ISI 99, Proceedings, *Tome LVIII*, *Book 2*, 197-200.

Daly, J.A., & Hexamer, A. (1983). Statistical power in research in English education. *Research in the teaching of English*, *17*, 157-164.

Daniel, L.G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools*, *5*, 23-32.

Daniel, L.G. (1998). Fight the Good Fight: A Response to Thompson, Knapp, and Levin. *Research in the Schools*, *5*, 59-62.

Daniel, L.G. (1998). The statistical significance controversy is definitely not over: a rejoinder to responses by Thompson, Knapp, and Levin. *Research in the Schools*, *5*, 63-66.

Dar, E. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist*, *42*, 145-151.

Dar, R. (1998). Null hypothesis tests and theory corroboration: Defending NHSTP out of context. *Behavioral and Brain Sciences*, *21*, 196-197.

Dar, R., R., Serlin, C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology*, *62*, 75-82.

Davidoff, F. (1999). Standing statistics right side up [Editorial]. *Annals of Internal Medicine*, *130*, 1010-1021.

Dawes, R.M. (1988). Probabilistic versus causal thinking. *In* D. Cichetti & Grove (Eds.), *Thinking Clearly About Psychology*, *Vol. 1 Matters of Public Interest: Essays in honor of Paul Everett Meehl*, Minneapolis: University of Minnesota Press, 235-264.

Dawid P. (2000). A word from the president. *The ISBA Bulletin*, *7*, 1-2.

De Cristofaro, R. (1996). The role of inductive inference in statistical analysis. *Metron*, *54*, 17-29.

De Cristofaro, R. (2002). The inductive reasoning in statistical inference. *Communications in Statistics*: *Theory and Methods*, *31*, 1079-1089.

De Cristofaro, R. (2004). On the foundations of the likelihood principle. *Journal of Statistical Planning and Inference*, *126*, 401-411.

deFinetti, B. (1974). Bayesianism: Its unifying role for both the foundations and applications of statistics. *International Statistical Review*, *42*, 117-130.

DeGroot, M. H. (1989). *Probability and Statistics*. Reading, MA: Addison-Wesley.

Deheuvels, P. (1984). How to analyze bio-equivalence studies? The right use of confidence intervals. *Journal of Organizational Behaviour and Statistics*, *1*, 1-15.

DeLong, J. B., & Lang, K. (1992). Are all economic hypotheses false? *Journal of Political Economy*, *100*, 1257-1272.

del Rosal, A.B., Costas, C.S., Bruno, J.A.S., & Osinski, I.C. (2001). The judgment against null hypothesis. Many witnesses and a virtuous sentence. *Psicothema*, *13*, 173-178.

Deming, W. E. (1975). On probability as a basis for action. *The American Statistician*, *29*, 146-152.

Denhière, G., & Lecoutre, B. (1983). Mémorisation de récits: Reconnaissance immédiate et différée d'énoncés par des enfants de 7, 8 et 10 ans. *L'Année Psychologique*, *83*, 345-376.

Denis, D.J. (2005). The modern hypothesis testing hybrid: R. A. Fisher's fading Influence. *Journal de la SFdS* [With comments], to appear.

Dérozières, A. (1985). Histoire de formes: statistiques et sciences sociales avant 1940. *Revue Française de Sociologie*, *26*, 277-310.

Detsky, A.S., & Sackett, D.L. (1985). When is a negative clinical trial big enough. *Archives of Internal Medicine*, *145*, 709-712.

Diamond, G.A., & Forrester, J.S. (1983). Clinical trials and statistical verdicts: Probable grounds for appeal. *Annals of Internal Medicine*, *98*, 385-394.

Dignam, J.J., Bryant, J, Wieand, HS, *et al.* (1998). Early stopping of a clinical trial when there is evidence of no treatment benefit: protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. *Controlled Clinical Trials*, *19*, 575-588.

Dixon, P. (1998). Why scientists value *p* values. *Psychonomic Bulletin Research*, *5*, 390-396.

Dixon, P., & O'Reilly, T. (1999). Scientific versus statistical inference. *Canadian Journal of Experimental Pychology*, *53*, 133-149.

Dodd, D.H., & Schultz, R.F., Jr., (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin*, *79*, 391-395.

Dracup, C. (1995). Hypothesis testing: What it really is. The *Psychologist*, *8*, 359-362.

Duggan, T.J., & Dean, C.W. (1968). Common misinterpretations of significance levels in sociological journals. *The American Sociologist*, *3*, 45-46 [Reprinted in Morrison & Henkel, 1970, 161-165].

DuMouchel, W. (1989). Bayesian metaanalysis. *In* D.A. Berry (Ed.), *Statistical Methodology in the Pharmaceutical Science*, New York: Marcel Dekker, 509-529.

Duncan, D.B. (1965). A Bayesian approach to multiple comparisons. *Technometrics 7*, 171-222.

Dunlap, W.P., & May J.G. (1989). Judging statistical significance by inspection of error bars. *Journal of the Psychonomic Society*, *27*, 67-68.

Dunne A., Pawitan, Y., & Doody, L. (1996). Two-sided P-values from discrete asymmetric distributions based on uniformly most powerful unbiased tests. *Statistician*, *45*, 397-405.

Dunnet, C.W., & Gent, M. (1977). Significance testing to establish equivalence betweeen treatments with special reference to treatment in form of 2x2 tables. *Biometrics*, *33*, 593-602.

Durand, J.-L. (1997). Analyse de l'ouvrage de N. Guéguen, *Manuel de statistique pour psychologues*, Paris: Dunod, 1997. *L'Année Psychologique*, sous presse.

Dwyer, J.H. (1974). Analysis of variance and the magnitude of effects: A general approach. *Psychological Bulletin, 81*, 731-737.

E

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. *In* D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge, England: Cambridge University Press, 249-267.

Edgeworth, F.Y. (1885). Methods of Statistics. *Royal Statistical Society Jubilee Volume*, 181-217.

Edgington, E.S. (1964). A tabulation of inferential statistics used in psychology journals. *American Psychologist*, *19*, 202-203.

Edgington, E.S. (1974).A new tabulation of statistical procedures used in APA journals. *American Psychologist*, *29*, 25-26.

Edwards, A.W.F. (1972). *Likelihood*. Cambridge, England: Cambridge University Press.

Edwards, A.W.F. (1974). A History of Likelihood. *International Statistical Review*; *42*, 9-15

Edwards, A.W.F. (1976). Fiducial probability. *The Statistician*, *25*, 15-35.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, *63*, 400-402.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193-242.

Edwards, W. (1995). Number magic, auditing acid and materiality: a challenge for auditing research. *Auditing*, *14*, 176-187. Efron B. (1978). Controversies in the foundations of statistics. *The American Mathematical Monthly*, *85*, 231-246. Efron, B. (1996). Empirical Bayes methods for combining likelihoods. *Journal of the American Statistical Association*, *91*, 538-550.

Efron, B. (1996). Why isn't everyone a Bayesian? [With discussion]. *The American Statistician*, *40*, 1-11.

Efron, B. (1998). R.A. Fisher in the 21st century [With discussion]. *Statistical Science*, *13*, 95-122.

Eizner Favreau, O. (1997). Sex and gender comparisons: Does null hypothesis testing create a false dichotomy? *Feminism and Psychology*, *7*, 63.

Elifson, K.W., Runyon, R.P., & Haber, A. (1990). *Fundamentals of Social Statistics*. New York: McGraw-Hill.

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

Ellis, N. (2000). Editorial. *Language Learning*, *50*.

Elmore, P.B., & Woehlke, P.L. (1988). Statistical methods employed in *American Educational Research Journal, Educational Researcher, and Review of Educational Research* from 1978 to 1987. *Educational Researcher*, *17*, 19-20.

Ely, M. (1999). The importance of estimates and confidence intervals rather than p values. *Sociology*, *33*, 185-190.

Erhardt, C. (1959). Statistics, a trap for the unwary. *Obstetrics and Gynecology*, *14*, 549-554.

Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science*, *8*, 18-20.

Etzioni, R.D., Kadane, J.B. (1995). Bayesian statistical methods in public health and medicine. *Annual Review of Public Health*, *16*, 23-41.

Evans, S.J.W, Mills, P., & Dawson, J. (1988). The end of the p-value? *British Heart Journal*, *60*, 177-180.

Eysenck, H.J. (1960). The concept of statistical significance and the controversy about one-tailed effects. *Psychological Review*, *67*,

269-271.

Falissard, B., & Landais, P. (1995). Les statistiques en médecine: et s'il était temps de prendre un peu de distance? *Médecine Thérapeutique*, *1*, 775-781.

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

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist*, *53*, 798-799.

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

Fan, X. (2001). Statistical significance and effect size in educational research: Two sides of a coin. <>*Journal of Educational Research*, *94*, 275-282.

Fan X., Thompson B. (2001). Confidence Intervals About Score Reliability Coefficients, Please: An EPM Guidelines Editorial. *Educational and Psychological Measurement*, *61*, 517-531.

Favreau, O.E. (1993). Do the Ns justify the means? Null hypothesis testing applied to sex and other differences. *Canadian Psychology*, *34*, 64-78.

Fayers, P.M., Ashby, D., & Parmar, M.K. (1997). Tutorial in biostatistics Bayesian data monitoring in clinical trials. *Statistics in Medicine*, *16*, 1413-1430.

Feinstein, A. R. (1977). *Clinical Biostatistics*. St. Louis, MO: C. V. Mosby.

Fan X., Thompson B. (2001)Feinstein, A.R. (1978). Clinical biostatistics: stochastic significance, apposite data, and some remedies for the intellectual polluants of statistical vocabulary. *Clinical Pharmaceutical Therapy*, *22*, 113-123.

Feinstein, A. R. (1985). *Clinical Epidemiology: The Architecture of Clinical Research*. Philadelphia, PE: W. B. Saunders Co.

Felson, D.T., Anderson, J.J, & Meenan, R.F. (1990). Time for changes in the design, analysis, and reporting of rheumatoid arthritis clinical trials. *Arthritis and Rheumatism*, *33*, 140-149.

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

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C., & Schmitt, R. (2005). Evaluating the effectiveness of editorial policy to improve statistical practice: The case of the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology*, in press.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119-126.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed and random-effects effect sizes. *Educational and Psychological Measurement*, *61*, 575-604.

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

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments & Computers*, *36*, 312-324.

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future APA guidelines for statistical practice. *Theory & Psychology*, *12*,

825-853.

Finney, D. J. (1988). Was this in your statistics textbook? III. Design and analysis. *Experimental Agriculture*, *24*, 421-432.

Finney, D. J. (1989). Was this in your statistics textbook? VI. Regression and covariance. *Experimental Agriculture*, *25*, 291-311.

Finney, D. J. (1989). Is the statistician still necessary? *Biom. Praxim.*, *29*, 135-146.

Fisher, L.D. (1996). Comments on Bayesian and frequentist analysis and interpretation of clinical trials. *Controlled Clinical Trials*. 1996, 17, 423-434.

Fisher, R.A. (1990/1925). *Statistical Methods for Research Workers*. London: Oliver and Boyd. (Reprint, 14th edition, *in* Fisher, 1990).

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, Series A*, *222*, 309-368.

Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophic Society*, *22*, 700-725.

Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *33*, 503-513.

Fisher, R.A. (1929). The statistical method in psychical research. *Proceedings of the Society for Psychical Research*, *39*, 189-192.

Fisher, R.A. (1990/1935). *The Design of Experiments*. London: Oliver and Boyd. (Reprint, 8th edition, *in* Fisher, 1990).

Fisher, R.A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*, 39-54.

Fisher, R.A. (1935). Statistical tests. *Nature*, *136*, 474.

Fisher, R. A. (1943). Note on Dr Berkson's criticisms of tests of significance. *Journal of the American Statistical Association*, *38*, 103-104

Fisher, R. A. (1948). Conclusions fiduciaires. *Annales de l'Institut Henri Poincaré*, *10*, 191-213.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B*, *17*, 69-78.

Fisher, R. A. (1990/1956). *Statistical Methods and Scientific Inference*. London: Oliver and Boyd. (Reprint, 3rd edition, *in* Fisher, 1990).

Fisher, R. A. (1959). Mathematical probability in the natural sciences. *Technometrics*, *1*, 21-29.

Fisher, R.A. (1962). Some examples of Bayes's method of the experimental determination of probabilities *a priori*. *Journal of the Royal Statistical Society, Series B*, *24*, 118-124.

Fisher, R. A. (1990). *Statistical Methods, Experimental Design and Scientific Inference* [Re-issue edited by J.H. Bennet with a foreword by F. Yates]. Oxford: Oxford University Press.

Fisher, R.A., & MacKenzie, W.A. (1923). Studies in crop variation: 2. The manurial response of different potato varieties. *Journal of Agricultural Science*, *13*, 311-320.

Fiske, D.W., & Fogg, L. (1990). But the reviewers are making different criticism of my paper! Diversity and uniqueness in reviewer comments. *American Psychologist*, *45*, 591-598.

Fleishman, A. E. (1980). Confidence interval for correlation ratios. *Educational and Psychological Measurement*, *40*, 659-670.

Fleiss, J.L. (1969). Estimating the magnitude of experimental effects. *Psychological Bulletin*, *72*, 273-276.

Fleiss, J.L. (1986). Significance tests have a role in epidemiologic research: reactions to A. M. Walker. *American Journal of Public Health*, *76*, 559-560.

Fleiss, J.L. (1986). Confidence intervals vs. significance tests: Quantitative interpretation (Letter). *American Journal of Public Health*, *76*, 587.

Fleiss, J. (1986). Dr. Fleiss responds (Letter). *American Journal of Public Health*, *76*, 1033-1044.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin, 106*, 155-160.

Ford, J. (1975). *Paradigms and Fairly Tables* (two volumes), Chapter 19. London: Routledge & KeganPaul.

Forge, R.L. (1967). Confidence intervals or tests of significance in scientific research. *Psychological Bulletin, 68*, 446-447.

Fowler, R.L. (1984). Approximating probability levels for testing null hypotheses with noncentral *F* distributions. *Educational and Psychological Measurement, 44*, 275-281.

Fowler, R.L. (1985). Testing for substantive significance in applied research by specifying nonzero effect nullhypotheses. *Journal of Applied Statistics, 70*, 215-218.

Fraser, D.A.S. (1996). Some remarks on pivotal models and the fiducial argument in relation to structural models. *International Statistical Review, 64*, 231-236.

Frederick, B.N. (1999). Fixed-, random-, and mixed-effects ANOVA models: A user-friendly guide for increasing the generalizability of ANOVA results. *In* B. Thompson (Ed.). (1999). *Advances in social science methodology*, vol. 5. Stamford, CT: JAI Press, 111-122.

Freedman, D. (1999). From association to causation: Some remarks on the history of statistics. *Statistical Science, 14*, 243-258.

Freedman, D., Pisani, R., &Purves, R. (1997). *Statistics* (3rd edition). New York: Norton.

Freedman L. (1996). Bayesian statistical methods [Editorial]. *British Medical Journal, 313*, 569-570.

Freedman, L.S., Spiegelhalter, D.J., & Parmar, M.K.B. (1994). The what, why and how of Bayesian clinical trials monitoring. *Statistics in Medicine, 13*, 1371-1383.

Freeman, P.R. (1993). The role of *p*-values in analysing trial results. *Statistics in Medicine, 12*, 1443-1452.

Freiman, J.A., Chalmers, T.C., Smith, H., & Kueber, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *New England Journal of Medicine, 299*, 690-694.

Frías, Ma.D., Pascual, J., & Garcia, J.F. (2000). Tamaño del efecto del tratamiento y significación estadística [Effect size and statistical significance]. *Psicothema, 12*, 236-240.

Frick, R.W. (1995). Accepting the null-hypothesis. *Memory and Cognition, 23*, 132-138.

Frick, R.W. (1995). A problem with confidence intervals [Comment]. *American Psychologist, 50*, 1102-1103.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379-390.

Frick, R. W. (1998). Interpreting statistical testing: Processes, not populations and random sampling. *Behavior Research Methods, Instruments, and Computers, 30*, 527-535.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers, 30*, 690-697.

Frick, R.W. (1999). Defending the statistical status quo. *Theory & Psychology, 9*, 183-189.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin, 70*, 245-251.

Friedman, M. (1988). Money and the stock market. *Journal of Political Economy, 96*, 221-239.

Friedman, S. B., & Phillips, S. (1981). What's the difference? Pediatric residents and their inaccurate concepts regarding statistics. *Pediatrics, 68*, 644-646.

Fry, T.C. (1965). *Probability and its Engineering Uses* (2nd edition). Princeton, NJ: D. Van Nostrand.

G

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Gaito, J., & Firth, J. (1973). Procedures for estimating the magnitude of the effects. *Journal of Psychology*, *83*, 151-161.

Galtung, J. (1967). On the use of statistical tests. *In Theory and Methods of Social Research*. New York: Columbia University Press, 358-388.

Garbe, E., Rohmel, J., & Gundert-Remy, U. (1993). Clinical and statistical issues in therapeutic equivalence trials. *European Journal of Clinical Pharmacology*, *45*, 1-7.

Gardner, M.J., & Altman, D.G. (1986). Confidence intervals rather than *P* values: estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746-750.

Gardner, M.J., & Altman, D.G. (1989). Estimation rather than hypothesis testing: confidence intervals rather than *P* values. *In* M.J. Gardner & D.G. Altman (Eds.), *Statistics with Confidence - Confidence Intervals and Statistical Guidelines*. London: British Medical Association, 6-19.

Gardner, M.J., & Altman, D.G. (Eds.) (1989). *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. London: British Medical Association.

Gauch Jr., H. G. (1988). Model selection and validation for yield trials with interaction. *Biometrics*, *44*, 705-715.

Gavarret, J. (1840). *Principes Généraux de Statistique Médicale, ou Développement des Règles Qui Doivent Présider à Son Emploi*. Paris: Bechet jeune et Labe.

Geerstma, J.C. (1983). Recent views on the foundational controversy in statistics. *South African Statistical Journal*, *17*, 121-146.

Geary, R. C. (1947). Testing for normality. *Biometrika*, *34*, 209-242.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Gendreau, P. (2002). We must do a better job of cumulating knowledge. *Canadian Psychology*, *43*, 205-210.

Gerard, P. D., Smith, D. R., & Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management*, *62*, 801-807.

Gibbons, J.D., & Pratt, J.W. (1975). *P*-values: Interpretation and methodology. *The American Statistician*, *29*, 20-25.

Giere, R.N. (1972). The significance test controversy. *The British Journal for the Philosophy of Science*, *23*, 170-181.

Gigerenzer, G. (1991). From tools to theories: a heuristic of discovery in cognitive psychology. *Psychological Review*, *98*, 254-267.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "Heuristics and Biases". *In* N. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology*, *2*, 83-115.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. *In* G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Hillsdale, NJ: Erlbaum, 311-339.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199-200.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as Intuitive Statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J.,& Krüger, L. (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life?* Cambridge, England: Cambridge University Press.

Gill, M. (1993). The significance of "significance". *Edinburgh Working Papers in Applied Linguistics*, *4*, 63-80.

Glaser, D.N. (1976). The controversy of significance testing. *American Journal of Critical Care*, *8i(5)*.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Metaanalysis in Social Research*. Beverly Hills, CA: Sage.

Glenberg, A. M. (1988). *Learning from data: An introduction to statistical reasoning*. San Diego: Harcourt Brace Jovanovich.

Gliner, J.A., Leech, N.L., & Morgan, G.A. (2002). Problems With Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say? *The Journal of Experimental Education, 71*, 83-92.

Glymour, C. (1981). Why I am not a Bayesian. *In* D. Papineau (Ed.), *Philosophy of Science*. Oxford: Oxford University Press, 290-313.

Godambe, V.B., & Sprott, D.A. (Eds.) (1971). *Foundations of Statistical Inference*. Toronto: Holt, Rinehart & Winston of Canada.

Gold, D. (1958). Comment on "A critique of tests of significance". *American Sociological Review, 23*, 85-86 [Reprinted in Morrison & Henkel, 1970, 107-108].

Gold, D. (1964). Some problems in generalizing aggregate associations. *The American Behavioral Scientist, 8*, 16-18 [Reprinted in Morrison & Henkel, 1970, 172-181].

Gold, D. (1969). Statistical tests and substantive significance. *The American Sociologist, 4*, 42-46 [Reprinted in Morrison & Henkel, 1970, 172-181].

Goldberger, A.S. (1991). *A course in Econometrics*. Cambridge, MA: Harvard University Press.

Goldstein, H., & Healy, M.J.R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A, 158*, 175-177.

Good, I.J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association, 53*, 799-813.

Good, I.J. (1973). *In* V.P. Godambe & D.A. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart & Winston.

Good, I.J. (1981). Some logic and history of hypothesis testing. *In* J.C. Pitt (Ed.), *Philosophy in Economics*, Dordrecht, Netherlands: D. Reidel, 149-174.

Good, I.J. (1983). *Good Thinking. The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.

Good, I.J. (1984). An error by Neyman noticed by Dickey (C209). *Journal of Statistical Computation and Simulation, 20*, 159-160.

Goodman, S.N. (1992). A comment on replication, P-values and evidence. *Statistics in Medicine, 11*, 875-869.

Goodman, S.N. (1993). *P* values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology, 137*, 485-496.

Goodman, S.N. (1993). Author's response to "Invited commentary: *p* values, hypothesis tests, and likelihood". *American Journal of Epidemiology, 137*, 500-501.

Goodman, S.N. (1989). Meta-analysis and evidence. *Controlled Clinical Trials, 10*, 188-204, 435.

Goodman, S.N. (1998). Multiple comparisons, explained. *American Journal of Epidemiology*, 147, 807-812

Goodman, S.N., & Berlin, J.A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121*, 200-206.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine, 130*, 995-1004.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 1: The Bayes factor. *Annals of Internal Medicine, 130*, 1005-1013.

Goodman, S.N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health, 78*, 1568-1574.

Goodwin, L.D., & Goodwin, W.L. (1985). Statistical techniques in *AERJ* articles, 1979-1983: The preparation of graduate students to read the educational research literature. *Educational Researcher*, *14*, 5-11.

Gordon, H.R.D. (2001). American Vocational Education Research Association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Education Research*, *26*(2)

Gore, S.M. (1981). Assessing clinical trials - trial size. *British Medical Journal*, *282*, 1687-1689.

Gore, S.M. (1981). Statistics in question: Assessing methods - confidence intervals. *British Medical Journal*, *283*, 660-662.

Gower, J. C. (1983). Data analysis: multivariate or univariate and other difficulties. *In* H. Martens and H. Russwarm, Jr. (Eds.), *Food Research and Data Analysis*, London, U.K.: Applied Science, 39-67.

Graham, J.M. (2001). Review of Statistics with Confidence. *Educational and Psychological Measurement*, *61*, 668-674.

Granaas, M. (2002). Hypothesis testing in psychology: Throwing the baby out with the bathwater? Cape Town, South-Africa : ICOTS 6 [http://icots6.haifa.ac.il/PAPERS/3M1_GRAN.PDF]

Granger, C.W.J., King, M.L., & White, H. (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics*, *67*, 173-187.

Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *69*, 54-61.

Gray, M.W. (1983). Statistics and the law . *Mathematics Magazine*, *56*, 67-81.

Graybill, F.A. (1976). *Theory and Application of the Linear Model*. Duxbury Press, Mass. 704pp.

Graybill, F. A., & Iyer, H. K. (1994). *Regression Analysis: Concepts and Applications*. Belmont, CA: Duxbury Press.

Green, C.D. (2002). Comment on Chow's "Issues in Statistical Inference". *History and Philosophy of Psychology Bulletin*, *14*, 42-46.

Greenfield, M. L. V. H., Kuhn, J. E., & Wojtys, J. E. (1996). Current concepts. A statistics primer. *P* values: probability and clinical significance. *American Journal of Sports Medicine*, *24*, 863-865.

Greenhouse, J.B. (1992). On some applications of Bayesian methods in cancer clinical trials. *Statistics in Medicine*, *11*, 37-53.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*, *79*, 340-349.

Greenland, S., & Robins J.M. (1991). Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology*, *2*, 244-251.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20.

Greenwald, A. G. (1993). Consequences of prejudice against the null hypothesis. *In* G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Hillsdale, NJ: Erlbaum, 419-448.

Greenwald, A.G., Gonzalez, R., Harris, R.J., & Guthrie, D. (1996). Effect sizes and *p* values. What should be reported and what should be replicated? *Psychophysiology*, *33*, 175-183.

Gregson, R.A.M. (1998). Understanding Bayesian procedures. *Behavioral and Brain Sciences*, *21*, 201-202.

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

Grouin, J.-M., & Lecoutre, B. (1996). Probabilités prédictives: Un outil pour la planification des expériences, *Revue de Statistique appliquée*, *XLIV*, 21-35.

Grunkemeier, G.L. & Payne, N. (2002). Bayesian analysis: A new statistical paradigm for new technology. *The Annals of Thoracic Surgery*, *74*, 1901-1908.

Guilford, J.P. (1942). *Fundamentals of Statistics in Psychology and Education*. New York: Basic Books.

Guthery, F.S., Lusk, J.J., & Peterson, M.J. (2001). The fall of the null hypothesis: liabilities and opportunities. *Journal of Wildlife Management*, *63*, 379-384.

Guttman, L. (1977). What is not what in statistics? *The Statistician*, *26*, 81-107.

Guttman, L. (1979). Cyril Burt and the careless star worshippers. *L'Echo des Messaches*, *9*, 7-8.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis*, *1*, 3-10.

H

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Haase, R.F. (1974). Power analysis of research in counselor education. *Counselor Education and Supervision*, *14*, 124-132.

Haase, R.F., Waechter, D.M., & Solomon, G.S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology*, *29*, 58-65.

Hacking, I. (1965). *The Logic of Statistical Inference*. Cambridge, England: Cambridge University Press.

Hacking I. (1975). *The Emergence of Probability. A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge, England: Cambridge University Press.

Hagen, R.L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15-24.

Hager, W. (1996). On testing a priori hypotheses about quantitative and qualitative trends. *Methods of Psychological Research Online*, *1*(4).

Hager, W. (2000). About some misconceptions and the discontent with statistical tests in Psychology. *Methods of Psychological Research Online*, *5*(1).

Hager, W., & Westermann, R. (1983). Zur wahl und prüfung psychologischer hypothesen in psychologischen untersuchungen. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *30*, 67-94.

Hagood, M.J. (1941). The notion of a hypothetical universe. *In* M. J. Hagood, *Statistics for Sociologists*. New York: Reynal & Hitchcock, 612-616. [Reprinted in Morrison & Henkel, 1970, 65-78].

Hahn, G.J. (1974). Don't let statistical significance fool you! *Chemtech*, *4*, 16-18.

Hahn, G.J. (1990). Commentary. *Technometrics*, *32*, 257-258.

Hahn, G.J., & Meeker, W.Q. (1991). *Statistical Intervals: A Guide for Practitioners*. New York: Wiley.

Hall, P., & Selinger, B. (1986). Statistical significance: balancing evidence against doubt. *Australian Journal of Statistics*, *28*, 354-370.

Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures and applications. *Behavioral Research and Therapy*, *34*, 489-499.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, *7*(1).

Hammond, G. (1996). The objections to null hypothesis testing as a means of analysing psychological data. *Australian Journal of Psychology*, *48*, 104-106.

Hancock, G.R., & Freeman M.J. (2001). Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educational and Psychological Measurement*, *61*, 741-758.

Hand, D.J., & Taylor, C. (1987). *Multivariate Analysis of Variances and Repeated Measures: A practical Approach for Behavioural Scientists*. London and New York: Chapman and Hall.

Hansen, M.H., & Edwards, W.E. (1950). On the important limitation to the use of data from samples. *Bulletin de l'Institut International de Statistique*, 214-219.

Harcum, E.R. (1990). Methodological versus empirical literature: Two views on casual acceptance of the null hypothesis. *American Psychologist, 45*, 404-405.

Hardy, A., Harvie, P., & Koestler, A. (1973). *The challenge of Chance.* New York: Random House.

Hardy, R.J., & Thompson, S.G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine, 15*, 619-629.

Harlow, L. L. (1997). Significance Testing Introduction and Overview. *In* L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What If There Were No Significance Tests?* Hillsdale, NJ: Erlbaum., 1-17.

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

Harris, E.K. (1993). On *p* values and confidence intervals (why can't we *p* with more confidence?). *Clinical Chemistry, 39*, 927-928.

Harris, M.J. (1991). Significance tests are not enough: The role of effect size estimation in theory corroboration [Comment on Chow, 1991a]. *Theory and Psychology, 1*, 337-360, 375-382.

Harris, M.J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin, 97*, 363-386.

Harris, R.J. (1994). *ANOVA: An Analysis of Variance Primer*. Itasca, IL: F.E. Peacock.

Harris, R.J. (1997). Significance tests have their place. *Psychological Science, 8*, 8-11.

Harris, R.J. (1997). Reforming significance testing via three-valued logic. *In* L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What If There Were No Significance Tests?* Hillsdale, NJ: Erlbaum, 145-174.

Harper, W.L., & Hooker, C.A. (Eds.) (1976). *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II Foundations and Philosophy of Statistical Inference*, Dordrecht, Netherlands: D. Reidel, 149-174.

Hauck, W.W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics, 12*, 83-91.

Hauschke, D., & Steinijans, V.W. (1996). A note on conventional null hypothesis testing in active control equivalence studies. *Controlled Clinical Trials, 17*, 347.

Hays, W.L. (1963). *Statistics for Psychologists*. New York: Holt, Rinehart & Winston.

Healy, M.J.R. (1978). Is statistics a science? *Journal of the Royal Statistical Society, Series A, 141*, 385-393.

Healy, M.J.R. (1989). Comments on the paper by McPherson. *Journal of the Royal Statistical Society, Series A, 152*, 232-234.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L.V. (1987). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*, 490-499.

Hedges, L.V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*, 443-455.

Hedges, L.V. & Olkin I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press.

Heldref Foundation (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.

Salsburg, D. (1985). The religion of statistics as practiced in medical journals. *The American Statistician, 39*, 220-223.

Salsburg, D. (1994). Intent to treat: The *reductio ad absurdum* that became gospel. *Pharmacoepidemiology and Drug Safety, 3*, 329-335.

Salsburg, D.S. (1986). *Statistics for Toxicologists*. New York: Marcel Dekker.

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: W.H. Freeman.

Samaniego, F.J., & Reneau, D.M. (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association, 89*, 947-957.

Samurçay, R., & Hoc, J.-M. (1996). Causal versus topographical support for diagnosis in a dynamic situation. *Le Travail Humain, 59*, 45-68.

Sánchez, J., Valera, A., Velandrino, A, & Marin, F. (1992). Un estudio de la potencia estadística en Anales de Psicología (1984-1991) [A study of statistical power in the journal Anales de Psicología]. *Anales de Psicología, 8*, 19-32.

Savage, L. (1954). *The Foundations of Statistical Inference*. New York: John Wiley & Sons.

Savage, L.J. (1957). Nonparametric Statistics. *Journal of the American Statistical Association, 52*, 332-333.

Savage, L.J. (1976). On rereading R.A. Fisher [With discussion]. *Annals of Statistics, 4*, 441-500.

Savitz, D.A. (1993). Is statistical significance testing useful in interpreting data? *Reproductive Toxicology, 7*, 95-100.

Savitz, D.A., & Olshan, A.F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology, 14*, 904-908.

Savitz, D.A., Tolo, K.-A., & Poole, C. (1994). Statistical significance testing in the *American Journal of Epidemiology*, 1970-1990. *American Journal of Epidemiology, 139*, 1047-1052.

Sayn-Wittgenstein, L. (1965). Statistics - salvation or slavery? *Forestry Chronicle, 41*, 103-105.

Sawyer, A.G., & Ball, A.D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research, 18*, 275-290.

Sawyer, A.G., & Peter, J.P. (1983). The significance of statistical significance tests in marketing research. *Journal of Marketing Research, 20*, 122-133.

Scarr, S. (1997). Rules of evidence: A larger context for the statistical debate. *Psychological Science, 8*, 16-17.

Schafer, W.D. (1993). Interpreting statistical significance and nonsignificance, *Journal of Experimental Education, 61*, 383-387.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Schenker, N. & Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55*, 182-186.

Schervish, M.J. (1992). Bayesian analysis of linear models. *In* J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith (Eds.), *Bayesian Statistics IV*, Oxford: Oxford University Press, 419-434.

Schervish, M.J. (1995). *Theory of Statistics*. New York: Springer Verlag.

Schervish, M.J. (1996). *P* values: what they are and what they are not. *The American Statistician, 50*, 203-206.

Scheutz, F., Andersen, B., & Wulff, H.R. (1988). What do dentists know about statistics? *Scandinavian Journal of Dental Research, 96*, 281-287.

Schield, M. (1998). Using Bayesian strength of belief to teach classical statistics. *In* L. Pereira-Mendoza, L. Seu, T. Wee & W.K. Wong (Eds.), *Statistical Education - Expanding the Network*, Proceedings of the Fifth International Conference on Teaching of Statistics, Vol. 1, Vooburg, Netherlands: ISI Permanent Office, 245-2251.

Schlaiffer, R. (1959). *Probability and Statistics for Business Decisions*. New York: McGraw-Hill.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173-1181.

Schmidt, F.L. (1996). Board of scientific affairs action on significance testing. *Industrial-Organizational Psychologist*, *33*, 110.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.

Schmidt, F.L., & Hunter, J.E. (1995). The impact of data-analysis methods on cumulative research knowledge: statistical significance testing, confidence intervals, and meta-analysis. *Evaluation and the Health Professions*, *18*, 408-427.

Schmidt, F.L., & Hunter, J.E. (1997)[284]. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. *In* L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What If There Were No Significance Tests?* Hillsdale, NJ: Erlbaum., 37-64.

Schmidt, K. (1995). Statistical tests and estimations [Background paper]. *Drug Information Journal*, *29*, 483-491.

Schmitt, S.A. (1969). *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics*. Reading, MA: Addison Wesley.

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. *Evaluation Review*, *8*, 573-582.

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657-680.

Schulman, J.L., Kupst, M.J., & Suran, B.G. (1976). The worship of "*p*": Significant yet meaningless research results. *Bulletin of the Menninger Clinic*, *40*, 134-143.

Schwartz, D. (1984). Statistique et vérité. *Journal de la Société Statistique de Paris*, *125*, 74-83.

Schweder, T. (1988). A significance version of the basic Neyman-Pearson theory for scientific hypothesis testing. *Scandinavian Journal of Statistics*, *15*, 225-242.

Schweder, T., & Hjort, N.L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics*, *29*, 309-332.

Schwertman, N.C. (1996). A connection between quadratic-type confidence limits and fiducial limits. *The American Statistician*, *50*, 242-243.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin*, *105*, 309-316.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online*, *1*(4), 41-63

Sedlmeier, P. (2002). Beyond uncritical significance testing: Contrasts and effect sizes.*Contemporary Psychology*, *47*, 430-432.

Seeman, J. (1973). On supervising student research. *American Psychologist*, *28*, 900-906.

Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference*. Boston: D. Reidel.

Selvin, H.C. (1957). A Critique of tests of significance in survey research. *American Sociological Review*, *22*, 519-527 [Reprinted *in* Morrison and Henkel, 1970, 94-106].

Selvin, H.C. (1958). Reply to Beshers. *American Sociological Review*, *23*, 199-200 [Reprinted *in* Morrison & Henkel, 1970, 113-115].

Selvin, H.C., & Stuart, A. (1966). Data dredging procedures in survey analysis. *The American Statistician*, *20*, 20-23.

Selvin, S., & White, M.C. (1993). Description and reporting of statistical methods. *American Journal of Infection Control*, *21*, 210-215.

Selwyn, W.J., Dempster, A.P., & Hall, N.R. (1981). A Bayesian approach to bioequivalence for the 2x2 changeover design. *Biometrics*, *37*, 11-21.

Selwyn, W.J., & Hall, N.R. (1984). On Bayesian methods for bioequivalence. *Biometrics*, *40*, 1103-1108.

Selwyn, W.J., Hall, N.R., & Dempster, A.P. (1985). Letter to the Editor. *Biometrics*, *41*, 561.

Serlin, R.C. (1987). Hypothesis testing, theory building, and the philosophy of science. *Journal of Counseling Psychology*, *34*, 365-371.

Serlin, R.C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education*, *61*, 350-360.

Serlin, R.C., & Lapsley, D.K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 77-83.

Serlin, R.C., & Lapsley, D.K. (1993). Rational appraisal of psychological research and the good-enough principle. *In* G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences, Vol. 1: Methodological Issues*, Hillsdale: NJ: Erlbaum, 199-228.

Shafer, G. (1986). Savage revisited. *Statistical Science*, *1*, 463-501.

Share, D.L. (1984). Interpreting the outcome of multivariate analysis: A discussion of current approaches. *British Journal of Psychology*, *75*, *349-362.*

Shaver, J.P. (1985). Chance and nonsense: A conversation about interpreting tests of statistical significance, Part 1. *Phi Delta Kappa*, *67*, 57-60.

Shaver, J.P. (1985). Chance and nonsense: a conversation about interpreting tests of statistical significance, Part 2. *Phi Delta Kappa*, *67*, 138-141. *Erratum*, 1986, *67*, 624.

Shaver, J. (1992). What significance testing is, and what it isn't. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA., April 1992.

Shaver, J.P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education*, *61*, 293-316.

Shea, C. (1996). Psychologists debate accuracy of "significance tests". *Chronicle of Higher Education*, *42*, 12-16.

Shrout, P.E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, *8*, 1-2.

Shulman, L.S. (1970). Reconstruction of educational research. *Review of Educational Research*, *40*, 371-393.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Signorelli, A. (1974). Statistics: tool or master of the psychologist? *American Psychologist*, *11*, 221-223.

Sim, J., & Reid, N. (1999). Statistical inference by confidence intervals: Issues of interpretation and utilization. *Physical Therapy*, *79*, 186-195.

Simberloff, D. (1990). Hypotheses, errors, and statistical assumptions. *Herpetologica*, *46*, 351-357.

Simon, R. (1986). Confidence intervals for reporting results of clinical trials. *Annals of Internal Medicine*, *105*, 429-435.

Simon, R., & Altman, D.G. (1994). Statistical aspects of prognostic factor studies in oncology [Editorial]. *British Journal of Cancerology*, *69*, 979-985.

Simon, R., & Wittes, R.E. (1985). Methodologic guidelines for reports of clinical trials. *Cancer Treatment Reports*, *69*, 1-3.

Skinner, B.F. (1956). A case history in scientific method. *American Psychologist*, *11*, 221-223.

Skipper, Jr, J.K., Guenther, A.L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist*, *2*, 16-18 [Reprinted *in* Morrison & Henkel, 1970, 1155-160].

Slakter, M.J., Wu, Y., & Suzuki-Slakter, N.S. (1991). *, **, ***; statistical nonsense at the .00000 level. *Nursing Research*, *40*, 248-249.

Smeeton, N.C. (Ed.) (1994). Conference on practical Bayesian statistics [Special issue]. *The Statistician*.

Smith, A. (1995). A conversation with Dennis Lindley. *Statistical Science*, *10*, 305-319.

Smith, C.A.B. (1960). Book review of Norman T. J. Bailey: Statistical Methods in Biology. *Applied Statistics*, *9*, 64-66.

Smith, K. (1983). Tests of significance: some frequent misunderstandings. *American Journal of Orthopsychiatry*, *53*, 315-321.

Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, *25*, 970-975.

Smithson, M. (2000). *Statistics with Confidence*. London: Sage.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605-632.

Smithson, M. (2002) . *Confidence intervals*. London: Sage Publications, Inc.

Snyder, P. (2000). Reporting results of group quantitative investigations. *Journal of Early Intervention*, *23*, 145-150.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, *61*, 334-349.

Snyder, P.A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly*, *13*, 335-348.

Sohn, D. (1993). Psychology of the scientist: LXVI. The idiots savants have taken over the psychology labs! Or why in science using the rejection of the null hypothesis as the basis for affirming the research hypothesis is unwarranted. *Psychological Reports*, *73*, 1167-1175.

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

Soric, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, *84*, 608-610.

Spiegelhalter,D.J., Freedman, L.S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, *5*, 1-13.

Spiegelhalter, D.J., Freedman, L.S. (1988). Bayesian approaches to clinical trials [With discussion]. *In* J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith (Eds.), *Bayesian Statistics 3*, Oxford: Oxford University Press, 453-477.

Spiegelhalter, D.J., Freedman, L.S., & Blackburn, P.R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, *7*, 8-17.

Spiegelhalter, D.J., Freedman, L.S., & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials [With discussion]. *Journal of the Royal Statistical Society, Series A*, *157*, 357-416.

Spiegelhalter D.J., Myles J.P., Jones D.R. & Abrams K.R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, 2000, *4*, 1-142.

Spiegelhalter, D., Thomas, A., Best, N., & Gilks, W. (1998). BUGS Bayesian Inference Using Gibbs Sampling. Cambridge, UK MRC Biostatistics Unit.

Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science*, *41*, 211-226.

Spielman, S. (1978). Statistical dogma and the logic of significance testing. *Philosophy of Science*, *45*, 120-135.

Spriet, A., & Bieler, D. (1979). When can "non significantly different" treatments be considered as equivalent. *British Journal of Clinical Pharmacology*, *7*, 623-624.

Standards of reporting trials group (1994). A proposal for structured reporting of randomized controlled trials. *Journal of the*

American Medical Association, *272*, 1926-1931.

Stangl, D. (1998). Classical and Bayesian paradigms: Can we teach both. In L. Pereira-Mendoza, L. Seu, T. Wee & W.K. Wong (Eds.), Statistical Education - Expanding the Network, Proceedings of the Fifth International Conference on Teaching of Statistics, Vooburg, Netherlands: ISI Permanent Office, Vol. 1, 251-258.

Statistics in Medicine (1993). Papers from the Conference on Methodological and Ethical Issues in Clinical Trials, special issue on Bayesian inference (D. Ashby, Ed.), *12*.

Steger, J.A. (Ed.) (1971). *Readings in Statistics for the Behavioral Scientists*. New York: Holt, Rinehart & Winston.

Steidl, R.J., Hayes, J.P., & Schauber, E. (1997). Statistical power analysis in wildlife research. *Journal of Wildlife Management*, *61*, 270-279.

Steiger, J.H., & Fouladi, R.T. (1992). *R2*: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers*, *4*, 581-582.

Steiger, J.H., & Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models. *In* L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What If There Were No Significance Tests?* Hillsdale, NJ: Erlbaum., 221-257.

Steinfatt, T.M. (1990). Ritual versus logic in significance testing in communication research. *Communication Research Reports*, *7*, 90.

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, *54*, 30-34.

Sterling, T.D. (1960). What is so peculiar about accepting the null hypothesis? *Psychological Reports*, *7*, 363-364.

Sterling, T.D., Rosenbaum, W.L., & Weinkam, J.J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108-112.

Sterne, J.A.C., & Davey Smith, G. (2001). Sifting the evidence—what's wrong with significance tests? *British Medical Journal*, *322*, 226–231.

Sterne, J.A.C. (2002). Teaching hypothesis tests - time for significant change? *Statistics in Medicine*, *21*, 985-994.

Stevens, S.S. (1968). Measurement, statistics, and the schemapiric view. *Science*, *161*, 849-856.

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard University Press.

Stone, M. (1969). The role of significance testing: Some data with a message. *Biometrika*, 56, 495-493.

Street, D.J. (1990). Fisher's contributions to agricultural statistics. *Biometrics*, *46*, 937-945.

Student (1908). The probable error of a mean. *Biometrika*, *6*, 1-25.

Suen, H. K. (1992). Significance testing: Necessary but insufficient. *Topics in Early Childhood Special Education*, *12*, 66-81.

Sullivan J.R. (2000). A review of post-1994 literature on whether statistical significance tests should be banned. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX, January 29, 2000 (Texas A&M University 77843-4225).

Summers, L.H. (1991). The scientific illusion in empirical macroeconomics. *Scandinavian Journal of Economics*, *93*, 129-148.

Suter, G.W. (1996). Abuse of hypothesis testing statistics in ecological risk assessment. *Human and Ecological Risk Assessment*, *2*, 331-347.

Sutlive, V.H., & Ulrich, D.A. (1998). Interpreting statistical significance and meaningfulness in adapted physical activity research. *Adapted Physical Activity Quarterly*, *15*, 103–118.

Sverdrup, E. (1975). Tests without power. *Scandinavian Journal of Statistics*, *2*, 158-160.

Svyantek, D. J., & Ekeberg, S. E. (1995). The earth is round (So we can probably get there from here). *American Psychologist*, *50*,

1101.

Sylvester, R.J. (1988). A Bayesian approach to the design of phase II clinical trials. *Biometrics*, *44*, 823-836.

T

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Tamhane, A. C. (1996). Review of R. E. Bechhofer, T. J. Santner and D. M. Goldsman, Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons, New York: Wiley, 1995. *Technometrics*, *38*, 289-290.

Tannock, I.F. (1996). False-positive results in clinical trials. multiple significance tests and the problem of unreported comparisons. *Journal of the National Cancer Institute*, 1996, *88*; 206-207.

Tatsuoka, M. (1993). Effect size. *In* G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Hillsdale, NJ: Erlbaum, 461-479.

Taube, A. (1980). Significance, importance and equality: Three basic concepts in the analysis of a difference. *Upsala Journal of Medical Science*, *85*, 97-102.

Taylor, D.J., & Muller, K.E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *The American Statistician*, *49*, 43-47.

The Statistician (1993). Vol. 42, special issue: *Conference on Practical Bayesian Statistics* (1992).

Thomas, D.C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M., & Armstrong, B.G. (1985). The problem of multiple inference in studies designed to generate hypotheses. *American Journal of Epidemiology*, *122*, 1080-1095.

Thomas, L., & Juanes, F. (1996). The importance of statistical power analysis: An example from Animal Behaviour. *Animal Behaviour*, *52*, 856-859.

Thompson, B. (1988). A note about significance testing. *Measurement and Evaluation in Counseling and Development*, *20*, 146-148.

Thompson, B. (1989). Asking "what if" questions about significance tests. *Measurement and Evaluation in Counseling and Development*, *22*, 66-68.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361-377.

Thompson, B. (Guest Ed.). (1993). Statistical significance testing in contemporary practice [Special issue]. *The Journal of Experimental Education*, *61*(4).

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Thompson, B. (1994). The concept of statistical significance testing. *Measurement Update*, *4*, 5-6 [ERIC Document Reproduction Service No. ED 366 654].

Thompson, B. (1994). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: The neo-classical perspective. *In* B. Thompson (Ed.), *Advances in Social Science Methodology, Vol. 3*, Greenwich, CT: JAI Press, 3-27.

Thompson, B. (1995). Publishing your research results: Some suggestions and counsel. *Journal of Counseling and Development*, *73*, 342-345.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, *55*, 525–534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher*, *26*, 29-32.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist*, *53*, 799-800.

Thompson, B. (1998). Review of What if there were no significance tests? by L. Harlow, S. Mulaik & J. Steiger (Eds.). *Educational and Psychological Measurement*, *58*, 332-344.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools*, *5*, 39-42.

Thompson, B. (1998). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas. Invited address presented at the annual meeting of the American Educational Research Association, San Diego [ERIC Document Reproduction Service No. ED 419 023].

Thompson, B. (1999). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children*, *65*, 329-338.

Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. *Educational Psychology Review*, *11*, 157-169.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*, *9*, 165-181.

Thompson, B. (1999). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology*, *9*, 191-196.

Thompson, B. (1999). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *Journal of Psychology*, *133*, 133-140.

Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*, 11, 157-169.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80-93.

Thompson, B. (2001). Editor's Note on the "Colloquium on Effect Sizes: The Roles of Editors, Textbook Authors, and the Publication Manual. *Educational and Psychological Measurement*, *61*, 211-332.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, *80*, 64-71.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 24-31.

Thompson, B., & Kieffer, K.M. (2000). Interpreting statistical significance test results: A proposed new "What if" method. *Research in the Schools*, *7*, 3-10.

Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in The Journal of Experimental Education. *Journal of Experimental Education*, *66*, 75-83.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development*, *76*, 436-441.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174–195.

Thompson, W.D. (1987). Statistical criteria in the interpretation of epidemiologic data. *American Journal of Public Health*, *77*; 191-194.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, *110*, *526-535*.

Tryon, W.W. (1998). The inscrutable null hypothesis. *American Psychologist*, *53*, 796.

Tryon, W.W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological-methods*, *6*, 371-386.

Tullock, G. (1959). Publication decisions and tests of significance: A comment. *Journal of the American Statistical Association*, *54*, 593.

Tukey, J.W. (1960). Conclusions *vs* decisions. *Technometrics*, *2*, 1-11 [Reprinted *in* L.V. Jones (Ed.), 1986, *The Collected Works of John W. Tukey, Volume III, Philosophy and Principles of Data Analysis: 1949-1964*, Monterey, CA: Wadsworth & Brooks/Cole, 127-142].

Tukey, J.W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, *33*, 1-67.

Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83-91.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison Wesley.

Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100-116.

Tukey, J.W. (1993). Where should multiple comparisons go next? *In* F.M. Hoppe (Ed.), *Multiple Comparisons, Selection, and Applications in Biometry*, New York: Marcel Dekker, Inc., 187-208.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 237-251.

Tyler, R. (1931). What is statistical significance? *Educational Research Bulletin*, *10*, 118-142.

U

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Upton, G.J.G. (1992). Fisher's exact test. *Journal of the Royal Statistical Society, Series A*, *155*, 395-402.

Utts, J. (1988). Successful replication versus statistical significance. *Journal of Parapsychology, 52,* 305-320.

V

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20.

Vacha-Haase, T., & Ness, C.M. (1999). Statistical significance testing as it relates to practice: Use within Professional Psychology: Research and practices. *Professional Psychology: Research and practices*, *30*, 104-105.

Vacha-Haase, T., & Nilsson, J.E. (1998). Statistical significance reporting: Current trends and usages within MECD. *Measurement and Evaluation in Counseling and Development*, *31*, 46-57.

Vacha-Haase, T., Nilsson, J.E., Reetz, D.R., Lance, T.S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology*, *10*, 413-425.

Valera, A., Sánchez, J., & Marin, F. (1997). Pruebas de significación y magnitud del efecto: Reflexiones y propuestas [Significance tests and effect magnitude: Reflections and proposals]. *Anales de Psicología*, *13*, 85-90.

Valera, A., Sánchez, J., & Marin, F. (2000). Hypothesis testing and Spanish psychological research: Analyses and proposals [in Spanish]. *Psicothema*, *12*, 549-552.

Valera, A., Sánchez, J., Marin, F., & Velandrino, A. (1998). Potencia estadística de la Revista de Psicología General y Applicada (1990-1992) [Statistical power in the journal Revista de Psicología General y Applicada]. *Revista de Psicología General y Applicada*, *51*, 233-246.

Vallecillos, A. (1995). Comprension de la logica del contraste de hipotesis en estudiantes universitarios. *Recherches en Didactique des Mathématiques*, *15*, 53-81.

Vallecillos, A. (1996). Students' conceptions of the logic of hypothesis testing. *Hiroshima Journal of Mathematics Education*, *4*, 43-61.

Vallecillos, A. (1998). Research and teaching of statistical inference. *Proceeding of the International Conference on the Teaching of Mathematics*, Boston: John Wiley & Sons, Inc., 296-298.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. *Bulletin of the International Statistical Institute*, ISI 99, Proceedings, Tome LVIII, Book 2, 201-204.

VanVoorhis, W.C., & Morgan, B.L. (2001). Statistical rules of thumb: What we don't want to forget about sample sizes. *Psi Chi Journal*, *6* (4).

Vardeman, S.B. (1987). Comment. *Journal of the American Statistical Association*, *82*, 130-131.

Vargha, A., & Delaney, H.D. (2000). A critique and improvement of the *CL* common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*, 101-132.

Vaughan, G.M., & Corballis, M.C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, *72*, 204-213.

Vaughn, G.M., & Corballis, M.C. (1969). Beyond tests of significance: Estimating strengths of effects in selected ANOVA designs. *Psychological Bulletin*, *72*, 204-223.

Venables, W. (1975). Calculation of confidence intervals for noncentrality parameters. *Journal of the Royal Statistical Society, Series B*, *37*, 406-412.

Venn (1888). Cambridge anthropometry. *Journal of the Anthropological Institute* , *18*, 140-154.

Victor, N. (1987). On clinically relevant differences and shifted nullhypotheses. *Methods of Information in Medicine*, *26*, 109-116.

Vokey, J.R. (1998). Statistics without probability: Significance testing as typicality and exchangeability in data analysis. *Behavioral and Brain Sciences*, *21*, 225-226.

Vollset, S.E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, *12*, 809-824.

Wade, O.L., & Waterhouse, J.A.H. (1977). Significant or important? [Editorial]. *British Journal of Clinical Pharmacology*, *4*, 411-412.

Wade, P.R. (2000). Bayesian methods in conservation biology. *Conservation Biology*, *4*, 411-412.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *6*, 212-213.

Wainer, H., & Robinson, D.H. (2003). Shaping up the practice of Null Hypothesis Significance Testing. *Educational Researcher*, 32, 22-42.

Wald, W. (1947). *Sequential analysis*. New York: Dover.

Walker, A.M. (1986). Reporting the results of epidemiologic studies. *American Journal of Public Health*, *76*, 556-558.

Walker, H.M. (1929). *Studies in the History of Statistical Method*. Baltimore: Williams & Wilkins Company.

Walley, P. (1991 – *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Walley, P. (1996). Inferences from multinomial data: Learning about a bag of marbles [with discussion]. *Journal of the Royal Statistical Society B*, *58*, 3-57.

Wallis, W.A., & Roberts, H.V. (1956). *Statistics: A New Approach*. New York: MacMillan Publishing Company.

Walster, G.W., & Cleary, T.A. Statistical significance as a decision rule. *In* E.F. Borgatta & G.W. Bohrnstedt (Eds.), *Sociological Methodology*, San Francisco, CA: Jossey-Bass, 246-254.

Wampold, N.E., Davis, B., & Good, R.H. (1990). Hypothesis validity of clinical research. *Journal of Consulting and Clinical Psychology*, *58*, 360-367.

Wang, C. (1993). *Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety*. New York: Marcel Dekker.

Wang, Y.H. (2000). Fiducial intervals: What are they?. *The American Statistician*, *54*, 105-111.

Ward, R. C., Loftis, J. C., & McBride, G. B. (1990). *Design of Water Quality Monitoring Systems*. New York: Van Nostrand Reinhold.

Warren, W.G. (1986). On the presentation of statistical analysis: reason or ritual. *Canadian Journal of Forest Research*, *16*, 1185-1191.

Walster, G., & Cleary, T. (1970). Statistical significance as a decision rule. *In* E. Borgatta & G. Bohrnstedt (Eds.), *Sociological Methodology*, San Francisco: Jossey-Bass, 246-254.

Watson, J.M., & Moritz, J.B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*, 145-168.

Weinbach, R.W. (1989). When is statistical significance meaningful? A practice perspective. *Journal of Sociology and Social Welfare*, *16*, 31-37.

Weitzman, R.A. (1984). Seven treacherous pitfalls of statistics illustrated. *Psychological Reports*, *54*, 355-363.

Wellek, S., & Michaelis, J. (1991). Elements of significance testing with equivalence problems. *Methods of Information in Medicine*, *30*, 194-198.

Welsh, A.H. (1996). *Aspects of Statistical Inference*. New York: Wiley.

Wendell, J. P. (1991). More on Jahn's statistics. *Skeptical Inquirer*, *16*, 89-90.

Wendell, J. P. (1992). Jahn's statistics again. *Skeptical Inquirer*, *16*, 330.

West, L.J. (1990). Distinguishing between statistical and practical significance. *Delta Pi Epsilon Journal*, *32*, 1-4.

Westermann, R., & Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics*, *11*, 117-146.

Westfall, P.H., Johnson, W.O. & Utts, J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 84, 419-427.

Westlake, W.J. (1976). Symmetrical confidence intervals in analysis of comparative bioavailability trials. *Biometrics*, *32*, 741-744.

Westlake, W.J. (1981). Response to bioequivalence testing: A need to rethink (reader reaction response). *Biometrics*, *37*, 591-593.

White, A.L. (1980). Avoiding errors in educational research. *In* R. J. Shumway (Ed.), *Research in Mathematics Education*, Reston, Va: NCTM, 47-65.

Whitmore, G.A., & Xekalaki, E. (1990). *P*-values as measures of predictive validity. *Biometrical Journal*, *32*, 977.

Windeler, J., & Conradt, C. (2000). How can "significance" and "relevance" be combined? [in German]. *Medizinische Klinik*, *95*, 68-71.

Wiens, J. A. (1989). *The Ecology of Bird Communities*. Cambridge, England: Cambridge University Press.

Wietzman, R.A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports*, *54*, 355-363.

Wilcox, R.R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*, 300-314.

Wilcox, R.R., & Muska, J. (1999). Measuring effect size: A non-parametric analogue of *omega-square*. *British Journal of Mathematical and Statistical Psychology*, *52* , 93-110.

Wilkinson, L. and Task Force on Statistical Inference, *APA Board of Scientific Affairs* (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, *54*, 594-604.

Willer, D. (1967). *Scientific Sociology*. Englewood Cliffs, N.J.: Prentice-Hall.

Willer, D., & Willer, J. (1973). *Systematic Empiricism: Critique of a Pseudoscience*. Englewood Cliffs, N.J.: Prentice-Hall.

Williams, A.M. (1997). Students' understanding of hypothesis testing: the case of the significance concepts. *In* F. Biddulph & K. Karr (Eds.), *People in Mathematics Education, Proceedings of the MERGA 20*, Aotearoa, Australia, 585-59.

Williams, A.M. (1998). Students' understanding of the significance level concept. *In* L. Pereira-Mendoza, L. Seu, T. Wee & W.K. Wong (Eds.), *Statistical Education - Expanding the Network*, Proceedings of the Fifth International Conference on Teaching of Statistics, Vooburg, Netherlands: ISI Permanent Office, *Vol. 2*, 743-749.

Williams, V.S.L., Jones, L.V., & Tukey, J.W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, *24*, 42-69.

Willson, V.L. (1980). Research techniques in *AERJ* articles: 1969 to 1978. *Educational Researcher*, *9*, 5-10.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209-212.

Wilson, G. (2003). Tides of change: Is Bayesianism the new paradigm in statistics? *Journal of Statistical Planning and Inference*, *113*, 371-374.

Wilson, K.V. (1961). Subjective statistics for the current crisis. *Contemporary Psychology*, *6*, 229-231.

Wilson, W.R., & Miller, H.L. (1964). A note on the inconclusiveness of accepting the null hypothesis. *Psychological Review*, *71*, 238-242.

Wilson, W.R., Miller, H.L., & Lower, J.S. (1967). Much ado about the null hypothesis. *Psychological Bulletin*, *67*, 188-196.

Winch, R.F., & Campbell, D.T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist*, *4*, 140-143.

Winer, B.J. (1962). *Statistical Principles in Experimental Designs*. New York: McGraw-Hill.

Winkler, R. L. (1972). *Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart & Winston.

Winkler, R.L. (1974). Statistical analysis: Theory versus practice. *In* C.-A.S. Staël Von Holstein (Ed.), *The Concept of Probability in Psychological Experiments*. Dordrecht, Netherlands: D. Reidel, 127-140.

Winkler, R.L. (1993). Bayesian Statistics: An overview. *In* G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Statistical Issues*, Hillsdale, NJ: Erlbaum, 201-232.

Witehead, J. (1993). The case for frequentism in clinical trials. *Statistics in Medicine*, *12*, 1405-1413 [With discussion: The case against frequentism, 1415-1419].

Wolfowitz, J. (1967). Remarks on the theory of testing hypotheses. *New York Statistician*, *18*, 439-441.

Wolins, L. (1982). *Research Mistakes in the Social and Behavioral Sciences*. Ames: Iowa State University Press.

Wonnacott, R. J., & Wonnacott, T. H. (1985). *Introductory Statistics* (4th edition). New York: John Wiley & Sons.

Woolley, T.W. (1983). A comprehensive power-analytic investigation of research in medical education. *Journal of Medical Education*, *58*, 710-715.

Woolley, T.W., & Dawson, G.O. (1983). A follow-up power analysis of the statistical tests in the Journal of Research in Science Teaching. *Journal of Research in Science Teaching*, *20*, 673-681.

Woolson, R.F., & Kleinman, J.C. (1989). Perspectives on statistical significance. *Annual Review of Public Health*, *10*, 423-440.

Wright , A., & Ayton, P. – (1994). *Subjective probability*. Chichester: Wiley.

Wulff, H.R. (1973). Confidence limits in evaluating controlled therapeutic trials. *Lancet*, *2*, 969-970.

Wulff, H.R., Andersen, B., Brandenhoff, P., & Guttler, F. (1987). What do doctors know about statistics? *Statistics in Medicine*, *6*, 3-10.

Y

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Yates, F.(1939). An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters. *Proceedings of the Cambridge Philosophical Society*, *35*, 579-591.

Yates, F.(1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *Journal of the American Statistical Association*, *46*, 32-33.

Yates, F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics*, *20*, 307-321.

Yeaton, W.H., & Sechrest, L. (1986). Use and misuse of no-difference findings in eliminating threats to validity. *Evaluation Review*, *10*, 836-852.

Yoccoz, N.G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, *72*, 106-111.

Young, M.A. (1993). Supplementing tests of statistical significance: Variation accounted for. *Journal of Speech and Hearing Research*, *36*, 644-656.

Yule, G.U., & Greenwood, M. (1915). The statistics of anti-typhoid and anti-cholera inoculations and the interpretation of such statistics in general. *Proceedings of the Royal Society of Medicine*, *8*, 113-194.

Z

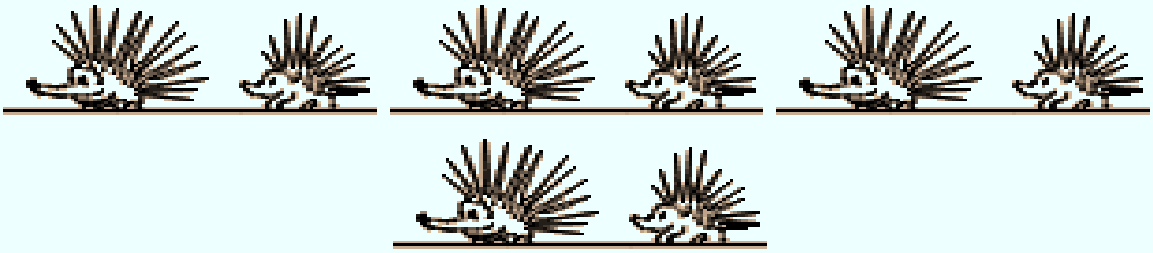
A B C D E F G H I J K L M N O P Q R S T U V W Y Z

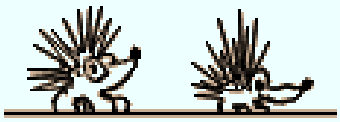
Zabell, S.L. (1989). R.A. Fisher on the history of inverse probability. *Statistical Science*, *4*, 247-263.

Zabell, S.L. (1992). R.A. Fisher and the fiducial argument. *Statistical Science*, *7*, 369-387.

Zeisel, H. (1955). The significance of insignificant differences. *Public Opinion Quarterly*, *17*, 319-321 [Reprinted in Morrison & Henkel, 1970, 79-80]

Zuckerman, M., Hodgins, H., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49-53.





RÉSUMÉS / ABSTRACTS

Abdi, H. (1987) **[*Exemple d'assimilation de l'hypothèse nulle à l'hypothèse d'une valeur de zéro pour le paramètre testé*]** "[...] l'effet est nul dans la Population, de là le terme d'Hypothèse Nulle." (page 74)

Abelson, R.P. (1997) Recognizing that hindsight often provides the clearest vision, this article examines the current significance testing controversy from a unique perspective - the future! It is the year 2006, significance tests have been banned since 1999, and already the pendulum of public opinion is swinging back in their favor: At this point, the author rediscovers a long-lost manuscript written in 1996, and finds that his views on the significance testing controversy have renewed relevance. Specifically: (1) Although bad practice certainly has characterized some significance testing, many of the critics of significance tests overstate their case by concentrating on such bad practice, rather than providing a balanced analysis; (2) Proposed alternatives to significance testing, especially meta-analysis, have flaws of their own; (3) Significance tests fill an important need in answering some key research questions, and if they did not exist they would have to be invented.

Aczel, A.D. (1995) **[*Example of interpretation of frequentist confidence intervals in terms of probabilities about parameters*]** "[the confidence level] a measure of the confidence we have that the interval does indeed contain the parameter of interest." (page 205)

Albert, J. (1995) Teaching elementary statistical inference from a traditional viewpoint can be hard, due to the difficulty in teaching sampling distributions and the correct interpretation of statistical confidence. Bayesian methods have the attractive feature that statistical conclusions can be stated using the language of subjective probability. Simple methods of teaching Bayes' rule are described, and these methods are illustrated for inference and prediction problems for one and two proportions. We discuss the advantages and disadvantages of traditional and Bayesian approaches in teaching inference and give texts that provide examples and software for implementing Bayesian methods in an elementary class.

Algina J., Moulder B.C. (2001) The increase in the squared multiple correlation coefficient (Delta*R*^2 associated with a variable in a regression equation is a commonly used measure of importance in regression analysis. The probability that an asymptotic confidence interval will include DeltaRho^2 was investigated. With sample sizes typically used in regression analyses, when DeltaRho^2=0.00 and the confidence level is .95 or greater, the probability will be at least .999. For DeltaRho^2=.01 and a confidence level of .95 or greater, the probability will be smaller than the nominal confidence level. For DeltaRho^2=.05 and a confidence level of .95, tables are provided for the sample size necessary for the probability to be at least .925 and to be at least .94.

Altham, P.M.E. (1969) A relationship is derived between the posterior probability of negative association of rows and columns of a 2x2 contingency table and Fisher's "exact" probability, as given in existing tables for testing the hypothesis of no association of rows and columns. The result for the 2x2 table is generalized to provide the posterior probability that one discrete-valued random variable is stochastically larger than another.

Amorim, M.A. (1999) [Example of use of ANOVA fiducial Bayesian procedures].

Amorim, M.A., Glasauer, S., Corpinot, K., & Berthoz, A. (1997) [Example of use of ANOVA fiducial Bayesian procedures].

Amorim, M.A., Loomis, J.M., & Fukusima, S.S. (1998) [Example of use of ANOVA fiducial Bayesian procedures].

Amorim, M.-A., & Stucchi, N. (1997) [Example of use of ANOVA fiducial Bayesian procedures].

Amorim, M.-A., Trumbore, B., & Chogyen, P.L. (2000). [Example of use of ANOVA fiducial Bayesian procedures].

Anderson, D.R., Burnham, K.P., & Thompson, W.L. (2000) This paper presents a review and critique of statistical null hypothesis testing in ecological studies in general, and wildlife studies in particular, and describes an alternative. Our review of Ecology and the journal of Wildlife Management found the use of null hypothesis testing to be pervasive. The estimated number of P-values appearing within articles of Ecology exceeded 8,000 in 1991 and has exceeded 3,000 in each year since 1984, whereas the estimated number of P-values in the Journal of Wildlife Management exceeded 8,000 in 1997 and has exceeded 3,000 in each year since 1991. We estimated that 47% (SE=3.9%) of the P-values in the Journal of Wildlife;fe 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. Towards this objective, we describe a relatively new paradigm of data analysis based on Kullback-Leibler information. This paradigm is an extension of likelihood theory and, when used correctly, avoids many of the fundamental limitations and common misuses of null hypothesis testing. Information-theoretic methods focus on providing a strength of evidence for an a priori set of alternative hypotheses, rather than a statistical test of a null hypothesis. This paradigm allows the following types of evidence for the alternative hypotheses: the rank of each hypothesis, expressed as a model; an estimate of the formal likelihood of each model, given the data; a measure of precision that incorporates model selection uncertainty; and simple methods to allow the use of the set of alternative models in making formal inference. We provide an example of the information-theoretic approach using data on the effect of lead on survival in spectacled elder ducks (Somateria fischeri). Regardless of the analysis paradigm used, we strongly recommend inferences based on a priori considerations be clearly separated from those resulting from some form of data dredging.

Anderson, D.R., Link, W.A., Johnson, D.H., & Burnham, K.P. (2001) We give suggestions for tile presentation of research results fi om frequentist, information-theoretic, and Bayesian analysis paradigms, followed by several general suggestions. The information-theoretic and Bayesian methods offer alternative approaches to data analysis and inference compared to traditionally used methods. Guidance is lacking on the presentation of results under these alternative procedures and on nontesting aspects of classical frequentist methods of statistical analysis. Null hypothesis testing has come under intense criticism. We recommend less reporting of the results of statistical tests of null hypothesis in casts where the null is surely false anyway, or where the null hypothesis is of little interest to science or management.

Atkins, L., & Jarrett, D. (1981) Significance tests perform a vital function in the social sciences because they appear to supply an objective method of drawing conclusions from quantitative data. Sometimes they are used mechanically, with little comment, and with even less regard for whether or not the required assumptions are satisfied. Often, too, they are used in a way that distracts attention from consideration of the practical importance of the questions posed or that disguises the inadequacy of the theoretical basis for the investigation conducted. We shall show how these tests developed historically from methodological ideas imported from the natural sciences and from ideological commitments inherent in nineteenth century social thought. We shall use the results of a recent investigation to present and criticise tests of significance. And in describing alternative approaches to evaluating research we shall argue that the central status of these tests in social science is by no means based on a consensus, even amongst statisticians, as to their appropriateness.

Azar, B. (1999) If implemented, a new set of recommendations for analyzing and reporting data will encourage researchers to be more rigorous and detailed in their reporting, and also open them up to using a broader group of methods and statistical techniques, says Robert Rosenthal, PhD, co-chair of APA's Task Force on Statistical Inference, which penned the recommendations.
[*Example of misinterpretations of null hypothesis significance tests*] "[a significant result] indicates that the chances of the finding being random is only 5 percent or less"

Bailar, J.C., & Mosteller, F. (1988) Provides 15 directions on manuscript preparation for reporting scientific statistics including essential elements needed for specific statistics. Provides detail on parts of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals.

Bakan, D. (1967/1966) I will attempt to show that the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; and that, furthermore, a great deal of mischief has been associated with its use. [...] 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. [...] Most important, 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.

Bartko, J.J. (1991) There is a body of literature in biostatistics (e.g., Blackwelder, 1982; Blackwelder & Chang, 1984; Detsby & Sackett, 1985; Dunnett & Gent, 1977; Makuh & Simon, 1978) that discusses "proving" the null hypothesis in an attempt at establishing bioequivalence. Some readers may be interested in a selection of articles along these lines.

Bassok, M., Wu, L.L., & Olseth, K.L. (1995) **[*Example of misinterpretations of null hypothesis significance tests*]** "In addition to the overall interpretative bias there was a very strong interaction between the training and the transfer problems [chi-square(1) = 14.71, $p<0.001$]." "Subject's performance *was not affected* by differences in the size of the assigned and the receiving sets [chi-square(1) = 0.08, n.s.], so we combined the results of subjects [...]"

Bayarri, M. J. & Berger, J. O. (2004) Statistics has struggled for nearly a century over the issue of whether the Bayesian or frequentist paradigm is superior. This debate is far from over and, indeed, should continue, since there are fundamental philosophical and pedagogical issues at stake. At the methodological level, however, the debate has become considerably muted, with the recognition that each approach has a great deal to contribute to statistical practice and each is actually essential for full development of the other approach. In this article, we embark upon a rather idiosyncratic walk through some of these issues.

Batanero, C. (2000) In spite of the widespread use of significance testing in empirical research, its interpretation and researchers' excessive confidence in its results have been criticised for years. In this paper, we first describe the logic of statistical testing in the Fisher and Neyman-Pearson approaches, review some common misinterpretations of basic concepts behind statistical tests, and analyse the philosophical and psychological issues that can contribute to these misinterpretations. We then revisit some frequent criticisms against statistical tests and conclude that most of them refer not to the tests themselves, but to the misuse of tests on the part of researchers. We agree with Levin (1998a) that statistical tests should be transformed into a more intelligent process that helps researchers in their work, and finally suggest possible ways in which statistical education might contribute to the better understanding and application of statistical inference.

Beauchamp, K.L., & May, R.B. (1964) In replicating the study by Rosenthal and Gaito (1963), [...] subjects were asked to express their "degree of belief in research findings as a function of associated *p* levels". [...] The results generally confirm those of Rosenthal and Gaito although [...] in the replication no significant "cliff effect" was found in intervals following the .05, .01 or any other *p* levels.

Berger, V.W. (2000) Hypothesis testing, in which the null hypothesis specifies no difference between treatment groups, is an important tool in the assessment of new medical interventions. For randomized clinical trials, permutation tests that reflect the actual randomization are design-based analyses for such hypotheses. This means that only such design-based permutation tests can ensure internal validity, without which external validity is irrelevant. However, because of the conservatism of permutation tests, the virtues of permutation tests continue to be debated in the literature, and conclusions are generally of the type that permutation tests should always be used or permutation tests should never be used. A better conclusion might be that there are situations in which permutation tests should be used, and other situations in which permutation tests should not be used. This approach opens the door to broader agreement, but begs the obvious question of when to use permutation tests. We consider this issue from a variety of perspectives, and conclude that permutation tests are ideal to study efficacy in a randomized clinical trial which compares, in a heterogeneous patient population, two or more treatments, each of which may be most effective in some patients, when the primary analysis does not adjust for covariates. We propose the p-value interval as a novel measure of the conservatism of a permutation test that can be defined independently of the significance level. This p-value interval can be used to ensure that the permutation test have both good global power and an acceptable degree of conservatism.

Bernard, J.-M. (1996) In considering the inference about the unknown proportion of a Bernoulli process, it is shown that the choices involved in the frequentist approach are equivalent, from a Bayesian viewpoint, to the choice of a particular ignorance prior within a restricted *ignorance zone*. This link sheds light on the nature of both kinds of choices, an on undesirable properties that go with null variance data.

Bernard, J.-M. (2000) Section 1 deals with the inference on one frequency, that is with binary data, under either an hypergeometric or a binomial sampling model; it will enable us to introduce the key concepts involved in the Bayesian approach and to compare it to the frequentist one. From this point on, we shall focus on Bayesian inference without further attempting to provide a systematic comparison

with frequentist inference. The predictive approach to inference, again on one frequency, is presented in Section 2. We then give, through concrete and real examples, an insight on how the Bayesian approach can be extended to situations involving several frequencies, first considering simple designs (Section 3), and then more cmplex ones (Section 4). The computational aspects, left aside in the first sections, are sketched in Section 5. Finally, Section 6 summarizes the major points put forward in the chapter.

Berry, D.A. (1987) The classical design of clinical trials is dictated by the eventual analysis. If the design varies from that planned then classical analysis is impossible. The Bayesian approach on the other hand is completely flexible and is therefore ideal for addressing questions and practical decision problems. I contrast these two approaches in two types of clinical trials: (i) those that strive to treat patients as effectively as possible and (ii) those sponsored by pharmaceutical companies attempting to maximise their expected profit.

Berry, D.A. (1991) The Bayesian approach to inference and decision making provides an integrated way of addressing the various aspects of drug development, from the early preclinical study of compounds through the clinical and postmarking phases. In particular, it provides a natural, convenient way for choosing among experimental designs. An essential aspect of the process of evaluating design strategies is the ability to calculate predictive probabilities of potential results. I describe a Bayesian approach to experimental design and illustrate it by considering a particular type of clinical trial. Also, I compare Bayesian and classical statistical attitudes toward design.

Berry, D.A. (1993) This paper describes a Bayesian approach to the design and analysis of clinical trials, and compares it with the frequentist approach. Both approaches addresses learning under uncertainty. But they are different in a variety of way. The Bayesian approach is more flexible. For example, accumulating data from a clinical trial can be used to update Bayesian measures, independent of the design of the trial. Frequentist measures are tied to the design, and interim analyses must be planned for frequentist measures to have meaning. Its flexibility makes the Bayesian approach ideal for analysing clinical trials. In carrying out a Bayesian analysis for inferring treatment effect, information from the clinical trial and other sources can be combined and used explicitly in drawing conclusions. Bayesians and frequentists address making decisions very differently. For example, when choosing or modifying the design of a clinical trial, Bayesians use all available information, including that which comes from the trial itself. The ability to calculate predictive probabilities for the future observations is a distinct advantage of the Bayesian approach to designing clinical trials and other decisions. An important difference between Bayesian and frequentist thinking is the role of randomization.

Berry, D.A. (1996) This is an introduction to statistics for general students. It differs from standard texts in that it takes a Bayesian perspective. It views statistics as a critical tool of science and so it has a strong scientific overtone. While my outlook is conventional in many ways, its foundation is Bayesian. There are several advantages of the Bayesian perspective:
- It allows for direct probability statements, such as the probability that an experimental procedure is more effective than a standard procedure.
- It allows for calculating probabilities of future observations.
- It allows for incorporating evidence from previous experience and previous experiments into overall conclusions.
- It is subjective. This is a standard objection to the Bayesian approach; different people reach different conclusions from the same experiment results.
- There would be comfort in giving an answer that others would also give. But differences of opinion are the norm in science and an approach that explicitly recognizes such differences is realistic.
- Despite differences in focus between the standard and Bayesian approaches, there are more similarities than differences. Many of the principles illustrated in the examples and exercises of this text are not peculiar to either approach.

Berry, D.A. (1997) University courses in elementary statistics are usually taught from a frequentist perspective. In this sarticle I suggest howsuch courses can be taught using a Bayesian approach and I undicatewhy students in a Bayesian course are well served. A principal focus of any good elementary course is the application of statistics to real and important scientific problems. The Bayesian approach fits neatly with a scientific focus. Bayesians take a larger view and one onot limited to data analysis. In particular, the Bayesian approach is subjective and requires assessing prior probabilities. This requirement forces users to relate current experimental evidence to other available information - including previous experiments of a related nature, where "related" is judged subjectively. I discuss difficulties faced by instructors and students in elementary Bayesian courses and I provide a sample syllabus for an elementary Bayesian,course.

Berry, D.A. & Hochberg, Y. (1999) We discuss Bayesian attitudes towards adjusting inferences for multiplicities. In the simplest Bayesian view, there is no need for adjustments and the Bayesian perspective is similar to that of the frequentist who makes inferences on a per-comparison basis. However, as we explain, Bayesian thinking can lead to making adjustments that are in the same spirit as those made by frequentists who subscribe to preserving the familywise error rate. We describe the differences between assuming independent prior distributions and hierarchical prior distributions. As an example of the latter, we illustrate the use of a Dirichlet process prior distribution in the context of multiplicities. We also discuss some quasi-Bayesian procedures which combine Bayesian and frequentist ideas. This shows the potential of Bayesian methodology to yield procedures that can be evaluated using "objective" criteria. Finally, we comment on the role of subjectivity in Bayesian approaches to the complex realm of multiple comparisons problems, and on robust vs. informative priors.

Berry, G. (1986) 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 [...] 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).

Beshers J (1958) Hanan Selvin (1957) has confused statistical inference with causal inference. All this statements to the effect that sociologists need not employ significance tests in survey research are based upon this confusion. [...] Significance tests are of little value for surveys which (1) ignore the principle of sampling, and (2) are not guided by theory. Perhaps such exploratory surveys may generate hypotheses to be verified by a subsequent well-designed survey utilizing significance tests.

Bezeau, S.; Graves, R. (2001) Cohen, in a now classic paper on statistical power, reviewed articles in the 1960 issue of one psychology journal and determined that the majority of studies had less than a 50-50 chance of detecting an effect that truly exists in the population, and thus of obtaining statistically significant results. Such low statistical power, Cohen concluded, was largely due to inadequate sample sizes. Subsequent reviews of research published in other experimental psychology journals found similar results. We provide a statistical power analysis of clinical neuropsychological research by reviewing a representative sample of 66 articles from the Journal of Clinical and Experimental Neuropsychology, the Journal of the International Neuropsychology Society, and Neuropsychology. The results show inadequate power, similar to that for experimental research, when Cohen's criterion for effect size is used. However, the results are encouraging in also showing that the field of clinical neuropsychology deals with larger effect sizes than are usually observed in experimental psychology and that the reviewed clinical neuropsychology research does have adequate power to detect these larger effect sizes. This review also reveals a prevailing failure to heed Cohen's recommendations that researchers should routinely report a priori power analyses, effect sizes and confidence intervals, and conduct fewer statistical tests.

Bhattacharyya & Johnson (1997) **[*Example of interpretation of frequentist confidence intervals in terms of probabilities about parameters*]**"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." (page 243)>

Bird, K.D. (2002) Although confidence interval procedures for analysis of variance (ANOVA) have been available for some time, they are not well known and are often difficult to implement with statistical packages. This article discusses procedures for constructing individual and simultaneous confidence intervals on contrasts on parameters of a number of fixed-effects ANOVA models, including multivariate analysis of variance (MANOVA) models for the analysis of repeated measures data. Examples show how these procedures can be implemented with accessible software. Confidence interval inference on parameters of random-effects models is also discussed.

Blaich, C.F. (1998) If the NHSTP [Null Hypothesis Significance Test Procedure] procedure is essential for controlling for chance, why is very little, if any, discussion of the nature of chance by Chow [Chow, 1996] and other advocates of the procedure. Also, many criticisms that Chow takes to be aimed against the NHSTP procedure are actually directed against the kind of theory that is tested by the procedure.

[37] **[*Example of misinterpretation of significance levels*]** "[...] 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." (page 639)>

Borenstein, M. (1997) **Learning objectives**: This paper provides the reader with an overview of several key elements in study planning and analysis. In particular, it highlights the differences between significance tests (statistical significance) and effect size estimation (clinical significance). **Data sources**: This paper focuses on methodologic issues, and provides an overview of trends in research. **Paper selection**: References were selected to provide a cross-section of the approaches currently being used. The paper also discusses a number of logical fallacies that have been cited as examples in earlier papers on research design. **Conclusions**: 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. This approach incorporates all the information normally included in a test of significance but in a format that highlights the element of interest (clinical significance rather than statistical significance). This approach should also have an impact on study planning--a study should have enough power to reject the null hypothesis and also to yield a precise estimate of the treatment effect.

Boring, E.G. (1919) So it happens that the competent scientist does the best the can in obtaining unselected samples, makes his observations, computes a difference and its "significance", and then – absurd as it may seem – very often discards his mathematical result, because in his judgment the mathematically "significant" difference is nevertheless not large compared with that he believes is the discrepancy between his samples and the larger groups which they represent. [...] The case is one of many where statistical ability,

divorced from a scientific intimacy with the fundamental observations, leads nowhere.

Box, G.E.P., & Tiao, G.C. (1973) The object of this book is to explore the use and relevance of Bayes' theorem to problems such as arise in scientific investigation in which inferences must be made concerning parameter values about which little is known *a priori*. In Chapter 1 we discuss some important general aspects of the Bayesian approach, including: the role of Bayesian inference in scientific investigation, the choice of prior distributions (and, in particular, of noninformative prior distributions), the problem of nuisance parameters, and the role and relevance of sufficient statistics. In Chapter 2, a number of standard problems concerned with the comparison of location and scale parameters are discussed. Bayesian methods, for the most part well known, are derived there which closely parallel the inferential techniques of sampling theory associated with *t*-tests, *F*-tests, Bartlett's tests, the analysis of variance, and with regression analysis. [...] Now, practical employment of such techniques has uncovered further inferential problems, and attempts to solve these, using sampling theory, have had only partial success. One of the main objective of this book, pursued from Chapter 3 onwards, is to study some of these problems from a Bayesian viewpoint.
The following are examples of the further problems considered: 1. How can inferences be made in small samples about parameters for which no parsimonious set of sufficient statistics exists? 2. To what extent are inferences about means and variances sensitive to departures from assumptions such as error Normality, and how can such sensitivity be reduced? 3. How should inferences be made about variance components? 4. How and in what circumstances should mean squares be pooled in the analysis of variance? 5. How can information be pooled from several sources when its precision is not exactly known, but can be estimated as, for example, in the "recovery of interblok information" in the analysis of incomplete block designs? 6. How should data be transformed to produce parsimonious parametrization of the model as well as to increase sensitivity of the analysis?
The main body of the text is an investigation of these and similar questions with appropriate analysis of the mathematical results illustrated with numerical examples. We believe that this (1) provides evidence of the value of the Bayesian approach, (2) offers useful methods for dealing with the important problems specifically considered and (3) equips the reader with techniques which he can apply in the solution of new problems.

Braitman, L.E. (1988) The statistical descriptors known as confidence intervals can increase the ability of readers to evaluate conclusions drawn form small trials. Fortunately, an increasing number of journals are asking authors to add confidence intervals to the reporting of data in their papers.

Braitman, L.E. (1991) In this editorial, I use hypothetical examples to illustrate point estimates and confidence intervals of the differences between the percentages of patients responding to two treatments for a cancer. These examples show how confidence intervals can help assess the clinical and statistical significance of such differences.

Brandstätter, E. (1999) The article argues to replace null hypothesis significance testing by confidence intervals. Correctly interpreted, confidence intervals avoid the problems associated with null hy-pothesis statistical testing. Confidence intervals are formally valid, do not depend on a-priori hypotheses and do not result in trivial knowledge. The first part presents critique of null hypothesis significance testing; the second part replies to critique against confidence intervals and tries to demonstrate their superiority to null hypothesis significance testing.

Breslow, N. (1990) Attitudes of biostatisticians toward implementation of the Bayesian paradims have changed during the past decade due to the increased availability of computational tools for realistic problems. Empirical Bayes' methods, already widely used in the analysis of longitudinal data, promise to improve cancer incidence maps by accounting for overdispersion and spatial correlation. Hierarchical Bayes' methods offer a natural framework in which to demonstrate the bioequivalence of pharmacologic compounds. Their use for quantitative risk assessment and carcinogenesis bioassay is more controversial, however, due to uncertainty regarding specification of informative priors. Bayesian methods simplify the analysis data from sequential clinical trials and avoid certain paradoxes of frequentist inference. They offer a natural setting for the synthesis of expert opinion in deciding policy matter. Both frequentist and Bayes' methods have a place in biostatistical practice.

Bristol, D.R. (1995) Sample size determination is a very important aspect of planning a clinical trial. The actual calculation is usually the responsibility of the project statistician, but fruitful communications with the clinical monitor are required. When the variable of interest follows a normal distribution, the statistician must have specified values of the error variance and a difference. Some consequences of mispecification of these values, with emphasis on the difference, are presented. Some discussion of the role of communication between the project statistician and the clinical monitor is also presented.

Brown, F.L. (1973) **[*Example of misinterpretation of significance levels*]** Pages 522-523 [Quoted by Seldmeier & Gigerenzer, 1989, page 314]

Camilleri, S.F. (1962) In attempting to clarify the role of probability in sociological research we have been led into a discussion of the

nature of scientific theory and induction. We have tried to articulate the principle that science induction is accomplished through the construction and verification of deductive theories, the primary concern of the social scientist ought to be the development of such theories. [...] We have tried to show that the hypothetical character of the risk probabilities associated with the level of significance and the pragmatic ambiguities of the rationale for choosing any particular level of significance seriously undermine its value in the evaluation of statistical hypotheses. 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.

Capraro R.M., Capraro M.M. (2002) The dialog surrounding effect sizes and statistical significance tests often places the two ideas into separate camps amid controversy. In light ofrecommendations by the Task Force on Statistical Inference and the fifth edition of the *American Psychological Association Publication Manual* calling for the reporting of effect sizes, a review of treatments of effect sizes in textbooks may be quite timely. This study reviews textbooks published since 1995 and as regards treatments of effect sizes and statistical significance tests. Of the textbooks examined, every textbook (*n*= 89) included the topic of statistical significance testing (2,248 pages), whereas only a little more than two thirds of the textbooks (*n*= 60) included information on effect sizes (789 pages).

Charron, C. (2002) [Example of use of Bayesian methods: ANOVA fiducial Bayesian procedures and procedures for implication hypotheses in 2x2 tables (Lecoutre & Charron, 2000)]

Chatfield, C. (1988) Based on a teaching course for final-year undergraduates, and on wide consultancy experience, this readable book provides a wealth of information and valuable insight for the student statistician and practioner alike.
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." (page 51)
"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." (page 51)

Chow, S.L. (1988) I describe and question the argument that in psychological research, the significance test should be replaced (or, at least, supplemented by a more informative index (viz., effect size or statistical power) in the case of theory-corroboration experimentation because it has been made on the basis of some debatable assumptions about the rationale of scientific investigation. The rationale of theory-corroboration experimentation requires nothing more than a binary decision about the relation between two variables. This binary decision supplies the minor premise for the syllogism implicated when a theory is being tested. Some metatheoretical considerations reveal that the magnitude of the effect-size estimate is not a satisfactory alternative to the significance test.

Chow, S.L. (1989) Shows that agreeing with Folger's (1989) methodological observations does not mean that it is incorrect to use significance tests. This contention is based on the dynamics of theory corroboration, with reference to which the following distinction are illustrated, namely, the distinction between (a) statistical hypothesis testing, theory corroboration, and syllogistic argument, (b) a responsible experimenter and a clinical experimenter, (c) logical validity and methodological correctness, and (d) warranted assertability and truth.

Chow, S.L. (1991) This is Chow's response to four comments on his critique of the view that research conclusions be based on multiple context-dependent criteria. Five themes could be identified in the comments. In reply, it is argued that care should be taken not to use the alpha level whimsically because the continuum *similarity*, is being used as a dichotomy in theory corroboration. The superiority of effect-size estimates to statistical significance is more apparent than real. Chow's assessment of meta-analysis is illustrated with the *apples and oranges* issues.

Chow, S.L. (1991) In sum, two putative advantages of basing theoretical conclusions on statistical power can be questioned. A null hypothesis is not a categorical proposition descriptive of the world, but a prescriptive statement. Using tests of significance is not incompatible with making rational judgment.

Chow, S.L. (1996) The null-hypothesis significance-test procedure (NHSTP) is defended in the context of the theory-corroboration experiment, as well as the following contrasts: (a) substantive hypotheses versus statistical hypotheses, (b) theory corroboration versus statistical hypothesis testing, (c) theoretical inference versus statistical decision, (d) experiments versus nonexperimental studies, and (e) theory corroboration versus treatment assessment. The null hypothesis can be true because it is the hypothesis that errors are randomly distributed in data. Moreover, the null hypothesis is never used as a categorical proposition. Statistical significance means only that chance influences can be excluded as an explanation of data; it does not identify the nonchance factor responsible. The experimental conclusion is drawn with the inductive principle underlying the experimental design. A chain of deductive arguments gives rise to the theoretical conclusion via the experimental conclusion. The anomalous relationship between statistical significance and the effect size often used to criticize NHSTP is more apparent than real. The absolute size of the effect is not an index of evidential support for the

substantive hypothesis. Nor is the effect size, by itself, informative as to the practical importance of the research result. Being a conditional probability, statistical power cannot be the *a priori* probability of statistical significance. The validity of statistical power is debatable because statistical significance is determined with a single sampling distribution of the test statistic based on H0, whereas it takes two distributions to represent statistical power or effect size. Sample size should not be determined in the mechanical manner envisaged in power analysis. It is inappropriate to criticize NHSTP for nonstatistical reasons. At the same time, neither effect size nor confidence interval estimate nor posterior probability can be used to exclude chance as an explanation of data. Nor can any of them fulfill the nonstatistical functions expected of them by critics.

Chow, S. L. (1998) Sohn (1998) presents a good argument that neither statistical significance nor effect size is indicative of the replicability of research results. His objection to the Bayesian argument is also succinct. However, his solution of the "replicability belief" issue is problematic, and his verdict that significance tests have no role to play in empirical research is debatable. The strengths and weaknesses of Sohn's argument may be seen by explicating some of his assertions.

Ciancia, F., Maitte, M., Honoré, J., Lecoutre, B., & Coquery, J.-M. (1988) **[*Illustration of standard Bayesian methods*]**"Analysis of variance was extended by standard Bayesian inferences. Whereas *F* ratio is only a test of the null hypothesis *delta*= 0, standard Bayesian inferences enabled us to investigate the magnitude of *delta*. To a significant result, a statement of this kind is added: *P*(*delta*<X) = 0.95 or *P*(*delta*>X) = 0.95, indicating which value (X) from the *delta* parameter has a 0.95 probability of being exceeded. In the case of a nonsignificant effect a statement of the following kind is calculated: *P*(|*delta*|<X) = 0.95, giving the interval centered on 0 and containing with a probability of 0.95 the true effect: if the value of *X* is small, the effect must, with reason, be considered negligible."

Clément, E., & Richard, J.-F. (1997)[Example of use of standard Bayesian methods].

Cohen, J. (1962) The purpose of the study was to survey the articles of the *Journal of Abnormal and Social Psychology*, 1960, 61, from the point of view of the power of their statistical tests to reject their major null hypotheses, for defined levels of departure of population parameters from null conditions, i.e., size of effect. Conventional tests conditions were employed in power determination: nondirectional tests at the .05 level. [...] It was found that the average power (probability of rejecting false null hypotheses) over the 70 research studies was .18 for small effects, .48 for medium effects, and .83 for large effects. These values are deemed to be far too small and suggest that much research in the abnormal-social area has lead to the failure to reject null hypotheses which are in fact false. [...] 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, J. (1988) The power of a statistical test is the probability that it will yield statistically significant results. Since statistical significance is so earnestly sought and devoutly wished for by behavioral scientists, one would think that the *a priori* probability of its accomplishment would be routinely determined and well understood. Quite surprisingly, this is not the case. Instead, if we take as evidence the research literature, we find that statistical power is only infrequently understood and almost never determined. The immediate reason for this is not hard to discern – the applied statistics textbooks aimed at behavioral scientists, with few exceptions, give it scant attention.
The purpose of this book is to provide a self-contained comprehensive treatment of statistical power analysis from an "applied" viewpoint. The purpose of Chapter 1 is to present the basic conceptual framework of statistical hypothesis testing, giving emphasis to power, followed by the framework within which this book is organized. Each of the succeeding chapters present a different statistical test. They are similarly organized as follows: 1. The test is introduced and its use described. 2. The ES [effect size] index is described and discussed in detail. 3. The characteristics of the power tables and the method of their use are described and illustrated with examples. 4. The characteristics of the sample size tables and the method of their use are described and illustrated with examples. 5. The use of the power tables for significance tests is described and illustrated with examples.

Cohen, J. (1994) After 4 decades of severe criticism, the ritual of null hypothesis significance testing - mechanical dichotomous decisions around a sacred .05 criterion - still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of *p* as the probability that H0 is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects H0 one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods is suggested. For generalization, psychologists must finally rely, as has been done in all the older science on replication.

Cooper & Topher (1994) **[*Example of incorrect definition of a p-value*]** "The *p*-value is the probability that the results could have occurred by pure chance given that the null (conventional) hypothesis is true."

Corroyer, D., Devouche, E., Bernard, J.-M., Bonnet, P., & Savina, Y. (2003) Nous comparons six logiciels statistiques (EyeLID-2, PAC, SPSS, Statistica, Statview, Var3) pour l'analyse de données relevant de l'ANOVA (plan S<A2*B2> déséquilibré) sur les aspects descriptif et inductif et de plusieurs points de vue: 1/ accès à diverses options de comparaisons (équipondérer ou non, spéccifique ou non); 2/ intégration d eprocédures liées à des avancées méthodologiques récentes définies en particulier sous l'égide de l'*APA* (évaluation de la taille des effets, inférence bayésienne); 3/ mode d'accès aux procédures. Nous constatons que toutes les options ou procédures souhaitables ne sont pas toujours disponibles. Il apparaît donc nécessaire de recourir à plusieurs logiciels. Pour certains logiciels, on constate parfois un déficit d'information dans l'affichage, voire des incohérences entre divers résultats produits, ceci risquant de conduire le chercheur à des conclusions erronées.

Cox, D.R. (1977) The main object of the paper is to give a general review of the nature and importance of significance tests. Such tests are regarded as procedures for measuring the consistency of data with a null hypothesis by the calculation of a *p*-value (tail area). A distinction is drawn between several kinds of null hypothesis. The ways of deriving tests, namely via the so-called absolute test, via implicit consideration of alternatives and via explicit consideration of alternatives are reviewed. Some of the difficulties of the multidimensional alternatives are outlined and the importance of the diagnostic ability of a test is stressed. Brief examples include tests of distributional form including multivariate normality. The effect of modifying statistical analysis in the light of the data is discussed, four main cases being distinguished. Then a number of more technical aspects of significance tests are outlined, including the role of the continuity correction, Bayesian tests and the use of tests in the comparison of alternative models. Finally the circumstances are reviewed under which significance tests can provide the main summary of a statistical analysis.

Cox, D.R. (2001) Comment on Sterne, J.A.C., &: Davey Smith , G. (2001). Sifting the evidence—what's wrong with significance tests? *BMJ*, *322*, 226–231.

Cox, D.R., & Snell, E.J. (1981) P values, or significance levels, measure the strength of the evidence against the null hypothesis; the smaller the P value, the stronger the evidence against the null hypothesis An arbitrary division of results, into “significant” or “non-significant” according to the P value, was not the intention of the founders of statistical inference A P value of 0.05 need not provide strong evidence against the null hypothesis, but it is reasonable to say that P < 0.001 does. In the results sections of papers the precise P value should be presented, without reference to arbitrary thresholds Results of medical research should not be reported as “significant” or “non-significant” but should be interpreted in the context of the type of study and other available evidence. Bias or confounding should always be considered for findings with low P values To stop the discrediting of medical research by chance findings we need more powerful studies.

Craig, J.R., Eison, C.L., & Metze, L.P. (1976) The issue of the interpretation of significance tests is addressed. An argument is presented that some measure of association such as *omega-square* should be provided as an interpretation/decision-making aid for scientific consumers and journal editors. Published research articles were examined regarding use of measures of association and the relationship between sample size and the amount of variance shared by the independent and the dependent variable. The results indicated that no articles reported measures of association and that many published studies are based upon small degrees of relationship between the independent and dependent variable. A change in report-writing and journal edition practices is suggested.

Crow, E.L. (1991) Rosenthal (1990) is wrong in advocating the use of his and Rubin's binomial effect size display (BESD) "to index the practical value of our research results [...]". He is wrong because the BESD corresponds to no real population of interest.

Cumming, G., & Finch, S. (2001) Reform of statistical practice in the social and behavioural sciences requires wider use of confidence intervals (CIs), and of effect size measures and meta-analysis. In this context we discuss four reasons for promoting use of CIs: (i) they give useful, interpretable information; (ii) they are linked to statistical significance tests with which researchers are already familiar; (iii) they can encourage meta-analytic thinking that focuses on estimation; and (iv) CI width gives information about precision that may be more useful than a statistical power value. We focus on a basic standardised effect size measure, Cohen’s delta (also referred to as Cohen’s *d*). We give methods and examples for the calculation of CIs for delta, which require use of noncentral t distributions, and contrast these with the familiar CIs for original score means. We discuss noncentral t distributions, unfamiliar to many social scientists, and apply these also to statistical power and to simple meta-analysis of standardised effect sizes. We provide the *ESCI* graphical software, which runs under Microsoft Excel, to illustrate the discussion. Wider use of CIs for delta and other effect size measures should help promote highly desirable reform of statistical practice in the social sciences.

Cumming, G., & Finch, S. (2005). Wider use in psychology of confidence intervals (CIs), especially as error bars in figures, is a desirable development. However, psychologists seldom use CIs and may not understand them well. The authors discuss the interpretation of figures with error bars, and analyze the relationship between CIs and statistical significance testing. They propose 7 rules of eye to guide the inferential use of figures with error bars. These include general principles: Seek bars that relate directly to effects of interest, be sensitive to experimental design, and interpret the intervals. They also include guidelines for inferential

interpretation of the overlap of CIs on independent group means. Wider use of interval estimation in psychology has the potential to improve research communication substantially.

Cumming, G., Williams, J., & Fidler, F. (2004) Confidence intervals (CIs) and standard error bars give information about replication, but do researchers have an accurate appreciation of that information? Authors of journal articles in psychology, behavioural neuroscience, and medicine were invited by email to visit a website and indicate on a figure where they judged replication means would plausibly fall. Responses from 263 researchers suggest that many leading researchers in the three disciplines under-estimate the extent that future replications will vary. A 95% CI will on average capture 83.4% of future replication means. A majority of respondents, however, hold the confidence level misconception (CLM) that a 95% CI will on average capture 95% of replication means. Better understanding of CIs is needed if they are to be successfully used more widely in psychology.

D'Agostini, G. (1999) Subjective probability is based on the intuitive idea that probability quantifies the degree of belief that an event will occur. A probability theory based on this idea represents the most general framework for handling uncertainty. A brief introduction to subjective probability and Bayesian inference is given, with comments on typical misconceptions which tend to discredit it and with comparisons to other approaches.

D'Agostini, G. (2000) Criticisms of so called 'subjective probability' come on the one hand from those who maintain that probability in physics has only a frequentistic interpretation, and, on the other, from those who tend to 'objectivise' Bayesian theory, arguing, e.g., that subjective probabilities are indeed based 'only on private introspection'. Some of the common misconceptions on subjective probability will be commented upon in support of the thesis that coherence is the most crucial, universal and 'objective' way to assess our confidence on events of any kind.

D'Agostini, G. (2000) This contribution to the debate on confidence limits focuses mostly on the case of measurements with 'open likelihood', in the sense that it is defined in the text. I will show that, though a prior-free assessment of *confidence* is, in general, not possible, still a search result can be reported in a mostly unbiased and efficient way, which satisfies some desiderata which I believe are shared by the people interested in the subject. The simpler case of 'closed likelihood' will also be treated, and I will discuss why a uniform prior on a sensible quantity is a very reasonable choice for most applications. In both cases, I think that much clarity will be achieved if we remove from scientific parlance the misleading expressions 'confidence intervals' and 'confidence levels'.

D'Agostini, G. (2003) This report introduces general ideas and some basic methods of the Bayesian probability theory applied to physics measurements. Our aim is to make the reader familiar, through examples rather than rigorous formalism, with concepts such as: model comparison (including the automatic Ockham's Razor filter provided by the Bayesian approach); parametric inference; quantification of the uncertainty about the value of physical quantities, also taking into account systematic effects; role of marginalization; posterior characterization; predictive distributions; hierarchical modelling and hyperparameters; Gaussian approximation of the posterior and recovery of conventional methods, especially maximum likelihood and chi-square tests under well defined conditions; conjugate priors, transformation invariance and maximum entropy motivated priors; Monte Carlo estimates of expectation, including a short introduction to Markov Chain Monte Carlo methods.

Daniel, L.G. (1998) Statistical significance tests (SSTs) have been the object of much controversy among social scientists. Proponents have hailed SSTs as an objective means for minimizing the likelihood that chance factors have contributed to research results; critics have both questioned the logic underlying SSTs and bemoaned the widespread misapplication and misinterpretation of the results of these tests. The present paper offers a framework for remedying some of the common problems associated with SSTs via modification of journal editorial policies. The controversy surrounding SSTs is overviewed, with attention given to both historical and more contemporary criticisms of bad practices associated with misuse of SSTs. Examples from the editorial policies of Educational and Psychological Measurement and several other journals that have established guidelines for reporting results of SSTs are overviewed, and suggestions are provided regarding additional ways that educational journals may address the problem.

Dar, R. (1998) Chow’s [Chow, 1996] account of Bayesian inference logic and procedures is replete with fundamental misconceptions, derived from secondary sources and not adequately informed by modern work. The status of subjective probabilities in Bayesian analyses is misrepresented and the cogent reasons for the rejection by many statisticians of the curious inferential hybrid used in psychological research are not presented.

De Cristofaro, R. (1996) In this paper we support the idea that statistical inference can be worked out as a branch of inductive inference. Indeed; unless the assumptions are changed from the very beginning, we must find the solution to the problem of inference inside probability calculus. Moreover, we are not allowed to reach a conclusion (as the choice of a hypothesis) that is outside the scope of probability theory. In this connection, the importance of an appropriate analysis of the complete posterior distribution about the parameters in question is underlined, particularly where there are several parameters.

De Cristofaro, R. (2002) In this article, we support the idea that inductive reasoning can be worked out within probability theory, by means of a logical solution to the old problem of prior probabilities, and that accepting or rejecting hypotheses is a pragmatic choice, which does not belong to inductive reasoning. Many authors that solve statistical inference by simply examining the likelihood function do not follow Bayes theorem. We are consistent with Bayes theorem, and we think that the piece of information about the design is potentially contained in the prior. Later on, we provide a justification to the Jeffreys-rule to assign prior probabilities in the version supported by Box and Tiao. Our conclusion is that the best method to communicate the conclusions of a statistical research in an objective way consists in a probabilistic statement. On the contrary, the significance level is not often a good method of summarizing the information in the posterior distribution.

De Cristofaro, R. (2003) According to the likelihood principle, if the designs produce proportional likelihood func-tions, one should make an identical inference about a parameter from the data irrespective of the design, which yields the data. If it comes to that, there are several counter-examples, and/or paradoxical consequences to likelihood principle. Besides, as we will see, contrary to a widely held opinion, such a principle is not a direct consequence of Bayes theorem. In particular, the piece of information about the design is one part of the evidence, and it is relevant for the prior. Later on, a justification to Jeffreys-rule to assign prior probabilities in the version supported by Box and Tiao is provided. Another basic idea of the present paper is that (apart from other information) the equiprobability as-sumption is to be linked to the idea of the impartiality of design with respect to the pa-rameter under consideration. The whole paper has remarkable implications on the founda-tions of statistics from the notion of sufficiency, the relevance of the stopping rule and of the randomization in survey sampling and in the experimental design, the difference be-tween ignorable and non-ignorable designs, until a reconciliation of different approaches to the inductive reasoning in statistical inference.

Deheuvels, P. (1984) We describe test procedures enabling to decide whether bioequivalence is true or not from the study of the results of a comparative analysis. We prove that a correct use of Student confidence intervals gives a test uniformly more powerful that the corresponding methods based on Westlake [Westlake, 1976] confidence intervals.

del Rosal, A.B., Costas, C.S., Bruno, J.A.S., & Osinski, I.C. (2001) Null hypothesis significance testing has been a source of debate within the scientific community of behavioral researchers for years, since inadequate interpretations have resulted in incorrect use of this procedure. In this paper, we present a revision of the latest contributions of methodologists of different opinions, for and against, and we also set out the guidelines to research within behavioral science recently issued by the A.P.A. (American psychological Association) Task Force in Statistical Inference (Wilkinson, 1999).

Denhière, G., & Lecoutre, B. (1983) [*Illustration des méthodes bayésiennes standard*] Trois groupes de 60 enfants âgés de 7, 8 et 10 and ont été soumis à une expérience de reconnaissance immédiate et différée (une semaine) d'énoncés appartenant à des récits, et de distracteurs sémantiquement proches et lointains. L'expérience tentait de répondre à quatre questions: 1. Constate-t-on un "effet de niveau" en reconnaissance comme en rappel? 2. Observe-t-on un effet de l'âge comparable à celui obtenu en rappel ? 3. L'information est-elle stockée sous forme lexicale et/ou conceptuelle? 4. Des récits différents par leur contenu conduisent-ils à des performances différentes? L'analyse bayésienne des comparaisons (extension bayésienne de l'analyse de la variance) permet de répondre négativement aux question 1, 2 et 4. L'absence d'effet de niveau en reconnaissance immédiate et différée et les faibles différences de performance en fonction de l'âge conduisent à privilégier les modèles qui prévoient une représentation hiérarchique de l'information en mémoire et un processus de recherche et de récupération de l'information du type haut-bas.
Story memory: immediate and delayed recognition of statements by 7, 8 and 10 years old children
[*Example of use of standard Bayesian methods*] Three groups of 60 children (7, 8 and 10 years old) participated in an immediate and delayed (8 days) recognition experiment. Children had to identify original statements (segments of the story) and to reject statements which were semantically closed tp and distant from the original ones. The experiment intended to answer four questions: 1. Is there a level-effect present in recognition as there in recall? 2. Is there an effect of age similar in recognition and in recall? 3. Is the information stored in conceptual and/or lexical form? 4. Is the influence of the content of the stories different in recognition and in recall? The Bayesian Analysis of Comparisons (Bayesian extensions of ANOVA) leads to a negative answer to question 1, 2 and 4. The absence of a level-effect in immediate and delayed recognition and the small differences between the three age groups are in agreement with memory models which predict a hierarchical representation of information and a top down retrieval process.

Duggan, T.J., & Dean, C.W. (1968) [...] two elementary safeguards can be exercised in reporting results. One is routinely to compute and report a measure of degree of association in addition to the statistical test whenever this is possible. The second safeguard is the introduction of care and caution in the verbal interpretation of data tables and the inferred association of variables.

DuMouchel, W. (1989) In this chapter we provide step-by-step instructions for setting up a Bayesian hierarchical in order to combine statistical summaries from several studies into a single super analysis which integrates the results from each study. A discussion of the data requirements of the methodology is followed by a specification of a particular Bayesian Model designed to be both flexible and easy

to use. A set of formulas define all the computations necessary to obtain the posterior distributions of the relevant parameters. An example metaanalysis shows how different specifications of the prior distribution can affect the results.

Dunne A., Pawitan, Y., & Doody, L. (1996) In statistical practice P-values are regularly used to express the amount of evidence in the data, but there is no agreement on how to compute two-sided P-values when the sampling distributions are discrete and asymmetric. Doubling the one-sided P-value, or adding the probabilities less than or equal to the probability of the observed data, has been suggested in practice. However, since P-values are associated with a test, it is not clear what tests correspond to those suggested P-values. In this paper we suggest a way to compute the two-sided P-value as the smallest significance level for which, given the data, we would reject the null hypothesis in favour of a two-sided alternative by using the appropriate uniformly most powerful unbiased test. The method is illustrated using the small sample testing of a binomial proportion and the exact analysis of 2x2 tables as examples. The resulting P-value is compared with the previous two methods and a general discussion on the nature of P-values and two-sided tests is given.

Edwards, W., Lindman, H., & Savage, L.J. (1963) Bayesian statistics, a currently controversial viewpoint concerning statistical inference, is based on a definition of probability as a particular measure of the opinions of ideally consistent people. Statistical inference is modification of these opinions in the light of evidence, and tools of Bayesian statistics include the theory of specific distributions and the principle of stable estimates , which specifies when actual prior opinions may be satisfactorily approximated by a uniform distribution. A common feature of many classical significance tests is that a sharp null hypothesis is compared with a diffuse alternative hypothesis. Often evidence which, for a Bayesian statistician, strikingly supports the null hypothesis leads to rejection of that hypothesis by standard classical procedures. The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.

Efron, B. (1996) This article attempts to answer the following question: why is most scientific data analysis carried out in a non-Bayesian framework? The argument consists mainly of some practical examples of data analysis, in which the Bayesian approach is difficult but Fisherian/frequentist solutions are relatively easy. There is a brief discussion of objectivity in statistical analyses and of the difficulties of achieving objectivity within a Bayesian framework. The article ends with a list of practical advantages of Fisherian/frequentist methods, which so far seem to have outweighted the philosophical superiority of Bayesianism.

Efron, B. (1998) Fisher is the single most important figure in 20th century statistics. This talk examines his influence on modern statistical thinking, trying to predict how Fisherian we can expect the 21st century to be. Fisher's philosophy is characterized as a series of shrewd compromises between the Bayesian and frequentist viewpoints, augmented by some unique characteristics that are particularly useful in applied problems. Several current research topics are examined with an eye toward Fisherian influence, or the lack of it, and what this portends for future statistical developments.

Elifson, K.W., Runyon, R.P., & Haber, A. (1990) **[*Example of interpretation of frequentist confidence intervals in terms of probabilities about parameters*]** "[...] we assert that the population mean probably falls within the interval that we have established." (page 367)

Ellis, N. (2000) *Language Learning*, like many journals that publish research using quantitative and statistical methods, is increasingly influenced by the advantages of the reporting of effect sizes. Submitting authors to this journal have to date been referred to the statement in the Publication Manual of the American Psychological Association (4th edition) which emphasizes that statistical significance *p* values are not acceptable indices of effect because they depend on sample size and that "you are [therefore] encouraged to provide effect size information." (APA, 1994, p. 18). Unfortunately, empirical studies of this and other journals (Wilkinson & the American Psychological Association Task Force on Statistical Inference, 1999) indicate that this encouragement has had negligible impact. The reporting of effect sizes is essential to good research. It enables readers to evaluate the stability of results across samples, operationalizations, designs, and analyses. It allows evaluation of the practical relevance of the research outcomes. It provides the basis of power analyses and meta-analyses needed in future research. This role of effect sizes in meta-analysis is clearly illustrated in the article by Norris and Ortega which follows this editorial statement. Submitting authors to Language Learning are therefore required henceforth to provide a measure of effect size, at least for the major statistical contrasts which they report. [...] Always present effect sizes and their confidence intervals for primary outcomes. These effect sizes might be of various forms. If the units of measurement are meaningful on a practical level (e.g., reading rate, normed proficiency test scores), then unstandardized measures (regression coefficient or mean difference) are appropriate. If not, standardized differences (*d*) or uncorrected (e.g., *r*, *R-square*, *eta-square*) or corrected (e.g., adjusted *R-square*, *omega-square*) variance-accounted-for-statistics should be reported. These effect sizes are required in addition to the usual inferential statistical tests of significance, they do not replace them. It is also appropriate in the textual argument of the results section to place these effect sizes in their practical and theoretical context.

Ely, M. (1999) The importance of reporting estimates and confidence intervals for statistical analyses has been well publicised in the

arena of medical studies for some years now. the requirement to give confidence intervals for the main results of a study has been included in the statistical guidelines for contributors to medical journals since the 1980s and methodological points such as this are discussed in the Statistical Notes section of the ;British Medical Journal. If the use of quantitative methods in British sociology is to be encouraged, as Frank Bechhofer (1996) suggests is needed, it is important to have a forum for the dissemination of basic methodological issues which is accessible to researchers within the discipline. This note aims to achieve such dissemination by using an example from current research to illustrate this fundamental, but often overlooked, aspects of quantitative analysis.

Falk, R., & Greenbaum, C.W. (1995) We present a critique showing the flawed logical structure of statistical significance tests. We then attempt to analyze why, in spite of the faulty reasoning, the use of significance tests persists. We identify the illusion of probabilistic proof by contradiction as a central stumbling block, because it is based on a misleading generalization of reasoning from logic to inference under uncertainty. We present new data from a student samples and examples from the psychological literature showing the strength and prevalence of this illusion. We identify some intrinsic cognitive mechanisms (similarity to *modus tollens* reasoning; verbal ambiguity in describing the meaning of significance tests; and the need to rule out chance findings) and extrinsic social pressures which help to maintain the illusion. We conclude by mentioning some alternative methods for presenting and analyzing psychological data, none of which can be considered the ultimate method.

Fan X., Thompson B. (2001) Confidence intervals for reliability coefficients can be estimated in various ways. The present article illustrates a variety of these applications. This guidelines editorial also promulgates a request that *EPM* authors report confidence intervals for reliability estimates whenever they report score reliabilities and note what interval estimation methods they have used. This will reinforce reader understanding that all statistical estimates, including those for score reliability, are affected by sampling error variance. And these requirements may also facilitate understanding that tests are not impregnated with invariant reliability as a routine part of printing.

Fidler, F. (2002) The fifth edition of the *Publication Manual of the American Psychological Association* (APA) draws on recommendations for improving statistical practices made by the APA Task Force on Statistical Inference (TFSI). The manual now acknowledges the controversy over null hypothesis significance testing (NHST) and includes both a stronger recommendation to report effect sizes and a new recommendation to report confidence intervals. Drawing on interviews with some critics and other interested parties, the present review identifies a number of deficiencies in the new manual. These include lack of follow-through with appropriate explanations and examples of how to report statistics that are now recommended. At this stage, the discipline would be well served by a response to these criticisms and a debate over needed modifications.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C., & Schmitt, R. (2005) In 1997, Philip Kendall's editorial encouraged authors in JCCP to report effect sizes and clinical significance. The present authors assessed the influence of that editorial, and other APA initiatives to improve statistical practices, by examining 239 JCCP articles published from 1993 to 2001. For ANOVA, reporting of means and standardized effect sizes increased over that period, but the rate of effect size reporting for other types of analyses surveyed remained low. Confidence interval reporting increased little, reaching 17% in 2001. By 2001 the percentage of articles considering clinical (not only statistical) significance was 40%, compared with 36% in 1996. In a follow-up survey of JCCP authors (N=62), many expressed positive attitudes toward statistical reform, but gave little indication that they understood what was involved. Substantially improving statistical practices may require stricter editorial policies and further guidance for authors on reporting and interpreting measures.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004) Since the mid-1980s, confidence intervals (CIs) have been standard in medical journals. We sought lessons for psychology from medicine's experience of statistical reform by investigating two attempts by Kenneth Rothman to change statistical practices. We examined 594 American Journal of Public Health (AJPH) and 110 Epidemiology articles. Rothman's editorial instruction to report CIs and not p values was largely effective: in AJPH sole reliance on p values dropped from 63% to 5%, and CI reporting rose from 10% to 54%; Epidemiology showed even stronger compliance. However compliance was superficial: very few authors referred to CIs when discussing results. These results support what other research has indicated: editorial policy alone is not a sufficient statistical reform mechanism. Achieving substantial, desirable change will entail considerable guidance for full use of CIs and appropriate effect size measures. This will require study of researchers' understanding of CIs, improved education, and development of empirically-justified recommendations for improved statistical practice.

Fidler, F., & Thompson, B. (2001) Most textbooks explain how to compute confidence intervals for means, correlation coefficients, and other statistics using "central" test distributions (e.g., t, F that are appropriate for such statistics. However, few textbooks explain how to use "noncentral" test distributions (e.g., noncentral t, noncentral F to evaluate power or to compute confidence intervals for effect sizes. This article illustrates the computation of confidence intervals for effect sizes for some ANOVA applications; the use of intervals invoking noncentral distributions is made practical by newer software. Greater emphasis on both effect sizes and confidence intervals was recommended by the APA Task Force on Statistical Inference and is consistent with the editorial policies of the 17 journals that now explicitly require effect size reporting.

Finch, S., Cumming, G., & Thomason, N. (2001) Reformers have long argued that misuse of Null Hypothesis Significance Testing (NHST) is widespread and damaging. We analyzed 150 papers from the *Journal of Applied Psychology (JAP)* covering 1940 to 1999. We examined statistical reporting practices related to misconceptions about NHST, APA guidelines, and reform recommendations. Our analysis reveals (a) inconsistency in reporting alpha and p-values, (b) use of ambiguous language in describing NHST, (c) frequent acceptance of null hypotheses without consideration of power, (d) that power estimates are rarely reported, (e) virtually no confidence intervals. APA guidelines have been followed only selectively. Research methodology reported in *JAP* has increased greatly in sophistication over 60 years, but inference practices have shown remarkable stability. There is little sign that decades of cogent critiques by reformers had by 1999 led to changes in statistical reporting practices in *JAP*.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2004) Geoffrey Loftus, Editor of Memory & Cognition from 1994 to 1997, strongly encouraged presentation of figures with error bars and avoidance of null hypothesis significance testing (NHST). The authors examined 696 Memory & Cognition articles published before, during and after the Loftus editorship. Use of figures with bars increased to 47% under Loftus' editorship then declined. Bars were rarely used for interpretation, and NHST remained almost universal. Analysis of 309 articles in other psychology journals confirmed that Loftus' influence was most evident in the articles he published, but was otherwise limited. An email survey of authors published by Loftus revealed some support for his policy but also allegiance to traditional practices. Reform of psychologists' statistical practices would require more than editorial encouragement.

Finch, S., Thomason, N., & Cumming, G. (2002) We review the publication guidelines of the American Psychological Association (APA) since 1929 and document their advice for authors about statistical practice. Although the advice has been extended with each revision of the guidelines, it has largely focussed on Null Hypothesis Significance Testing (NHST) to the exclusion of other statistical methods. In parallel, we review over 40 years of critiques of NHST in psychology. The critiques have had little impact on the APA guidelines. The guidelines are influential in broadly shaping statistical practice, although in some cases they are not closely followed. They have an important role to play in reform of statistical practice in psychology. Following the report of the APA's Task Force on Statistical Inference, we propose that revisions of the guidelines reflect a broader philosophy of analysis and inference, provide detailed statistical requirements for reporting research, and directly address concerns about NHST. In addition the APA needs to develop ways to ensure that its editors succeed in their leadership role in achieving essential reform.

Fisher, R. A. (1959) "The subject of a probability statement if we know what we are talking about, is singular and unique."
"We must therefore specify that if a Bayesian probability *a priori* is available we shall usr the method of Bayes, and that the first condition for the applicability of the fiducial argument is that no probability *a priori* of the form needed for Bayes' theorem shall be available."
"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, R.A. (1962) Some further examples are given of Bayes' method of determining probabilities *a priori* by an experiment.

Fisher, R. A. (1990) This book brings together as a single volume three of Fisher's most influential textbooks: *Statistical Methods for Research Workers*, *The Design of Experiments*, and *Statistical Methods and Scientific Inference*. Whilst the text of each is unchanged save for the correction of minor misprints, in this new edition Dr Franck Yates has provided a foreword which sheds fresh light on Fisher's thinking and on the writing and reception of each of the books. Dr Yates discusses some of the key issues tackled in the three books and reflects on how the ideas expressed have come to permeate modern statistical practice.

Folger, R. (1989) Presents a logical justification for the following statements and discusses their implications: It is duplicitous (misleading) to use significance tests for making binary (either/or) decisions regarding the validity of a theory; the binary choice between calling results significant or not significant should not govern the confidence placed in a theory, because such confidence cannot be gained in the either/or fashion characterizing deductive certainty. The implications include grounds for describing ways that effect size estimates become useful in making judgments about the value of theories.

Freedman, L.S., Spiegelhalter, D.J., & Parmar, M.K.B. (1994) We discuss the advantages and limitations of group sequential methods for monitoring clinical trials data. We describe a Bayesian approach, based upon the use of sceptical prior distributions, that avoids some of the limitations of group sequential methods. We illustrate its use with data from a trial of levamisole plus 5-Fluorouracil for colorectoral cancer.

Freeman, P.R. (1993) The current widespread practice of using *p*-values as the main means of assessing and reporting the results of clinical trials cannot be defended. Reasons for grave concern over the present situation range from the unsatisfactory nature of *p*-values themselves, their very common misunderstanding by statisticians as well as by clinicians and their serious distorting influence on our perception of the very nature of clinical trials. It is argued, however, that only fully understanding the reasons why they have become so universally popular can we hope to change opinion and introduce more sensible ways of summarizing and reporting results. Some of the ways in which this might happen are discussed.

Freiman, J.A., Chalmers, T.C., Smith, H., & Kueber, R.R. (1978) Seventy-one "negative" randomized control trials were re-examined to determine if the investigators had studied large enough samples to give a high probability (>0.90) of detecting a 25 per cent and 50 per cent therapeutic improvement in the response. Sixty-seven of the trials had a greater than 10 per cent risk of missing a true 25 per cent therapeutic improvement, and with the same risk, 50 of the trials could have missed a 50 per cent improvement. Estimates of 90 per cent confidence intervals for the true improvement in each trial showed that in 57 per cent these "negative" trials, a potential 25 per cent improvement was possible, and 34 of the trials showed a potential 50 per cent improvement. Many of the therapies labeled as "no different from control" in trials using inadequate samples have not received a fair test. Concern for the probability of missing an important therapeutic improvement because of small sample sizes deserves more attention in the planning of clinical trials.

Frías, Ma.D., Pascual, J., & Garcia, J.F. (2000) Currently, there is a growing interest in the study of the sensitive and validity of the statistical conclusions of experimental design. Although most of books on experimental design stress these issues, many students on applied psychology still do not take advantage of these advances, as can be deduced by low statistical power. The goal of this article is to examine the impact of the guidelines of the editorial Board of peer reviewed respect to the computation and interpretation of the measures of effect size as well as the values of statistical significance.

Frick, R.W. (1995) This article concerns acceptance of the null hypothesis that one variable has no effect on another. Despite frequent opinions to the contrary, this null hypothesis can be correct in some situations. Appropriate criteria for accepting the null hypothesis are (1) that the null hypothesis is possible; (2) that the null hypothesis is possible; and (3) that the experiment was a good effort to find an effect. These criteria are consistent with the meta-rules for psychology. The good-effort criterion is subjective, which is somewhat undesirable, but the alternative – never accepting the null hypothesis – is neither desirable nor practicable.

Frick, R.W. (1996) The many criticisms on null hypothesis testing suggest when it is not useful and what is should not be used for. This article explores when and why its use is appropriate. Null hypothesis testing is insufficient when size of effects is important, but it is ideal for testing ordinal claims relating the order of conditions, which are common in psychology. Null hypothesis testing also is insufficient for determining beliefs, but it is ideal for demonstrating sufficient evidential strength to support an ordinal claim, with sufficient evidence being 1 criterion for a finding entering the corpus of legitimate findings in psychology. The line between sufficient and insufficient evidence is currently set at $p<.05$; there is little reason for allowing experimenters to select their own value of alpha. Thus null hypothesis testing is an optimal method for demonstrating sufficient evidence for an ordinal claim.

Fry, T.C. (1965) Contains discussions of Byesian point of view.

Gendreau, P. (2002) It is argued that our attempts at knowledge cumulation have been flawed in four ways. They are the eroding of "empiricism" in clinical practice, the tendency towards paradigm passion and ethnocentricism, the failure to attend to "simple" measures of effect size, and the misuse of significance testing. It is recommended that speciality designations, the replacement of significance testing with point estimates and confidence intervals, the use of practical effect size statistics, the establishment of data repositories, and a renewed focus on replication would help resolve some of these problems.

Gigerenzer, G. (1998) What Chow [Chow, 1996] calls NHSTP is an inconsistent hybrid of Fisherian and Neyman-Pearsonian ideas. In psychology, it has been practiced like ritualistic handwashing and sustained by wishful thinking about its utility. Chow argues that NHSTP is an important tool for ruling out chances as an explanation for data. I disagree. This ritual discourages theory development by providing researchers with no incentive to specify hypotheses.

Glaser, D.N. (1976) Despite the rumblings and ominous overtones of the proposed banning of the significance test, a more temperate solution has been offered by a wide array of researchers. Significance testing will always have its advocates and opponents. At this time, however, more than any other, researchers are considering the import of instituting effect sizes, confidence intervals, and power analysis alongside the traditional mode of significance testing. The recommendation is not that clinical researchers disavow significance testing, but rather that they incorporate additional information that will supplement their findings.

Gold, D. (1958) Since, at a given level of significance, statistical significance demands a greater *degree* of relationship from a small

sample than a large sample, it might appear that the researcher can more easily treat substantively important differences by selecting small samples rather than large samples. This, of course, is not true. It is simply that 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, D. (1969) It has been contended that a test of significance can be viewed as an indication of the probability that an observed association could be generated in a given set of data by a random process model, without respect o sampling considerations. Statistical significance, in these terms, provides an explicit criterion for attributing substantive importance to the observed association. However, statistical significance is only the minimal criterion, necessary but not sufficient. In addition, the analyst must attend to the size of the association and must also make this criterion for the acceptance of the importance of the association reasonably clear. Some rules of thumb along these lines, especially useful in assessing associations among mixed variables (qualitative and quantitative), have been suggested as illustrative of a general approach to be taken.

Good, I.J. (1984) See Neyman, Scott and Smith, 1969.

Goodman, S.N., & Berlin, J.A. (1994) Although there is a growing understanding of the importance of statistical power considerations when designing studies and of the value of confidence intervals when interpreting data, confusion exists about the reverse arrangement: the role of confidence intervals in study design and of power in interpretation. Confidence intervals should play an important role when setting sample size, and power should play no role once the data have been collected, but exactly the opposite procedure is widely practiced. 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, S.N. (1999) An important problem exists in the interpretation of modern medical research data: Biological understanding and previous research play little formal role in the interpretation of quantitative results. This phenomenon is manifest in the discussion sections of research articles and ultimately can affect the reliability of conclusions. The standard statistical approach has created this situation by promoting the illusion that conclusions can be produced with certain "error rates," without consideration of information from outside the experiment. This statistical approach, the key components of which are *P* values and hypothesis tests, is widely perceived as a mathematically coherent approach to inference. There is little appreciation in the medical community that the methodology is an amalgam of incompatible elements, whose utility for scientific inference has been the subject of intense debate among statisticians for almost 70 years. This article introduces some of the key elements of that debate and traces the appeal and adverse impact of this methodology to the *P* value fallacy, the mistaken idea that a single number can capture both the long-run outcomes of an experiment and the evidential meaning of a single result. This argument is made as a prelude to the suggestion that another measure of evidence should be used-the Bayes factor, which properly separates issues of long-run behavior from evidential strength and allows the integration of background knowledge with statistical findings.

Goodman, S.N. (1999) Bayesian inference is usually presented as a method for determining how scientific belief should be modified by data. Although Bayesian methodology has been one of the most active areas of statistical development in the past 20 years, medical researchers have been reluctant to embrace what they perceive as a subjective approach to data analysis. It is little understood that Bayesian methods have a data-based core, which can be used as a calculus of evidence. This core is the Bayes factor, which in its simplest form is also called a *likelihood ratio.* The minimum Bayes factor is objective and can be used in lieu of the *p* value as a measure of the evidential strength. Unlike *p* values, Bayes factors have a sound theoretical foundation and an interpretation that allows their use in both inference and decision making. Bayes factors show that *p* values great I y overstate the evidence against the null hypothesis. Most important, Bayes factors require the addition of background knowledge to be transformed into inferences-probabilities that a given conclusion is right or wrong. They make the distinction clear between experimental evidence and inferential conclusions while providing a framework in which to combine prior with current evidence.

Graham, J.M. (2001) The present book review of Statistics With Confidence is framed in terms of both the recent report of the APA Task Force on Statistical Inference and ongoing movements in the field. The review is structured in terms of two major issues: the interpretation of confidence intervals (null hypothesis significance testing [NHST] versus non-NHST) and the ethics of statistics.

Granaas, M. (2002) For many years null hypothesis testing (NHT) has been the dominant form of statistical analysis.

Gray, M.W. (1983) **[*Example of misinterpretation of a p-value*]** "For the 2×>85 table linking departments and admission rates, *chi-square* = 2121, and the probability that the admission rate is the same in all departments is approximately zero." (page 77) [Quoted by Falk and Greenbaum, 1995, page 83]

Greenwald, A.G., Gonzalez, R., Harris, R.J., & Guthrie, D. (1996) Despite publications of many well-argued critiques of null hypothesis testing (NHT), behavioral science researchers continue to rely heavily on this set of practices. Although we agree with most critics' catalogs of NHT flaws, this article also takes the unusual stance of identifying virtues that may explain why NHT continues to be so extensively used. These virtues include providing results in the form of a dichotomous (yes/no) hypothesis evaluation and providing an index (*p*-value) that has a justifiable mapping onto confidence in repeatability of a null hypothesis rejection. The most criticized flaws of NHT can be avoided when the importance of a hypothesis, rather that the *p* value of its test, is used to determine that a finding is worth of report, and when $p>.05$ is treated as insufficient basis for confidence in the replicability of an isolated non-null finding. 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.

Gregson, R.A.M. (1998) Chow's [Chow, 1996] defense of NHSTP [Null Hypothesis Significance Test Procedure] ignores the fact that in psychology it is used to test substantive hypotheses in theory-corroborating research. In this role, NHSTP is not only inadequate, but damaging to the progress of psychology as a science. NHSTP does not fulfill the Popperian requirement that theories be tested severely. It also encourages nonspecific predictions and feeble theoretical formulations.

Grissom, R.J., & Kim, J.J. (2001) Estimation of the effect size parameter, D, the standardized difference between population means, is sensitive to heterogeneity of variance (heteroseedasticity), which seems to abound in psychological data. Pooling s(2)s assumes homoscedasticity, as do methods for constructing a confidence interval for D, estimating D from t or analysis of variance results, formulas that adjust estimates for inflation by main effects or covariates, and the Q statistic. The common language effect size statistic as an estimate of Pr(X1>X2), the probability that a randomly sampled member of Population I will outscore a randomly sampled member of Population 2, also assumes normality and homoscedasticity. Various proposed solutions are reviewed, including measures that do not make these assumptions, such as the probability of superiority estimate of Pr(X1>X2). Ways to reconceptualize effect size when treatments may affect moments such as the variance are also discussed.

Grunkemeier, G.L. & Payne, N. (2002) Full Bayesian analysis is an alternative statistical paradigm, as opposed to traditionally used methods, usually called frequentist statistics. Bayesian analysis is controversial because it requires assuming a prior distribution, which can be arbitrarily chosen; thus there is a subjective element, which is considered to be a major weakness. However, this could also be considered a strength since it provides a formal way of incorporating prior knowledge. Since it is flexible and permits repeated looks at evolving data, Bayesian analysis is particularly well suited to the evaluation of new medical technology. Bayesian analysis can refer to a range of things: from a simple, noncontroversial formula for inverting probabilities to an alternative approach to the philosophy of science. Its advantages include: (1) providing direct probability statements — which are what most people wrongly assume they are getting from conventional statistics; (2) formally incorporating previous information in statistical inference of a data set, a natural approach which we follow in everyday reasoning; and (3) flexible, adaptive research designs allowing multiple looks at accumulating study data. Its primary disadvantage is the element of subjectivity which some think is not scientific. We discuss and compare frequentist and Bayesian approaches and provide three examples of Bayesian analysis: (1) EKG interpretation, (2) a coin-tossing experiment, and (3) assessing the thromboembolic risk of a new mechanical heart valve.

Hager, W. (2000) The function and potential importance of statistical tests in examining and evaluating substantive and psychological hypotheses is discussed. Psychological hypotheses are sharply distinguished from statistical hypotheses. Decisions on statistical hypotheses must be separated from decisions on psychological hypotheses. Some differences between both kinds of hypotheses are addressed, and the question is attacked whether they are complementary or not. The answer to this question is negative. The use of the modus tollens in theory corroboration is discussed and it is argued that evaluations of substantive hypotheses always is accompanied by some inductive aspects. A further reason for the discontent with statistical tests is identified: most of these tests are rather insensitive the differential patterns of predictions and of data, whereas differential patterns of data can be derived from nearly every psychological hypothesis or theory. To test for these differential patterns statistical tests should be applied thoughtfully, not routinely.

Hagood, M.J. (1941) For developments in statistical theory of the last decade or two have shown the tests formerly used to be incorrect, and those who are using as guides texts published 10 years or more ago are likely to be using unacceptable tests of significance for their correlation coefficients. The most common test of significance answers for the universe the question as to whether or not association *exists* in the universe – that is, it investigates for the universe the first aspect of association.

Hancock, G.R., & Freeman M.J. (2001) Targeted toward the applied modeler, this article provides select power and sample size tables and interpolation strategies associated with the root mean square error of approximation test of not close fit under standard assumed conditions. It is hoped that researchers conducting structural equation modeling will be better informed as to power limitations when testing a model given a particular available sample size or, better yet, that they will heed the sample size recommendations contained herein when planning their study to ensure the most accurate assessment of the degree of close fit between data and model.

Hardy, A., Harvie, P., & Koestler, A. (1973)] **[*Examples of misinterpretation of a p-value*]** "Taken altogether, the receivers scored significantly beyond chance [...] with a calculated probability of 3,000 to 1 against it being just chance." (page 117) "The lady passed the ordeal with flying colors: she correctly identified the method of pouring for all eight cups, with odds against chance of one in seventy" (page 236) [Quoted by Falk and Greenbaum, 1995, page 82]

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.) (1997) This book is the result of a spirited debate stimulated by a recent meeting of the Society of Multivariate Experimental Psychology. Although the viewpoints span a range of perspectives, the overriding theme that emerges states that significance testing may still be useful if supplemented with some or all of the following -- Bayesian logic, caution, confidence intervals, effect sizes and power, other goodness of approximation measures, replication and meta-analysis, sound reasoning, and theory appraisal and corroboration. The book is organized into five general areas. The first presents an overview of significance testing issues that synthesizes the highlights of the remainder of the book. The next discusses the debate in which significance testing should be rejected or retained. The third outlines various methods that may supplement current significance testing procedures. The fourth discusses Bayesian approaches and methods and the use of confidence intervals versus significance tests. The last presents the philosophy of science perspectives. Rather than providing definitive prescriptions, the chapters are largely suggestive of general issues, concerns, and application guidelines. The editors allows readers to choose the best way to conduct hypothesis testing in their respective fields.
***Contents*:** Preface -- Part I, Overview; Harlow, L.L., Significance testing introduction and overview -- Part II, The debate: against and for significance testing; Cohen, J., The earth is round (p<.05); Schmidt, F.L., & Hunter, J., Eight common but false objections to the discontinuation of significance testing in the analysis of research data; Mulaik, S.A., Raju, N.S., & Harshman, R., There is a time and place for significance testing; Abelson, R.P., A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented) -- Part III, Suggested alternatives to significance testing; Harris, R.J., Reforming significance testing via three-valued logic; Rossi, J.S., A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning; McDonald, R.P., Goodness of approximation in the linear model; Steiger, J.H., & Fouladi, R.T., Noncentrality interval estimation and the evaluation of statistical models; Reichardt, C.S., & Gollob, H.F., When confidence intervals should be used instead of statistical significance tests, and vice versa -- Part IV, A Bayesian approach to hypothesis testing; Pruzek, R.M., An introduction to Bayesian inference and its application; Rindskopf, D., Testing "small", not null, hypotheses: Classical and Bayesian approaches -- Part V, Philosophy of science issues; Rozeboom, W.W., Good science is abductive, not hypothetico-deductive; Meehl, P.E., The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions.

Harris, M.J. (1991) Chow (1991) distinguishes between "practical impact" and "conceptual rigor" research, and he concludes that effect size estimation is useful only in practical impact research. I argue that significance tests do not answer substantive questions about the data and are useful only as a check that the results are unlikely to have occurred by chance. Chow's decision to regard the similarity between data and prediction as being a dichotomous judgment made on the basis of significance testing is therefore unwise. I conclude that effect sizes are the single best index of the relationship between theoretical predictions and the obtained data. The role of replications and meta-analysis in advancing theory is also discussed.

Harris, R.J. (1997) The many and frequent misinterpretations of null hypothesis significance testing (NHST) are fostered by continuing to present NHST logic as a choice between only two hypotheses. Moreover, the proposed alternatives to NHST are just as susceptible to misinterpretation as is (two-valued) NHST: Misinterpretations could, however; be great I y reduced by adopting Kaisers (1960) proposal of three-alternative hypothesis testing in place of the traditional two-alternative presentation. The real purpose of significance testing (NHST) is to establish whether we have enough evidence to be confident of the sign (direction) of the effect we 're testing in the population. This is an important contribution to the cumulation of scientific knowledge and should be retained in any replacement system. Confidence intervals (CIs - the proposed alternative to significance tests in single studies) can provide this control, but when so used they are subject to exactly the same Type 1, Type II, and Type III (statistical significance in the wrong direction) error rates as significance testing. There are still areas of research where NHST alone would be a considerable improvement over the current lack of awareness of error variance. Further; there are two pieces of information (namely, maximum probability of a Type III error and probability of a successful exact replication) provided by p values that are not easily gleaned from confidence intervals. Suggestions are offered for greatly increasing attention to power considerations and for eliminating the positive bias in estimates of effect-size magnitudes induced when we make statistical significance a necessary condition of publication.

Hauschke, D., & Steinijans, V.W. (1996) The purpose of this communication is to point out that it is generally not correct to use the conventional approach for testing therapeutic equivalence of two treatments.

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

Henri, V. (1898)] **[*Exemples de formulations fallacieuses à propos des seuils de signification*]** "Nous affirmons avec une probabilité égale à 0.95 que la différence n'est pas due au hasard." "Nous dirons que cette différence est produite par le hasard, et notre affirmation a

une probabilité d'exactitude supérieure à 98%." [citées par Rouanet & Bru, 1994]

Hirsch, L.S., & O'Donnell, A.M. (2001) **[*Example of interpretation of a nonsignificant result as a proof of the null hypothesis*]** "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)." (page 10)

Hodges, J.L., & Lehmann, E.L. (1954) The distinction between statistical significance and material significance in hypotheses testing is discussed. Modifications of the customary tests, in order to test for the absence of material significance, are derived for several parametric problems, for the chi-square test of goodness of fit, and for Student's hypothesis. The latter permits one to test the hypothesis that the means of two normal populations of equal variance, do not differ by more than a stated amount.

Hogben, L.T. (1957) ***The contemporary crisis of the uncertainty of uncertain inferences*** – We may have to reinstate statistics as continental demographers use the term. Laboratory experiments will have to stand on their own without protection from a façade of irrelevant computations. Sociologists will have to use their brains. In my view, science will not suffer.
Statistical prudence and statistical inference – At the core of contemporary controversy in statistical theory is the following question: What bearing, if any, has the rarity of an observable occurrence as prescribed by an appropriate stochastic hypothesis on our legitimate grounds for accepting or rejecting the latter when we have already witnessed the former? The form of the answer we deem to be appropriate will define what we here conceive to be the proper terms of reference of a Calculus of Judgments, i.e. *statistical inference* as some contemporary writers use the term. Such is the theme of this chapter.
Significance as interpreted by the school of R.A. Fisher – Of three widely divergent views [Bayes, Neyman-Pearson-Wald, and Fisher] about the nature of statistical inference, two have hitherto attracted little attention except among professional mathematicians, and have had few protagonists among practical statisticians [...]. Contrariwise, the overwhelming majority of research workers in the biological field (including medicine and agriculture), as also a growing body of workers in the social sciences, rely largely on rule of thumb procedures set forth in a succession of manuals modeled on *Statistical Methods for Research Workers* by R.A. Fisher.

Howard, G.S., Maxwell, S.E., & Fleming, K.J. (2000) Some methodologists have recently suggested that scientific psychology's overreliance on null hypothesis significance testing (NHST) impedes the progress of the discipline. In response, a number of defenders have maintained that NHST continues to play a vital role in psychological research. Both sides of the argument to date have been presented abstractly. The authors take a different approach to this issue by illustrating the use of NHST along with 2 possible alternatives (meta-analysis as a primary data analysis strategy and Bayesian approaches) in a series of 3 studies. Comparing and contrasting the approaches on actual data brings out the strengths and weaknesses of each approach. The exercise demonstrates that the approaches are not mutually exclusive but instead can be used to complement one another.

Hresko, W. (2000) The APA *Publication Manual* cites the need for including effect-size information in manuscripts utilizing quantitative data analysis techniques [...] If authors do not include this information in submitted manuscripts (and the manuscript is based on a quantitative research design), the author(s) will be asked to provide this information should the manuscript be recommended for publication or revision and publication.

International Committee of Medical Journal Editors (1991) Updated author guidelines from the International Committee of Medical Journal Editors including a section on writing statistics.

Iversen, G.R. (1998) The second half of this century has witnessed a very fruitful debate among statisticians about the relative merits of Bayesian and classical statistical inference. Neither side can claim a victory in this debate, since there is no way of proving that one approach is more correct than the other. But the debate has served the purpose of illuminating the strengths and weakness of each approach. The students we teach are to a very large extent exposed only to classical statistical inference. This is a choice made by their instructors, meaning all of us. This spring, in a small group of students studying both approaches I probed for their opinions of the two approaches. Not surprisingly, the Bayesian approach was well received by the students, even though they also had some misgivings about Bayesian statistics, at least early on>.

Iversen, G.R. (2000) Statistical methods have an impact on the results of any statistical study. We do not always realize that the statistical methods act in such a way as to create a construction of the world. We should therefore be more aware of the role of statistics in research, and the question is not so much about what we teach researchers but that we train them to be aware of the impact of the methods they use. This becomes particularly important in statistical inference where we have the choice between the classical, frequentist approach and the Bayesian approach. The two approaches create very different views of the world. The word probability carries with it a notion of uncertainty, and it is tempting to think that the uncertainty refers to parameters and not simply data.

Jaynes, E.T. (2003) The following material is addressed to readers who are already familiar with applied mathematics at the advanced undergraduate level or preferably higher; and with some field, such as physics, chemistry, biology, geology, medicine, economics, sociology, engineering, operations research, etc., where inference is needed. A previous acquaintance with probability and statistics is not necessary; indeed, a certain amount of innocence in this area may be desirable, because there will be less to unlearn.
We are concerned with probability theory and all of its conventional mathematics, but now viewed in a wider context than that of the standard textbooks. Every Chapter after the first has "new" i.e. not previously published) results that we think will be found interesting and useful. Many of our applications lie outside the scope of conventional probability theory as currently taught. But we think that the results will speak for themselves, and that something like the theory expounded here will become the conventional probability theory of the future.

Jefferys, H. (1990) Data from experiments that use random event generators are usually analyzed by classical (frequentist) statistical tests, which summarize the statistical significance of the test statistic as a *p*-value. However, classical statistical tests are frequently inappropriate to these data, and the resulting *p*-values can grossly overestimate the significance of the result. Bayesian analysis shows that a small *p*-value may not provide credible evidence that an anomalous phenomenon exists. An easily applied alternative methodology is described and applied to an example from the literature.

Jefferys, H. (1992) Dobyns' article [*Journal of Scientific Exploration, 6,* no 1] suggests some reasons why orthodox statistics might be superior to Bayesian statistics when discussing random event generator statistics. Several of his main arguments are examined and discussed.

Jefferys, H. (1995) In a recent column in this journal, Cooper (1994) stated that "The *p*-value is the probability that the results could have occurred by pure chance given that the null (conventional) hypothesis is true." This definition is incorrect and highly misleading, although similar statements are often found in the literature… A correct definition of the *p*-value is that it is the probability of obtaining the actual result we did, *or any more extreme result*, given that the null (conventional) hypothesis is true.

Jefferys, H. (1995) In response to my letter (Jefferys 1995), Dobyns and Jahn (1995) responded that my objection to their incorrect definition of *p*-values is "trivial" and mere "pedantic quibbling." It is easy to convince oneself that this is not the case.

Johns, D., & Andersen, J.S. (1990) Predictive probability is particularly useful in aiding a decision-making process related to drug development. This is especially true for decisions occurring as the result of interim analysis of clinical trials. Examples of clinical trial applications of Bayesian predictive probability and the use of the beta-binomial distribution are described.

Johnson, D.H. (1998) Wildlife biologists recently have been subjected to the credo that if you're not testing hypotheses, you're not doing real science. To protect themselves against rejection by journal editors, authors cloak their findings in an armor of *P* values. I contend that much statistical hypothesis testing is misguided. Virtually all null hypotheses tested are, in fact, false; the only issue is whether or not the sample size is sufficiently large to show it. No matter if it is or not, one then gets led into the quagmire of deciding biological significance versus statistical significance. Most often, parameter estimation is a more appropriate tool than statistical hypothesis testing. Statistical hypothesis testing should be distinguished from scientific hypothesis testing, in which truly viable alternative hypotheses are evaluated in a real attempt to falsify them. The latter method is part of the deductive logic of strong inference, which is better-suited to simple systems. Ecological systems are complex, with components typically influenced by many factors, whose influences often vary in place and time. Competing hypotheses in ecology rarely can be falsified and eliminated. Wildlife biologists perhaps adopt hypothesis tests in order to make what are really descriptive studies appear as scientific as those in the "hard" sciences. Rather than attempting to falsify hypotheses, it may be more productive to understand the relative importance of multiple factors.

Johnson, D.H. (1999) 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. This paper describes how statistical hypothesis tests are often viewed, and then contrasts that interpretation with the correct one. I discuss the arbitrariness of P-values, conclusions that the null hypothesis is true, power analysis, and distinctions between statistical and biological significance. Statistical hypothesis testing, in which the null hypothesis about the properties of a population is almost always known a priori to be false, is contrasted with scientific hypothesis testing, which examines a credible null hypothesis about phenomena in nature. More meaningful alternatives are briefly outlined, including estimation and confidence intervals for determining the importance of factors, decision theory for guiding actions in the face of uncertainty, and Bayesian approaches to hypothesis testing and other statistical practices.

Johnstone, D.J. (1988) [...] Oakes [Oakes, 1986] misrepresents Fisher's position on points of logic. There is also some overstatement of the case for confidence intervals. More interesting is the author's positive explanation for the widespread acceptance of significance tests

among applied researchers, for there is no less settled logic or scheme of inference within theoretical statistics, as instantiated by the current papers of Casella and Berger (1987) and Berger and Sellke (1987) in the *Journal of the American Statistical Association*. That research workers in applied fields continue to use significance tests routinely may be explained by forces of supply and demand in the market for statistical evidence where the commodity traded is not so much evidence, but "statistical significance".

Johnstone, D.J., & Lindley, D.V. (1995) In empirical research in the social sciences expressions of statistical significance are meant to capture and summarise the evidence implied by data. To evaluate the evidential content of statements such as "the difference between means is significant at alpha = 5%", we consider the Bayesian probability of the hypotheses tested, where the conditioning event is an announcement of general form significant at alpha. By proceeding as if neither observed effects nor their exact P-values are reported, the meaning of such descriptions of themselves is revealed. It is demonstrated, for large samples particularly, that a report merely that data are significant at alpha has no objective meaning, and under some conditions should be interpreted not as evidence against the null hypothesis, as is usually supposed, but as strong evidence in its favor. This conclusion is supported by both algebraic arguments and example calculations for the special, but important case of the normal mean. It is also found that significance at one level tends to imply significance at much lower levels, the more strongly the larger the sample.

Jones, L.V., & Tukey, J.W. (2000) The conventional procedure for null hypothesis significance testing has long been the target of appropriate criticism. A more reasonable alternative is proposed, one that not only avoids the unrealistic postulation of a null hypothesis but also, for a given parametric difference and a given error probability, is more likely to report the detection of that difference.

Kadane, J.B. (1995) This paper reviews that Bayesian statistics is and gives pointers to the literature. The need for a subjectively determined prior distribution, likelihood, and loss function is often taken to be the principal disadvantage of Bayesian statistics. This paper argues that the requirement that these be explicitly stated is a distinct Bayesian advantage. Advances in Bayesian technology make it ready now to be the main inferential tool for clinical trials.

Kahneman, D., & Tversky, A. (1972) This paper explores a heuristic – *representativeness* – according to which the subjective probability of an event, or a sample, is determined by the degree to which it: (i) is similar in essential characteristics to its parent population; and (ii) reflects the salient features of the process by which it is generated. This heuristic is explicated in a series of empirical examples demonstrating predictable and systematic errors in the evaluation of uncertain events. In particular, since sample sizes does not represent any property of the population, it is expected to have little or no effect on judgment of likelihood. This prediction is confirmed in studies showing that subjective sampling distributions and posterior probability judgments are determined by the most salient characteristic of the sample (e.g., proportion, mean) without regard to the size of the sample. The present heuristic approach is contrasted with the normative (Bayesian) approach to the analysis of the judgment of uncertainty.

Kendall, P. (1957) 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, P.C. (1997) Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the required effect size but also a consideration of clinical significance. (page> 3).

Kieffer, K.M., Reese, R.J., & Thompson, B. (2000) The authors of the present methodological review investigated the patterns of statistical usage and reporting practices in 756 articles published in the American Educational Research Journal (AERJ) and in the Journal of Counseling Psychology (JCP) over a 10-year period. First, some findings from other similar reviews are summarized. Second, the authors present a framework for characterizing selected research practices that emphasizes, in part, elements of the recent report of the American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson and APA Task Force on Statistical Inference, 1999). Third, characterizations of 10 years of analytic practices in 2 journals are presented >and evaluated within that framework. The article concludes with a discussion of the changes that may be necessary to improve the statistical state of affairs in behavioral research.

Kirk, R.E. (1995) **[*Example of interpretation of frequentist confidence intervals in terms of probabilities about parameters*]** "We can be 95% confident that the population mean is between 110.06 and 119.94." (page 43)". [*Ambiguous formulation*] "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." (page 42).

Kirk, R.E. (1996) Statistical significance is concerned with whether a research result is due to chance or sampling variability; practical significance is concerned with whether the result is useful in the real world. A growing awareness of the limitations of null hypothesis significance tests has led to a search for ways to supplement these procedures. A variety of supplementary measures of effect magnitude

have been proposed. The use of these procedures in four APA journals is examined, and an approach to assessing the practical significance of data is described.

Kish, L. (1959) I intend to touch on several problems dealing with the interplay of statistics with the more general problems of scientific inference. [...] The aim of this paper is not a profound analysis, but a clear elementary treatment of several related problems. The literature references contain more thorough treatments. Moreover, these are not *all* problems in this area, nor even necessarily the most important ones; the reader may find that his favorite, his most problem, has been omitted. The problems selected are a group with a common core, they arise frequently, yet they are widely misunderstood [Statistical tests of survey data; Experiments, survey, and other investigations; Some misuses of statistical tests].

Kotrlick, J.W. (2000) Authors should report effect sizes in the manuscript and tables when reporting statistical significance.

Krantz, D.H. (1999) A controversy concerning the usefulness of "null" hypothesis tests in scientific inference has continued in articles within psychology since1960 and has recently come to a head, with serious proposals offered for a test ban or something close to it. This article sketches some of the views of statistical theory and practice among different groups of psychologists, reviews a recent book offering multiple perspectives on null hypothesis tests, and argues that the debate within psychology is a symptom of serious incompleteness in the foundations of statistics.

Krueger, J. (1998) Can research on social-perceptual biases benefit from improved and diversified statistical methods? Having reached the brink of nihilism, I conclude that (a) any point-hypothesis can be rejected by null hypothesis significance testing (NHST), (b) any such hypothesis can be accepted by Bayesian inference, (c) effect size estimates are meaningful only if that meaning is imported from extra-statistical considerations, and (d) taxonomies of biases and their causes will be messy because most biases are overdetermined.

Krueger, J. (1998) Social psychology has painted a picture of human misbehavior and irrational thinking. For example, prominent social cognitive biases are said to distort consensus estimation, self perception, and causal attribution. The thesis of this target article is that the roots of this negativistic paradigm lie in the joint application of narrow normative theories and statistical testing methods designed to reject those theories. Suggestions for balancing the prevalent paradigm include (a) modifications to the ruling rituals of Null Hypothesis Significance Testing, (b) revisions of what is considered a normative response, and (c) increased emphasis on individual differences in judgment.

Krueger, J., & Funder, D.C. (2001) Mainstream social psychology focuses on how people characteristically violate norms of action through social misbehaviors such as conformity with false majority judgments, destructive obedience, and failures to help those in need. Likewise, they are seen to violate norms of reasoning through cognitive errors such as misuse of social information, self-enhancement, and an over-readiness to attribute dispositional characteristics. The causes of this negative research emphasis include the apparent informativeness of norm violation, the status of good behavior and judgment as unconfirmable null hypotheses, and the allure of counter-intuitive findings. The shortcomings of this orientation include frequently erroneous imputations of error, findings of mutually contradictory errors, incoherent interpretations of error, an inability to explain the sources of behavioral or cognitive achievement, and the inhibition of generalized theory. Possible remedies include increased attention to the complete range of behavior and judgmental accomplishment, analytic reforms emphasizing effect sizes and Bayesian inference, and a theoretical paradigm able to account for both the sources of accomplishment and of error. A more balanced social psychology would yield not only a more positive view of human nature, but also an improved understanding of the bases of good behavior and accurate judgment, coherent explanations of occasional lapses, and theoretically grounded suggestions for improvement.

Lecoutre, B. (1981) Cet article montre, à partir d'exemples concrets, comment les *procédures fiducio-bayésiennes* permettent l'investigation des mécanismes individuels, en fournissant des résultats inférentiels, non seulement sur l'*effet moyen*, mais aussi sur les *effets individuels*. Techniquement, ces procédures sont développées pour l'effet associé à un contraste et pour l'effet associé à une comparaison (à un nombre quelconque de degrés de liberté) dans un plan du type S*T (Sujets*Traitements).
Bayes-fiducial procedures for investigating individual mechanisms in Psychology
Illustrates, with concrete examples, how *Bayes-fiducial procedures* allow the investigation of individual mechanisms by yielding inferential results, not only about the *mean effect*, but also about *individual effects*. Technically, these procedures have been developed for the effect associated with a contrast and for the effect associated with a comparison (with any number of freedom) in a S*T (Subjects*Treatments) design.

Lecoutre, B. (1984) Cet ouvrage se situe dans un courant de recherche, né en France dans les années 1970, à partir des travaux de H. Rouanet et D. Lépine, qui consiste à refondre, à partir d'une formalisation algébrique, les méthodes traditionnelles s'analyse statistique des données expérimentales. Les données des chercheurs sont en règle générale des données *structurées*; la *formalisation* des structures, étroitement liée au plan de recueil des données (plan d'expérience ou plan d'enquête) fournit un cadre aux questions que le chercheur se

pose à propos de ses données. Le problème de la *généralisabilité* des conclusions est incontournable; l'idée d'*inférence spécifique* permet, à l'intérieur de chaque situation, d'appliquer des procédures inférentielles adaptées pour les structures qui interviennent dans cette situation. Il s'agit de fournir des procédures, répondant aux objectifs réels de l'induction; les *procédures bayésiennes*, envisagées comme un prolongement des *tests de signification* usuels, permettent notamment de se prononcer sur l'importance de chaque effet examiné, et non seulement sur son existence; en particulier les *procédures fiducio-bayésiennes* expriment, pour chaque question posée par le chercheur, "ce que les données ont à dire", indépendamment de toute information extérieure. Il en résulte une construction nouvelle, de plus en plus autonome par rapport aux développements traditionnels de l'analyse de la variance à l'anglo-saxonne: l'*Analyse Bayésienne des Comparaisons*, parce que la notion formalisée de comparaison y joue un rôle central.

Lecoutre, B. (1985) It is shown, in the case of the inference on a contrast between means, how Bayes-fiducial analyses can be carried out, given only the observed effect and the significance level; Bayes-fiducial limits can be obtained immediately by mean of tables, in order to establish whether an effect is negligible or notable. The role of significance testing in experimental methodology is thus discussed as far as the generalizibility of descriptive conclusions about the magnitude of effects is concerned.

Lecoutre, B. (1985) The usual F-test of the analysis of variance is reconsidered within the Bayesian framework, in terms of predictive distributions. This leads to the notion of semi-Bayesian significance test, so called because it consists in only probabilizing the space of nuisance parameters, thus bringing a general principle for "eliminating" nuisance parameters, or more exactly incorporating information about these parameters. The approach is shown to extend the F-tests, by allowing the testing of hypotheses of non-zero effects.

Lecoutre, B. (1994) On examine l'utilisation et les apports de l'inférence statistique dans l'étude des raisonnements inductifs. On montre que certains aspects de ne sont pas toujours clairement pris en compte. En particulier on a souvent utilisé une approche exclusivement normative de l'inférence bayésienne, alors que celle-ci est en fait une construction beaucoup plus souple et beaucoup plus élaborée qu'il peut apparaître. On insiste sur la nécessité de l'articulation d'une approche normative et d'une approche descriptive, visant à étudier la cohérence des réponses, plutôt que leur exactitude.
Statistical inference and inductive reasoning
The use and the contribution of statistical inference in studying inductive reasoning is investigated. It is shown that some aspects are not always clearly taking into account. In particular, an exclusive normative use of Bayesian inference has often been involved. Bayesian inference is in fact a more flexible and elaborated construction that it can appear. Furthermore, the need for articulating a normative approach and a descriptive approach, in order to study the coherence of the responses rather than their accuracy, is stressed.

Lecoutre, B. (1996) Cet ouvrage propose à l'utilisateur de l'analyse de variance une approche pratique, réaliste et constructive de l'inférence statistique, qui lui apporte un regard nouveau sur ses données. Les procédures bayésiennes standard sont aussi objectives et simples à utiliser que les procédures traditionnelles (tests de signification *t* ou *F* familiers, intervalles de confiance). Intégrant ces dernières, elles en éclairent les difficultés et les insuffisances, et renouvellent en profondeur la méthodologie du traitement statistique des données expérimentales.
Des réponses concrètes sont ainsi apportées à des questions essentielles dans la pratique: *Interprétation* - Comment interpréter correctement les procédures d'inférence statistique? *Importance des effets* - Comment juger de l'importance d'un effet: ("significativité clinique, psychologique [...]" et "significativité statistique"? Peut-on "prouver l'hypothèse nulle" d'absence d'effet quand c'est l'hypothèse de recherche? *Apport réel des données* - Comment apprécier "ce que les données ont à dire" et examiner dans quelle mesure des informations supplémentaires remettraient en cause les conclusions? *Plans d'expérience complexes* - Comment analyser les plans expérimentaux complexes largement utilisés, tels que les dispositifs avec mesures répétées ou croisés (cross-over)? *Conditions de validité* - Comment comparer des moyennes sans supposer l'égalité des variances? *Choix des effectifs* - Comment déterminer les effectifs nécessaires pour "avoir de bonnes chances" d'obtenir une conclusion donnée?
La présentation des méthodes est effectuée à partir d'exemples réels. Les programmes informatiques sous Windows, didactiques et conviviaux, permettent la mise en œuvre interactive très simple de toutes les procédures au fur et à mesure de leur exposé. L'ouvrage présente ainsi une conception originale, qui en fait pour le plus grand nombre de lecteurs un outil précieux, utilisable aussi bien pour une initiation au traitement des données expérimentales que pour des applications sophistiquées.
***Sommaire*:** Quelques éléments de réflexion -- Du *t* de Student aux procédures fiducio-bayésiennes -- De l'ANOVA aux procédures fiducio-bayésiennes -- Prise en compte d'informations extérieures aux données -- Analyse spécifique: Plans S<G>*O -- Analyse spécifique: Illustrations -- Comparer des moyennes sans supposer l'égalité des variances -- Compléments sur les procédures fréquentistes -- Déterminer les effectifs nécessaires -- Distributions utiles -- Solutions bayésiennes -- Annexes: les programmes Windows -- Références bibliographiques -- Index.

Lecoutre, B. (1996) En fait nous n'avons pas seulement (ou même nous n'avons pas...) besoin d'une procédure de décision brutale, qui ne concerne que la valeur zéro et ne nous renseigne pas sur l'importance réelle de la corrélation. Mais nous devons aussi pouvoir "tester" d'autres valeurs, et plus simplement obtenir une "fourchette" qui nous permette d'apprécier réellement l'information apportée par les données. Nous allons rappeler que, dans les cas les plus courants de traitements de données numériques, il est immédiat de passer du test usuel à cette fourchette. Bien entendu il faudra justifier et interpréter celle-ci; on pourra se réjouir de savoir qu'elle peut être regardée

comme un intervalle de confiance (*fréquentiste*), comme un intervalle *fiduciaire*, ou comme un intervalle de crédibilité *bayésien* standard. Dans la suite nous l'appellerons simplement "intervalle", laissant le lecteur libre de choisir son *cadre de justification et d'interprétation*.

Lecoutre, B. (1997) L'objet de cet article est de guider le lecteur peu familiarisé dans la découverte de l'inférence bayésienne. Quatre idées pourront motiver cette découverte: l'inférence bayésienne n'est pas récente; elle apparaît supérieure sur le plan théorique; elle est une inférence naturelle; elle va devenir de plus en plus facilement utilisable. L'exposé sera très partiel (et partial), avec tous les oublis et toutes les insuffisances inévitables s'agissant d'un sujet aussi débattu que l'inférence statistique.
Nous prendrons comme point de départ le fait que les interprétations spontanées des résultats des procédures statistiques traditionnelles (seuils de signification, intervalles de confiance), même par des utilisateurs "avertis", sont le plus souvent en termes de probabilités sur les paramètres, qui sont en fait les probabilités *naturelles*: "celles qui vont du connu vers l'inconnu".

Lecoutre, B. (1998) The innumerable articles denouncing the deficiencies of significance testing urge us to reform the teaching of statistical inference for experimental data analysis. Bayesian methods are a promising alternative. However, teaching the Bayesian approach should not introduce an abrupt changeover from the current frequentist procedures: at the very least, the two approaches should co-exist for many years to come. According to this fact, we have developed statistical computer program, that incorporate both current practices and standard Bayesian procedures. These programs are used in the graduate statistics course in psychology, where Bayesian methods are especially introduced for inferences about effect sizes in the analysis of variance framework

Lecoutre, B. (1999) The *K*-prime and *K*-square distributions, involved in the Bayesian predictive distributions of standard *t* and *F* tests are investigated. They generalize the classical *noncentral t* and *noncentral F* distributions and can receive different characterizations. Their moments and their probability density and distribution functions are made explicit.

Lecoutre, B. (1999) The purpose of this paper is to argue that a widely accepted objective Bayesian methods, with the Fisher's fiducial motivation, are not only desirable but also feasible. These methods bypass the common misuses of null hypothesis significance testing and offer promising *new ways* in statistical methodology.

Lecoutre, B. (1999) The specific analysis approach allows the traditional analysis of variance procedures to be taught as direct extensions of the basic procedures of comparisons of means by Student's *t* tests. An appealing feature of this approach is that statistical inference procedures that provide genuine information about the magnitude of effects become easy to implement. This can be done in the usual frequentist framework as well as in the Bayesian framework. The opportunity of teaching these methods in the context of realistic complex experimental designs involving several factors is a compelling argument for the specific analysis approach.

Lecoutre, B. (2000) In this chapter we shall examine how, when analyzing experimental data, the researcher can call on intuitive knowledge to understand the principles and methodological implications of two of the main statistical inference procedures, namely, the traditional significance test and fiducial Bayesian inference. The underlying general problem will be the comparison of means in experimental designs. This problem is usually considered in an analysis of variance framework. In fact, it can be amply illustrated here in the case of a simple situation of inference concerning a mean.

Lecoutre, B. (2001) In recent years many authors have stressed the interest of the Bayesian predictive approach for designing ("how many subjects?") and monitoring ("when to stop?") experiments. The predictive distribution of a test statistic can be used to include and extend the frequentist notion of power in a way that has been termed predictive power or expected power. More generally Bayesian predictive procedures give the researcher a very appealing method to evaluate the chances that the experiment will end up showing a conclusive result, or on the contrary a non-conclusive result. The prediction can be explicitly based on either the hypotheses used to design the experiment, expressed in terms of prior distribution, or on partial available data, or on both.

Lecoutre, B. (2004) On se situe dans le cadre de l'analyse causale de données d'expériences "randomisées" (les traitements sont affectés à chaque unité expérimentale par tirage au sort). Les apports de quelques fondateurs de l'inférence statistique sont rapidement examinés. On considère ensuite les travaux récents, et notamment ceux sur les *modèles graphiques structuraux* de Pearl, qui visent à unifier sous une interprétation unique un certain nombre d'approches, incluant notamment les *analyses contrefactuelles*, les *modèles graphiques*, les *modèles d'équations structurelles*. La plupart de ces travaux reposent sur une approche contrefactuelle (invoquant des *résultats potentiels*: "si un autre traitement avait été affecté à l'unité expérimentale..." de l'inférence causale. Dans un article provocateur, Dawid (2000) soutient que cette approche est essentiellement métaphysique, et pleine de tentations de faire des inférences qui ne peuvent pas être justifiées sur la base de données empiriques. Concernant plus particulièrement les modèles graphiques structuraux, la critique de Dawid est que les "variables latentes" en jeu dans de tels modèles ne sont pas de véritables variables concomitantes (variables mesurables, qui peuvent être supposées non affectées par le traitement appliqué) et qu'il n'y a alors aucun moyen, même en principe, de vérifier les suppositions ("assomptions") faites - qui affecteront néanmoins les inférences qui en découlent. Dawid qualifie en

conséquence ces modèles de *pseudo-déterministes* et les considère comme *non scientifiques*. Les différents arguments et les solutions proposées sont examinés et discutés.

Experimentation, statistical inference and causal analyis

The causal analysis of "randomised" experimental data (treatments are randomly assigned to each experimental unit) is considered here. The contributions of some founders of statistical inference are briefly examined. Recent works, and especially Pearl's *graphical structural models*, are then considered. These models include *counterfactual analyses*, *graphical models*, *structural equations models*. Most of these models are based on a counterfactual approach (involving potential response : "if another treatment had been allocated to the experimental unit...") to causal inference. In a provocative article, Dawid (2000) argues that this approach is essentially metaphysical, and full of temptations to make inferences that cannot be justified on the basis of empirical data. Regarding graphical structural models, Dawid's major criticism is that "latent variables" involved in such models are not genuine concomitant variables (measurable variables, that can be assumed unaffected by the treatment applied) and that there is no way, even in principle, of verifying the assumptions made - which will nevertheless affect the ensuing inferences. Dawid terms these models *pseudodeterministic* and regards them as *unscientific*. The arguments and solutions are reviewed and discussed.

Lecoutre, B. (2005) The most frequently proposed interval estimates procedures for both original and standardized units in ANOVA situations are reviewed and discussed. It is demonstrated that the use of interval estimates for the conventional ANOVA effect size measures based on the ANOVA *F* test and involving the noncentral *F* disribution must be discouraged. It is argued that usual interval estimates for contrasts, including the Scheffé simultaneous estimate and the TOST (two one-sided tests procedure) based confidence interval for assesssing smallness, make appropriate inferences. The frequentist and Bayesian interpretations are briefly discussed.

Lecoutre, B. (2006) The use of frequentist Null Hypothesis Significance Testing (NHST) is so an integral part of scientists' behavior that its uses cannot be discontinued by flinging it out of the window. Faced with this situation, the suggested strategy for training students and researchers in statistical methods for experimental data analysis involves a smooth transition towards the Bayesian paradigm. Its general outlines are as follows. (1) To present natural Bayesian interpretations of NHST outcomes to draw attention to their shortcomings. (2) To create as a result of this the need for a change of emphasis in the presentation and interpretation of results. (3) Finally to equip users with a real possibility of thinking sensibly about statistical inference problems and behaving in a more reasonable manner. The conclusion is that teaching the Bayesian approach in the context of experimental data analysis appears both desirable and feasible. This feasibility is illustrated for analysis of variance methods.

Lecoutre, B., & Charron, C. (2000) Procedures for prediction analysis in 2×2 contingency tables are illustrated by the analysis of successes to six types of problems associated with the acquisition of fractions. According to Hildebrand, Laing, and Rosenthal (1977), hypotheses such as "success to problem type *A* implies in most cases success to problem type *B*" can be evaluated from a numerical index. This index has been considered in various other frameworks and can be interpreted in terms of a measure of predictive efficiency of implication hypotheses. Confidence interval procedures previously proposed for this index are reviewed and extended. Then, under a multinomial model with a conjugate Dirichlet prior distribution, the Bayesian posterior distribution of this index is characterized, leading to straightforward numerical methods. The choices of "noninformative" priors for discrete data are shown to be no more arbitrary or subjective than the choices involved in the frequentist approach. Moreover, a simulation study of frequentist coverage probabilities favorably compares Bayesian credibility intervals with conditional confidence intervals.

Lecoutre, B., & Derzko, G. (2001) Statistical inference procedures dedicated to asserting the smallness of effects are commonly used in the field of bioequivalence studies in pharmacology. They are however still virtually ignored in psychology. One possible reason is that experimental investigations generally involve complex designs for which solutions have not been developed in detail. The focus here is precisely on the extension of these procedures to all the situations where the usual ANOVA *F* tests apply. Smallness test and confidence interval procedures, both for raw effects, such as contrasts between means and their several *df* extensions, and for standardized effect size measures similar to Cohen's *d* and *f*, are considered. They are illustrated and contrasted with alternative Bayesian procedures. From a practical viewpoint, the computations require no more than the observed effect size, the usual *F* ratio, and percent points of statistical distributions.

Lecoutre, B., Derzko, G., & Grouin, J.-M. (1995) This paper investigates the Bayesian procedures for comparing proportions. These procedures are especially suitable for accepting (or rejecting the equivalence of two population proportions. Furthermore the Bayesian predictive probabilities provide a natural and flexible tool in monitoring trials, especially for choosing a sample size and for conducting interim analyses. These methods are illustrated with two examples where antithrombotic treatments are administrated to prevent further occurrences of thromboses

Lecoutre, B., & ElQasyr, K. (2005) Adaptative designs for clinical trials that are based on a generalization of the "play-the-winner" rule are considered as an alternative to previously developed models. Theoretical and numerical results show that these designs perform better for the usual criteria. Bayesian methods are proposed for the statistical analysis of these designs.

Lecoutre, B., Lecoutre, M.-P., & Grouin, J.-M. (2001) The use of frequentist Null Hypothesis Significance Testing (NHST) is so an integral part of scientists' behavior that its uses cannot be discontinued by flinging it out of the window. Faced with this situation, our teaching strategy involves a smooth transition towards the Bayesian paradigm. Its general outlines are as follows. (1) To present natural Bayesian interpretations of NHST outcomes to call attention about their shortcomings. (2) In this way to create the need for a change of emphasis in the presentation and interpretation of results. (3) Finally to equip the students with a real possibility of thinking sensibly about statistical inference problems and behaving in a more reasonable manner. Our conclusion is that teaching the Bayesian approach in the context of experimental data analysis appears both desirable and feasible.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (2001) The current context of the "significance test controversy" is first briefly discussed. Then experimental studies about the use of null hypothesis significance tests by scientific researchers and applied statisticians are presented. The misuses of these tests are reconsidered as judgmental adjustments revealing the researchers' requirements towards statistical inference. Lastly alternative methods are considered. We come naturally to the point of asking "won't the Bayesian choice be unavoidable?"
Usages, abus et mésusages des test de signification dans la communauté scientifique: Le choix bayésien ne sera-t-il pas incontournable?
Nous discutons d'abord brièvement le contexte actuel de la "controverse sur le test de signification". Puis nous présentons des recherches expérimentales sur l'usage des tests de signification de l'hypothèse nulle par des chercheurs scientifiques et des statisticiens professionnels. Les mauvais usages de ces tests sont reconsidérées comme des jugements adaptatifs, qui révèlent les exigences des chercheurs envers l'inférence statistique. Finalement, nous envisageons les solutions de rechange.
Nous en venons naturellement à poser la question: "le choix bayésien ne sera-t-il pas incontournable?"

Lecoutre, B., Mabika, B., & Derzko, G. (2002) The comparison of two Weibull distributions with unequal shape parameters, in the case of right censored survival data obtained for several independent samples, is considered within the Bayesian statistical methodology. The procedures are illustrated with the example of a mortality study where a new treatment is compared to a placebo. The posterior distributions about relevant parameters allowing to search for a conclusion of clinical superiority of the treatment, and the predictive distributions used to obtain an early stopping rule at an interim analysis, are considered for a class of appropriate priors.

Lecoutre, B., & Poitevineau, J. (2000) Il y a de bonnes raisons de penser que le rôle des tests de signification usuels dans la recherche en psychologie sera considérablement réduit dans un proche avenir. Les résultats des analyses statistiques traditionnelles devraient être systématiquement complétés ("au delà des seuls seuils observés *p*") pour inclure systématiquement la présentation d'indicateurs de la grandeur des effets et leurs estimations par intervalles. Ces procédures pourraient rapidement devenir de nouvelles *normes* de publication. Dans cet article, nous passons d'abord en revue les principaux abus des tests de signification et les solutions de rechange proposées. Parmi celles-ci, des méthodes d'intervalle de confiance (*fréquentistes*) et des méthodes d'intervalles de crédibilité (*fiducio-bayésiens*) permettent d'estimer l'importance réelle des effets, et en particulier d'apprécier leur caractère négligeable ou notable. A partir d'un exemple numérique, nous illustrons ces méthodes pour l'analyse de contrastes entre moyennes dans un plan d'expérience complexe, en considérant à la fois les effets *bruts* et les effets *relatifs* (calibrés). Nous discutons les similitudes et les différences des approches fréquentistes et bayésiennes, leur interprétation correcte et leur utilisation pratique.
Beyond traditional significance tests: Prime time for new publication norms
There are good reasons to think that the role of usual null hypothesis significance testing in psychological research will be considerably reduced in the near future. Traditional statistical analysis results should be enhanced ("beyond simple *p* value statements") to systematically include effect sizes and their interval estimates. Quite soon, these procedures could become new publication *norms*. In this paper main abuses of significance tests and alternative available solutions are first reviewed. Among these solutions, both confidence interval (*frequentist*) methods and credibility interval (*fiducial Bayesian*) methods have been developed for assessing effect sizes, and especially for asserting the negligibility or the notability of effects. From a numerical example, these methods are illustrated for analysing contrasts between means in a complex experimental design. Both *raw* and *relative* (calibrated) effects are considered. The similarities and differences between the frequentist and Bayesian approaches, their correct interpretations, and their practical uses, are discussed.

Lecoutre, B., Poitevineau, J., Derzko, G., & Grouin, J.-M. (2000) L'objectif de cet exposé, est d'illustrer, pour reprendre l'expression de Lewis (1982), la *désirabilité* et la *faisabilité* des méthodes bayésiennes en analyse de variance. Il ne sera pas question ici de revenir sur les débats sur l'inférence statistique, mais simplement de montrer de manière constructive comment des procédures bayésiennes de routine peuvent être aisément mises en œuvre et apporter des réponses simples et directes aux critiques méthodologiques formulées à l'encontre de l'usage des tests de signification usuels.

Lecoutre, B., Poitevineau, J., & Lecoutre, M.-P. (2005) It is shown that an interval estimate for a contrast between means can be straightforwardly computed, given only the observed contrast and the associated *t* or *F* test statistic (or equivalently the corresponding *p*-value). This interval can be seen as a *frequentist* confidence interval, as a standard *Bayesian* credibility interval, or as a *fiducial interval*. This interval estimate can be viewed either as a frequentist confidence interval or a fiducial interval or a Bayesian credible interval. This

gives Null Hypothesis Significance Tests (NHST) users the possibility of an easy transition towards more appropriate statistical practices. Conceptual links between NHST and interval estimates are outlined.
Une raison pour ne pas abandonner les tests de signification de l'hypothèse nulle
On montre que l'on peut directement calculer un intervalle pour un contraste entre moyennes, étant donné seulement la valeur observée du contraste et la statistique du test *t* ou *F* associé (ou encore, de manière équivalente le seuil observé correspondant ("*p-value*"). Cet intervalle peut être vu comme un intervalle de confiance *fréquentiste* ou comme un intervalle de crédibilité *bayésien* ou comme un intervalle *fiduciaire*. Cela donne aux utilisateurs des tests de signification usuels la possibilité d'une transition facile vers des pratiques statistiques plus appropriées. On met en avant les liens conceptuels entre les tests et les intervalles de confiance ou de crédibilité.

Lecoutre, B., Poitevineau, J., & Lecoutre, M.-P. (2005) When reading Denis' paper the feeling is that Fisher cannot be judged responsible for the "problems associated with today's model". Even if we agree that current uses of NHST are farm from being pure Fisherian, our analysis is somewhat different. In order to understand Fisher's real contribution, it is of direct importance to recall his statistical ideas about causality and probability. In particular his works, not only on the fiducial theory, but also on the Bayesian method in his last years, are a fundamental counterpart to his emphasis on significance tests. In conclusion, while the Fisher's responsibility in the today's practices cannot be discarded, the verdict imposes oneself: "responsible, not guilty".
Fisher: Responsable, non coupable
La lecture de l'article de Denis donne l'impression que Fisher ne peut pas être jugé responsable des "problèmes associés au modèle d'aujourd'hui". Même si nous sommes d'accord que les usages actuels des tests de signification de l'hypothèse nulle sont loin d'être purement fishériens, notre analyse est sensiblement différente. Pour comprendre la contribution réelle de Fisher, il est essentiel de rappeler ses idées statistiques sur la causalité et la probabilité. En particulier ses travaux, non seulement sur la théorie fiduciaire, mais aussi sur la méthode bayésienne dans ses dernières années, constituent une contrepartie fondamentale à son insistance sur l'usage des tests de signification. En conclusion, tandis que la responsabilité de Fisher dans les pratiques actuelles ne peut pas être rejetée, le verdict s'impose de lui même: "responsable, non coupable".

Lecoutre, B., Rouanet, H., & Denhière, G. (1988) L'Analyse des Comparaisons, constituée à partir de 1968 a fait l'objet de plusieurs exposés d'ensemble: Rouanet et Lépine (1977), Hoc (1983), Lecoutre (1984); on peut la voir comme une restructuration de l'analyse de variance comportant notamment les innovations suivantes. – Elaboration d'un *langage d'interrogation de données* permettant de formuler les questions du chercheur dans le cadre des facteurs du plan expérimental. – Principe d'*inférence spécifique* (détaillé dans Rouanet et Lépine, 1983), qui consiste à fonder l'inférence sur un modèle posé, non plus au niveau du protocole de base mais de chaque protocole dérivé pertinent correspondant à chaque demande d'analyse. – *Techniques bayésiennes*. 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. [...] Très tôt, l'Analyse Bayésienne des Comparaisons a été appliquée au problème de la validation des modèles: Rouanet, Lépine et Holender, 1978.

Lecoutre, M.-P. (1982) Studied the behaviors of psychologists spontaneously developed in conflictual situations of statistical data processing ; what is intended is not a normative aim, but the look for coherence lines. Three usual conflicting problems (for example, a same procedure applied to an experiment and to its replicate yielding discrepant results) were presented to 27 psychologists from several laboratories ; the responses were recorded as semi-directive interviews. Two main findings. First, while behaviors were well differentiated -mostly according to the weights of several criteria such as the observed results, significance tests, reference theories and so on- it was possible to infer global attitudes that are common to almost all searchers. Secondly, it appeared that the issue of data grouping is a critical one for many searchers.

Lecoutre, M.-P. (1992) [Example of use of standard Bayesian methods].

Lecoutre, M.-P. (2000) This chapter presents the findings of an experimental research project aimed at describing and analyzing the judgments made in situations of statistical inference by researchers (in this case, researchers in psychology) who, having completed an experiment, proceed to a statistical analysis of their data. The rest of the chapter is divided into two parts, one for each stage in the study. The first part examines the role of the various ingredients which contribute to the formulation of a statistical conclusion, along with the interpretations they give rise to. The second part, the study of statistical prediction situations, leads us to examine how a statistical conclusion is understood and interpreted: what significance do researchers attach to conclusions such as, for example "there is a difference between two treatments", or "there is an effect of such and such a factor", *etc.*

Lecoutre M.-P., Clément E., Lecoutre B. (2004) [Example of use of standard Bayesian methods].

Lecoutre, M.-P., Poitevineau, J., & Lecoutre, B. (2003) We investigated the way experienced users interpret Null Hypothesis Significance Testing (NHST) outcomes. An empirical study was designed to compare the reactions of two populations of NHST users,

psychological researchers and professional applied statisticians, when faced with contradictory situations.
The subjects were presented with the results of an experiment designed to test the efficacy of a drug by comparing two groups (treatment/*placebo*). Four situations were constructed by combining the outcome of the *t* test (significant vs nonsignificant) and the observed difference between the two means *d* (large vs small). Two of these situations appeared as conflicting (*t* significant/*d* small and *t* nonsignificant/*d* large). Three fundamental aspects of statistical inference of statistical inference were investigated by means of open questions: drawing inductive conclusions about the magnitude of the true difference from the data in hand, making predictions for future data, and making decisions about stopping the experiment. The subjects were 25 statisticians from pharmaceutical companies in France, subjects well versed in statistics, and 20 psychological researchers from various laboratories in France, all with experience in processing and analyzing experimental data.
On the whole, statisticians and psychologists reacted in a similar way and were very impressed by significant results. It must be outlined that professional applied statisticians were not immune to misinterpretations, especially in the case of nonsignificance. However, the interpretations that accustomed users attach to the outcome of NHST can vary from one individual to another, and it is hard to conceive that there could be a consensus in front of seemingly conflicting situations. In fact beyond the superficial report of “erroneous” interpretations, it can be seen in the misuses of NHST intuitive judgmental “adjustments”, that try to overcome its inherent shortcomings. These findings encourage the many recent attempts to improve the habitual ways of analyzing and reporting experimental data.
Même les statisticiens ne sont pas à l’abri des erreurs d’interprétation des Tests de Signification de l’Hypothèse Nulle
Nous avons étudié la manière dont des utilisateurs expérimentés interprètent les résultats des Tests de Signification de l’Hypothèse Nulle. Une étude empirique a été menée pour comparer les réactions de deux populations d’utilisateurs, des chercheurs en psychologie et des statisticiens professionnels, face à des situations conflictuelles.
On présentait aux sujets les résultats d’une expérience planifiée pour tester l’efficacité d’un médicament en comparant deux groupes (traitement/*placebo*)>. Quatre situations étaient construites en combinant l’issue du test *t* (significatif vs non-significatif) et la différence observée *d* entre les deux moyennes (grande vs petite). Deux de ces situations apparaissaient conflictuelles (*t* significatif/*d* petite et *t* non-significatif/*d* grande). Trois aspects fondamentaux de l’inférence statistique étaient examinés au moyen de questions ouvertes: tirer une conclusion inductive sur la grandeur de la vraie différence, faire une prédiction relative à des données futures, et prendre une décision sur l’arrêt de l’expérience. Les sujets étaient 25 statisticiens de l’industrie pharmaceutique en France, donc experts en statistique, et 20 chercheurs en psychologie de différents laboratoires français, ayant tous une expérience de l’analyse des données expérimentales.
Dans l’ensemble, les statisticiens et les psychologues se sont comportés d’une manière similaire et ont été très influencés par les résultats significatifs. Un résultat important est que les statisticiens ne sont pas à l'abri des abus d'interprétation des tests, en particulier quand le résultat est non significatif. Cependant l’interprétation des tests peut varier considérablement d’un individu à l’autre et est loin de donner lieu à un consensus face à des situations en apparence conflictuelles. En fait au delà du constat superficiel de l’existence d’interprétations “erronées”, on peut voir dans les mésusages des tests des “ajustements” de jugement intuitifs, pour tenter de surmonter leurs insuffisances fondamentales. Ces résultats encouragent les nombreuses tentatives récentes d’améliorer les procédures habituelles pour analyser les données expérimentales et présenter les résultats.

Lecoutre, M.-P., & Rouanet, H. (1993) Probabilistic judgments made by researchers in psychology were investigated in statistical prediction situations. From these situations, it is possible to test the "representativeness hypothesis" (Tversky & Kahneman, 1971) and the "significance hypothesis" (Oakes, 1986). The predictive judgments concerned both an elementary descriptive statistic and a significance test statistic. In the first case, the predictive judgments were generally coherent and it comparatively well to Bayesian standard predictive probabilities. As for the two hypotheses tested, our findings are compatible with the significance hypothesis, but go against the representativeness hypothesis.

Lee, P. (1989) This book is concerned with estimating the values of unknown parameters and investigating the degree of confidence we can have in various hypotheses. The Bayesian approach is distinguished by giving a probability distribution to the unknown parameters and then modifying it in the light of experimental data. This is controversial because for a theory with no new data available, the statistician's own beliefs have to be incorporated into the analysis. The author presents the ideas behind Bayesian Statistics at a level suitable for advanced undergraduate or postgraduate students. -- The discrepancies between the conclusions of Bayesian and "classical" statistics are highlighted. -- Full treatment of Bayesian statistics is presented; easily accessible to students with some knowledge of statistics. -- Clear exposition of *where* and *why* theBayesian approach differs from the "classical" approach. -- Discusses how real *prior* information can be incorporated into statistical analyses and explains the difficulties which can arise when "conventional priors" are used. -- Excellent appendix includes tables useful in Bayesian statistics (and not readily available elsewhere).

Lehmann, E.L. (1993) The Fisher and Neyman-Pearson approaches to testing statistical hypotheses are compared with respect to their attitudes to the interpretation of the outcome, to power, to conditioning, and to the use of fixed significance levels. It is argued that despite basic philosophical differences, in their main practical aspects the two theories are complementary rather than contradictory and that a unified approach is possible that combines the best features of both. As applications, the controversies about the Behrens-Fisher problem and the comparison of two binomials (2x2 tables) are considered from the present point of view.

Lépine, D., & Rouanet, H. (1975) La méthode traditionnelle du t de Student peut être utilisée pour mettre à l'épreuve un modèle d'hypothèse nulle à 1 degré de liberté dans une grande variété de plans d'expérience. Mais souvent cette méthode ne suffit pas à répondre

à l'attente du chercheur, qui voudrait inférer, à partir des seules données expérimentales, sur l'importance, dans la population parente, de l'écart à l'hypothèse nulle. Les auteurs proposent ici, à partir des conceptions fiduciaires de Fisher, une méthode qui répond à cette attente en autorisant, dans les cas où le test "*t*" est valide, une inférence probabiliste portant sur le paramètre d'écart à l'hypothèse nulle. Des exemples concrets illustrent la méthode, qui est utilisée ici principalement pour répondre à la question: "existe-t-il, en un sens qui est à préciser par le chercheur, un écart "négligeable" ou au contraire "notable" à l'hypothèse nulle habituelle?". La méthode apparaît ainsi comme un prolongement du test de signification usuel, en vue d'une inférence plus précise, et permet ainsi de discriminer les cas où les données expérimentales peuvent valablement conduire à une conclusion dans les termes souhaités de ceux où l'information n'est pas suffisante pour permettre une conclusion inférentielle ferme.
Introduction to fiducial methods: Inference about a contrast between means
In many experimental design, the traditional Student *t*-test may be used to test a model of null hypothesis with one degree of freedom. But often this method does not satisfy the aim of the researcher who wishes to make inferences about the deviation from the null hypothesis within the parent population. On the basis of Fisher's notion of fiducial inference, the authors propose a method that satisfies this aim: it enables a probabilistic inference about the parameter of discrepancy from the null hypothesis, that can be used whenever the *t*-test is valid. Concrete examples illustrate this method which is used, in the present case, to answer the question: "does there exist, in a sense to be specified by the experimenter, a negligible or an important discrepancy from the null hypothesis?". This method is an extension of the common test of significance and allows a more precise inference; it also permits the discrimination of cases in which the information contained in the data is sufficient to yield a firm conclusion, from those in which it does not.

Levy, P. (1967) **[*Example of misinterpretation of a p-value*]** "Statistical significance refers only to... the confidence with which a null hypothesis may be rejected." (page 37) [Quoted by Falk and Greenbaum, 1995, page 82]

Lewis, C. (1993) In this chapter, it is assumed that the reader has some familiarity with the basic concepts of Bayesian inference and of conventional analysis of variance. This allows attention to be focused on what happens when the two are brought together. For this purpose, extensive use is made of results presented by Box and Tiao (1973). This source provides, by far, the most extensive treatment of analysis of variance from a Bayesian point of view, and the interested reader will find in it proofs and generalizations of most of the material appearing here. [...] the emphasis is on laying out, as clearly as possible, a Bayesian approach to analysis of variance.

Lindley, D.V. (1998) It is argued that the determination of bioequivalence involves a decision, and is not purely a problem of inference. A coherent method of decision-making is examined in detail for a simple trial of bioequivalence. The result is shown to differ seriously from the inferential method, using significance tests, ordinarily used. The reason for the difference is explored. It is show how the decision-analytic method can be used in more complicated and realistic trials and the case for its general use presented.

Lindley D.V. (2000). This paper puts forward an overall view of statistics. It is argued that statistics is the study of uncertainty. The many demonstrations that uncertainties can only combine according to the rules of the probability calculus are summarized. The conclusion is that statistical inference is firmly based on probability alone. Progress is therefore dependent on the construction of a probability model; methods for doing this are considered. It is argued that the probabilities are personal. The roles of likelihood and exchangeability are explained. Inference is only of value if it can be used, so the extension to decision analysis, incorporating utility, is related to risk and to the use of statistics in science and law. The paper has been written in the hope that it will be intelligible to all who are interested in statistics.

Lipset, S.M., Trow, M.A., & Coleman, J.S. (1956) In this book, no statistical tests of significance have been used. [...] It can be defended, and we shall defend it at length because there seems to be no good statement of our position in print. Statistical tests of hypotheses, however, seem to be of quite limited aid in building theoretical science.

Little, J. (2001) Few concepts in the social sciences have wielded more discriminatory power over the status of knowledge claims than that of statistical significance. Currently operationalized as alpha=0.05, statistical significance frequently separates publishable from nonpublishable research, renewable from nonrenewable grants, and, in the eyes of many, experimental success from failure. If literacy is envisioned as a sort of competence in a set of social and intellectual practices, then scientific literacy must encompass the realization that this cardinal arbiter of social scientific knowledge was not born out of an immanent logic of mathematics but socially constructed and reconstructed in response to sociohistoric conditions.

Locascio, J.J. (1999) Limitations and inappropriate uses of null hypothesis statistical significance testing (NHST) in behavioral research have been widely cited. Critics recommend alternative data analysis approaches and even outright "banning" of it from professional journals. I agree with most criticisms, but would stop short of supporting a ban.

Loftus, G.R. (1991) [Review of Gigerenzer *et* al.,1989] The *Empire of Chances* is about the history and current use of probability theory and statistics. [...] Because this review is for psychologists, I will organize it around the book's insights into a question that I believe is at

the heart of much malaise in psychological research: How has the virtually barren technique of hypothesis testing come to assume such importance in the process by which we arrive at our conclusions from our data? I will first describe why this question is timely and important; I will then provide a brief synopsis of the book; and finally, I will detail the book's answers to the question.

Loftus, G.R. (1993) 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, standard deviations*. 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, G.R. (2002) This chapter has two main purposes: the first is to discuss serious problems with two generally accepted and widely used foundations of current statistical analysis in psychology, the linear model and null hypothesis significance testing; the second purpose is to describe some alternative valuable techniques. The alternatives are grouped into six categories: (1) use of sophisticated pictorial and graphical techniques for data display, (2) use of confidence intervals in numerous situations, (3) use of planned comparisons, emphasizing *contrasts*, (4) use of percent total variance accounted for, (5) representation of theoretical fits to data, and (6) use of *equivalence techniques* to investigate interactions. Detailed hypothetical numerical examples along with associated calculations and graphs are constructed to illustrate each of the techniques.

Loftus, G.R., & Masson, M.E.J. (1994) We argue that to best comprehend many data sets, plotting judiciously selected sample statistics with associated confidence intervals can usefully supplement, or even replace, standard hypothesis-testing procedures. We note that most social science statistics textbooks limit discussion of confidence intervals to their use in between-subject designs. Our central purpose in this article is to describe how to compute an analogous confidence interval that can be used in within-subject designs. This confidence interval rests on the reasoning that because between-subject error term – that is, on the variability due to the subject condition*interaction. Computation of such a confidence interval is simple and is embodied in equation 2 on p. 482 of this article. This confidence interval has two useful properties. First, it is based on the same error term as is the corresponding analysis of variance, and hence leads to comparable conclusions. Second, it is related by a known factor (square-root of 2) to a confidence interval of the difference between sample means: accordingly, it can be used to infer the faith one can put in some pattern of sample means as a reflection on the underlying pattern of population means. These two properties correspond to analogous properties of the more widely used between-subject confidence interval.

Loredo, T.J. (1990) The Bayesian approach to probability theory is presented as an alternative to the currently used long-run relative frequency approach, which does not offer clear, compelling criteria for the design of statistical methods. Bayesian probability theory offers unique and demonstrably optimal solutions to well-posed statistical problems, and is historically the original approach to statistics. The reasons for earlier rejection of Bayesian methods are discussed, and it is noted that the work of Cox, Jaynes, and others answers earlier objections, giving Bayesian inference a firm logical and mathematical foundation as the correct mathematical language for quantifying uncertainty. The Bayesian approaches to parameter estimation and model comparison are outlined and illustrated by application to a simple problem based on the gaussian distribution. As further illustrations of the Bayesian paradigm, Bayesian solutions to two interesting astrophysical problems are outlined: the measurement of weak signals in a strong background, and the analysis of the neutrinos detected from supernova SN 1987A. A brief bibliography of astrophysically interesting applications of Bayesian inference is provided.

Ludbrook, J. (2000) In a recent review article, the problem of making false-positive inferences as a result of making multiple comparisons between groups of experimental units or between experimental outcomes was addressed.2. It was concluded that the most universally applicable solution was to use the Ryan-Holm step-down Bonferroni procedure to control the family-wise (experiment-wise) type 1 error rate. This procedure consists of adjusting the P values resulting from hypothesis testing. It allows for correlation among hypotheses and has been validated by Monte Carlo simulation. It is a simple procedure and can be performed by hand.3. However, some investigators prefer to estimate effect sizes and make inferences by way of confidence intervals rather than, or in addition to, testing hypotheses by way of P values and it is the policy of some editors of biomedical journals to insist on this. It is not generally recognized that confidence intervals, like P values, must he adjusted if multiple inferences are made from confidence intervals in a single experiment.4. In the present review, it is shown how confidence intervals can be adjusted for multiplicity by an extension of the Ryan-Holm step-down Bonferroni procedure. This can be done for differences between group means in the case of continuous variables and for odds ratios or relative risks in the case of categorical variables set out as 2 x 2 tables.

Lutz, W., & Nimmo, I.A. (1977) 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.

Lykken, D. (1968) 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. The value of any research can be determined, not from the statistical results, but only by skilled, subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied, and so on. [...] 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.

Macdonald, R.R. (1997)] This paper argues that a Fisherian approach to statistical inference, which views statistical testing as determining the chance probability of an event, is coherent, consistent with modern statistical methods and forms a sound theoretical basis for the use of statistical tests in psychology. It is argued that Fisherian statistical tests are concerned with establishing the direction of tested effects, give rise to confidence intervals and are quite consistent with power analyses. Researchers are encouraged to report chance probabilities, and to interpret them according to the prevailing conditions rather than using a fixed decision rule. The contrasting Newman Pearson approach, which views statistical testing as a quality control procedure for accepting hypotheses, posits unreasonable research practices which psychologists do not and should not be expected to follow. Newman Pearson theory has caused confusion in the psychological literature and criticisms have been levelled at statistical testing in general that ought to have been directed specifically at Newman Pearson testing. Statistical inference, of any sort, is held to be insufficient to characterize the process of testing scientific hypotheses. Data should be seen as evidence to be used in psychological arguments and statistical significance is just one measure of its quality. It restrains researchers from making too much of findings which could otherwise be explained by chance.

Markus, K.A. (2001) Critics have put forth several arguments against the use of tests of statistical significance (TOSSes). Among these, the converse inequality argument stands out but remains sketchy, as does criticism of it. The argument states that we want $P(H|D)$ (where H and D represent hypothesis and data, respectively), we get $P(D|H)$, and the 2 do not equal one another. Each of the terms in '$P(D|H)<>(H|D)$' requires clarification. Furthermore, the argument as a whole allows for multiple interpretations. If the argument questions the logic of TOSSes, then defenses of TOSSes fall into 2 distinct types. Clarification and analysis of the argument suggest more moderate conclusions than previously offered by friends and critics of TOSSes. Furthermore, the general method of clarification through formalization may offer a way out of the current impasse.

Mauk, A-M.K (2000) The present paper summarizes the recommendation that statistical significance testing be replaced or at least accompanied by the reporting of effect sizes and confidence intervals and discusses, in particular, confidence intervals. The recent report of the APA Task Force on Statistical Inference suggested that confidence intervals should always be reported.

McCloskey, D.N., & Ziliak, S.T. (1996) "In a survey of papers published in the American Economic Review, the authors found that 59% use the word ‘significance' in ambiguous ways at one point meaning ‘statistically significantly different from the null,' at another ‘practically important' or ‘greatly changing our scientific opinion,' with no distinction."

McGraw, K.O. (1991) In light of the data distorsionsintroduced by BESDs [Binomial Effect Size Displays], I fail to see how they can be touted as "intuitively appealing" and "perfectly transparent" ways of representing treatment effects on dichotomously measured outcomes.

McLean, J.E. (2001) Hypothesis testing is widely regarded as an essential part of statistics, but its use in research has led to considerable controversy in a number of disciplines, especially psychology, with a number of commentators suggesting it should not be used at all. A root cause of this controversy was the overenthusiastic adoption of hypothesis testing, based on a greatly exaggerated view of its role in research. A second cause was confusion between the two forms of hypothesis testing developed by Fisher on the one hand and Neyman and Pearson on the other. This paper discusses these two causes, and also proposes that there is a more general misunderstanding of the role of hypothesis testing. This misunderstanding is reflected in vocabulary such as 'the true value of the parameter'.

McLean, J.E., & Kaufman, A.S. (1998) The research methodology literature in recent years has included a full frontal assault on statistical significance testing. The purpose of this paper is to promote the position that, while significance testing as the sole basis for result interpretation is a fundamentally flawed practice, significance tests can be useful as one of several elements in a comprehensive interpretation of data. Specifically, statistical significance is but one of three criteria that must be demonstrated to establish a position empirically. Statistical significance merely provides evidence that an event did not happen by chance. However, it provides no information about the meaningfulness (practical significance) of an event or if the result is replicable. Thus, we support other researchers who recommend that statistical significance testing must be accompanied by judgments of the event’s practical significance and replicability.

McLean, J.E., & Kaufman, A.S. (2000) The Publication Manual of the American Psychological Association (1994), the style guide required by *Research in the Schools*, provides little guidance. The Manual discusses both statistical significance (pp. 17-18) and effect

size (p. 18), but only "encourages" (p. 18) authors to provide effect size information.
Since the [*RITS*] Special Issue [on statistical significance], we have also "encouraged" authors to provide effect size information. In fact, we have required authors to provide effect size information to accompany statistical significance tests unless they could provide a compelling reason not to. Since that time, we have had no author make a compelling case to omit effect size information. Thus, we have decided to make this policy explicit and require that authors accompany the reporting of statistical significance tests with effect size information. Specifically, the following line has been added to the Research in the Schools "Information for Authors" section: "All reporting of statistical significance must include an estimate of effect size."
We are hopeful that this change will encourage educational researchers to consider effect size and practical significance when evaluating the results of a study. We are also encouraging our Editorial Board members to consider effect size and the practical significance of a study in their recommendations. In the end, we hope this change supports the movement towards the reporting of more complete and accurate results of research studies. Since this change in policy merely formalizes what we have been practicing for at least two years, we do not expect that it will have an impact on the number of manuscripts we receive, but we do hope it will have an impact on the quality of the manuscripts.

Meehl, P.E. (1967) The purpose of the present paper is not so much to propound a doctrine or defend a thesis (especially as I should be surprised if either psychologists or statisticians were to disagree with whatever in the nature of a "thesis" it advances), but to call the attention of logicians and philosophers of science to a puzzling state of affairs in the currently accepted methodology of the behavior sciences which I, a psychologist, have been unable to resolve to my satisfaction. The puzzle, sufficiently striking (when clearly discerned) to be entitled to the designation "paradox", is the following: *In the physical sciences, the usual result of an improvement in experimental design, instrumentation, or numerical mass of data is to increase the difficulty of the "observational hurdle" which the physical theory of interest must successfully surmount; whereas, in psychology and some of the allied behavioral sciences, the usual effect of such improvement in the experimental precision is to provide an easier hurdle for the theory to surmount.*

Melton, A.W. (1962) [*About his "misinterpretation of significance levels"*] Melton has been often blamed for having claimed erroneously that the significance level determines the probability that a significant result will be found in a replication (pages 553-554: see, for instance, Bakan, 1966, Seldmeier & Gigerenzer, 1989). But, Melton has never asserted that the probability of reproducing results was 1-*p*. He considered only that the smaller the level the more secure was the reproductibility, which is justified whatever the theoretical statistical framework.

Mendoza J.L., Stafford K.L. (2001) In this article, the authors introduce a computer package written for Mathematica, the purpose of which is to perform a number of difficult iterative functions with respect to the squared multiple correlation coefficient under the fixed and random models. These functions include, among others, computation of confidence interval upper and lower bounds, power calculation, calculation of sample size required for a specified power level, and providing estimates of shrinkage in cross validating the squared multiple correlation under both the random and fixed models. Attention is given to some of the technical issues regarding the selection of, and working with, these two types of models as well as to issues concerning the construction of confidence intervals.

Mialaret, G. (1996) **[*Exemple d'abus d'interpretation d'un intervalle de confiance*]** "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." (page 112).

Mittag, K.C., & Thompson, B. (2000) Almost as soon as statistical significance tests were popularized near the turn of this century, critics emerged (Berkson, 1938; Boring, 1919). The criticism since then has been fairly continual (e.g., Carver, 1978; Meehl, 1978; Rozeboom, 1960), but recent commentary has been particularly striking (cf. Cohen, 1994; Kirk, 1996; Schmidt, 1996; Thompson, 1996, 1999a). Of course, statistical tests also have support from some, though even most advocates concur that the tests are sometimes misused or misunderstood (e.g., Frick, 1996; Robinson & Levin, 1997). Particularly thoughtful advocacy for continued reliance on statistical testing has been offered by Abelson (1997) and Cortina and Dunlap (1997). A balanced and comprehensive treatment of the controversies is provided by Harlow, Mulaik, and Steiger (1997; for detailed reviews of this book, see Levin, 1998, and Thompson, 1998). Huberty (1987, 1993) and Huberty and Pike (1999) provide the related historical perspective. However, as Tryon (1998) recently lamented, "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... (page 796)" Indeed, several *empirical* studies have shown that many researchers do not fully understand the statistical tests that they employ (Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993).
The present report was written to address two objectives. First, we wanted to explore current perceptions of AERA members regarding statistical significance tests. We also explored perceptions regarding other statistical issues, such as score reliability (e.g., Thompson & Vacha-Haase, 2000) and stepwise methods (e.g., Cliff, 1987, Huberty, 1989; Thompson, 1995), about which there has also been some controversy. The present investigation was particularly timely given the recent release of the related various recommendations of the

American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson & The APA Task Force on Statistical Inference, 1999). These recommendations will be considered soon in revising the previous 1994 edition of the APA publication manual, incorporated by many behavioral science journals into editorial requirements. Second, we also wanted our report to serve as a vehicle promoting further discussion of controversial statistical issues. Although we have arrived at reasoned positions regarding the merits of some research practices, reasonable people disagree over such issues. We hope our presentation will provide a framework prompting further discussion.

Morris, S.B., & Lobsenz, R.E. (2000) The two most common methods for assessing adverse impact, the four-fifths rule and the z-test for independent proportions, often produce discrepant results. These discrepancies are due to the focus on practical versus statistical significance, and on differing operational definitions of adverse impact. In order to provide a more consistent frame work for evaluating adverse impact, a new significance test is proposed, which is based on the same effect size as the four-fifths rule. Although this new test was found to have slightly better statistical power under some conditions, both tests have low power under the typical conditions where adverse impact is assessed. An alternative to significance testing would be to report an estimate of the adverse impact ratio along with a confidence interval indicating the degree of precision in the estimate.

Morrison, D.E., & Henkel, R.E. (1969) Apart from implication for improved *use*, however, our analysis, like Selvin's [Selvin, 1957], more basically questions the general *utility* of the tests in basic (*not* applied) scientific research. The test provides neither the necessary nor the sufficient scope or type of knowledge that basic scientific social research requires. [...] But how *is* scientific inference possible if significance tests are of little help? This question leads us beyond the scope of this paper, but we have offered some hints: replication over diverse samples as well as internally, the use of abstract concepts, and the incorporation of such concepts in deductive theories with the conditions of their validity specified. There are, of course, no computational formulas for scientific inference: the questions are must more difficult and the answers much less definite than those of statistical inference.
[*Example of misinterpretation of significance levels*] "[...] 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 [...]" (pages 195-209;196)

Mulaik, S.A., Raju, N.S., & Harshman, R.A. (1997) We expose fallacies in the arguments of critics of null hypothesis significance testing who go too far in arguing that we should abandon significance tests altogether: Beginning with statistics containing sampling or measurement error, significance tests provide prima facie evidence for the validity of statistical hypotheses, which may be overturned by further evidence in practical forms of reasoning involving defeasible or dialogical logics. For example, low power may defeat acceptance of the null hypothesis. On the other hand, we support recommendations to report point estimates and confidence intervals of parameters, and believe that the null hypothesis to be tested should be the value of the parameter given by a theory or prior knowledge. We also use a Wittgensteinian argument to question the coherence of concepts of subjective degree of belief underlying subjective Bayesian alternatives to significance testing.

Murphy, K.R. (1997) 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.

Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986) How do American psychologists use statistical results and related information to interpret research evidence? In previous studies it was found that confidence ratings increased with larger sample sizes at the same *p* values, and we found a similar relationship in this study. Perhaps confidence in research results is a two-step or three-step process for psychological researchers. First, confidence is earned by the rejectability of the null hypothesis, with p=.05 considered a critical level, but with greater confidence in general given to lower *p* values. Second, given a *p* low enough, psychologists researchers' confidence is increased by increases in obtained effect sizes, and especially so for younger investigators. Third, psychologists trust this effect size more when sample size is larger, because the effect size is, in general, more accurately estimated when sample sizes are larger.

Nester, M.R. (1996) Hypothesis testing, as performed in the applied sciences, is criticized. Then assumptions that the author believes should be axiomatic in all statistical analyses are listed. These assumptions render many hypothesis tests superfluous. The author argues that the image of statisticians will not improve until the nexus between hypothesis testing and statistics is broken.

Neyman, J., Scott, E.L., & Smith, J.A. (1969) **[*Example ofinterpretation of a p-value as Pr*(H|X)]** 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.

Nickerson, R. S. (2000) Null hypothesis significance testing (NHST) is arguably the most widely used approach to hypothesis evaluation among behavioral and social scientists. It is also very controversial. A major concern expressed by critics is that such testing is misunderstood by many of those who use it. Several other objections to its use have also been raised. In this article the author reviews

and comments on the claimed misunderstandings as well as on other criticisms of the approach, and he notes arguments that have been advanced in support of NHST. Alternatives and supplements to NHST are considered, as are several related recommendations regarding the interpretation of experimental data. The concluding opinion is that NHST is easily misunderstood and misused but that when applied with good judgment it can be an effective aid to the interpretation of experimental data.

Nunnally, J.C. (1975) **[*Example of misinterpretation of significance levels*]** "[...] 95 [chances] out of 100 that the observed difference will hold up in future investigations." (page 195) [Quoted by Carver, 1978]

Ogles, B.M., Lunnen, K.M., & Bonesteel K. (2001) The meaningfulness of psychotherapy outcome as measured in therapy research is a persistent and important issue. Following a period of emphasis on statistically significant findings for treated versus control groups, many researchers are renewing efforts to investigate the meaningfulness of individual change. Several statistical methods are available to evaluate the meaningfulness of clients' changes occurring as a result of treatment. This article reviews the history of the clinical significance concept; describes the various methods for defining improvement, recovery, and clinically significant change; examines current criticisms of the methods; and describes the current use of the methods in practice.

O'Grady, K.E. (1982) Measures of explained variance (e.g. proportion of variance accounted for) are often considered to indicate the importance of a statistical finding. Three potential limitations to this viewpoint are discussed: psychometric, methodological, and theoretical. A psychometric perspective suggests that errors of measurement produce an upper bound to any measure of explained variance, this upper bound being the product of the reliabilities of the variables whose association is under investigation. A methodological perspective suggests several factors that influence the magnitude of measures of explained variance, including the intentions of the researcher, the design of the research, and the population sampled. A theoretical perspective suggests that most behavior has multiple determinants, and thus the magnitude of measures of explained variance can me made when researchers examine the agreement between the magnitude that their theory would suggest and the empirical finding.

O'Hagan, T. (1996) First Bayes is a program intended to help with teaching and learning elementary Bayesian Statistics. It deals with quite simple and standard statistical models, with an emphasis on obtaining some understanding of how the Bayesian approach works. It is *not* a package for *doing* statistical analysis of practical data. Four standard one parameter models are offered: binomial data, gamma data, Poisson data and normal data with known variance. A major feature of First Bayes is that such data may be analysed using an arbitrary mixture of distributions from the conjugate family. By this means, essentially arbitrary prior distributions can be defined, allowing the user to obtain an excellent understanding of how the likelihood and prior distribution are combined by Bayes' Theorem. Prior, likelihood and posterior can be plotted on a single "triplot". Analysis of two simple kinds of linear model are also offered. One is the case of one or more normal samples with common but unknown variance (one-way analysis of variance), and the other is simple linear regression. Marginal distributions may be computed (and examined) for arbitrary linear combinations of the location parameters. In the case of regression, scatter and residual plots can be produced. Predictive distributions are available in a variety of forms for all analyses.

O'Rourke, K. (1996) I agree with many points Kadane makes in "Prime Time for Bayes" [Kadane, 1995] but question the conclusion that the Bayesian approach should become the mainstay of statistical analysis in randomized clinical trials and completely disagree with the author's concluding remark that this would provide a [the only?] "foundation for clinical trials that makes sense." In fact, there is no known rational basis (i.e., a basis that "makes sense") for empirical science or any process of induction.

Pagano, R.R. (1990) **[*Example of interpretation of frequentist confidence intervals in terms of probabilities about parameters*]** "an interval such that the probability is 0.95 that the interval contains the population value." (page 288)

Pascual, J., Frías, Ma.D., & Garcia, J.F. (2000) The standard hypothesis testing method has a number of well-known logical fallacies and the results of the procedures are often misinterpreted. Many scholars have suggested that, perhaps, NHST should be abandoned altogether in favor of other bases for conclusions such as confidence intervals and effect size estimates. Other researchers are often interested in testing the hypothesis that the effects of treatments, interventions, etc. are negligibly small rather than testing the hypothesis that treatments have no effects whatsoever. We argue that we must question the "old" procedures to stimulate the application of new statistical procedures in the progress of scientific inference.

Pascual, J., Garcia, J.F., & Frías, Ma.D. (2000) This paper analyses the relationship between the concepts of statistical significance (level of probability, p) and replicability. The level of statistical significance (for example, p = 0.01) indicates the probability of the data under the null hypothesis assumption, however, this does not mean that in a later replication the probability to obtain significant differences will be the complementary 0.99. if correctly understood, replicability is exclusively related to the reliability and consistency of the data. The only way to evaluate reliability is through repeated empirical tests.

Pearson, E.S. (1955) This paper contains a reply to some criticisms made by Sir Ronald fisher in his recent article on" Statistical Methods and Scientific Induction".

Pearson, K. (1900) The object of this paper is to investigate a criterion of the probability of any theory of an observed system of errors and to apply it to the determination of goodness of fit in the case of frequency errors.

Perlman, M.D., & Wu, L. (1999) In the past two decades, striking examples of allegedly inferior likelihood ratio tests (LRT) have appeared in the statistical literature. These examples, which arise in multiparameter hypothesis testing problems, have several common features. In each case the null hypothesis is composite, the size a LRT is not similar and hence biased, and competing size alpha tests can be constructed that are less biased, or even unbiased, and that dominate the LRT in the sense of being everywhere more powerful. It is therefore asserted that in these examples and, by implication, many other testing problems, the LRT criterion produces "inferior," "deficient," "undesirable," or "flawed" statistical procedures. This message, which appears to be proliferating, is wrong. In each example it is the allegedly superior test that is flawed, not the LRT. At worst, the "superior" tests provide unwarranted and inappropriate inferences and have been deemed scientifically unacceptable by applied statisticians. 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. Although the LR criterion is not infallible, we believe that it remains a. generally reasonable first option for non-Bayesian parametric hypothesis-testing problems.

Poitevineau, J. (1998) La thèse présentée est celle de l'inadaptation à la recherche expérimentale de la pratique du test de signification par les chercheurs en psychologie. La question de cette inadaptation est abordée selon trois approches, normative, prescriptive et descriptive qui constituent les trois parties de la thèse. La première partie est consacrée à l'étude, d'un point de vue méthodologique, de la norme statistique constituée par les théories du test statistique de Fisher et de Neyman et Pearson. Les principales caractéristiques de ces deux théories sont rappelées puis les nombreuses critiques dont les tests continuent d'être l'objet sont examinées, ainsi que les abus d'utilisation et des raisons possibles de la persistance de l'usage des tests et des abus. La deuxième partie aborde la question de la pertinence des prescriptions. Parmi les principales solutions de rechange aux tests qui sont passées brièvement en revue, seules les méthodes d'intervalle de confiance et les méthodes bayésiennes paraissent devoir s'imposer comme véritables "challengers" des tests traditionnels. De l'analyse de six des manuels d'inférence statistique à l'usage des psychologues parmi les plus connus, il ressort que les théories des tests statistiques y sont rarement rapportées fidèlement et qu'ils contiennent déjà des abus d'interprétation, particulièrement dans les exemples présentés. La troisième partie est consacrée aux attitudes des chercheurs en psychologie à l'égard des tests de signification. Des réanalyses statistiques de résultats déjà publiés ainsi qu'une réanalyse que nous avons menée au moyen d'outils fiducio-bayésiens sont présentées. Nous relatons aussi des d'expériences menées auprès de chercheurs, dont deux expériences que nous avons réalisées. Nous concluons à une pratique inadaptée au plan méthodologique, mais socialement adaptée, d'un outil inadéquat dont le mode d'emploi est trompeur. Nous évoquons aussi le probable changement d'attitude des psychologues vis-à-vis du test de signification, en conséquence de prochaines recommandations de l'*American Psychological Association*, et les possibilités d'une plus grande utilisation de l'analyse bayésienne qui en découlent.
Methodology of the analysis of experimental data: A study of the use of significance tests by psychologists, from normative, prescriptive, and descriptive approaches
The thesis presented is that the current use of significance tests by psychologists is unsuited for experimental research. This question is examined through three approaches, normative, prescriptive and descriptive, which constitute the three parts of the dissertation. The first part is devoted, from a methodological viewpoint, to the study of the theories of statistical test developed by Fisher and by Neyman and Pearson, and which now constitute the statistical norm. The main features of these theories are first described, then the numerous criticisms which are still directed at the statistical tests are examined. Misuses of tests are also examined, as well as possible reasons for the continued use of these tests.The second part deals with the pertinence of the prescriptions. Among the main alternatives to the tests reviewed, only confidence interval methods and Bayesian methods seem to be potential challengers to the traditional tests. From the analysis of six popular textbooks of statistical inference designed for psychologists, it appears that the theories of statistical test are rarely accurately reported and that those textbooks contain some misuses, particularly among the examples used. The third part is devoted to the attitudes of psychologists toward significance tests. Some statistical re-analyses of published results are presented, as long as a re-analysis we performed using standard Bayesian tools. Some experiments involving researchers as subjects are also reported, including the two we realized for this thesis. We conclude that the use of significance tests by psychologists is a socially adapted but methodologically unsuited use of an inadequate tool promoted through misleading guide-lines of standard textbooks. We also mention a probable change in psychologists' attitude toward significance tests, as a consequence of recommendations from the *American Psychological Association* that are likely to appear in the near future, and the possibility that Bayesian analysis will become more and more used.

Poitevineau, J. (1999) En 1962 Cohen a publié une recherche qui a servi de modèle à beaucoup d'autres. Son objectif premier était d'ordre méthodologique: mettre en évidence certains problèmes soulevés par l'usage des tests de signification, et éventuellement en tirer les conséquences pour une meilleure pratique. Il s'agissait pour lui de voir comment les psychologues, si soucieux de se prémunir contre l'erreur de première espèce, se gardaient de l'erreur de seconde espèce. Autrement dit, voir si la puissance des tests utilisés par les

psychologues était suffisante pour que l'hypothèse nulle ait de bonnes chances d'être rejetée quand elle est fausse. A cette fin il a analysé tous les articles parus dans le volume 61 (1960) du *Journal of Abnormal and Social Psychology*. Mais dans ce type d'étude, les données (les effets, les statistiques de test) apparaissant dans les articles n'ont aucun rôle: pour calculer la puissance du test utilisé il suffit de connaître la structure du plan d'analyse, les effectifs, et la valeur de l'effet vrai qui est fixée par hypothèse. Nous présentons une réanalyse que nous avons effectuée dans une perspective plus descriptive que celle en jeu dans les études de puissance comme celle de Cohen: il s'agit, d'une part de recenser quels sont les abus d'interprétation des tests explicitement commis, et d'autre part de chercher à préciser quelle est la portée réelle des conclusions autorisées en ce qui concerne l'importance des effets, en relation précisément avec les abus (ou les insuffisances) des interprétations fournies par les auteurs. En retour cela permettra d'examiner si les tailles d'échantillon sont suffisantes pour obtenir des conclusions satisfaisantes sur l'importance des effets (relativement à un certain critère). La méthode fiducio-bayésienne qui utilise une distribution *a priori* non informative nous servira de norme: c'est dans le cadre de cette méthode, et donc par rapport à elle, que nous tâcherons de répondre en examinant comment les conclusions tirées par les chercheurs à partir de tests statistiques usuels pourraient être prolongées ou modifiées. Cette méthode permet de choisir le type d'inférence *a posteriori*, ce que ne permettent pas, en toute rigueur, les méthodes fréquentistes (de test ou d'intervalle de confiance) de recherche de conclusion d'effet négligeable ou notable. Pour faciliter la comparaison avec les études antérieures, nous avons choisi de réanalyser des articles parus dans le *Journal of Abnormal Psychology*. Nous avons retenu le volume 103 (année 1994), c'est-à-dire le plus récent disponible au moment de ce travail.

Poitevineau J. (2004) La pratique des tests statistiques par les chercheurs en psychologie est abordée selon trois aspects. Du point de vue normatif les tests apparaissent inadaptés; les principales critiques sont présentées. Du point de vue descriptif, l'examen des manuels statistiques, les réanalyses d'articles publiés et les expériences auprès de chercheurs montrent l'existence de nombreux abus d'utilisation. Enfin, du point de vue prescriptif, des solutions de rechange sont envisagées, en particulier les méthodes bayésiennes qui apparaissent particulièrement prometteuses.
The use of significance tests by psychologists: normative, descriptive and prescriptive viewpoints
At a normative level, the significance tests appear to be ill-suited and the main criticisms are reported. At a descriptive level, both examination of statistical textbooks, re-analyses of published papers and experiments about the use of significance tests by psychologists clearly reveals many misuses. At a prescriptive level, alternative solutions are considered, especially the Bayesian methods which appear to be especially attractive.

Poitevineau, J., & Lecoutre, B. (1998) Chow's book [Chow, S.L. (1996). *Statistical Significance: Rationale, Validity and Utility*. London: Sage.] makes a provocative contribution to the debate on the role of statistical significance, but it involves some important misconceptions in the presentation of the Fisher and Neyman-Pearson's theories. Moreover, the author's caricature-like considerations about "Bayesianism" are completely irrelevant for discarding the Bayesian statistical theory. These facts call into question the objectivity of his contribution.

Poitevineau, J., & Lecoutre, B. (2001) Comments about previous studies indicate that the interpretation of significance levels by psychological researchers is unequivocally dictated by a binary decision-making framework. In particular confidence in a *p* level would drop abruptly just beyond the fateful .05 level ("cliff effect"). A replication of Rosenthal and Gaito's experiment on the degree of confidence in *p* levels shows that these claims should be moderated. Detailed analysis of individual curves reveals that the attitude of researchers towards *p*-values is far from being as homogeneous as might be expected. However most psychological researchers in our study rated graduated confidence judgments, as either exponential or linear. Only a minority of "all-or-none" respondents exhibited an abrupt drop in confidence.

Pratt, J.W. (1965) This paper is an attempt to present in an orderly way various ideas about the interpretation of standard inference statements from the Bayesian point of view. [...] The use of insufficient statistics will be considered first, in Section 2. It proves easy to assimilate them into the Bayesian framework. An example is given in which this leads to some progress on a problem of Bayesian non-parametric statistics. Estimation and confidence regions are taken up in Sections 3 and 4 respectively, and it is shown that classical properties give approximately, in a certain weak Bayesian sense, corresponding Bayesian properties. Classical anomalies no longer seem disturbing from this point of view. Ideas related to maximum likelihood are postponed to Section 5, and some general remarks concerning the approximation idea of Sections 3 and 4, including the application of the likelihood principle, are postponed to Section 6. tests of hypotheses are discussed in the next two Sections, significance levels and *P*-values in Section 7 and common uses of tests in Section 8. The general conclusion here is that only certain one-tailed *P*-values are interpretable Bayesianly and that, even when this interpretation is applicable, conventional tests are seldom well articulated to practical problems. Section 9 contains a few final comments.

Press, W.H. (1989) To understand their data better, astronomers need to use statistical tools that are more advanced than traditional "freshman lab" statistics. As an illustration, the problem of combining apparently incompatible measurements of a quantity is presented from both the traditional, and a more sophisticated Bayesian, perspective. Explicit formulas are given for both treatments. Results are shown for the value of the Hubble Constant, and a 95% confidence interval of 66<H0<82 (km/s/Mpc) is obtained.

Pruzek, R.M. (1997) Students in the social and behavioral sciences tend generally to learn inferential statistics from tests and materials that emphasize significance tests or confidence intervals. Bayesian statistical methods support inferences without reference to either significance tests or confidence intervals. This chapter provides an introduction to Bayesian inference. It is shown that this class of methods entails use of prior information and empirical data to generate posterior distributions that in turn serve as the basis for statistical inferences. Two relatively simple examples are used to illustrate the essential concepts and methods of Bayesian analysis and to contrast inference statements made within this subjectivist framework with inference statements derived from classical methods. In particular, the role of posterior distributions in making formal inferences is described and compared with inferences based on classical methods. It also is argued that Bayesian thinking may help to improve definitions of inferential problems, especially in the behavioral and social sciences where the complex nature of applications often may require special strategies to make it realistic for investigators to attempt rigorous formal inferences. Sequentially articulated studies as seen as having special virtue in using results from previous studies to inform inferences about later ones. Numerous references are briefly described to aid the reader who seeks to learn more about Bayesian inference and its applications.

Racine, A., Grieve, A.P., Flühler, H., & Smith, A.F.M. (1986) Four typical applications of Bayesian methods in pharmaceutical research are outlined. The implications of the use of such methods are discussed, and comparisons with traditional methodologies are given. Although a great deal has been written on the comparative merits and demerits of different approach to statistical inference, this debate has very largely been conducted by theoreticians. Indeed, one of the recurring criticisms of the Bayesian approach seems to have been that it is not "practical". Against this background, it seemed to us – from the perspective of an applied statistics section in a major pharmaceutical company – of some interest to give a review of a variety of day-to-day problems which have been analysed for non-statistical clients within the company using Bayesian methods. For the most part, we shall present a straightforward account of the models, methodology and inference summaries employed, but potentially controversial issues will be clearly signposted. Our hope is that the shift of the focus of the debate from the theoretical to the practical domain will stimulate a more productive discussion of these issues, and one which the "practical statistician" will feel less able to ignore.

Richardson, J.T.E. (1996) Two different approaches have been used to derive measures of effect size. One approach is based on the comparison of treatment means. The standardized mean difference is an appropriate measure of effect size when one is merely comparing two treatments, but there is no satisfactory analogue for comparing more than two treatments. The second approach is based on the proportion of variance in the dependent variable that is explained by the independent variable. Estimates have been proposed for both fixed-factor and random-factor designs, but their sampling properties are not well understood. Nevertheless, measures of effect size can allow quantitative comparisons to be made across different studies, and they can be a useful adjunct to more traditional outcome measures such as test statistics and significance levels.

Rindskopf, D. (1997) Critical attacks on null hypothesis testing over the years have not greatly diminished its use in the social sciences. This chapter tells why the continued use of hypothesis tests is not merely due to ignorance on the part of data analysts. In fact, a null hypothesis that an effect is exactly zero should be rejected in most circumstances; what investigators really want to test is whether an effect is nearly zero, or whether it is large enough to care about. Although relatively small sample sizes typically used in psychology result in modest power, they also result in approximate tests that an effect is small (not just exactly zero), so researchers are doing approximately the right thing (most of the time) when testing null hypotheses. Bayesian methods are even better, offering direct opportunities to make statements such as "the probability that the effect is large and negative is .01; the probability that the effect is near zero is .10; and the probability that there is a large positive effect is .89."

Rindskopf, D. (1998) [...] my preferred solutions in the "controversy" about null-hypothesis testing is (1) recognize that we really want to test the hypothesis that an effect is "small", not null, and (2) use Bayesian methods, which are much more in keeping with the way humans naturally think than are classical statistical methods.

Robert, C.P. (1994) This book is a translation of a French book written to supplement the gap in the French statistical literature about Bayesian Analysis and Decision Theory. As a result, its scope is wide enough to cover most graduate programs. It builds on very little prerequisites in Statistics and only requires basic skills in calculus, measure theory, and probability. In terms of level and existing literature, this book starts at a level similar to those of the introductory books of Lee (1989) and Press (1989), but it also goes further and keeps up with most of the recent advances in Bayesian Statistics, while motivating the theoretical appeal of the Bayesian approach on decision-theoretic justifications. Nonetheless, this book differs from the reference book of Berger (1985a) by including the more recent developments of the Bayesian field (the Stein effect for spherically symmetric distributions, multiple shrinkage, loss estimation, decision theory for testing and confidence regions, hierarchical developments, Bayesian computation, mixture estimation, etc.).
The plan of the book is as follows: Chapter 1 is an introduction to statistical models, including the Bayesian model and some connections with the Likelihood Principle. The book then proceeds with Chapter 2 on Decision Theory, considered from a classical point of view, this approach being justified through the axioms of rationality and the need to compare decision rules in a coherent way. It also includes a presentation of usual losses and a discussion of the Stein effect. Chapter 3 gives the corresponding analysis for prior distributions and deals in detail with conjugate priors, mixtures of conjugate priors, and noninformative priors, including a concluding section on prior

robustness. Classical statistical models are studied in Chapter 4, paying particular attention to normal models and their relations with linear regression. This chapter also contains a section on sampling models that allows us to include the pedagogical example of capture-recapture models. Tests and confidence regions are considered separately in Chapter 5, since we present the usual construction through "0-1" losses, but also include recent advances in the alternative decision-theoretic evaluations of testing problems. The second part of the book dwells on more advanced topics and can be considered as providing a basis for a more advanced graduate course. Chapter 6 covers complete class results and sufficient/necessary admissibility conditions. Chapter 7 introduces the notion of invariance and its relations with Bayesian Statistics, including a heuristic section on the Hunt--Stein theorem. Hierarchical and empirical extensions of the Bayesian approach, including some developments on the Stein effect, are treated in Chapter 8. Chapter 9 is rather appealing, considering the available literature, as it incorporates in a graduate textbook an introduction to state-of-the-art computational methods (Laplace, Monte Carlo and, mainly, Gibbs sampling). In connection with this chapter, a short appendix provides the usual pseudorandom generators. Chapter 10 is a more personal conclusion on the advantages of Bayesian theory, also mentioning the most common criticisms of the Bayesian approach.

[255] **[*Exemple d'interprétation de l'intervalle de confiance fréquentiste en termes de probabilité sur les paramètres*]** "Par exemple, si dans un sondage de taille 1000, on trouve *P* [fréquence] = 0,613, la proportion *pi* à estimer a une probabilité 0,95 de se trouver dans la fourchette [0,58 ; 0,64]" (pages 221-222).

Robert, M. (1994) **[*Exemple de "formulation ambiguë"*]** "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." (page 66)

Robinson, D.H., & Wainer, H. (2002) Recent criticisms of null hypothesis significance testing (NHST) have appeared in wildlife journals (Cherry 1998; Johnson 1999; Anderson et al. 2000, 2001; Guthery et al. 2001). In this essay, we discuss these criticisms with regard to both current usage of NHST and plausible future use. We suggest that the historical use of such procedures was reasonable and that current users might spend time profitably reading some of Fisher's applied work. However, modifications to NHST, and to the interpretations of its outcomes, might better suit the needs of modern science. Our primary conclusion is that NHST most often is useful as an adjunct to other results (e.g., effect sizes) rather than as a stand-alone result. We cite some examples, however, where NHST can be profitably used alone. Last, we find considerable experimental support for a less dogmatic attitude toward the interpretation of the probability yielded from such procedures.

Rogers, J.L., How>ard, K.I., & Vessey, J. (1993) Equivalency testing, a statistical method often used in biostatistics to determine the equivalence of 2 experimental drugs, is introduced to social scientists. Examples of equivalency testing are offered, and the usefulness of the method to the social scientists is discussed.

Rosenthal, R. (1991) When used appropriately, the BESD [Binomial Effect Size Displays] has been used to excellent advantage by methodologically sophisticated behavioral researchers and by experienced mathematical statisticians.

Rosenthal, R., & Gaito, J. (1963) A total of 19 psychologists (graduate students and faculty) rated their degree of confidence in a variety of *p* levels for each of two assumed sample sizes. The relationship between degree of confidence and magnitude of *p* levels appeared to be exponential regardless of sample size assumed and type of Ss employed. Ss have greater confidence in a given *p* level when it was associated with a larger sample size suggesting that investigators use the probability of both Type I and Type II errors as criteria of "belief". Graduate students Ss tended to place more confidence in given *p* levels than did faculty members. For 84 per cent of the Ss, the .05 level had cliff characteristics manifested by a relatively more precipitous loss of confidence in moving from the .05 to the .10 level than was true at either higher or lower levels of significance.

Rosenthal, R., & Gaito, J. (1964) In a careful and probably improved replication of our study, Beauchamp and May (1964) [...] could find "no evidence" for any .05 cliff effect. [...] Additional evidence for the existence of an .05 cliff was to be found in their extended report. [...] For *p* values based on large samples, their 11 graduate student Ss expressed a *greater* average degree of confidence in the .05 level than did in the .03 level! This interesting finding even if not statistically significant certainly is consistent with our hypothesis that the .05 level has rather special characteristics.
[*Example of interpretation of significance levels in terms of probabilities about parameters*] "In summary, the probability that the .05 level of significance possesses cliff characteristics 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'."

Rosnow R.L., & Rosenthal, R. (2002) we introduce students to statistical methods that, although enormously useful, do not yet generally appear in undergraduate methods texts: meta-analysis, contrast analysis, interval estimates of effect sizes and their practical

interpretation, and so on. The emphasis of these discussions is intended to resonate with the spirit and substance of the guidelines recommended by the American Psychological Association's Task Force on Statistical Inference (Wilkinson et al., 1999).

Rothman, K.J. (1978) 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.

Rouanet, H. (1996) In experimental data analysis when it comes to assessing the importance of effects of interest, 2 situations are commonly met. In situation 1, asserting largeness is sought: "The effect is large in the population." In situation 2, asserting smallness is sought: "The effect is small in the population." In both situations, as is well known, conventional significance testing is far from satisfactory. The claim of this article is that Bayesian inference is ideally suited to making adequate inferences. Specifically, Bayesian techniques based on "noninformative" priors provide intuitive interpretations and extensions of familiar significance tests. The use of Bayesian inference for assessing importance is discussed elementarily by comparing 2 treatments, then by addressing hypotheses in complex analysis of variance designs.

Rouanet, H. (1998) Chow's efforts towards a methodology of theory-corroboration and the plea for significance testing are welcome, but there are many risky claims. A major omission is a discussion of significance testing in the Bayesian framework. We sketch here the Bayesian reinterpretation of the significance level for assessing direction of effects.

Rouanet, H. (2000) In this introductory chapter, we will discuss descriptive and inductive procedures, then logical and statistical inference, lastly hypotheses and assumptions. Next we will describe the background of current statistical practice, building on the opposition between Mathematical Statistics and Statistics for Researchers. An outline of the book will close the chapter. In Appendix A, we will sketch the historical relationship between Probability and Statistics. In Appendix B, we will discuss the current issue of mind as an intuitive statistician.

Rouanet, H. (2000) In this chapter we will revisit, from a methodological standpoint, the current statistical practice. We will concentrate our examination on statistical tests, or *tests* (for short), also called *hypothesis tests* (in mathematical statistics), or *significance tests* (by Fisher and in statistics for researchers). Confidence method, which are also used, but in a more modest scale, will be reviewed more briefly. We will first review the familiar Student's-test and the chi-square test, exemplified on the classical Student data and Mendel data, in connection with two common research paradigms. Then; enlarging the discussion, we will proceed to a step-by-step examination of the current statistical practice. Then, we will discuss the perverse effects of the official doctrine and the real problems that the researcher is faced with. An overall reassessment of current statistical practice and the outline of alternative frameworks will close the chapter. In the Appendix, we will discuss further false problems and wrong tracks, namely the two-tailed vs one-tailed quarrel and the chimera of power.

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., & Le Roux, B. (2000) This book, with a Foreword by the outstanding philosopher of science and mathematical psychologist Patrick Suppes of Stanford University, is the outgrowth of the work developed within the *Groupe Mathématiques et Psychologie*, a research unit of the University René Descartes and C.N.R.S. (the French National Center for Scientific Research). New ways in statistical methodology are presented, which complement the familiar significance tests by new methods better suited to the researchers' objectives, in the first place Bayesian methods. In mathematical statistics, Bayesian methods have made a breakthrough in the last few years, but those developments are still ignored by the current statistical methodology and practice. The present book is really the first one to fill this gap. This book is written for a large audience of researchers, statisticians and users of statistics in behavioral and social sciences, and contains both an analysis of the attitude of researchers toward statistical inference, and concrete proposals for improving statistical practice. The statistical consulting experience of the authors is centered around psychology and covers a broad range of subjects from social sciences to biostatistics. All methods developed by the authors are implemented in software.

Rouanet, H., Bernard, J.-M., & Lecoutre, B. (1986) The familiar sampling procedures of statistical inference can be recast within a purely set-theoretic (ST) framework, without resorting to probabilistic prerequisites. This article is an introduction to the ST approach of statistical inference, with emphasis on its attractiveness for teaching. The main points treated are unsophisticated ST significance testing and ST inference for a relative frequency (proportion).

Rouanet, H., & Bert, M.-C. (2000) This chapter is an introduction to Combinatorial Inference, or Set-theoretic Inference, an alternative to frequentist inference that our Math & Psy Group has been developing since the early eighties: see Rouanet, Bernard, Lecoutre (1986) and Rouanet, Bernard, Le Roux (1990). Its motivation is to provide researchers with a framework that can be used when the "validity assumptions" of the common procedures are not met. [...] Roughly speaking, the algorithms of combinatorial procedures coincide with those of conventional tests, while the random framework is discarded. As a result, in data analysis, it will be possible to keep many

familiar algorithms, while the conclusions of Combinatorial Inference are stated in terms of new concepts, such as typicality and homogeneity, formalized in a nonprobabilistic way. We first present Typicality tests, and Homogeneity tests. Then we outline the making of combinatorial inference and discuss related viewpoints.

Rouanet, H., & Bru, B. (1994) C'est avec émotion que nous avons découvert que Victor Henri, cofondateur de *L'Année Psychologique* avec "son maître" Alfred Binet, fut aussi le premier "statisticien des psychologues" français [Henri, 1895, 1898]. [...] Les notes de lecture qui suivent n'ont pas d'autre but que d'inviter les chercheurs actuels à lire à leur tour ce pionnier de la "statistique des chercheurs", c'est-à-dire cette activité qui, bien distincte de la "statistique mathématique", cherche avant tout à répondre aux besoins méthodologiques des chercheurs.

Rouanet, H., & Lecoutre, B. (1983) Whenever in a complex design inferences on separate effects are sought, the (overall) distributional assumptions of a general model are irrelevant. The *specific inference approach* is examined as a useful alternative to the conventional general model approach. The specific inference for a particular effect, based only on data relevant to this effect, is valid regardless of the complexity of the design. Specific inference is first discussed in terms of significance testing. It is argued that the usual ANOVA table can be regarded as a system of specific analyses, each on resting on a separate specific model in its own right. Then specific inference is discussed within a Bayesian framework. A standard Bayesian ANOVA is suggested as a direct extension of the usual *F*-test ANOVA. Technical developments and methodological implications are outlined.

Rouanet, H., Lecoutre, M.-P., Bert, M.-C., Lecoutre, B., & Bernard, J.-M. (1991) Cet ouvrage pluridisciplinaire est à la charnière de la psychologie cognitive et de la statistique. On y trouvera d'une part une analyse de la démarche et des attitudes des chercheurs face à l'inférence statistique, d'autre part des propositions pour renouveler les pratiques statistiques, en les complétant par des méthodes (combinatoires et bayésiennes) mieux adaptées et aujourd'hui accessibles grâce à l'informatique. Cet ouvrage s'adresse aux chercheurs, enseignants et utilisateurs de la statistique, tout particulièrement en psychologie et en sciences humaines, qui y trouveront un texte de référence pour l'approche de la nouvelle école statistique de Paris.
***Sommaire*:** Rouanet H., Avant-propos, 5-7 -- Rouanet H., Les pratiques statisticiennes en question, 9-22 -- Rouanet H., Les tests statistiques revisités, 23-45 -- Lecoutre M.-P., Et... le point de vue des chercheurs? Quelques éléments de réflexion, 47-77 -- Rouanet H., L'approche de l'inférence ensembliste, 79-86 -- Bert M.-C., Inférence sur une moyenne: Test de signification ensembliste et test d'hypothèse, 87-93 -- Lecoutre B., Du test de signification à l'inférence fiducio-bayésienne, 95-120 -- Bernard J.-M., Inférence bayésienne et prédictive sur les fréquences, 121-153.

Rouanet, H., Lépine, D., & Holender, D. (1978) The use of significance tests for validating models is basically inadequate, since nonsignificant results only pertain to the compatibility of the data with the model. In this paper an alternative form of data analysis is proposed, whose objective is to assess the acceptability of the model given the data. The use of Bayes-fiducial inference is suggested in this connection; the approach is exemplified through the analysis of an experiment planned for investigating a model of successive stages of information processing in binary choice reaction.

Rouanet, H., Lépine, D., & Pelnard-Considère, J. (1976) There are various things one may put forward to arouse sympathy for the Bayesian approach; the key ideas in this paper are: (i) The Bayesian approach leads to conclusions directly interpretable in terms of the psychological research objectives, in contrast to the conclusions from the usual, often misused, significance tests. (ii) Technically, the Bayesian approach may lead to surprisingly simple procedures not requiring sophisticated calculations or machine programs; this is especially true for the simplest of Bayesian methods, which we call "Bayes-fiducial methods", to which this paper is devoted. (iii) The Bayesian approach can be used to analyse real data with educational psychologists handle everyday. To show this, we have worked out a complete Bayes-fiducial analysis of a set of educational data; the paper is organized around the analysis of this example.

Rouanet, H., Le Roux, B., Bernard, J.-M., & Lecoutre, B. (2000) This final chapter provides an opening along two directions. Firstly, it deals with *geometric data*. Secondly, data are *structured*, that is to say there is a design to investigate several sources of variation [...]. To analyze such data,one may contemplate two lines of approach. Along the line of Geometric Data Analysis (GDA), as developed in France, one would start by representing data as *clouds of points* and proceed to the descriptive exploration of these clouds. Along the line of the Anglo-Saxon MANOVA tradition, one would start with a *statistical model* and proceed to inductive analyses. In this chapter, we present a statistical strategy which combines both approaches. As in GDA>, we conceptualize statistical procedures as geometric operations on clouds of points, and as in MANOVA, we carry out inductive analyses. For each analysis, the procedures will be performed along the following three phases: (i) Descriptive analysis and observed effects; (ii) MANOVA significance testing and existence of effects; (iii) Bayesian MANOVA and importance (largeness) of effects.

Royall, R.M. (1986) Contradictory interpretations of how the meaning of a significance test depends on the sample size are examined.

Royall, R.M. (1999) Current frequentist methods use probabilities to measure both the chance of errors and the strength of observed

evidence (for discussion see Royall, 1997, ch.5). The Law of Likelihood explains that it is likelihood ratios, not probabilities, that measure evidence. The concept of statistical evidence embodied in the Law of Likelihood, and represented in the terms "weak evidence" and "misleading evidence" that are central to the evidential paradigm, can lead to a body of statistical theory and methods that: (1) Requires a probability model for the observable random variables only (and is in that sense frequentist, not Bayesian). (2) Contains a valid, explicit, objective measure of the strength of statistical evidence. (3) Provides for explicit, objective measure (and control) of the probabilities of observing weak or misleading evidence.

Rozeboom, W.W (1960) 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*. While several serious objections to the 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. It is further argued that the proper application of statistics to scientific inference is irrevocably committed to extensive consideration of inverse probabilities, and to further this end, certain suggestions are offered, both for the development of statistical theory and for more illuminating application of statistical analysis to empirical data.

Rozeboom, W.W. (1991) Conceptual rigor is indeed a desideratum worth dedicated pursuit; in fact one might wish that Chow [Chow, 1991a] had pursued it somewhat more diligently in his present essay. I suggest that the approach to data interpretation he advocates here is an etch-a-sketch draft whose prospect for refinement into an operational logic of inference that professional scientists can live by appears minuscule.

Salsburg, D. (1994) The ‘Intent to Treat' paradigm for the analysis of a controlled randomized clinical trial is a direct result of applying the Neyman-Pearson formulation of hypothesis testing. If other formulations are used, the ‘Intent to Treat' paradigm makes no sense. Criticisms of the Neyman-Pearson formulation and whether it is applicable to scientific investigations have appeared in the statistical and philosophical literature since it was proposed. This paper reviews the nature of that criticism and notes why the Neyman-Pearson formulation, and with it the ‘Intent to Treat' paradigm, is inappropriate for use in the analysis of clinical trials.

Sánchez, J., Valera, A., Velandrino, A, & Marin, F. (1992) The aim of this paper was to determine the statistical power of the research published in the journal *Anales de Psicología* across its life (1984-1991). The sixteen studies available for this calculation were used for analyzing their statistical power. The results do not seem to differ from those originally obtained by Cohen (1962), showing average powers of .13, .47, and .76 for small, medium, and large effect sizes respectively. Also, the mean power for the estimated effect sizes from the proper studies increased to .71, very close to the minimum of .80 recommended by Cohen (1988). Finally, the results are compared to other recent power studies.

Schield, M. (1998) Previous papers by the author have argued that the Bayesian strength of belief can be used in interpreting classical hypothesis tests and classical confidence intervals. In hypothesis tests, one's strength of belief in the truth of the alternate upon rejecting the null was argued to be equal to (1-*p*) under certain conditions. In confidence intervals, being 95% confident was argued as being operationally equivalent to a willingness to bet on a 95% chance. These interpretations were taught in an introductory class of non-majors. Students found this approach to be extremely natural for confidence intervals. But in hypothesis testing, students had difficulty relating the quality of the test (*p*-value) to the quality of the decision. The underlying problem is student difficulty with related conditionals. To overcome this problem, we should teach more about conditionality – not less.

Schmidt, F.L., & Hunter, J.E. (1997) Logically and conceptually, the use of statistical significance testing in the analysis of research data has been thoroughly discredited. However, reliance on significance testing is strongly embedded in the minds and habits of researchers, and therefore proposals to replace significance testing with point estimates and confidence intervals o/ten encounter strong resistance. This chapter examines eight of the most commonly voiced objections to reform of data analysis practices and shows each of them to be erroneous. The objections are: (a) Without significance tests we would not know whether a finding is real or just due to chance; (b) hypothesis testing would not be possible without significance tests; (c) the problem is not significance tests but failure to develop a tradition of replicating studies; (d) when studies have a large number of relationships, we need significance tests to identify those that are real and not just due to chance; (e) confidence intervals are themselves significance tests; (f) significance testing ensures objectivity in the interpretation of research data; (g) it is the misuse, not the use, of significance testing that is the problem; and (h) it is futile to try to reform data analysis methods, so why try? Each of these objections is intuitively appealing and plausible but is easily shown to be logically and intellectually bankrupt. The same is true of the almost 80 other objections we have collected. 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, K. (1995) So why not aim directly at CIs [confidence intervals] even when it is more complex to du so? In most cases, a CI can be determined at least by Monte Carlo simulation and trial and error methods trying to find alternative null hypotheses can be accepted

given the observed value of the test statistic.

Schuirmann, D.J. (1987) The statistical test of the hypothesis of no difference between the average bioavailabilities of two drug formulations, usually supplemented by an assessment of what the power of the statistical test would have been if the true averages had been inequivalent, continue to be used in the statistical analysis of bioavailability/bioequivalence studies. In the present article, this Power Approach (which in practice usually consists of testing the hypothesis of no difference at level 0.05 and requiring an estimated power of 0.80) is compared to another statistical approach, the Two One-Sided Tests Procedure, which leads to the same conclusion as the approach proposed by Westlake [1981] based on the usual (shortest) 1-2*alpha* confidence interval for the true average difference. It is found that for the specific choice of *alpha*=0.05 as the nominal level of the one-sided tests, the two one-sided tests procedure has uniformly superior properties to the power approach in most cases. The only cases where the power approach has superior properties when the true averages are equivalent correspond to cases where the chances of concluding equivalence with the power approach when the true averages are not equivalent exceeds 0.05. With appropriate choice of the nominal level of significance of the one-sided tests, the two one-sided tests procedure always has uniformly superior properties to the power approach. The two one-sided tests procedure is compared to the procedure propose by Hauck and Anderson [1984].

Sedlmeier, P., & Gigerenzer, G. (1989) The long-term impact of studies of statistical power is investigated using J. Cohen's (1962) pioneering work as an example. We argue that the impact is nil; the power of studies in the same journal that Cohen reviewed (now the *Journal of Abnormal Psychology*) has not increased over the past 24 years. [...] Low power seems to go unnoticed: only 2 out of 64 experiments mentioned power, and it was never estimated. Nonsignificance was generally interpreted as confirmation of the null hypothesis (if this was the research hypothesis), although the median power was as low as .25 in these cases. We discuss reasons for the ongoing neglect of power.

Sedlmeier, P. (2002) APA review of books: Rosenthal, R., Rosnow, R.L., & Rubin, D.B. (1999). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York: Cambridge University Press. “This is a great book that everybody who uses ANOVA techniques should read”.

Selvin, H.C. (1957) Statistical tests are unsatisfactory in nonexperimental research for two fundamental reasons: It is almost impossible to design studies that meet the conditions for using the tests, and the situations in which the tests are employed make it difficult to draw correct inferences. [...] Even if studies could be designed so that the correlated biases were controlled, there would remain the problem of correctly interpreting the tests. Many users of tests confuse statistical significance with substantive importance or with size of association. Sociologists would do better to re-examine their purposes in using the tests and to try to devise better methods of achieving these purposes than to continue to resort to techniques that are at best misleading for the kinds of empirical research in which they are principally engaged.

Serlin, R.C., & Lapsley, D.K. (1993) After first examining the Meehlian complaints against psychological research in more detail, we then propose a number a remedies. Against the claim that the significance test cannot be made to threaten a theory with refutation, we propose a "good-enough" methodology that claims to do precisely that. Against the claim that psychological research is not cumulative, we argue, following Lakatos, that progress in research is never plainly evident but must instead be excavated from historical reconstructions of the various literatures. Along the way we provide examples of how one uses "good-enough" hypothesis testing. We also argue that the comparison with physics is not always to our disadvantage when the good-enough methodology and certain Lakatos considerations are kept in mind. Finally, we conclude with a discussion of what rational appraisal of psychological research might look like, and how this might have an impact on graduate training in psychology.

Shrout, P.E. (1997) Significance testing of null hypotheses is the standard epistemological method for advancing scientific knowledge in psychology, even though it has drawbacks and it leads to common inferential mistakes. These mistakes include accepting the null hypothesis when it fails to be rejected, automatically interpreting rejected null hypothesis as theoretically meaningful, and failing to consider the likelihood of Type II errors. Although these mistakes have been discussed repeatedly for decades, there is no evidence that the academic discussion has had an impact. A group of methodologists is proposing a new approach: simply ban significance tests in psychology journals. The impact of a similar ban in public-health and epidemiology journals is reported.

Sim, J., & Reid, N. (1999) This article examines the role of the confidence interval (CI) in statistical inference and its advantages over conventional hypothesis testing, particularly when data are applied in the context of clinical practice. A CI provides a range of population values with which a sample statistic is consistent at a given level of confidence (usually 95%). Conventional hypothesis testing serves to either reject or retain a null hypothesis. A CI, while also functioning as a hypothesis test, provides additional information on the variability of an observed sample statistic (ie, its precision) and on its probable relationship to the value of this statistic in the population from which the sample was drawn (ie, its accuracy). Thus, the CI focuses attention on the magnitude and the probability of a treatment or other effect. It thereby assists in determining the clinical usefulness and importance of, as well as the statistical significance of, findings.

The CI is appropriate for both parametric and nonparametric analyses and for both individual studies and aggregated data in meta-analyses. It is recommended that, when inferential statistical analysis is performed, CIs should accompany point estimates and conventional hypothesis tests wherever possible.

Skipper, Jr, J.K., Guenther, A.L., & Nass, G. (1967) There is a need for social scientists to choose levels of significance with full awareness of the implications of Type I and Type II error for the problem under investigation. The current use of arbitrary levels of alpha, while appropriate for some designs, detracts from interpretive power in others. Moreover, the tendency to dichotomy resulting from judging some results "significant" and other "nonsignificant" can be misleading both to professional and lay audiences. It is suggested that a more rational approach might be to report the actual level of significance, placing the burden of interpretive skill upon the reader. Such a policy would also encourage social scientists to give higher priority to selecting appropriate levels of significance for a given problem.

Smithson, M. (2001) The advantages that confidence intervals have over null-hypothesis significance testing have been presented on many occasions to researchers in psychology. This article provides a practical introduction to methods of constructing confidence intervals for multiple and partial R^2 and related parameters in multiple regression models based on "noncentral" F and Chi^2 distributions. Until recently, these techniques have not been widely available due to their neglect in popular statistical textbooks and software. These difficulties are addressed here via freely available SPSS scripts and software and illustrations of their use. The article concludes with discussions of implications forthe interpretation of findings in terms of noncentral confidence intervals, alternative measures of effect size, the relationship between noncentral confidence intervals and power analysis, and the design of studies.

Smithson, M. (2002) ***Table of contents*****:** Ch 1 Introduction and overview. Ch 2 Confidence statements and interval estimates; Why confidence intervals? Ch 3 Central confidence intervals; Central and standardizable versus noncentral distributions; Confidence intervals using the central t and normal distributions; Confidence intervals using the central chi-square and f distributions; Transformation principle. Ch 4 Noncentral confidence intervals for standardized effect sizes; Noncentral distributions;Computing noncentral confidence intervals. Ch 5 Applications in anova and regression; Fixed-effects ANOVA; A priori and post-hoc contrasts; Regression: multiple, partial, and semi-partial correlations; Effect-size statistics for MANOVA and setwise regression; Confidence interval for a regression coefficient; Goodness of fit indices in structural equations models. Ch 6 Applications in categorical data analysis; Odds ratio, Difference between proportions and relative risk; Chi-square confidence intervals for one variable; Two-way contingency tables; Effects in log-linear and logistic regression models. Ch 7 Significance tests and power analysis; Significance tests and model comparison; Power and precision; Designing studies using power analysis and confidence intervals; Confidence intervals for power. Concluding remarks. References.

Snyder, P. (2000) These guidelines are designed to provide authors who submit manuscripts to the journal and reviewers with a uniform set of expectations regarding the reporting of results from statistical investigations. The information contained in this editorial is not exhaustive in relation to issues that might be addressed. As research designs and methods continue to evolve, authors, reviewers, and editors associated with *JEI* will need to engage periodically in dialogue about how inquiry submitted to the journal will be evaluated. We invite your input about issues that you think should be addressed in future editorials.

Spiegelhalter, D.J., Freedman, L.S. (1988) We sumarise current statistical practice in clinical trials, and review Bayesian influence over the past 25 years. It is argued that insufficient attention has been paid to the dynamic context in which development of therapeutic innovations take place, in which *experimenters*, *reviewers* and *consumers* form different interest groups, and may well process the same evidence in different ways. We illustrate the elicitation of quantitative prior opinion in trial design and show how graphical expression of current belief can be related to regions of possible benefit with different clinical implications. Such displays may be used both for ethical monitoring of trials and to predict the consequences of further sampling.

Spiegelhalter, D.J., Freedman, L.S., & Blackburn, P.R. (1986) At an interim point in a clinical trial, trial organisers may wish to use the data on the initial series of patients to judge the likely consequences of further patient accrual. Halperin and colleagues [*Controlled Clinical Trials*, 1982, *3*, 311-323] have suggested calculating the power of a continued trial, *conditional* on the data observed so far and the null and alternative hypothesis specified at the start of the trial, derived by averaging the conditional power with respect to the current belief about the unknown parameters. Although numerical methods are generally required for evaluating the necessary integrals, the results may be presented graphically and enable the statistician to answer the question: "With the data so far, what is the chance that the trial will end up showing a conclusive result?"

Spiegelhalter, D.J., Freedman, L.S., & Parmar, M.K.B. (1994) Statistical issues in conducting randomized trials include the choice of a sample size, whether to stop a trial early and the appropriate analysis and interpretation of the trial results. At each of these stages, evidence external to the trial is useful, but generally such evidence is introduces in an unstructured and informal manner. We argue that a Bayesian approach allows a formal basis for using external evidence an in addition provides a rational way for dealing with issues such

as the ethics of randomization, trials to show treatment equivalence, the monitoring of accumulating data and the prediction of the consequences of continuing a study. The motivation for using this methodology is practical rather than ideological.

Sterne, J.A.C., & Davey Smith, G. (2001) P values, or significance levels, measure the strength of the evidence against the null hypothesis; the smaller the P value, the stronger the evidence against the null hypothesis An arbitrary division of results, into "significant" or "non-significant" according to the P value, was not the intention of the founders of statistical inference A P value of 0.05 need not provide strong evidence against the null hypothesis, but it is reasonable to say that P < 0.001 does. In the results sections of papers the precise P value should be presented, without reference to arbitrary thresholds Results of medical research should not be reported as "significant" or "non-significant" but should be interpreted in the context of the type of study and other available evidence. Bias or confounding should always be considered for findings with low P values To stop the discrediting of medical research by chance findings we need more powerful studies.

Sterne, J.A.C. (2002) Confusion in the teaching of statistical inference dates back to the conflict of Fisher's P-values and significance tests with the Neyman-Pearson hypothesis testing approach. To avoid the well-known pitfalls arising from over-reliance on significance tests and the division of results into 'significant' or 'not significant', many medical journals now insist that presentation of statistical analyses includes confidence intervals as well as or instead of P-values. The confusion over how to report statistical analyses which is evident in the recent medical literature is matched by divergent teaching of hypothesis tests between the 16 U.K. medical schools represented at the April 2000 Burwalls meeting. Suggested guidelines for the teaching of statistical inference to medical students are presented, and possible future developments are discussed.

Student (1908) **[*Example of interpretation of significance levels in terms of probabilities about parameters*]** "From the table the probability is .9985 or the odds are about 666 to 1 that 2 [the second soporific] is the better soporific." (page 21)

Sylvester, R.J. (1988) A new strategy for the design of Phase II clinical trials is presented which utilizes the information provided by the prior distribution of the response rates, the costs of treating a patient, and the losses or gains resulting from the decisions taken at the completion of the study. A risk function is derived from which one may determine the optimal Bayes sampling plan. The decision theoretic/Bayesian approach is shown to provide a formal justification for the sample sizes used in practice and shows the conditions under which such sample sizes are clearly inappropriate.

Taube, A. (1980) By means of data from ficticious cross over trials, it is first demonstrated that a statistically significant difference is not necessarily of a practically important order of magnitude. This fact is of special interest when the number of observations is large. Second, a statistically non significant difference does not prove the hypothesis about equality between, say, treatment effects. This fact is of special interest when the number of observations is small. For investigating whether equality is possible, confidence intervals are more useful than non significant results from tests of significance.

Thompson, B. (1994) 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-square*, *R-square*, *eta-square*) or corrected (e.g., adjusted *R-square*, *omega-square*) variance-accounted-for statistics.

Thompson, B. (1994) Too few researchers understand what statistical significance testing does and doesn't do, and consequently their results are misinterpreted. Even more commonly, researchers understand elements of statistical significance testing, but the concept is not integrated into their research. For example, the influence of sample size on statistical significance may be acknowledged by a researcher, but this insight is not conveyed when interpreting results in a study with several thousand subjects.
This Digest will help you better understand the concept of significance testing. The meaning of probabilities, the concept of statistical significance, arguments against significance testing, misinterpretation, and alternatives are discussed.

Thompson, B. (2001) The author asserts that editors should publicly declare their expectations and expose the rationales for editorial policies to public scrutiny. He argues that editorial policies ought to require effect size reporting, as those at 17 journals now do. He also argues (a) that score reliabilities should be reported; (b) that stepwise methods should not be used; (c) that structure coefficients should be interpreted; and (d) that if used wisely, confidence intervals differ from hypothesis tests in important ways. The use of noncentral t and F distributions to create confidence intervals about effect sizes also is appealing.

Thompson, B. (2001) The following three brief articles extend the discussion in the previous article regarding future prospects for progress in researchers' reporting and interpreting effect sizes. The authors of these brief pieces represent diverse views. [...] These three articles also serve as a useful precursor to the series of articles by Australian scholars that will be published in the August issue. These articles treat various issues in computing "noncentral" confidence intervals about effect sizes and related estimates.

Trafimow, D. (2003) Because the probability of obtaining an experimental finding given that the null hypothesis is true [$p(F|H0)$] is not the same as the probability that the null hypothesis is true given a finding [$p(H0|F)$], calculating the former probability does not justify conclusions about the latter one. As the standard null-hypothesis significance-testing procedure does just that, it is logically invalid (J. Cohen, 1994). Theoretically, Bayes's theorem yields $p(H0|F)$, but in practice, researchers rarely know the correct values for 2 of the variables in the theorem. Nevertheless, by considering a wide range of possible values for the unknown variables, it is possible to calculate a range of theoretical values for $pH0|F)$ and to draw conclusions about both hypothesis testing and theory evaluation.

Tryon, W.W. (1998) "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 [...]" (page 796)

Tryon, W.W. (2001) Null hypothesis statistical testing (NHST) has been debated extensively but always successfully defended. The technical merits of NHST are not disputed in this article. The widespread misuse of NHST has created a human factors problem that this article intends to ameliorate. This article describes an integrated, alternative inferential confidence interval approach to testing for statistical difference, equivalence, and indeterminacy that is algebraically equivalent to standard NHST procedures and therefore exacts the same evidential standard. The combined numeric and graphic tests of statistical difference, equivalence, and indeterminacy are designed to avoid common interpretive problems associated with NHST procedures. Multiple comparisons, power, sample size, test reliability, effect size, and cause-effect ratio are discussed. A section on the proper interpretation of confidence intervals is followed by a decision rule summary and caveats.

Vacha-Haase, T., Nilsson, J.E., Reetz, D.R., Lance, T.S., & Thompson, B. (2000) The recent fourth edition of the American Psychological Association Publication Manual emphasized that p values are not acceptable indices of effect and "encouraged" effect-size reporting. However, empirical studies of reporting practices of diverse journals unequivocally indicate that this new encouragement has to date been ineffective. Here two additional multi-year studies of APA journals are reported. Additionally, all 50 APA editorials that have been published since 1990 were reviewed to determine how many editors with approval have articulated policies more forceful than the APA Publication Manual's vague and seemingly self-canceling encouragement. It is suggested that changes in editorial policies will be required before improved reporting will become routine.

Valera, A., Sánchez, J., & Marin, F. (1997) Several proposals that enable to complement the information offered in statistical hypothesis testing are described. using these proposals reduce the most hard critic that significance tests have suffered. Significance tests do not offer information about the magnitude of the relationship among the involved variables. the proposals that are discussed in this paper are confidence intervals, effect size, binomial effect size display, counter-null value and common language effect size indicator.

Valera, A., Sánchez, J., & Marin, F. (2000) The purpose of this paper was to analyse the application of the most common statistical procedure for studying relationships among variables and empirical phenomena in psychology: The null hypothesis statistical test. In order to determine whether its use in psychological Spanish research is adequate, we carried out a power study of the papers published in Spanish journals. The analysis of the 169 experiments selected, with a total of 5,480 statistical tests, showed power values of 0.18, 0.58, 0.83, and 0.59 to low, medium, high, and estimated effect sizes, respectively. These values drastically decreased in about a 20% when the calculations were repeated controlling the Type I error inflation through Bonferroni adjustment. The results were very similar to those obtained in other international power studies and lead us to think about the need for a special attention for controlling the statistical power in designing a research. On the other hand, we discuss several complementary proposals to the use of significance tests that may improve the information obtained.

Valera, A., Sánchez, J., Marin, F., & Velandrino, A. (1998) Although the purpose of hypothesis testing is to reject the null hypothesis and to detect relationships among variables, the inadequate control of statistical power is very common in psychological research. In this way, the statistical power of the papers published in the *Revista de Psicología General y Applicada* since 1990 to 1992 was analyzed. The power for low, medium, high, and estimated effect sizes were computed. The values we found were .17, .57, .83, and .55, respectively. Moreover, the distribution of effect magnitudes and sample sizes in the journal papers was also analyzed. Finally, the results are discussed and compared with those of the other power studies in the literature.

VanVoorhis, W.C., & Morgan, B.L. (2001) In this article we highlight the statistical rules of thumb guiding the selection of sample sizes for detecting differences, associations, chi-square, and factor analyses.

Vargha, A., & Delaney, H.D. (2000) McGraw and Wong (1992) described an appealing index of effect size, called *CL*, which measures

the difference between two populations in terms of the probability that a score sampled at random from the first population will be greater than a score sampled at random from the second. McGraw and Wong introduced this "common language effect size statistic" for normal distributions and then proposed an approximate estimation for any continuous distribution. In addition, they generalized CL to the *n*-group case, the correlated sample case, and the discrete value case.
In the current paper a different generalization of *CLI*, called the *A* measure of stochastic superiority, is proposed, which may be directly applied for any discrete or continuous variable that is at least ordinally scaled. Exact methods for point and interval estimation as well as the significance tests of the *A*=.5 hypothesis are provided. New generalizations of *CL* are provided for the multi-group and correlated samples cases.

Victor, N. (1987) The currently usual one value for the judgement of the clinical relevance of therapeutic effects frequently does not suffice to adequately formulate the problems of clinical studies, and the statistical standard procedure (the testing of the classical nullhypothesis) fails to take this value duly into account. Therefore, 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. Methodological problems resulting from the shifting of the nullhypothesis are discussed, and other possibilities to take into account the clinically relevant difference (introduction of criteria of success) are considered.

Wade, O.L., & Waterhouse, J.A.H. (1977) 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 *statistically significant* 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.

Wang, Y.H. (2000) A general explanation of the fiducial confidence interval and its construction for a class of parameters in which the distributions are stochastically increasing or decreasing is provided. Major differences between the fiducial interval and Bayesian and frequentist intervals are summarized. Applications of fiducial inference in evaluating pre-data frequentist intervals and general post-data intervals are discussed.

Wellek, S., & Michaelis, J. (1991) The paper outlines an approach to the general methodological problem of equivalence assessment which is based on the classical theory of testing statistical hypotheses. Within this frame of reference it is natural to search for decisions rules satisfying the same criteria of optimality which are customarily applied in deriving solutions to one- and two-sided testing problems. For three standard situations very frequently encountered in medical applications of statistics, a concise account of such an optimal test for equivalence is presented. It is pointed out that tests based on the well-known principle of confidence interval inclusion are valid in the sense of guaranteeing the prespecified level of significance, but tend to have an unnecessarily low efficiency.

Windeler, J., & Conradt, C. (2000) Clinical trials are aimed at providing results which enable improvements in patient care. It is widely criticized, however, that the characterization of results as "significant" or "non-significant" does not allow any assessment of their clinical relevance. To counter this criticism 2 biostatistical concepts are available: the use of confidence intervals and the application of statistical tests with shifted null-hypotheses. Possibilities and limitations of these concepts are discussed in this contribution.

Wilkinson, L. and Task Force on Statistical Inference In the light of continuing debate over the applications of significance testing in psychology journals and following the publication of Cohen's (1994) article, the Board of Scientific Affairs (BSA) of the American Psychological Association (APA) convened a committee called the Task Force on Statistical Inference (TFSI) whose charge was "to elucidate some of the controversial issues surrounding applications of statistics including significance testing and its alternatives; alternative underlying models and data transformation; and newer methods made possible by powerful computers" (BSA, personal communication, February 28, 1996). Robert Rosenthal, Robert Abelson, and Jacob Cohen (cochairs) met initially and agreed on the desirability of having several types of specialists on the task force: statisticians, teachers of statistics, journal editors, authors of statistics books, computer experts, and wise elders. Nine individuals were subsequently invited to join and all agreed. These were Leona Aiken, Mark Appelbaum, Gwyneth Boodoo, David A. Kenny, Helena Kraemer, Donald Rubin, Bruce Thompson, Howard Wainer, and Leland Wilkinson. In addition, Lee Cronbach, Paul Meehl, Frederick Mosteller and John Tukey served as Senior Advisors to the Task Force and commented on written materials.
The TFSI met twice in two years and corresponded throughout that period. After the first meeting, the task force circulated a preliminary report indicating its intention to examine issues beyond null hypothesis significance testing. The task force invited comments and used this feedback in the deliberations during its second meeting.
After the second meeting, the task force recommended several possibilities for further action, chief of which would be to revise the statistical sections of the *American Psychological Association Publication Manual* (1994). After extensive discussion, the BSA recommended that "before the TFSI undertook a revision of the *APA Publication Manual*, it might want to consider publishing an article in *American Psychologist*, as a way to initiate discussion in the field about changes in current practices of data analysis and reporting".

This report follows that request. The sections in italics are proposed guidelines that the TFSI recommends could be used for revising the APA publication manual or for developing other BSA supporting materials. Following each guideline are comments, explanations, or elaborations assembled by Leland Wilkinson for the task force and under its review. This report is concerned with the use of statistical methods only and is not meant as an assessment of research methods in general. Psychology is a broad science. Methods appropriate in one area may be inappropriate in another.
Power and sample size. *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. Once the study is analyzed, confidence intervals replace calculated power in describing results.*
Hypothesis tests. *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.*
Effect sizes. *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 (r or d). It helps to add brief comments that place these effect sizes in a practical and theoretical context.*
Interval estimates. *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.*

Williams, A.M. (1998) Throughout introductory tertiary statistics subjects, students are introduced to a multitude of new terms for statistical concepts and procedures. One such term, significance level, has been considered in the statistical literature. Three themes of discussion relate to this concept – the problem of interpretation (and misinterpretation), the selection of an appropriate level, and the evaluation of results based on significance level. However, empirical research regarding this concept is very limited. This paper reports on a qualitative study which used concept maps and standard hypothesis tests to investigate student's conceptual and procedural knowledge of the significance level concept. Eighteen students completing an introductory tertiary statistics subject were interviewed after their final exam in statistics. Results showed that many students did not have a good understanding of the concept.

Williams, V.S.L., Jones, L.V., & Tukey, J.W. (1999) Three alternative procedures to adjust significance levels for multiplicity are the traditional Bonferroni technique, a sequential Bonferroni technique developed by Hochberg (1988), and a sequential approach for controlling the false discovery rate proposed by Benjamini and Hochberg (1995). These procedures are illustrated and compared using examples from the National Assessment of Educational Progress (NAEP). A prominent advantage of the Benjamini and Hochberg (B-H) procedure, as demonstrated in these examples, is the greater invariance of statistical significance for given comparisons over alternative family sizes. Simulation studies show that all three procedures maintain a false discovery rate bounded above, often grossly, by "alpha" (or "alpha"/2). For both uncorrelated and pairwise families of comparisons, the B-H technique is shown to have greater power than the Hochberg or the Bonferroni procedures, and its power remains relatively stable as the number of comparisons becomes large, giving it an increasing advantage when many comparisons are involved. We recommend that results from NAEP State Assessments be reported using the B-H technique rather than the Bonferroni procedure.

Wilson, G. (2003) For the last 50 years Bayesians and frequentists have disputed the appropriate way to do statistics. Bayesian methods have grown in popularity and acceptance, but how is the con1ict between Bayesians and frequentists likely to play out in the future? This article uses theories advanced by Thomas Kuhn and Lawrence Grossberg to offer a framework for understanding possible futures and to pose questions about the future of the field of statistics.

Winch, R.F., & Campbell, D.T. (1969) To do or not to do a test of significance – that is a question that divides men of good will and sound competence. 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.

Winkler, R.L. (1974) [...] The gap between theory and practice in statistical analysis is investigated, with particular attention given to the Bayesian approach to statistical analysis. [...] Current statistical practice in experimental psychology and various factors contributing to the theory-practice gap in statistical analysis are considered. Finally, some general questions involving scientific reporting and the use of Bayesian procedures in statistical inference are discussed.

Witehead, J. (1993) The title of this paper was chosen for me by the organizers of the conference. [...] A mote accurate title for my paper

would be "the case for frequentism in definitive phase III clinical trials". In fact, part of the paper could even be entitled "the case for Bayesianism in early phase clinical trials". Each approach has its place. In keeping with the requirements for robust arguments, I shall simplify my opinions into clear terms of black and white, rather than writing at greater length in various shades of grey. I shall begin with an account of the frequentist statements which can be made at the end of any scientific investigation. This will be made in an abstract setting. Then I shall briefly explain my own understanding of Bayesian approaches. Implementation of the two philosophies in the different phases of clinical research will be discussed, and I shall end with some remarks concerning the preservation of scientific rigour in clinical research.

Yates, F.(1951) "The emphasis given to formal tests of significance throughout [R.A. Fisher's] Statistical Methods [...] has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating." [...] "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."

Yoccoz, N.G. (1991) "In marked contrast to what is advocated by most statisticians, most evolutionary biologists and ecologists overemphasize the potential role of significance testing in their scientific practice. Biological significance should be emphasized rather than statistical significance. Furthermore, a survey of papers showed that the literature is infiltrated by an array of misconceptions about the use and interpretation of significance tests." [...] "By far the most common error is to confound statistical significance with biological, scientific significance. [...] " [...] "Statements like 'the two populations are significantly different relative to parameter X (P=.004)' are found with no mention of the estimated difference. The difference is perhaps statistically significant at the level .004, but the reader has no idea is if it is biologically significant." [...] "Most biologists and other users of statistical methods still seem to be unaware that significance testing by itself sheds little light on the questions they are posing."

Zeisel, H. (1955)] There is now, in the social sciences no greater need than the development of theoretical insights guided by empirical data. At such times, to provide this guidance and serve as a stimulant is the significance of statistically insignificant data. Even if the probability is great that an inference will have to be rejected later, the practical risk of airing is small. Subsequent and more elaborate studies may disprove some of these inferences; but for those that survive social science will be the richer.

Zuckerman, M., Hodgins, H., Zuckerman, A., & Rosenthal, R. (1993) We asked active psychological researchers to answer a survey regarding the following data-analytic issues: (a) the effect of reliability on Type I and Type II errors, (b) the interpretation of interaction, (c) contrast analysis, and (d) the role of power and effect size in successful replications. Our 551 participants (a 60% response rate) answered 59% of the questions correctly; 46% accuracy would be expected according to participants' response preferences alone. Accuracy was higher for respondents with higher academic ranks and for questions with "no" as the right answer. It is suggested that although experienced researchers are able to answer difficult but basic data-analytic questions at better than chance levels, there is also a high degree of misunderstanding of some fundamental issues of data analysis.