How Do We Do Hypothesis Testing?

Jeff Gill, jgill@ucdavis.edu

1 The Current Paradigm: Null Hypothesis Significance Testing

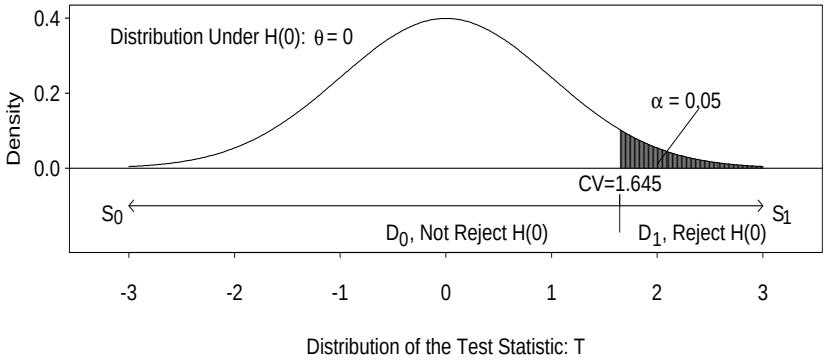
The current, nearly omnipresent, approach to hypothesis testing in all of the social sciences is a synthesis of the Fisher *test of significance* and the Neyman-Pearson *hypothesis test*. In this "modern" procedure, two hypotheses are posited: a null or restricted hypothesis (H_0) which competes with an alternative or research hypothesis (H_1) describing two complementary notions about some phenomenon. The research hypothesis is the probability model which describes the author's belief about some underlying aspect of the data, and operationalizes this belief through a parameter: θ. In the simplest case, described in every introductory text, a null hypothesis asserts that $\theta = 0$ and a complementary research hypothesis asserts that $\theta \neq 0$. More generally, the test evaluates a parameter vector: $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$, and the null hypothesis places restrictions on some subset ($\ell \leq m$) of the theta vector such as: $\theta_i = k_1\theta_j + k_2$ with constants k_1 and k_2.

A test statistic (T), some function of θ and the data, is calculated and compared with its known distribution under the assumption that H_0 is true. Commonly used test statistics are sample means ($\bar{X}$), chi-square statistics (χ^2), and t-statistics in linear (OLS) regression analysis. The test procedure assigns one of two decisions (D_0, D_1) to all possible values in the sample space of T, which correspond to supporting either H_0 or H_1 respectively. The p-value ("associated probability") is equal to the area in the tail (or tails) of the assumed distribution under H_0 which starts at the point designated by the placement of T on the horizontal axis and continues to infinity. If a predetermined α level has been specified, then H_0 is rejected for p-values less than α, otherwise the p-value itself is reported. Formally, the sample space of T is segmented into two complementary regions (S_0, S_1) whereby the probability that T falls in S_1, causing decision D_1, is either a predetermined null hypothesis cumulative distribution function (CDF) level: the probability of getting this or some lower value given a specified parametric form such as normal, F, t, etc. (α = size of the test, Neyman and Pearson), or the cumulative distribution function level corresponding to the value of the test statistic under H_0 is reported (p-value = $\int_{S_1} P_{H_0}(T = t)dt$, Fisher). Thus decision D_1 is made if the test statistic is sufficiently atypical given the distribution under H_0. This process is illustrated for a one tail test at $\alpha = 0.05$ in Figure 1.

2 Historical Development

The current null hypothesis significance test is synthesis of two highly influential but incompatible schools of thought in modern statistics. Fisher developed a procedure that produces significance levels from the data whereas Neyman and Pearson posit an intentionally rigid decision process which seeks to confirm or reject specified a priori hypotheses. The null hypothesis significance testing procedure is not influenced by the third major intellectual

Figure 1: Null Hypothesis Significance Testing Illustrated



stream of the time: Bayesianism, except as a reaction against this approach.

2.1 Fisher Test of Significance

Fisher (1925a, 1934, 1955) posited a single hypothesis, H_0, with a known distribution of the test statistic T. As the test statistic moves away from its conditional expected value, $E(T|H_0)$, H_0 becomes progressively less plausible (less likely to occur by chance). The relationship between T and the level of significance produced by the test is established by the density outside the threshold established by T (one or two tailed), going away from the density region containing the expected value of T given H_0. The outside density is the p-value, also called the *achieved* significance level. Fisher hypothesis testing is summarized by the following steps:

1. Identify the null hypothesis.
2. Determine the appropriate test statistic and its distribution under the the assumption that the null hypothesis is true.
3. Calculate the test statistic from the data.
4. Determine the achieved significance level that corresponds to the test statistic using the distribution under the assumption that the null is true.
5. Reject H_0 if the achieved significance level is sufficiently small. Otherwise reach no conclusion.

This construct naturally leads to the question of what p-value is sufficiently small as to warrant rejection of the null hypothesis. Although Fisher wrote in later years that this this threshold should be established by the context of the problem his influential work is full of phrases such as: "The value for Q is therefore significant on the higher standard (1 per cent) and that for N_2 at the lower standard (5 per cent)." (1971, p.152-3). Furthermore, this determination of significance levels at 0.01 or 0.05 was made by Fisher in the context of agricultural and biological experiments. This was partly defensive on Fisher's part as rigid significance threshold's are more conducive to Neyman-Pearson hypothesis testing: "In an acceptance procedure, on the other hand, acceptance is irreversible, whether the evidence for it was strong or weak. It is the result of applying mechanically rules laid down in advance;

no *thought* is given to the particular case, and the tester's state of mind, or his capacity for *learning*, is inoperative." (Fisher 1955, p.73-4).

2.2 Neyman and Pearson Hypothesis Testing

Neyman and Pearson (1928a, 1928b, 1933b, 1936a) reject Fisher's idea that only the null hypothesis needs to be tested. They argue that a more useful procedure is to propose two complementary hypotheses: Θ_A and Θ_B (or a class of Θ_{Bi}), which need not be labeled "null" or "alternative" but often are purely for convenience. Furthermore, Neyman and Pearson (1933b) point out that that one can posit a hypothesis and consecutively test multiple admissible alternatives against this hypothesis. Since there are now two competing hypotheses in any one test, Neyman and Pearson can define an a priori selected α, the probability of falsely rejecting Θ_A under the assumption that H_0 is true, and β, the probability of failing to reject Θ_A when H_0 is false. By convention, the first mistake is called a Type I error, and the second mistake is called a Type II error. Note that α and β are probabilities *conditional* on two mutually exclusive events: α is conditional on the null hypothesis being true, and β is conditional on the null hypothesis being false. A more useful quantity than β is $1-\beta$, which Neyman and Pearson (1933a, 1936a) call the *power* of the test: the long run probability of accurately rejecting a false null hypothesis given a point alternative hypothesis.

In this construct it is desirable to develop the test which has the highest power for a given a priori α. To accomplish this goal, the researcher considers the fixed sample size, the desired significance level, and the research hypothesis, then employs the test with the greatest power. Neyman and Pearson's famous lemma (1936b) shows that under certain conditions there exists a "uniformly most powerful" test which has the greatest possible probability of rejecting a false null hypothesis in favor of a point alternative hypothesis, compared to other tests. A sufficient condition is that the probability density tested has a monotone likelihood ratio. Suppose we have family of probability density functions $h(t|\theta)$ in which the random variable t is conditional on some unknown θ value to be tested. This family has a monotone likelihood ratio if for every $\theta_1 > \theta_2$, then: $\frac{h(t|\theta_1)}{h(t|\theta_2)}$ is a non-decreasing function of the random variable t. Suppose further that we perform a test such as H_0: $\theta_1 \leq \theta_2$ versus H_1: $\theta_1 > \theta_2$ (θ_2 a known constant), where t is a sufficient statistic for θ_1, and $h(t|\theta_1)$ has a monotone likelihood ratio. The Karlin-Rubin Theorem (1956) states that if we set $\alpha = P(t > t_0)$ and reject H_0 for an observed $t > t_0$ (t_0 a known constant), then this test has the most power relative to any other possible test of H_0 with this α level (Casella and Berger 1990: 366-70, Lehmann 1986: 78).

To contrast the Neyman-Pearson approach with Fisher's test of significance, note how different the the following steps are from Fisher's:

1. Identify a hypothesis of interest, Θ_B, and a complementary hypothesis, Θ_A.
2. Determine the appropriate test statistic and its distribution under the assumption that Θ_A is true.
3. Specify a significance level (α), and determine the corresponding critical value of the test statistic under the assumption that Θ_A is true.
4. Calculate the test statistic from the data.

5. Reject Θ_A and accept Θ_B if the test statistic is further than the critical value from the expected value of the test statistic (calculated under the assumption that Θ_A is true). Otherwise accept Θ_A.

The Neyman-Pearson approach is important in the context of decision theory where the decision in the final step above is assigned a *risk function* computed as the expected loss from making an error.

2.3 Producing The Synthesis

The null hypothesis significance test attempts to blend the two approaches described above producing the "synthesis." With Fisher hypothesis testing, no explicit complementary hypothesis to H_0 is identified, and the p-value that results from the model and the data is evaluated as the strength of the evidence for the research hypothesis. Therefore there is no notion of the power of the test nor of accepting alternate hypotheses in the final interpretation. Conversely, Neyman-Pearson tests identify complementary hypotheses: Θ_A and Θ_B in which rejection of one implies acceptance of the other and this rejection is based on a predetermined α level. Some people are surprised, but Neyman and Pearson actually do use the word "accept", see 1933b.

Neyman and Pearson's hypothesis test defines the significance level a priori as a function of the *test* (i.e. before even looking at the data), whereas Fisher's test of significance defines the significance level afterwards as function of the *data*. The current paradigm in the social sciences straddles these two approaches by pretending to select α a priori, but actually using p-values (or asterisks next to test statistics indicating ranges of p-values) to evaluate the strength of the evidence. This allows inclusion of the alternate hypothesis but removes the search for a more powerful test.

The synthesized test also attempts to reconcile the two differing perspectives on how the hypotheses are defined. It adopts the Neyman-Pearson convention of two explicitly stated rival hypotheses, but one is always labeled as the null hypothesis as in the Fisher test. In some introductory texts the null hypothesis is presented only as a null relationship: $\theta = 0$ (no effect), whereas Fisher intended the null hypothesis simply as something to be "nullified". The synthesized test partially uses the Neyman-Pearson decision process except that failing to reject the null hypothesis is treated as a quasi-decision: "modest" support for the null hypothesis assertion. There is also confusion in the synthesized test about p-values and long-run probabilities. Since the p-value, or range of p-values indicated by stars, is not set a priori, it is not the long-run probability of making a Type I error but is typically treated as such. The synthesized test thus straddles the Fisher interpretation of p-values from the data and the Neyman-Pearson notion of error probabilities from the test. It is very interesting to note that with the synthesized modern hypothesis test there is no claim of authorship. The acrimony, both intellectual and personal, between Fisher and Neyman & Pearson is legendary and continued until Fisher's death. So it is curious that no one was willing to claim responsibility for a potentially bridging approach and it appeared in the textbooks anonymously (Gigerenzer 1987: 21). The timidity of these authors led them to try and accommodate both perspectives by denying that there were substantive differences.

Many problems with the current paradigm result from the mixture of these two essentially incompatible approaches (Gigerenzer et.al. 1989, Gigerenzer 1993, Gigerenzer and Murray

1987, MacDonald 1997). While both approaches seek to establish that some observed relationship is attributable to effects distinct from sampling error, there are important differences as noted above. *Neither Fisher nor Neyman and Pearson would have been satisfied with the synthesis.* Fisher objected to preselection of the significance level as well as the mandatory two-outcome decision process. Neyman and Pearson disagreed with interpreting p-values (or worse yet, ranges of p-values indicated by "stars") as the probability of Type I errors since they do not constitute a long-range probability of rejection. Neyman and Pearson also considered the interpretation of data-derived p-values to be subjective and futile (Neyman and Pearson 1933b, footnote 1).

3 Status and Importance

Led in the social sciences by psychology, many are challenging the basic tenets of the way that nearly all social scientists are trained to develop and test empirical hypotheses. It has been described as a "strangle-hold" (Rozeboom 1960), "deeply flawed or else ill-used by researchers" (Serlin and Lapsley 1993), "a terrible mistake, basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (Meehl 1978), "an instance of the kind of essential mindlessness in the conduct of research" (Bakan 1960), "badly misused for a long time" (Cohen 1994), and that it has "systematically retarded the growth of cumulative knowledge" (Schmidt 1996). Or even more bluntly: "The significance test as it is currently used in the social sciences just does not work." (Hunter 1997).

Statisticians have long been aware of the limitations of null hypothesis significance testing as currently practiced in political science research. Jeffreys (1961) observed that using p-values as decision criteria is backward in its reasoning: "a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred." Another common criticism notes that this interpretation of hypothesis testing confuses inference and decision making since it "does not allow for the costs of possible wrong actions to be taken into account in any precise way" (Barnett 1973). The perspective of many statisticians toward null hypothesis significance testing is typified by the statement: "a P-value of 0.05 essentially does not provide any evidence against the null hypothesis (Berger, Boukai, and Wang 1997), and the observation that the null versus research hypothesis is really an "artificial dichotomy" (Gelman et.al. 1995). Berger and Sellke (1987) show that evidence against the null given by correctly interpreting the posterior distribution or corresponding likelihood function "can differ by an order of magnitude."

The basic problem with the null hypothesis significance test in political science is that it often does not tell political scientists what they think it is telling them. Most of the problems discussed here are interpretive in that they highlight misconceptions about the results of the procedure. From the current presentation of null hypothesis significance testing in published work it is very easy to confuse statistical significance with theoretical or substantive importance. It is also possible to have data which tells us something about an important political question, but which does not pass an arbitrary significance level threshold. In this circumstance, one would still want to know that there is empirical evidence of some phenomenon of interest. There is nothing mathematically wrong with p-values in these circumstances: they

are simply not sufficient statistics for actual theoretical importance.

It is important to know that there exist effective alternatives which require only modest changes in empirical methodology: confidence intervals, Bayesian estimation, and meta-analysis. Confidence intervals are readily supplied by even the simplest of statistical computing packages, and require little effort to interpret. Bayesian estimation eliminates many of the pathologies described, albeit with a greater setup cost. Meta-analysis is sometimes a complex process, but it offers the potential benefit of integrating and analyzing a wider scope of work on some political question.

4 References

Bakan, David. 1960. "The Test of Significance in Psychological Research." *Psychological Bulletin* 66, 423-37.

Barnett, Vic. 1973. *Comparative Statistical Inference.* New York: John Wiley & Sons.

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

Berger, James O. 1984. "The Robust Bayesian Viewpoint (with discussion)." In Joseph B. Kadane, ed. *Robustness of Bayesian Analysis.* Amsterdam: North Holland.

Berger, James O., B. Boukai, and Y. Wang. 1997. "Unified Frequentist and Bayesian Testing of a Precise Hypothesis." *Statistical Science* 12, 133-60.

Berger, James O., and Thomas Sellke. 1987 "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence." *Journal of the American Statistical Society* 82, 112-22.

Bernardo, José M. 1984. "Monitoring the 1982 Spanish Socialist Victory: A Bayesian Analysis." *Journal of the American Statistical Society* 79, 510-5.

Brenner-Golomb, Nancy. 1993. "R. A. Fisher's Philosophical Approach to Inductive Inference." In G. Keren, and C. Lewis, eds. *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Box, George E. P., and George C. Tiao. 1992. *Bayesian Inference in Statistical Analysis.* New York: John Wiley & Sons.

Carver, Ronald P. 1978. "The Case Against Statistical Significance Testing." *Harvard Education Review.* 48, 378-99.

Casella, George, and Roger L. Berger. 1990. *Statistical Inference.* Belmont, CA: Wadsworth & Brooks/Cole.

Chow, Siu L. 1996. *Statistical Significance: Rationale, Validity and Utility.* London: Sage.

Cleveland, William. 1993. *Visualizing Data.* Summit, NJ: Hobart Press.

Cohen, Jacob. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* December, 12, 997-1003.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112, 115-59.

Cohen, Jacob. 1977. *Statistical Power Analysis for the Behavioral Sciences.* Second Edition. New York: Academic Press.

Cohen, Jacob. 1962. "The Statistical Power of Abnormal-Social Psychological Research: A Review." *Journal of Abnormal and Social Psychology* 65, 145-53.

Cooper, Harris 1984. *The Integrative Research Review: A Systematic Approach.* Beverly Hills: Sage.

Falk, R., and C. W. Greenbaum. 1995. "Significance Tests Die Hard." *Theory and Psychology* 5, 396-400.

Fearon, James D. 1991 "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43, No. 2. (January), 169-95.

Fisher, Sir Ronald A. 1971. *The Design of Experiments.* Ninth Edition. New York: Hafner Press.

Fisher, Sir Ronald A. 1955. "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society* B, 17, 69-78.

Fisher, Sir Ronald A. 1934. *The Design of Experiments.* First Edition. Edinburgh: Oliver and Boyd.

Fisher, Sir Ronald A. 1925a. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Fisher, Sir Ronald A. 1925b. "Theory of Statistical Estimation." *Proceedings of the Cambridge Philosophical Society* 22, 700-25.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis.* New York: Chapman & Hall.

Gigerenzer, Gerd. 1993. "The Superego, the Ego, and the Id in Statistical Reasoning." In G. Keren, and C. Lewis, eds. *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, Gerd. 1987. "Probabilistic Thinking and the Fight Against Subjectivity." In Krüger, Lorenz, Gerd Gigerenzer, and Mary Morgan, eds. *The Probabilistic Revolution.* Volume 2. Cambridge, MA: MIT.

Gigerenzer, Gerd., and D. J. Murray. 1987. *Cognition as Intuitive Statistics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, Gerd., Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, Lorenz Krüger. 1989. *The Empire of Chance.* Cambridge: Cambridge University Press.

Gill, Jeff, and James Thurber. 1999. "Congressional Tightwads and Spendthrifts: Measuring Fiscal Behavior in the Changing House of Representatives." *Political Research Quarterly*, Forthcoming.

Glass, Gene V. 1976. "Primary, Secondary and Meta-Analysis of Research." *Educational Researcher* 5, 3-8.

Glass, G..V., B. McGaw, and M. L. Smith. 1981. *Meta-Analysis in Social Research.* Beverly Hills: Sage.

Greenwald, Anthony G. 1975. "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin* 82, 1-20.

Howson, Colin, and Peter Urbach. 1993. *Scientific Reasoning: The Bayesian Approach.* Second Edition. Chicago: Open Court.

Hunter, John E. 1997. "Needed: A Ban on the Significance Test." *Psychological Science* January, Special Section 8, 3-7.

Hunter, John E., and Frank L. Schmidt. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Beverly Hills: Sage.

Jacoby, William. 1997. "Statistical Graphics for Univariate and Bivariate Data." Beverly Hills: Sage.

Jeffreys, Harold. 1961. *The Theory of Probability.* Oxford: Clarendon Press.

Karlin, S., and Rubin H. 1956. "The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio." *Annals of Mathematical Statistics* 27, 272-99.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* XXVIII, No. 3 (September) 443-499.

King, Gary. 1989 *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* New York: Cambridge. University Press, 1989.

King, Gary, Michael Tomz, and Jason Wittenberg. 1998. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." Political Methodology Working Paper Archive: `http://polmeth.calpoly.edu`.

Leamer, Edward E. 1983. "Lets Take the Con Out of Econometrics." *American Economic Review* 73, No. 1 (March), 31-43.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: John Wiley & Sons.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses.* Second Edition. New York: Springer.

Lindsay, R. M. 1995. "Reconsidering the Status of Tests of Significance: An Alternative Criterion of Adequacy." *Accounting, Organizations and Society* 20, 35-53.

Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65, No. 3 (September), 682-93.

Macdonald, Ranald R. 1997. "On Statistical Testing in Psychology." *British Journal of Psychology* 88, No. 2 (May), 333-49.

Meehl, Paul E. 1990. "Why Summaries of Research on Psychological Theories Are Often Uninterpretable." *Psychological Reports* 66, 195-244.

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Counseling and Clinical Psychology* 46, 806-34.

Meier, Kenneth. 1997. "The value of replicating social-science research." *The Chronicle of Higher Education* 43 (February 7), 22.

Miller, Alan J. 1990. *Subset Selection in Regression.* New York: Chapman & Hall.

Neyman, Jerzy, and Egon S. Pearson. 1928a. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part I" *Biometrika* 20A, 175-240.

Neyman, Jerzy, and Egon S. Pearson. 1928b. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part II" *Biometrika* 20A, 263-94.

Neyman, Jerzy, and Egon S. Pearson. 1933a. "On the Problem of the Most Efficient Test of Statistical Hypotheses." *Philosophical Transactions of the the Royal Statistical Society* A 231, 289-337.

Neyman, Jerzy, and Egon S. Pearson. 1933b. "The Testing of Statistical Hypotheses in Relation to Probabilities *a priori*." *Proceedings of the Cambridge Philosophical Society* 24, 492-510.

Neyman, Jerzy, and Egon S. Pearson. 1936a. "Contributions to the Theory of Testings Statistical Hypotheses." *Statistical Research Memorandum* 1, 1-37.

Neyman, Jerzy, and Egon S. Pearson. 1936b. "Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses." *Statistical Research Memorandum* 1, 113-37.

Oakes, M. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences.* New York: John Wiley & Sons.

Popper, Karl. 1968. *The Logic of Scientific Discovery.* New York: Harper and Row.

Pollard, P., and J. T. E. Richardson. 1987. "On the Probability of Making Type One Errors." *Psychological Bulletin* 102, (July) 159-63.

Press, S. James. 1989. *Bayesian Statistics: Principles, Models and Applications.* New York: John Wiley & Sons.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." In Peter V. Marsden ed. *Sociological Methodology.* Cambridge, MA: Blackwells.

Ripley, Brian D. 1987. *Stochastic Simulation.* New York: John Wiley & Sons.

Rosnow, Ralph L., and Robert Rosenthal. 1989. "Statistical Procedures and the Justification of Knowledge in Psychological Science." *American Psychologist* 44, 1276-84.

Rozeboom, William W. 1960. "The Fallacy of the Null Hypothesis Significance Test." *Psychological Bulletin.* 57, 416-28.

Schmidt, Frank L. 1996. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers." *Psychological Methods* 1, 115-129.

Schmidt, Frank L., and John E. Hunter. 1977. "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62, 529-40.

Sedlmeier, Peter, and Gerd Gigerenzer. 1989. "Do Studies of Statistical Power Have an Effect on the Power of Studies." *Psychological Bulletin* 105 (March), 309-16.

Serlin, Ronald C., and Daniel K. Lapsley. 1993. "Rational Appraisal of Psychological Research and the Good-enough Principle." In G. Keren, and C. Lewis, eds. *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Sterling, Theodore 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 (March), 30-4.

Tanner, Martin. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Third Edition. New York: Springer-Verlag.

Western, Bruce. 1995. "A Comparative Study of Working-Class Disorganization: Union Decline in Eighteen Advanced Capitalist Countries." *American Sociological Review* 60 (April), 179-201.

Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88, No. 2 (June), 412-23.

Wolf, Frederic M. 1986. *Meta-Analysis: Quantitative Methods for Research Synthesis.* Beverly Hills: Sage.[7pt]