BEHAVIORAL AND BRAIN SCIENCES (1998) **21,** 169–239
*Printed in the United States of America*

# Précis of *Statistical significance: Rationale, validity, and utility*

**Siu L. Chow**
*Department of Psychology, University of Regina, Regina, Saskatchewan, Canada S4S 0A2*
**Electronic mail:** *chowsl@leroy.cc.uregina.ca*

**Abstract:** 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 $H_0$, 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. Neither can any of them fulfill the nonstatistical functions expected of them by critics.

**Keywords:** Bayesianism; effect size; null hypothesis; statistical hypothesis testing; statistical significance; theory corroboration

This précis of *Statistical significance: Rationale, validity, and utility* (Chow 1996) begins with a description of the main themes of its eight chapters.As criticisms of the null-hypothesis significance-test procedure (NHSTP) are answered in the context of the theory-corroboration experiment, the rationale of theory corroboration and the logical foundation of experimentation are described after a description of NHSTP itself. It is argued that NHSTP can (and should) be defended when some conceptual or metatheoretical distinctions are made. "Theory"and "hypothesis" will be used interchangeably in the subsequent discussion even though the former has a more grandiose connotation.

To begin with, because the statistical hypothesis is not the substantive hypothesis (Meehl 1978), corroborating a substantive hypothesis is more than testing a statistical hypothesis. Similarly, drawing a theoretical conclusion is more than deciding whether or not the result is statistically significant (Tukey 1960). It further follows that research data and conclusions are not (and should not be) accepted or rejected merely on the basis of statistical significance. Some criticisms of NHSTP seem persuasive when these distinctions are not made. Other criticisms of NHSTP are based on criteria imported from domains outside statistics. A case will be made that the dissatisfaction with NHSTP stems from attempts to use it to fulfill functions that belong to the theory-corroboration or treatment-assessment process. The alternative numerical indices (viz., effect size, confidence interval estimate, and statistical power) proposed by critics of NHSTP (henceforth referred to as critics) cannot fulfill these nonstatistical functions.



Siu L. Chow received his B.A. from the University of Adelaide, South Australia and Ph.D. from the University of Toronto. He is Professor of Psychology at the University of Regina, Saskatchewan, Canada. He teaches research methods, introductory statistics, and human information processing. His research interest is on the interface between perception and memory, particularly its implications on the study of the mind. Believing that conducting research is more than an extension of everyday thinking, he has revisited in some of his publications some metatheoretical issues such as the social psychology of the experiment, and the relevance of ecological validity, as well as the validity of meta-analysis, in theory-corroboration research.

© *1998 Cambridge University Press* *0140-525X/98 $12.50*



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## 1. An overview of *Statistical significance*

*Statistical significance* begins in Chapter 1 by recounting the commonly known criticisms of NHSTP. Also described is the methodological paradox that psychologists may inadvertently find support for weaker theories when they improve their research methods (Meehl 1967). The basic structure and rationale of NHSTP is illustrated with a completely randomized 1-factor, 2-level quasi-experiment in Chapter 2. It is shown that the null hypothesis can be true, particularly in experimental studies with manipulated variables. Also defended is the hybrid nature of NHSTP.

To distinguish between a substantive and a statistical hypothesis, the quartet of hypotheses associated with the to-be-studied phenomenon in the theory-corroboration experiment is introduced in Chapter 3. It is shown that the null hypothesis appears twice in NHSTP, once as the consequent and once as the antecedent of two conditional propositions. That statistical hypothesis testing is not theory corroboration is seen from the role statistical significance plays in the chain of deductive reasoning discussed in Chapter 4. The outcome of NHSTP is to supply the minor premise for the innermost of the series of three embedding conditional syllogisms.

Two meanings of "effect" are identified in Chapter 5. The anomalous relationship between statistical significance and effect size is more apparent than real because, in terms of the technical meaning of "effect," the effect size is not indicative of the amount of evidential support for the substantive hypothesis offered by data. Nor is the effect size, by itself, informative about the practical importance of the research result. Some conceptual difficulties with power analysis are identified in Chapter 6. Being a conditional probability, statistical power cannot be the *a priori* probability of obtaining statistical significance. Some of the issues raised by power analysts are concerns about the stability of the data. It is argued that the stability issue is neither a numerical nor a mechanical one.

The methodological assumptions underlying Bayesian statistics are considered in Chapter 7. The applicability of the Bayesian approach is questioned because the prototype of empirical research congenial to the Bayesian argument is not typical of psychological research, particularly the theory-corroboration kind. Experimental data can be defended in a relativistic milieu. The main arguments in defense of NHSTP are summarized in Chapter 8 with reference to a set of questions suggested by criticisms of NHSTP.

## 2. Criticisms of NHSTP

NHSTP has been criticized since the 1960s (Morrison & Henkel 1970). The same litany of criticisms of NHSTP is repeated periodically by various critics, as is noted recently by Thompson (1996). Some of the commonly known difficulties of relying on NHSTP are that (1) statistical significance may be the result of the fortuitous choice of the sample size or the $\alpha$ level, (2) the null hypothesis is never true, (3) nothing can be learned from statistical significance about the inverse probability of the hypothesis (i.e., the probability that the hypothesis is true, given the data), (4) the binary nature of NHSTP is antithetical to the fact that knowledge advances in an incremental manner, (5) statistical significance is not informative about the values of parameters, (6) the Type II error is unjustifiably neglected, and (7) nothing about the practical impact of the research result can be learned from its statistical significance.

Critics find it puzzling that psychologists persist in using NHSTP. This state of affairs indicates that NHSTP users suffer from distorted statistical intuitions and conceptual confusion (Gigerenzer 1993). However, the resiliency of NHSTP is warranted. It can be shown that the criticisms of NHSTP are debatable. The frame of reference used in the present defence of NHSTP is suggested by Meehl (1990) and Cohen (1994), but they restrict their criticisms of NHSTP to nonexperimental studies. Meehl (1967) adds that his criticisms are more applicable to experiments using subject variables (e.g., sex, race, educational level, etc.) than to those using manipulated variables (e.g., stimulus duration, method of training, etc.). These caveats raise two interesting questions:

> Q1. Why should NHSTP be more problematic in the case of subject-variable experiments than manipulated-variable experiments?
>
> Q2. What renders NHSTP more satisfactory in an experiment than in a nonexperiment?

Questions Q1 and Q2 suggest that many criticisms of NHSTP are not statistical in nature. The real issue is whether or not the research result is brought about by procedural artifacts or confounding variables. That is, criticisms of NHSTP are actually concerns about inductive conclusion validity (see Campbell & Stanley 1963; Chow 1992; Cook & Campbell 1979).

## 3. The quartet of hypotheses underlying the theory-corroboration experiment

In view of Questions Q1 and Q2, it may be instructive to reconsider the criticisms of NHSTP in the context of the theory-corroboration experiment. Moreover, some hitherto neglected distinctions may be seen more readily when such a frame of reference is adopted. For such an end, consider first the quartet of hypotheses implicated in the theory-corroboration experiment with reference to Table 1. (Ignore the entries in italics for the moment, i.e., Propositions *[P1.1′]* through *[P1.5′]*.)

Consider the phenomenon of linguistic competence that native speakers of English can understand and generate an infinite number of grammatical utterances. A hypothesis that has been used to explain this phenomenon is Miller's (1962) rendition of Chomsky's (1957) transformational grammar (see [P1.1] in Table 1). This psychological analog of the transformational grammar is a substantive hypothesis, and it is an explanatory theory.

Many theoretical implications follow from the hypothesis that transformational grammar is psychologically real. One such implication is that nonkernel sentences (e.g., negative sentences) are more difficult to process than kernel sentences. Specifically, whereas the kernel sentence is generated with the phrase-structure rules, a negative sentence requires the additional step of applying a negative transformation to the kernel sentence. The relationship between the substantive hypothesis and the implication in question is represented by [P1.2] in Table 1. The consequent of the conditional proposition, [P1.2], is the research hypothesis. However, in such a form, the research hypothesis is not sufficiently well defined for experimentation. For example,



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Table 1. *The logical relations among the to-be-explained phenomenon, theory, research hypothesis, experimental hypothesis, and statistical hypotheses (alternative and null) in a theory-corroboration experiment*

| Level of discourse | What is said at the level concerned | |
|---|---|---|
| To-be-explained | The linguistic competence of native speakers of English. | |
| Substantive hypothesis | The linguistic competence of native speakers of English is an analog of the transformational grammar. | [P1.1] |
| *Complement of theory* | *The linguistic competence of a native speaker of English is not an analog of the transformational grammar.* | *[P1.1′]* |
| Research hypothesis | **If** [P1.1], **then** it is more difficult to process negative sentences than kernel sentences. | [P1.2] |
| *Complement of research hypothesis* | ***If*** *−[P1.1],* ***then*** *there is no difference in difficulty in processing negative and kernel sentences.* | *[P1.2′]* |
| Experimental hypothesis | **If** the consequent of [P1.2], **then** it is more difficult to remember extra words after a negative sentence than a kernel sentence. | [P1.3] |
| *Complement of experimental hypothesis* | ***If*** *not the consequent of [P1.2],* ***then*** *it is equally difficult to remember extra words after a negative and a kernel sentence.* | *[P1.3′]* |
| "Statistical alternative hypothesis" | **If** the consequent of [P1.3], **then** $H_1$.[1] | [P1.4] |
| *"Statistical null hypothesis"* | ***If*** *not the consequent of [P1.3],* ***then*** $H_0$.[2] | *[P1.4′]* |
| Sampling distribution of $H_1$ | **If** $H_1$, **then** the probability associated with a difference between kernel and negative sentences as extreme as 1.729 standard error ($t_{df=19}$) units from an **unknown** mean difference is not known. | [P1.5] |
| Sampling distribution of $H_0$ | ***If*** $H_0$, ***then*** *the probability associated with a difference between kernel and negative sentences as extreme as 1.729 standard error* ($t_{df=19}$) *units from a mean difference of zero is 0.05 in the long run.* | *[P1.5′]* |

**1.** $H_1$ = mean of extra-sentence words recalled after negative sentences < mean of extra-sentence words recalled after kernel sentences.

**2.** $H_0$ = mean of extra-sentence words recalled after negative sentences ≥ mean of extra-sentence words recalled after kernel sentences.

it is necessary to specify the nature of the processing involved.

The problem of vagueness with [P1.2] is resolved by stipulating (a) a well-defined experimental task in a specific setting, and (b) a dependent variable whose identity is independent of the substantive hypothesis. A simplified version of Savin and Perchonock's (1965) task may be used to illustrate the solution. Suppose that subjects are presented with eight words *after* being shown either a kernel or a negative sentence on any trial. Suppose further that the repeated-measures design is used. That is, the same subjects receive both types of sentences in the course of the experiment.

The subjects must first recall the sentence verbatim and then recall as many of the eight extra words as possible. In the context of this experimental situation and of the auxiliary assumption that the short-term store has a limited capacity (Miller 1956), an implication of the consequent of [P1.2] is that it is more difficult to remember extra words after a negative sentence than after a kernel sentence. This implication of the research hypothesis is the experimental hypothesis, which appears as the consequent of [P1.3] in Table 1.

Because the experimental hypothesis is not amenable to statistical analysis in its present form, it is necessary to derive its implication at the statistical level. Specifically, the implication is that the mean of extra-sentence words recalled after negative sentences is smaller than that after kernel sentences. This implication is more commonly known as the statistical alternative hypothesis ($H_1$), and it is the consequent of [P1.4].

Consider the logical complement of $H_1$, in Table 1. It is stated that the mean of extra-sentence words recalled after negative sentences is equal to or larger than that after kernel sentences (see the consequent of *[P1.4′]* in Table 1). This logical complement of $H_1$ is the statistical null hypothesis ($H_0$). Given that whatever is true under the "larger than" component of $H_0$ is subsumed under the "equal to" component, the "larger than" component serves no further purpose in the present discussion.

That this appeal to $H_0$ is neither contrived nor arbitrary may be seen from the entries in italics in Table 1. The steps of derivation of *[P1.3′]* from *[P1.1′]* are the same as those implicated in deriving [P1.3] from [P1.1]. Hence, *[P1.3′]* is not contrived if *[P1.1′]* is not an arbitrary assertion. Being the logical complement of [P1.1], *[P1.1′]* is not a whimsical statement. In other words, $H_0$ is not as arbitrary as it has been characterized to be (see, e.g., Fisher 1959; Rozeboom 1960; Thompson 1996).

The null hypothesis has two uses. First, it can specify the sampling distribution of differences required for the test of significance (see *[P1.5′]*). Second, a decision about $H_1$ may be made through making a decision about $H_0$ because these two statistical hypotheses are mutually exclusive and exhaustive (see the "$H_0$, data, and chance influences" discussion in sect. 12 for an explication).

In sum, underlying the theory-corroboration experiment is a quartet of hypotheses, namely, the substantive, research, experimental, and statistical alternative hypotheses. It can be seen that neither $H_0$ nor $H_1$ is the substantive, research or experimental hypothesis. Hence, it becomes necessary to distinguish between testing a substantive hy-



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

pothesis at the conceptual level with empirical data (i.e., theory corroboration) and testing a statistical hypothesis (viz., statistical hypothesis testing). At the same time, it is noted in [P1.5] in Table 1 that $H_1$ cannot be used to specify the to-be-used sampling distribution of differences that underlies the *t* test because the magnitude of the difference between the means of the kernel and negative sentences is not specified in $H_1$. The complement of $H_1$ (i.e., $H_0$) is used instead (hence, *[P1.5′]* in Table 1). This invites a closer examination of NHSTP, particularly in view of the generally accepted verdict that $H_0$ is never true.

## 4. The Null-hypothesis Significance-test Procedure (NHSTP)

A consideration of how theory corroboration differs from statistical hypothesis testing may begin with a brief recounting of the rationale and procedure of NHSTP. Suppose that Savin and Perchonock's (1965) task is used, and the statistical alternative hypothesis is that fewer words are recalled after recalling negative sentences than kernel sentences. $H_1$ and $H_0$ are commonly (but misleadingly) written as follows under such circumstances:

(1) $H_1$: $u_{negative} < u_{kernel}$
(2) $H_0$: $u_{negative} \geq u_{kernel}$

Suppose further that the repeated-measures design is used, and there are 20 subjects. This experiment will be referred to as the "kernel-negative experiment" in subsequent discussion. The usual $\alpha$ level is set at 0.05. Strictly speaking, the test is whether or not the associated probability, *p*, of the calculated *t* is smaller than 0.05. "Associated probability" means "the probability of [the calculated *t*] plus the probabilities of all more extreme possible values" under $H_0$ (Siegel 1956, p. 11). In actual practice, the *t* (dependent sample in this example) is calculated and compared to the critical value of *t* (i.e., −1.728, df = 19, $\alpha$ = .05) for this particular one-tailed test.

This critical value of −1.729 is given by the appropriate *t* distribution, which is the standardization of the sampling distribution of differences (Siegel 1956). The binary decision is to choose between "calculated $t \leq -1.729$" and "calculated $t > -1.729$." The outcome of this binary decision determines the choice between the two *modus ponens* arguments depicted in the two top panels in Table 2. If the calculated *t* is −1.729 or smaller, the decision is that the result is significant (i.e., the "not $H_0$" conclusion in the top left panel of Table 2). If the calculated *t* is larger than the critical value, it is decided that the result is not significant (i.e., the "$H_0$ conclusion in the top right panel of Table 2).

It is assumed that $H_1$and $H_0$ are mutually exclusive and exhaustive (see the "$H_0$, data, and chance influences" discussion in sect. 12). Hence, denying $H_0$ leads to accepting $H_1$ by virtue of the disjunctive syllogism depicted in the lower panel of Table 2. The experimental conclusion drawn from a statistically significant result is that fewer words are recalled after recalling negative sentences than kernel sentences.

Of interest is the fact that the experimental conclusion is about the relationship between two variables (viz., *sentence type* and *number of extra words recalled*). However, theoretical conclusions go beyond a mere functional relationship between the independent and dependent variables. The theoretical interest concerns the nature of the linguistic competence. This more sophisticated meaning of research data at the theoretical level is not informed by the NHSTP exercise depicted in Table 2. This consideration has not featured in the debate about the validity or utility of NHSTP because discussants have in mind a different type of experiment (a point to be discussed in sect. 21: "Differences between the utilitarian and theory-corroboration experiments"). To see how the theoretical meaning is extracted from experimental data, it is necessary to consider what constitutes the theory-corroboration process.

Table 2. *Two conditional syllogisms (upper panel) and the disjunctive syllogism (lower panel) implicated in the null-hypothesis significance testing procedure (NHSTP)*

*Upper Panel*

| | Criterion exceeded | Criterion not exceeded |
|---|---|---|
| Major premise | *If* calculated $t \leq$ (criterion = −1.729), *then* not $H_0$. | *If* calculated $t >$ (criterion = −1.729), *then* $H_0$. |
| Minor premise | $t \leq$ (criterion = −1.729) [e.g., calculated $t$ = −2.05] | $t >$ (criterion = −1.729) [e.g., calculated $t$ = −1.56] |
| Conclusion | Not $H_0$ | $H_0$ |

*Lower Panel*

| | Statistical significance obtained |
|---|---|
| Major premise: | $H_1$ or $H_0$ |
| Minor premise: | not $H_0$ |
| Conclusion: | Therefore, $H_1$. |

## 5. The rationale of the theory-corroboration experiment

To corroborate the substantive hypothesis experimentally is to show that the experimental data are consistent with the tenability of the substantive hypothesis. That is, there is "warranted assertibility" (Manicas & Secord 1983). This idea suggests that a crucial consideration in theory corroboration is the logical relationship between the substantive hypothesis and the evidential data. Such a consideration requires more than a statistical decision. Also implicated is the judicious application of deductive and inductive logic in different stages of the exercise.

## 6. The role of deductive logic in the theory-corroboration experiment

Table 1 shows that $H_1$ is three implicative steps from the substantive hypothesis. At the same time, there is a chain of deductive reasoning leading from experimental data to the substantive hypothesis via $H_1$, the experimental hypothesis and the research hypothesis. This series of deductive reasoning may be seen more readily if the logical relations among the quartet of hypotheses shown in Table 1 are expressed in the form of a series of three embedding conditional syllogisms, as in Table 3.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Table 3. *The series of three embedding syllogisms (in roman font, italic, and boldface, respectively) underlying the theory-corroboration procedure when the null hypothesis is rejected*

| | | |
|---|---|---|
| **Major premise 3** | **If [P1.1][1] in Table 1, then [P3.1].[2]** | **[MAJ-3.3][7]** |
| *Major premise 2* | *If [P3.1], then [P3.2].[3]* | *[MAJ-3.2][6]* |
| Major premise 1 | If [P3.2], then $H_1$.[4] | [MAJ-3.1][5] |
| Minor premise 1 | $H_1$ is true. | [MIN-3.1] |
| Conclusion 1 | Therefore, [P3.2] is true in the interim (by virtue of experimental controls). | [CON-3.1] |
| *Minor premise 2* | *[P3.2] is true in the interim.* | *[MIN-3.2]* |
| *Conclusion 2* | *Therefore, [P3.1] is true in the interim (by virtue of experimental controls).* | *[CON-3.2]* |
| **Minor premise 3** | **[P3.1] is true in the interim.** | **[MIN-3.3]** |
| **Conclusion 3** | **Therefore, [P1.1] in Table 1 is true in the interim (by virtue of experimental controls).** | **[CON-3.3]** |

**1.** [P1.1] in Table 1 — The linguistic competence of a native speaker of English is an analog of the transformational grammar.
**2.** [P3.1] — It is more difficult to process negative sentences than kernel sentences (i.e., the consequent of [P1.2] in Table 1).
**3.** [P3.2] — It is more difficult to remember extra words after a negative sentence than a kernel sentence (i.e., the consequent of [P1.3] in Table 1).
**4.** $H_1$ — mean of extra-sentence words recalled after negative sentences $<$ mean of extra-sentence words recalled after kernel sentences.
**5.** [MAJ-3.1] — is [P1.4] in Table 1.
**6.** *[MAJ-3.2]* — is [P1.3] in Table 1.
**7.** **[MAJ-3.3]** — is [P1.2] in Table 1.

The syllogisms in Table 3 are called "conditional syllogisms" because their major premises are conditional propositions (viz., [MAJ-3.1], *[MAJ-3.2]*, and **[MAJ-3.3]**). The first (or the innermost) syllogism is made up of [MAJ-3.1], [MIN-3.1], and [CON-3.1]. The second syllogism consists of *[MAJ-3.2]*, *[MIN-3.2]*, and *[CON-3.2]*. **[MAJ-3.3], [MIN-3.3],** and **[CON-3.3]** collectively make up the last syllogism.

The minor premise of the first syllogism (i.e., [MIN-3.1]) is the outcome of NHSTP. The example depicted is one in which the data permit the rejection of $H_0$. To have established statistical significance is to accept that $H_1$ is true. To assert that $H_1$ is true in the first syllogism is to affirm the consequent of the conditional proposition, [MAJ-3.1]. The tentative conclusion is drawn that the antecedent of [MAJ-3.1] is true. This conclusion is used as the minor premise of the second syllogism to affirm the consequent of *[MAJ-3.2]*. This leads to the tentative conclusion that the antecedent of *[MAJ-3.2]* is true. Lastly, the conclusion of the second syllogism serves as the minor premise of the third syllogism. The antecedent of **[MAJ-3.3]** is considered true tentatively when its consequent is affirmed by the antecedent of *[MAJ-3.2]*.

## 7. The modus tollens and affirming the consequent asymmetry

Note that all three conclusions in Table 3 (i.e., [CON-3.1], *[CON-3.2]*, and **[CON-3.3]**) are qualified with the caveat, "in the interim (by virtue of experimental controls)." The "in the interim" qualification is necessary because there are alternative substantive hypotheses at the conceptual level (see sect. 32, the "Alternative substantive hypothesis versus statistical alternative hypothesis," for an elaboration). The "by virtue of experimental controls" qualification is necessary because deductive logic does not permit accepting the antecedent of a conditional proposition when its consequent is affirmed (Copi 1982). Hence, the propriety of accepting the antecedents of [MAJ-3.1], *[MAJ-3.2]*, and **[MAJ-3.3]** in Table 3 has to be warranted by experimental controls, as discussed in section 8, "Induction, experimental design, and controls."

Suppose that the outcome of NHSTP does not permit rejecting $H_0$. The chain of reasoning is shown in Table 4, in which the propositions in Table 3 are given a different set of numbers for identification purposes. For example, [MAJ-3.1] in Table 3 becomes [MAJ-4.1] in Table 4.

The minor premise of the first conditional syllogism in Table 4, [MIN-4.1], is "Not-$H_1$." Hence, the antecedent of [MAJ-4.1] is rejected by *modus tollens*. The minor premise of the second syllogism [MIN-4.2] is, in such an event, the denial of the consequent of [MAJ-4.2]. The *modus tollens* rule leads to the rejection of the antecedent of [MAJ-4.2]. Hence, [MIN-4.3] is the negation of the antecedent of [MAJ-4.2]. Consequently, [MIN-4.3] is the denial of the antecedent of [MAJ-4.3]. The third application of the *modus tollens* rule leads to the rejection of the antecedent of [MAJ-4.3], namely, [P1.1].

Unlike the case of affirming the consequent, *modus tollens* (i.e., denying the consequent of a conditional proposition) permits the unambiguous rejection of the antecedent of the conditional proposition. The difference between the arguments in Tables 3 and 4 is the asymmetry between *modus tollens* refutation and *affirming the consequent* confirmation of theories identified by Meehl (1967; 1978). It is noted here that the asymmetry is not brought about by using NHSTP. Instead, it is the consequence of the deductive reasoning implicated in corroborating theories. Hence, it is necessary to consider why affirming the consequent of [MAJ-3.1] (i.e., rejecting $H_0$) does not guarantee the truth of its antecedent.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Table 4. *The series of three embedding syllogisms (in roman font, italic, and boldface, respectively) underlying the theory-corroboration procedure when the null hypothesis is not rejected*

| | | |
|---|---|---|
| **Major Premise 3** | **If [P1.1][1] in Table 1, then [P4.1].[2]** | **[MAJ-4.3][7]** |
| *Major Premise 2* | *If [P4.1], then [P4.2].[3]* | *[MAJ-4.2][6]* |
| Major premise 1 | If [P4.2], then $H_1$.[4] | [MAJ-4.1][5] |
| Minor premise 1 | $H_1$ is not true. | [MIN-4.1] |
| Conclusion 1 | Therefore, [P4.2] is not true. | [CON-4.1] |
| *Minor premise 2* | *[P4.2] is not true.* | *[MIN-4.2]* |
| *Conclusion 2* | *Therefore, [P4.1] is not true.* | *[CON-4.2]* |
| **Minor premise 3** | **[P4.1] is not true.** | **[MIN-4.3]** |
| **Conclusion 3** | **Therefore, [P1.1] in Table 1 is not true.** | **[CON-4.3]** |

**1.** [P1.1] in Table 1 — The linguistic competence of a native speaker of English is an analog of the transformational grammar.
**2.** [P4.1] — It is more difficult to process negative sentences than kernel sentences (i.e., the consequent of [P4.2] in Table 1).
**3.** [P4.2] — It is more difficult to remember extra words after a negative sentence than a kernel sentence (i.e., the consequent of [P4.3] in Table 1).
**4.** $H_1$ — mean of extra-sentence words recalled after negative sentences $<$ mean of extra-sentence words recalled after kernel sentences.
**5.** [MAJ-4.1] — is [P1.4] in Table 1.
**6.** *[MAJ-4.2]* — is [P1.3] in Table 1.
**7.** **[MAJ-4.3]** — is [P1.2] in Table 1.

## 8. Induction, experimental design, and controls

Boring (1954; 1969) and Campbell (1969; Campbell & Stanley 1963) pointed out that to consider experimental controls was to consider Mill's (1973) methods of scientific inquiry (with the exception of his method of agreement; see Cohen & Nagel 1934). That is to say, underlying a valid experimental design is one of Mill's (1973) inductive methods (viz., method of difference, joint method of agreement and difference, method of residue, and method of concomitant variations). This may be illustrated with Table 5, which depicts the repeated-measures 1-factor, 2-level design used in the kernel-negative experiment described earlier.

The design of the kernel-negative experiment is described in Table 5 in a way that reflects the inductive principle of Mill's (1973) method of difference (see Chow 1992). Suppose that fewer words are recalled after negative sentences than after kernel sentences, and that the difference is statistically significant. The control variables (C1, C2, C3, C4, C5, and C6) can be excluded as explanations of the significant difference because each of them (e.g., C1) is represented by the same value (viz., NI) at both levels of the independent variable. This is one of the "constancy of condition" meanings of the term "control" (Boring 1954; 1969).

The extraneous variables (*E1, E2, . . . En*) may also be excluded because each of them (e.g., *E1*) is assumed to be

Table 5. *The inductive basis of the repeated-measures 1-factor, 2-level design (Method of difference)*

| | | Control variables | | | | | | Extraneous variables | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | Independent variable (sentence-type) | C1 | C2 | C3 | C4 | C5 | C6 | *E1* | *E2* | . . . | *En* | Dependent variable |
| Control | Kernel sentence | NI | T | I | R | S | C | *ER* | *IT* | . . . | *M* | Number of extra words recalled |
| Experimental | Negative sentence | NI | T | I | R | S | C | *ER* | *IT* | . . . | *M* | Number of extra words recalled |

C1 = Normal intonation (NI)
C2 = Task presentation via recorded tape (T)
C3 = Interval between end of sentence and beginning of words (I)
C4 = Rate of word presentation; 3/4 second per word (R)
C5 = Structure of sentence; "Animal" subject, present perfect transitive verb (S)
C6 = Fixed categories of words used in "extra" words (C)
*E1 = Extracurricular reading (ER)*
*E2 = Individual interests (IT)*
*En = Kernel and negative sentences randomly mixed (M)*


https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

represented at the same level (viz., *ER*). This assumption is justified by the fact that the same subject is tested in both the experimental and control conditions. Consequently, the difference between the "Kernel" and "Negative" conditions is rendered unambiguous by the fact that the experimental and control conditions are identical in all aspects but one. The only difference is brought about by the difference between the two levels of the independent variable.

## 9. Conflating NHSTP with theory corroboration

NHSTP is misunderstood because no distinction is made between the substantive and statistical hypotheses. Specifically, Meehl (1967) notes that there is a tendency to conflate the substantive hypothesis with the statistical hypothesis. This practice seems to be condoned when it is said that "the critical distinction between a statistical hypothesis and a substantive theory often breaks down. *To perform a significance test a substantive theory is not needed at all"* (Oakes 1986, p. 42, emphasis added).

What is said in the italicized sentence is true, but not because the distinction between the substantive and statistical hypotheses is unimportant or not real. It is true simply because testing a hypothesis at the statistical level (see Table 2) and corroborating a substantive hypothesis with empirical data at the conceptual level (viz., Table 3) are radically different exercises. This issue will be dealt with further in the "Differences between the utilitarian and theory-corroboration experiments" discussion in section 21.

## 10. Answers to questions Q1 and Q2

It may be concluded from the foregoing argument that, to the extent that all recognized control variables and procedures are included in the experiment, the statistically significant result may be attributed to the independent variable (Campbell 1969). The experiment is said to have inductive conclusion validity under such circumstances (Chow 1987a; 1992). For this reason, the propriety of accepting the antecedent of a conditional proposition by affirming its consequent in Table 3 is justified with the "in the interim" proviso.

The answer to Question Q1 may be seen readily from Table 6. Suppose that the kernel-negative experiment is conducted to assess the differential linguistic competence of science students from two disciplines. Neither the repeated-measures nor the completely randomized design can be used. Hence, different selected groups of subjects have to be assigned to the two levels of the independent variable, *faculty of study*. Although it is possible to maintain the constancy of condition in the case of some control variables, such is not the case with the extraneous variables. An extraneous variable (e.g., *E1*) may be represented at different levels in the experimental and control conditions (viz., *ER* and *ER′*, respectively) as a result of some fundamental differences between students of the two disciplines.

In short, the design of an empirical research is a description of how the data collection conditions are arranged. The empirical study is an experiment if the arrangement of its data collection conditions satisfies the formal requirement of one of Mill's (1973) inductive principles. The formal requirement makes it possible to exclude as explanations those factors that have been incorporated in the design as control variables or procedures. Various aspects of the formal requirement give rise to the three technical meanings of "control": (a) a valid comparison baseline, (b) constancy of conditions, and (c) provisions for excluding procedural artifacts (Boring 1954; 1969; Chow 1987a; 1992). Data interpretation becomes unambiguous to the extent that all recognized alternative interpretations are excluded by the judicious application of experimental controls (Campbell 1969).

An empirical study is a quasi-experiment when its design satisfies only some parts of the formal requirement. A nonexperimental study (e.g., the correlational study) is one in which there is no formal provision for satisfying the formal requirement. Hence, there is no provision for excluding alternative interpretations of the results in nonex-

Table 6. *Violation of the formal requirement of* method of difference *when a subject variable is used*

| | Subject variable (faculty of study) | Control variables | | | | | | Extraneous variables | | | | Dependent variable |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | C1 | C2 | C3 | C4 | C5 | C6 | *E1* | *E2* | . . . | *En* | |
| **C** | Biological sciences | NI | T | I | R | S | C | *ER* | *IT′* | . . . | *M′′* | Number of extra words recalled |
| **E** | Physical sciences | NI | T | I | R | S | C | *ER′* | *IT* | . . . | *M′* | Number of extra words recalled |

C1 = Normal intonation (NI)
C2 = Task presentation via recorded tape (T)
C3 = Interval between end of sentence and beginning of words (I)
C4 = Rate of word presentation; 3/4 second per word (R)
C5 = Structure of sentence; "animal" subject, present perfect transitive verb (S)
C6 = Fixed categories of words used in "extra" words (C)
*E1 = Extracurricular reading (ER or ER′)*
*E2 = Individual interests (IT or IT′)*
*En = Kernel and negative sentences randomly mixed (M′ or M′′)*



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

perimental studies. Given the fact that experimental controls serve to exclude explanations, it can be seen that data from quasi-experimental and nonexperimental studies are more ambiguous than experimental data. This is the answer to Question Q2. The comparison between Tables 5 and 6 provides the answer to Question Q1. These answers to Questions Q1 and Q2 lead to the realization that some criticisms of NHSTP are motivated by ambiguities in data interpretation. At the same time, a few criticisms arise because the nature of $H_0$ is misunderstood or misrepresented.

## 11. The nature of $H_0$

What is clear from the discussion of Tables 2 and 3 is that NHSTP does *not* determine whether or not the experimental data support the substantive hypothesis. Supplying the minor premise for the first syllogism in Table 3 or 4 is the only contribution NHSTP has to theory corroboration. The theoretical meaning of the experimental data is conferred by their logical relation with the experimental, research, and substantive hypotheses. Although statistical significance does not confer any theoretical meaning to data, it does have an important function. Specifically, it provides a rational basis for excluding chance influences as an explanation of data. This important (but limited) role may be seen from a closer examination of the statistical null hypothesis, $H_0$.

## 12. $H_0$, data, and chance influences

One way to paraphrase the antecedent of *[P1.4′]* in Table 1 is to say that the subjects are indifferent to whether the to-be-remembered sentence is a kernel or a negative sentence. Consequently, under such circumstances any observed difference between the means of the "Negative" and the "Kernel" conditions is the result of chance influences (or errors). That is, actual measurements made during data collection may be affected by unintended nonsystematic influences (i.e., errors) of various kinds. Consequently, *[P1.4′]* in Table 1 may be represented as the conditional proposition, [P7.1], in Table 7. By the same token, [P1.4] in Table 1 may be represented by [P7.2] in Table 7.

The representation adopted for $H_0$ and $H_1$ in Table 7 serves three functions. First, it highlights the meaning of the null hypothesis. It is a hypothesis about the influence of nonsystematic chance factors on data in the form of distributing the unintended influences randomly between the two conditions. Moreover, the errors are normally distributed with a mean of zero in each condition. Consequently, a statistically significant result will be correctly interpreted to mean only that an explanation of the data in terms of chance influences can be excluded with the level of strictness stipulated by the significance level (viz., $\alpha$).

Second, Table 7 makes explicit the mutually exclusive and exhaustive relationship between $H_0$ and $H_1$. That is, the contrast between $H_0$ and $H_1$ is informed by neither the substantive hypothesis nor the to-be-studied phenomenon. Instead, the contrast is informed by the data-collection procedure. It is a contrast between *chance* and *not chance.* That NHSTP is actually mute at the level of the substantive hypothesis may be seen from the fact that, in the event the result is statistically significant, the nonchance factor responsible for the data is not informed by statistical significance.

The third function of the tabular representation of Table 7 is to make explicit the fact that $H_0$ is not used as a categorical proposition. It appears twice, once as the consequent of the conditional proposition [P7.1] and once as the antecedent of the conditional proposition [P7.3]. This state of affairs means that, even if "$H_0$ is never true" were true, its contribution to the statistical decision process would not be affected because the truth of either [P7.1] or [P7.3] is not determined by the truth value of $H_0$ alone, but by the truth values of both the antecedent and consequent (Copi 1982). At the same time, it is important to emphasize that $H_0$ can (and should) be true, the common belief to the contrary notwithstanding.

Table 7. *The statistical null hypothesis ($H_0$) and the statistical alternative hypothesis ($H_1$) as components of conditional propositions*

| Where in Table 1 | Conditional proposition | |
|---|---|---|
| *[P1.4′]* | **If** change, **then** $H_0$. | [P7.1] |
| [P1.4] | **If** not chance, **then** $H_1$. | [P7.2] |
| | **If** $H_0$, **then** the test statistic is distributed as a *sampling distribution of the difference* whose mean difference is zero. | [P7.3] |

## 13. "$H_0$ is never true" revisited

Consider the antecedent of the conditional proposition, *[P1.3′]* in Table 1. It says that there is no difference in difficulty in processing negative and kernel sentences. In other words, $H_0$ is a hypothesis about the relationship between two theoretical populations, "Kernel" and "Negative" (viz., the hypothesized population of *all* subjects presented with kernel sentences and that of *all* subjects presented with negative sentences). In view of the fact that two populations are implicated in $H_0$ (not just one), it is not clear what $H_0$ is about when only one population is acknowledged, as in the statement: "A null hypothesis is any precise statement about a state of affairs in *a population*, usually the value of a parameter, frequently zero" (Cohen 1990, p. 1307, emphasis added).

The assertion, "things get downright ridiculous when $H_0$ is to the effect that the effect size (ES) is 0 – that the *population mean difference* is 0" (Cohen 1994, p. 1000, emphasis added), is questionable for a different reason. Two theoretical populations are properly recognized in this statement if "population mean difference" refers to the mean of the sampling distribution of differences. It needs two population distributions to give rise to a sampling distribution of differences. However, it can be shown that it is *not* ridiculous to have a mean difference of zero for the sampling distribution of differences.

Recall the two theoretical populations, "Kernel" and "Negative," in the kernel-negative experiment They are



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

procedurally defined populations. Specifically, they are defined in terms of the two levels of the independent variable, *Sentence-type*. The data collection situation in experimental psychology can (and should always) be made to ensure that the two procedurally defined populations will be identical if the subjects are indeed unconcerned about the difference between the two levels of the independent variable. This is effected in different situations by using the repeated-measures design, the matched-pair design, or the completely randomized design.

As an example, consider the repeated-measures design. The two test conditions (viz., presenting kernel sentences and presenting negative sentences) are imposed on the *same* group of subjects. This group of subjects becomes two hypothetical samples when described in terms of the two respective levels of the independent variable. The two hypothetical samples are identical before being exposed to the experimental manipulation. They remain identical if what is said in the experimental hypothesis is false. Why should the "Kernel" and "Negative" populations not have the same mean if the complement of the experimental hypothesis is true? Why is it ridiculous to expect the difference between the "Kernel" and "Negative" populations to be zero at the statistical level if the subjects are indifferent to the experimental manipulation? In other words, critics have not taken into account the fact that the null hypothesis is about neither the to-be-studied phenomenon nor some actual substantive populations. The null hypothesis is about the relationship between two or more procedurally defined hypothetical populations.

It is important to emphasize that the truth of $H_0$ depends on assigning subjects randomly to the experimental and control conditions or using the same subjects in both conditions. This iteration is necessary in view of a recent attempt to question the assertion, "$H_0$ is never true," with the following debatable scenario: "We give a placebo to a control group and [the to-be-tested] drug to the experimental group. We then mix these participants into one group" (Hagen 1997, p. 16). The data collection procedure depicted is unsatisfactory because it does not guarantee that the formal requirement of Mill's (1973) method of difference is met. This example may also be used to make the case that the validity of NHSTP must be assessed in the context of research methods.

In short, $H_0$ can be true. More importantly, it ought to be true if the data collection procedure is set up and conducted properly (hence, the importance of Cohen's [1994] and Meehl's [1990] caveat identified in Question Q2). The assertion, "$H_0$ is never true," seems self-evident only when $H_0$ is used as a categorical proposition descriptive of an ill-defined state of affairs. On the contrary, it is actually a statement about how the data are collected, a point also noted by Bakan (1966) and Phillips (1973).

More importantly, $H_0$ is never used as a categorical proposition. At one level of discourse (viz., *[P1.4′]* in Table 1), $H_0$ is a description of the data when certain assumptions or conditions are satisfied in the data collection situation (a point emphasized by Falk & Greenbaum 1995). At a different level of discourse (i.e., *[P1.5′]* in Table 1 or [P7.1] in Table 7), $H_0$ is a criterion for rejecting chance influences as an explanation of data. What renders $H_0$ indispensable is that it stipulates the to-be-used sampling distribution of the test statistic required for making the decision about chance influences (see [P7.3] in Table 7 or *[P1.5′]* in Table 1).

## 14. The ambiguity-anomaly criticisms of NHSTP

A statistically significant result is considered ambiguous by critics. They also find the relationship between statistical significance and the effect size anomalous. The ambiguity and anomaly stem from the fact that statistical significance may be the fortuitous consequence of having chosen a particular sample size. Consider Studies A and B in Table 8.

Although the effect size is the same in both studies, the result is significant in Study A, but not Study B. At the same time, the sample size is larger in Study A than in Study B. This is the basis of the sentiment shared among critics that statistical significance is assured if a large enough sample is used (see Thompson 1996, for a recent expression of this view). By the same token, a result may be nonsignificant because too small a sample is used. This difficulty may be called the *sample size-dependence problem*.

Study A is significant and Study C is not significant. Yet, the effect size is larger in Study C than in Study A. This is considered an anomaly, and it may be called the *incommensurate significance-size problem*. This problem suggests to critics that statistical significance is misleading at best, harmful at worst. The harm NHSTP does to research is that it precludes researchers from using more profitably the quantitative information in the data. Specifically, if researchers are satisfied with the NHSTP result, they may neglect to determine the confidence interval estimate of the parameter.

Studies A and D in Table 8 jointly show that the *incommensurate significance-size problem* may assume the form of the *magnitude-insensitivity problem*. Their results are significant; but the effect in Study D is larger than that in Study A. This useful information is not put to good use. The same point may be illustrated with Studies B and C. Although their results are not significant, the effect is larger in Study C than B. Again, the magnitude of the effect should be used (e.g., in meta-analysis; Glass et al. 1981; Schmidt 1996).

A closer examination of the following issues shows that these criticisms themselves are debatable. First, the ambiguity is a conceptual or methodological problem, not a quantitative issue. Second, the effect size and NHSTP express the difference between the means of the experimental and control groups at different levels of abstraction. Third, parameter estimation is not theory corroboration. Fourth, nonstatistical concerns cannot be addressed with statistics indices. Fifth, the validity of meta-analysis, as a theory-corroboration tool, can be questioned.

Table 8. *The putative ambiguity and anomaly of significance tests illustrated with four fictitious studies*

| Study | $u_E$ | $u_C$ | Effect size[1] $d = \frac{(u_E - u_C)}{\sigma_E}$ | Statistical test (e.g., $t$) significant? | df |
|---|---|---|---|---|---|
| A | 6 | 5 | 0.1 | Yes | 22 |
| B | 25 | 24 | 0.1 | No | 8 |
| C | 17 | 8 | 0.9 | No | 8 |
| D | 8 | 2 | 0.5 | Yes | 22 |

**1.** Cohen (1987)



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## 15. The sample size-significance dependence problem revisited

A persistent theme found in criticisms of NHSTP is that the fortuitous choice of the sample size (e.g., an unjustifiably large sample) may be responsible for a statistically significant result. However, Questions Q1 and Q2 suggest that the issue may have nothing at all to do with the sample size. The real concern may be questions about the internal validity of the research (Campbell & Stanley 1963; Cook & Campbell 1979). Be that as it may, that statistical significance may be questioned suggests that there are good reasons why affirming the consequent of [MAJ-3.1], *[MAJ-3.2]*, or **[MAJ-3.3]** in Table 3 does not guarantee the truth of its antecedent. This may be seen more readily from the following nonexperimental study.

Suppose that the effects of institutional constraints on a rehabilitation programme is assessed with a correlational study. It is found that the efficacy of the rehabilitation programme varies inversely with the number of institutional constraints. What does it mean to dismiss the study for the simple reason that the sample size is unusually large (e.g., n = 1,000)?

Note that to question the statistically significant result in this example is to question the conclusion that institutional constraints are really related to the failure of the rehabilitation programme. That is, this is a question about data interpretation (a conceptual concern), not about the numerical value of the test statistic or the sample size. Hence, it is necessary to consider the "fortuitous sample size" argument more closely, not in quantitative terms, but in qualitative terms. That is, the issue is why it is more likely to introduce confounding variables when more participants are included in the correlational study.

To increase the sample size is to recruit more participants in the correlational study. Chances are that the participants would have to be recruited from more diverse settings. Consequently, not only does the chance of having a confounding variable increase, it also becomes more difficult to identify the confounding variable. The result (be it statistically significant or nonsignificant) becomes more ambiguous regarding the relationship between institutional constraints and the efficacy of the rehabilitation treatment of interest. More important, it would not be valid to apply the chain of reasoning depicted in Table 3 under such circumstances. As may be recalled from Table 5, the situation is very different in the case of the experiment because of experimental controls.

Why is the sample size-significance dependence problem not seen by critics as a concern about the internal validity of the research? The real source of the ambiguity is obscured by the suggestion that statistical significance may be manipulated by cynical researchers. Specifically, it is intimated by some critics that cynical researchers use excessively large samples if their interests are vested in a statistically significant result, but small samples if their vested interests are served by a nonsignificant result. That a tool may be misused speaks ill only of its users, however. It does not mean that the tool itself is unsatisfactory, particularly when nothing inherent in the tool invites its being misused.

It would be possible to dismiss the cynicism issue as irrelevant were there not the impression that psychologists accept (or do not accept) a research conclusion solely on the basis of statistical significance (or nonsignificance). The impression is misleading. For example, cognitive psychologists do not accept or reject a finding merely on the basis of statistical significance or nonsignificance (see, e.g., Coltheart's [1980] or Haber's [1983], discussion of the iconic store). Cognitive psychologists examine assiduously whether or not (a) a proper experimental design has been used in the experiment, (b) subjects have been given sufficient training, (c) all recognizable control variables or procedures are properly instituted, and (d) the correct statistical procedure is used.

In short, experimental psychologists are meticulous about the internal validity of experiments (viz., both the inductive conclusion validity and statistical conclusion validity). They are aware that a statistically significant result may be ambiguous at the conceptual level as a result of various features found in the data collection procedure or situation. In fact, experimental psychologists are so conscientious about the inductive conclusion validity issues that their attempts to eliminate conceptual or methodological ambiguities have recently been dismissed as "methodolatory" (Danziger 1990) or "scientific rhetoric" (Gergen 1991).

The realization that the ambiguity issue has nothing to do with NHSTP obviously has important implications for reducing ambiguity. For example, the ambiguity cannot be reduced by testing more subjects or by analyzing parts of the data (as envisaged in Hunter & Schmidt's [1990] psychometric meta-analysis). Nor can another numerical index be used to disambiguate the statistically significant result (be it the effect size or statistical power). It is instructive to recall the following observation:

> The sum total of the reasons which will weigh with the investigator in accepting or rejecting the [substantive] hypothesis can very *rarely* be expressed in numerical terms. All that is possible for him is to balance the results of a mathematical summary, formed upon certain assumptions, against other less precise impressions based upon *a priori* or *a posteriori* considerations. (Neyman & Pearson 1928, p. 176; emphasis and explication in square brackets added) (Quote 1)

Two obvious examples of Neyman and Pearson's (1928) numerical terms are statistical power and the effect size. An example of the *a priori* considerations is the choice between the repeated-measures and completely randomized designs. The consideration as to whether or not there is any confounding variable after the completion of the experiment is an example of the *a posteriori* considerations in question.

## 16. Two levels of abstraction: Statistical significance and effect size

An assumption must be made explicit before one can assess whether or not Studies A and C in Table 6 suggests that statistical significance is anomalously related to the effect size. Specifically, it is necessary to assume that statements about statistical significance and the effect size are at the same level of abstraction. A look at how *t* and the effect size are respectively defined in Equations 1 and 2 suggests otherwise.

$$[a]\ t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{\bar{X}_1 - \bar{X}_2}} \qquad \text{(Equation 1)}$$

$$[b]\ d = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_i} \qquad \text{(Equation 2)}$$



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

The ($\mu_1 - \mu_2$) component of the numerator of Equation 1 is zero if the implication of chance influences is that $\mu_1 = \mu_2$ (Kirk 1984). Consequently, the numerator is the same in both equations, namely, the difference between the two sample means. On the one hand, the denominator in Equation 1 is the standard error of differences. It is a property of a theoretical distribution, namely, the sampling distribution of differences. This distribution is at a level more abstract than the population of raw scores. The denominator in Equation 2, on the other hand, is the standard deviation of one of the two conditions in that equation. This is a property of the population of raw scores. It follows that the test statistic used in NHSTP and the effect size are indices belonging to two different levels of abstraction. It seems neither valid nor appropriate to say that the relationship between statistical significance and the effect size is anomalous under such circumstances. This issue of mixing two levels of abstraction will surface again in the discussion of power analysis.

## 17. Effect size, the binary NHSTP decision, and evidential support

Two points are emphasized in the anomaly critiques of NHSTP. First, the NHSTP result is a binary decision (i.e., significant versus nonsignificant). Second, the effect size is a continuous variable. However, the propriety of juxtaposing statistical significance and the effect size may also be questioned for the following reasons. First, these criticisms are made with the assumption that $H_1$ is the substantive hypothesis. However, critics have not taken into account the facts that $H_1$ is the complement of $H_0$, and that $H_0$ is a hypothesis about chance influences on data. In other words, $H_1$ is neither the substantive nor the experimental hypothesis. It is but a statement to the effect that chance influences may be ruled out as an explanation of data. Consequently, to say that the result is statistically significant is to say something about the data and their collection. Statistical significance does not say anything about the substantive hypothesis.

The second reservation about critics' juxtaposing statistical significance and the effect size is a meta-theoretical one. To suggest supplementing statistical significance with the effect size in the theory-corroboration experiment is to say that the effect size has something to contribute to the evidential support for the substantive hypothesis. The putative importance of the effect size can be discounted in view of the argument that the warranted assertibility offered by experimental data is conferred by the implicative relations among the quartet of hypotheses (see Table 3) and the inductive principle underlying the experimental design (see Table 5), not by statistics. The effect size has no role in either the deductive or the inductive reasoning depicted in Tables 3 and 4. It follows that a larger effect size does not mean a greater support for the substantive hypothesis (also see Chow 1988). At the same time, the binary NHSTP suffices to provide the minor premise for the first conditional syllogism depicted in Table 3.

## 18. Effect size and practical importance

Something seems amiss to critics when nothing can be learned about the practical impact of the statistically significant research result. It is suggested that this shortcoming is the result of relying on NHSTP. Moreover, it can be rectified by reporting the effect size, particularly when the binomial effect-size display (BESD) is used (Rosenthal & Rubin 1979; 1982). This may be called the "effect informs impact" claim. Of interest are (a) the fact that the argument in support of the claim is incomplete, and (b) the reason why the claim intrudes into the assessment of NHSTP. This discussion will clarify the unwarranted practice of conflating statistical hypothesis testing with theory corroboration.

## 19. The "effect informs impact" claim revisited

There is a conceptual gap in the "effect informs impact" claim. Consider the correlation coefficient, *r*, between medication (aspirin vs. placebo) and myocardial infarction, MI (absence or presence) in Rosnow and Rosenthal's (1989) illustration. The *r* is used as an index of the effect size. What BESD does effectively is to convert the Pearson $r = 0.034$ into the "change in success rate" in the form of a percentage, where "success" means the absence of MI in the illustration. The "success rates" for the aspirin and placebo conditions are given, respectively, by Equations 3 and 4, respectively, as follows:

[a] The success rate for the Aspirin Condition: $0.5 + r/2$ (Equation 3)
[b] The success rate for the Placebo Condition: $0.5 - r/2$ (Equation 4)

The change in success rate is simply the difference between [a] and [b]. It turned out to be 3.4%. The conclusion is drawn that the implications of an effect of this magnitude are "far from unimpressive" (Rosnow & Rosenthal 1989, p. 1279), despite the fact that an $r = 0.034$ is statistically nonsignificant.

The BESD is justified on the grounds that it is "intuitively appealing . . . [and] easily understood by researchers, students, and lay persons" (Rosenthal 1983, p. 11). The difficulty is that the validity of this justification itself is by no means self-evident. It is simply not clear why the said rate of 3.4% is impressive. Would the same rate of change be impressive if the research were about the attitude change of some obscure film critics? Would it be more impressive if the film critics were prominent ones? It seems that, at least in the "Aspirin-MI" example, a change in success rate of 3.4% owes its impressiveness to the nature of the to-be-monitored phenomenon (viz., incidents of MI), not to the magnitude of the change itself.

There is also the following question. To whom is the effect size impressive? A 3.4% change in the attitudes of film critics may not impress those who are interested in artistic issues. However, it may have a greater impact on film producers when they consider the monetary implications. In other words, impressiveness is in the eye of the beholder, not the size of the effect per se.

In short, by itself, the effect size says nothing about the practical impact of the result. What is required is some criteria that relate the effect size to the judgment about impressiveness or practical impact. These criteria are outside the domain of statistics. Moreover, these criteria are domain-specific. Consequently, the claim that BESD is the general purpose index of practical impact is questionable. At the same time, the propriety of criticizing NHSTP in terms of practical validity may also be questioned because statistics and practical impact belong to different domains (Chow 1991a; 1991c).



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## 20. The intrusion of nonstatistical issues

The kernel-negative experiment used to introduce the rationale and procedure of NHSTP is like neither the examples used to introduce NHSTP in statistics textbooks nor those used in criticisms of NHSTP. The commonly used examples are studies designed to ascertain the effectiveness of a course of action or treatment (e.g., using a new method to teach statistics). Typically, the new method is applied to one class of students, whereas the traditional method is used in another class of students. The mean performance of the two classes is tested with NHSTP. The only concern is whether or not the new method of teaching produces a better result. This is an issue about treatment assessment, not about *why* the new method produces a better result. Experiments of this type are tokens of the agricultural model experiments (Hogben 1957; Meehl 1978; Mook 1983). Given their pragmatic objective, these experiments may also be characterized as utilitarian experiments. To see why nonstatistical issues intrude on the discussion of the role of NHSTP in empirical research, it is necessary to consider the nature of the utilitarian experiment.

## 21. The differences between the utilitarian and theory-corroboration experiments

It may be recalled from Table 1 that experimental data in the theory-corroboration experiment are at increasing deductive distances from the experimental, research, and substantive hypotheses. As may be seen from Table 9, the same is not true of the utilitarian experiment for the following reason. Given the specificity of the objective, the choice of the independent and dependent variables in the utilitarian experiment is restricted by the research objective itself. This, in turn, determines the experimental and research hypotheses. Consequently, the statistical and substantive hypotheses are indistinguishable.

Additional differences between the utilitarian and the theory-corroboration experiments have been shown in Table 10. These differences may be used to understand, as well as to answer, some of the criticisms of NHSTP. To begin with, it has been noted that the impetus for the utilitarian experiment is primarily, if not exclusively, to find the solution to a practical problem (e.g., students' poor understanding of statistics; see Row 1 in Table 1). That is, the role of a theory, if there is one at all, is minimal in this kind of experiment (hence, the "atheoretical" characterization in Row 4).

Suggestive of this difference is the fact that, whereas unobservable hypothetical entities or processes (e.g., the language processor) are the concerns of the theoretical endeavor in the theory-corroboration experiment, the subject matters of utilitarian experiments are observable activities or events (e.g., students' test scores; see Row 2). The result of the utilitarian experiment is used to guide a particular course of action (e.g., whether or not to adopt the new method of teaching; see Row 3). Experimental data in the theory-corroboration experiment, on the other hand, are used to assess whether or not there is evidential support for an explanatory substantive hypothesis (see Row 3). No pragmatic course of action follows. Nor is any practical problem solved as a result of the theory-corroboration experiment.

The experimental manipulation in the utilitarian experiment is the to-be-assessed efficient cause itself (e.g., the new method of teaching versus the traditional teaching method; see Row 7). However, the independent variable used in the theory-corroboration experiment is not an efficient cause. For example, the presentation of a kernel or a negative sentence does not shape or constrain subjects' behaviour in the way a teaching method may shape students' learning. In presenting kernel and negative sentences, the experimenter provides the hypothetical linguistic processor different contexts or environments in which to exhibit its theoretical properties. In other words, the independent variable in the theory-corroboration experiment is either a formal or a material cause, not an efficient one.

Table 9. *The logical relations among the to-be-investigated phenomenon, pragmatic, research, and experimental hypotheses of the utilitarian experiment*

| | What is said at the level concerned | |
|---|---|---|
| *To-be-investigated phenomenon* | *A dissatisfaction with students' current understanding of statistics.* | |
| Substantive (pragmatic) hypothesis | Method E is more effective than Method C. | [P9.1] |
| Research hypothesis | *If* [P9.1], *then* Method E produces better understanding than Method C. | [P9.2] |
| Experimental hypothesis | *If* the consequent of [P9.2], *then* students taught with Method E have higher scores than those taught with Method C. | [P9.3] |
| "Statistical Alternative hypothesis" | *If* consequent of [P9.3], *then* $H_1$.[1] | [P9.4] |
| Sampling distribution of $H_1$ | *If* $H_1$, *then* the probability associated with a difference between Methods E and C as extreme as 1.729 standard error units from an **unknown** mean difference is not known (assuming df = 19). | [P9.5] |
| Sampling distribution of $H_0$ | *If* $H_0$,[2] *then* the probability associated with a difference between Methods E and C as extreme as 1.729 standard error units from a mean difference of zero is 0.05 in the long run (assuming df = 19). | [P9.5′] |

1. $H_1$ = mean of Method E > mean of Method C.
2. $H_0$ = mean of Method E ≤ mean of Method C.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## 22. "Effect": Vernacular and technical meanings

The contrast between the independent variable being the efficient cause in the utilitarian experiment versus the formal (or material) cause in the theory-corroboration experiment has important implications for how "effect" or "effective" is understood in the context of NHSTP. "Effect" is used in its vernacular sense in the ambiguity-anomaly and the insensitivity to effect size criticisms of NHSTP. This is also the sense assumed (as well as congenial to) the utilitarian experiment (see Rows 5 and 10 in Table 10). This is understandable in view of the fact that the experimental manipulation itself is substantively efficacious (e.g., method of teaching). This does not mean, however, that it is justified when the independent variable is not an efficient cause (e.g., sentence type). What is important is that it is not even justified when the experimental manipulation consists of two efficient causes, but for a different reason.

To adopt the vernacular meaning of "effect" is to use a statistically significant result to do something more than rejecting chance influences as an explanation. It is to assert that the research manipulation is *the* explanation (see also sect. 31, "The specificity of $H_1$ and related issues"). This assumption is justified only to the extent that the inductive conclusion validity is assured, however. In fact, as has been noted earlier in section 15, "The sample size-significance dependence problem revisited," questions about a statistically significant result arise because there are doubts about the inductive conclusion validity. More important, these questions are not statistical ones. Consequently, it is doubtful that specifying the effect size or determining the confidence interval estimate would allay the nonstatistical concerns that underlie the reservations about the statistically significant result.

Recall that $H_0$ is a statement about the consequence of chance influences on data collection. "Effect" at this level of discourse refers to the difference between the means of two data collection conditions. The NHSTP concern is whether or not the difference is large enough for the rejection of the explanation in terms of chance influences. This technical meaning of "effect" is different from its vernacular meaning. It does not implicate any assumption of efficacy. More importantly, by itself, NHSTP does not identify the reason for the sufficiently large difference that leads to the "statistically significant" decision. Nor should there be any reason to expect an answer coming from NHSTP when the issues implicated are nonstatistical ones.

In sum, critics' concern about the effect size may be represented by the questions tabulated in the left-hand column of Table 11 (see Rosnow & Rosenthal 1989). These questions are asked because "effective" is interpreted in its vernacular sense. However, Question [PV-2] does not directly lead to [PV-3] or [PV-4]. It is necessary to provide an independent set of criteria outside the domain of statistics to justify asking Question [PV-3] or [PV-4] in conjunction with Question [PV-2] (see sect. 19, "The 'effect informs impact' claim revisited"). Such a set of criteria is not available.

Suppose that the technical meaning of "effect" is adopted in discussing NHSTP. Although Question [CR-1] is literally the same as [PV-1], it leads to an entirely different set of questions relating to the difference between two data collection conditions brought about by the experimental manipulation. It may be seen readily that, with the excep-

Table 10. *Some differences between the agricultural (utilitarian) model and theory-corroboration experiments*

| | | Agricultural model (utilitarian) | Theory corroboration |
|---|---|---|---|
| 1. | Impetus | To solve a practical problem; reflexive data collection. | To explain a phenomenon; independent of data collection. |
| 2. | Subject matter | The practical problem involving observable events. | Unobservable hypothetical entity and its theoretical properties. |
| 3. | Consequence of research | Take a particular course of action; closure of investigation. | Accept tentatively, revise or reject the theory; no closure to the investigation. |
| 4. | Role of theory | Atheoretical. | To-be-tested theory explicitly stated; used to guide experimental design. |
| 5. | Substantive question | "Is the treatment effective?" "How effective is the treatment?" | "Why does the phenomenon occur?" |
| 6. | Experimental hypothesis | The practical question itself. | Qualitatively different from the to-be-assessed substantive hypothesis. |
| 7. | Experimental manipulation | The to-be-assessed efficient cause itself. | Different from the to-be-explained phenomenon. |
| 8. | Dependent measure | The practical problem itself. | Different from the to-be-explained phenomenon. |
| 9. | Statistical significance | To indicate that the explanation of data in terms of chance variations can be ruled out at the $\alpha$ level. | |
| 10. | Effect | Substantive efficacy (i.e., the consequence of an efficient cause). | The difference between the means of two conditions (i.e., the consequence of a formal or a material cause). |
| 11. | Ecological validity | Necessary. | Irrelevant, may even be detrimental. |



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Table 11. *Different sets of research questions pertinent to practical validity (PV) and conceptual rigor (CR) for the utilitarian and theory-corroboration experiments, respectively*

| | Practical validity concerns (utilitarian research) | | Conceptual rigor concerns (theory-corroboration research) |
|---|---|---|---|
| | The independent variable is the *efficient* cause. | | The independent variable is the *material* or *formal* cause. |
| [PV-1] | Is Treatment T effective? | [CR-1] | Is Treatment T effective? |
| [PV-2] | How effective is Treatment T? | [CR-2] | Is the independent variable a valid choice? |
| [PV-3] | How impressive is Treatment T? | [CR-3] | Do the data warrant the acceptance of Theory K, which underlines the choice of the dependent variable? |
| [PV-4] | Is Treatment T important? | [CR-4] | Is the implementation of the independent variable valid? |
| | | [CR-5] | Does the study have hypothesis validity? |

tion of Question [CR-5], these are questions about the data-collection conditions, particularly the inductive principle that underlies the experimental design.

## 23. Power analysis

The power of a statistical test has recently become an important consideration in the assessment of empirical studies in psychology. Cohen's (1987) power analytic approach to empirical research has the following themes. First, if Phenomenon P exists, its effects must be detectable. Second, the evidence for the truth of a substantive hypothesis about Phenomenon P is the detectability of the effect envisaged in the hypothesis. Third, the substantive hypothesis is represented by $H_1$ in NHSTP. Fourth, to detect the effect is to obtain statistical significance (i.e., to accept $H_1$ by rejecting $H_0$). Hence, statistical significance is indicative of the truth of $H_1$ or the fact that Phenomenon P exists. These four interconnecting themes may collectively be identified as the *existence-detectability-significance thesis.* For this reason, it is important for power analysts to know the *a priori* probability of obtaining statistical significance. That *a priori* probability is the power of the statistical test (Cohen 1987; see also Mosteller & Bush 1954).

The Type II error is assumed by critics to have real-life consequences. Hence, NHSTP users are faulted for ignoring it as a result of their exclusive obsession with the Type I error. With the advent of power analysis, the Type II error can now be controlled by specifying the level of statistical power desired for the investigation. This is possible because the power of a statistical test is $(1 - \beta)$, where $\beta$ is the probability of committing the Type II error. The value of can be controlled by setting the level of the power.

That power analysis is currently well received is understandable in view of the facts that critics are convinced that NHSTP is problematic and that power analysis is presented as a remedy for the difficulties of ambiguity and anomaly attributed to NHSTP. However, if the criticisms of NHSTP themselves are debatable, it should become easier to consider power analysis in a more judicious way. There are good reasons to question the existence-detectability-significance thesis of power analysis.

Consider its first theme, namely, that if $H_1$ is true, there is a detectable effect. This theme is contrary to the fact that the tenability of some hypotheses depends on *not* rejecting $H_0$ (i.e., not detecting any effect, in the parlance of power analysis). An example is Schneider and Shiffrin's (1977) study of automatic detection. The third theme of the thesis, that $H_1$ is the substantive hypothesis, is debatable in view of the quartet of hypotheses identified in Table 1 and the discussion in section 12, "$H_0$, data, and chance influences." Consequently, all power analytic assertions based on identifying $H_1$ with the substantive hypothesis are questionable.

The detectability of the effect is equated with statistical significance in the second theme of the existence-detectability-significance thesis. This makes explicit an implicit assumption in power analysis that NHSTP is the same as the theory of signal detection procedure (TSD). An examination of this NHSTP-TSD affinity assumption reveals additional conceptual difficulties in power analysis.

## 24. The NHSTP-TSD affinity in power analysis

Indicative of the NHSTP-TSD affinity envisaged in power analysis are assertions such as: "Since effects are appraised against a background of random variation" (Cohen 1987, p. 13), and "[the said appraisal consists of] *detecting* a difference between the means of populations A and B . . . " (Cohen 1987, p. 6, emphasis added). At the level of rationale, the appeal is made to Neyman and Pearson's (1928) emphasis on the posterior probability. It is believed that researchers first determine what a sample statistic is (e.g., $\bar{X}$). They then ask (or wish to ask) what the probability is that the sample has been selected from Population P with parameter $\mu$ (see Cohen 1994). An appeal to the *a posteriori* probability in this "from sample statistic to population parameter" manner is also found in a TSD analysis.

It is recognized in TSD that an observer's response bias is a function of the prior odds (viz., the probability of the noise event to that of the signal event) and the payoff matrix (i.e., the costs for committing errors and the gains caused by making a correct detection). Something very similar is suggested in power analysis. Specifically, it is suggested that the placement of the decision axis used to make the statistical decision should reflect a balance struck between statistical power and $\alpha$ (Cohen 1987, p. 5). This is achieved by taking into account the ratio of the probability of the Type II error to the probability of the Type I error. Researchers are further urged to pay attention to "the relationship between *n* and power *for* [their] *situation*, taking into account the increase in cost to achieve a given increase in power" (Cohen 1965, p. 98; Cohen's emphasis).

## 25. Issues raised by the NHSTP-TSD affinity

A correspondence between two sets of descriptive terms becomes obvious if the affinity between NHSTP and TSD



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

is recognized. Of particular interest is that between *statistical power* and *hit rate.* It renders questionable the following assertion: "The power of a statistical test is the probability that *it will yield* statistically significant results" (Cohen 1987, p. 1, emphasis added). (Quote 2)

## 26. Statistical power: A conditional probability

The Type I error is made when the researcher rejects a true $H_0$; this is analogous to committing a false alarm in TSD. Power analysts use [$H_1$ True] as a subcolumn heading in the upper left panel of Table 12. The Type II error is committed when the researcher fails to reject $H_0$ when $H_1$ is true. The logical complement of Type II error (viz., rejecting $H_0$ when $H_1$ is true) in NHSTP is equivalent to a *hit* in TSD (see the upper right panel). Note that a hit in TSD refers to a "Yes" response *contingent on* the presence of a signal event. That is, a *hit* is a characterization of the observer's behavior, given that the signal is present. It says nothing about the signal event per se. It follows that the hit rate in TSD is a conditional probability, namely, the probability of an observer saying "Yes" when a signal event does indeed occur. In other words, the hit rate says nothing about the exact probability of the presence of a signal event.

At the same time, as may be seen from the two lower panels of Table 12, the TSD analog of statistical power is the hit rate. Hence, the statistical power is a conditional probability (see also Chow 1991c). That is to say, knowing the power of a test (a conditional probability) is not the same as knowing the probability of obtaining statistical significance (an exact probability). More important, given the NHSTP-TSD affinity, the *statistical power* index says something about the researcher, not $H_1$, in much the same way the hit rate says something about the observer, not the signal event. In short, statistical power does not (and cannot) enlighten us as to the probability of obtaining statistical significance.

## 27. Statistical power: A misleading sense of efficacy

An efficacious capability is attributed to the statistical procedure in Quote 2. It suggests that statistical significance is reached by virtue of the numerical index, *statistical power* (see the emphasis in Quote 2). This assertion is misleading because, at the level of statistics, *statistical power* simply refers to the cumulative probability over a range of parameter values (viz., *all* values that are as extreme as the critical values of the test statistic). No efficacy of any sort is implicated at this level of discourse. A nonstatistical theoretical justification is required if an efficacious capability is attributed to statistical power. Because no such justification is offered, it is only proper *not* to attach any extra-statistical meaning to the term *statistical power.*

There is no *a priori* reason why the decision to reject $H_0$ in the event it is false should not simply be called *Type II correct decision.* Power analysis might not have been so readily accepted had a nonevocative term like *not-β* been used instead of *power.* Perhaps an excess and unwarranted meaning is attributed to a conditional probability as a result of its being labeled with the evocative term *power,* a connotative meaning of which is *being efficacious.* The same is also true of *statistical significance.*

Table 12. *The correspondence between some concepts (upper left) and their probabilities (lower left) in NHSTP and concepts (upper right) and their probabilites (lower right), given the NHSTP-TSD affinity in power analysis.*

*Upper Panel*

| NHSTP Concepts | | | TSD Concepts | | |
|---|---|---|---|---|---|
| Decision | State of affairs | | TSD response | State of affairs | |
| | $H_0$ True | $H_0$ False [$H_1$ True] | | Noise | Signal |
| "Not reject" | Correct acceptance | Type II error | "No" | Correct rejection | Miss |
| "Reject" | Type I error | Correct rejection | "Yes" | False alarm | Hit |

*Lower Panel*

| NHSTP Concepts | | | TSD Concepts | | |
|---|---|---|---|---|---|
| Decision | State of affairs | | TSD response | State of affairs | |
| | $H_0$ True | [$H_1$ True] | | Noise | Signal |
| "Not reject" | p(Correct acceptance) | p(Type II error) = $\beta$ | "No" | Correct rejection rate | Miss rate |
| "Reject" | p(Type I error) = $\alpha$ | Power = $(1 - \beta)$ | "Yes" | False alarm rate | Hit rate |



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## 28. Graphical representation of statistical power, effect, and NHSTP

It is taken for granted in the discussion so far that the concept, *statistical power,* is valid. The validity of power analysis becomes more questionable if there are reservations about the validity of *statistical power* itself. That Quote 2 is inconsistent with statistical power being a conditional probability is one such reservation. There are additional reservations.

## 29. Two levels of abstraction: Statistical significance and statistical power

Consider the assertion: "A salutary effect of power analysis is that it draws one forcibly to consider the magnitude of effects" (Cohen 1990, p. 1309). This assertion is made because of the functional relationship between statistical power and effect size (given $n$ and $\alpha$) envisaged in power analysis. This functional relationship is readily seen from Panels A and B of Figure 1. Before proceeding any further, it must be noted that Cohen (1965; 1987; 1992a; 1992b) does not use any graphical representation when he discusses statistical power, effect size or the functional relationship between the two. Nonetheless, Figure 1 is used for ease of exposition. Its use is justified by the fact that it is consistent with how *d* and *statistical power* are defined in power analysis.

The x-axis in both panels represents population scores (as stipulated by how *d* is defined in Equation 2 in sect. 16). The left and right distributions in either panel represent the control and experimental distributions, respectively. The effect size is shown by the distance between the two distributions, and statistical power by the area shaded with slanting lines. To power analysts, Panels A and B represent two situations in which the desired effect is larger in Panel B than in Panel A, and Panel B represents a more powerful test than Panel A. Of interest is whether or not research manipulations that are expected to be differentially efficacious would have a different impact on NHSTP.

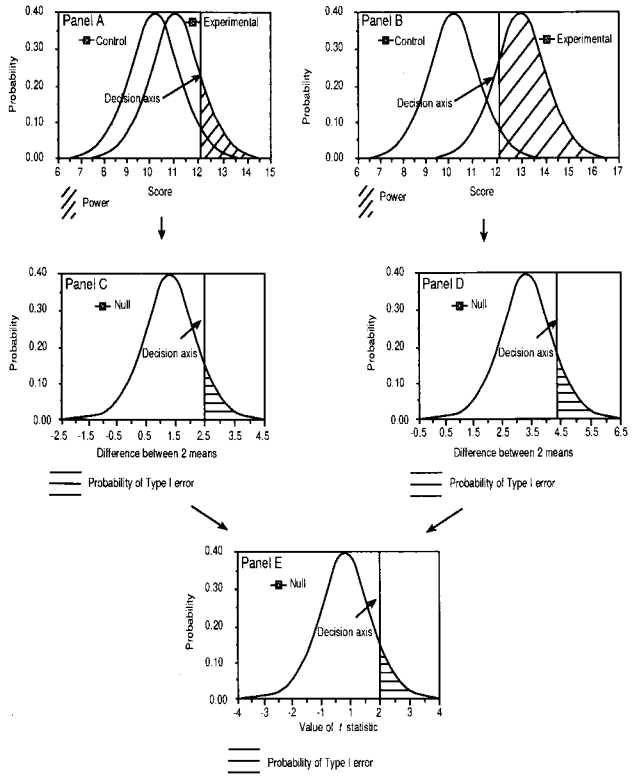



Figure 1. The graphical representation of two effect sizes (Panels A and B), and the corresponding differences between two means in raw-score units (Panels C and D), as well as in standard error units (Panel E).

## 30. $H_0$ and research manipulation efficacy

The pair of population distributions in the "small effect" situation (viz., Panel A in Fig. 1) gives rise to the lone sampling distribution of the difference depicted in Panel C of Figure 1. Similarly, the pair of population distributions in the "large effect" situation brings about another lone sampling distribution of the difference (i.e., the one depicted in Panel D of Fig. 1). The two sampling distributions of the difference in Panels C and D have the same standard error of the difference seen in the present example. However, the two sampling distributions cover different parts of the difference between two means continuum in raw-score units (viz., from -2.5 to 4.5 in Panel C versus from −0.5 to 6.5 in Panel D).

Consider the numerator used in calculating the test statistic, *t*. It is often written as $(\overline{X}_1 - \overline{X}_2)$. However, it is really a short-hand form for $[(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2) = 0]$. As has been noted before, the $(\mu_1 - \mu_2)$ component is left out when it is numerically equal to 0 (see Kirk 1984). The distribution in the top panel of Figure 2 represents a sampling distribution of the difference for a situation in which $\mu_1 - \mu_2 = 0$. That is, the mean difference in the sampling distribution of the difference between two means is zero.

Power analysts suggest that the desired difference, $(\overline{X}_1 - \overline{X}_2)$, may be 3.0 (or 1 or any definite value), rather than 0. The numerator now becomes $[(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2) = 3.0]$ or $[(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2) = 1.0]$. That is, the mean difference in the sampling distribution of the difference implicated in NHSTP is 3.0 (or 1.0), and it is graphically represented in the bottom (or middle) panel of Figure 2. The three sampling distributions in the three panels of Figure 2 have the same standard error of difference, but different values for the mean difference (viz., 0, 1, and 3.0). They represent the sampling distribution under $H_0$ in three different situations. Specifically, the bottom panel represents a research manipulation expected to be more efficacious than the one depicted in the middle or top panel.

Depicted on the x-axis of the graphical representation in any panel of Figure 2 is the range of possible values of the difference between two means. In other words, the three panels in Figure 2 collectively show that the difference in the expected efficacy of the research manipulation is represented by the spatial displacement of the sampling distribution of the differences between two means along the continuum of all possible values of the difference between two means. This state of affairs is different from the impression conveyed by Panels A and B in Figure 1.

In carrying out NHSTP, only one sampling distribution is used (viz, the one contingent on $H_0$ being true). Moreover, the researcher uses a standardized form of the sampling distribution depicted in either Panel C or D of Figure 1 (viz., the *z* or *t* distribution; see Siegel 1956). That is, regardless of the mean difference in raw-score units, the



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

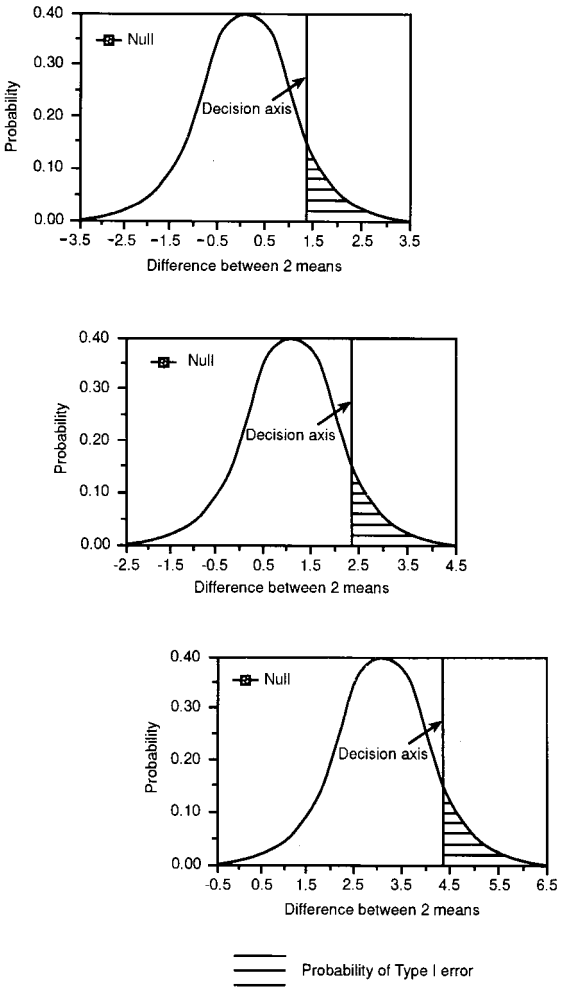



Figure 2. The sampling distribution of the difference in raw-score units when the mean difference is 0 (top panel), 1 (middle panel), and 3 (bottom panel).

standardized representation of the to-be-used sampling distribution of the difference remains the same (viz., Panel E in Fig. 1). More importantly, the location of the decision axis *vis-à-vis* the mean of the sampling distribution of the difference remains unchanged for the same $\alpha$ level. It follows that the outcome of NHSTP is *not* affected by the desired effect or expected efficacy of the research manipulation.

Figure 1 shows that two distributions of population scores converge on one standardized distribution via a lone sampling distribution of the test statistic. Panel A or B in Figure 1 shows that it takes two population distributions to depict statistical power, whereas Panel E shows that only one sampling distribution is used to depict NHSTP. Moreover, two different levels of discourse are implied in Panels A (or B) and E. This demonstrates that it is impossible to represent statistical power graphically without misrepresenting NHSTP. It casts doubt on the validity of the concept of statistical power.

Some important points may now be summarized. First, no distribution based on $H_1$ is implied in NHSTP (see Panel E of Fig. 1). Second, the mean difference in raw-score units of the sampling distribution of difference reflects the theoretical difference between two population means. When expressed in terms of the raw-score unit, this difference is graphically represented by the spatial displacement of the sampling distribution on the difference between two means continuum (see the three panels in Fig. 2).

Third, it is not possible to represent graphically the conditional probability, statistical power, if the rationale of NHSTP is properly represented with a single sampling distribution of the difference between two means. Fourth, the desired effect of the research manipulation (in the technical sense of the word) has no impact on NHSTP because the to-be-employed sampling distribution is standardized (e.g., in the form of the appropriate *t* distribution) before being used to make the "chance versus nonchance" decision.

## 31. The specificity of $H_1$ and related issues

For nonpower analysts, "Type II error" in the upper left panel of Table 12 refers to the error committed when a false $H_0$ is not rejected (i.e., ignore the [$H_1$ True] column heading). No mention is made of $H_1$ in this definition. It may be recalled from the lower panel of Table 2 that $H_0$ and $H_1$ are mutually exclusive and exhaustive. This is emphasized in Table 7 by depicting that $H_0$ is the implication of chance influences, and that $H_1$ is the implication of some ill-defined, nonchance influences. It follows that, although $H_0$ and $H_1$ are mutually exclusive and exhaustive alternatives, "$H_0$ False" is not synonymous with "$H_1$."

Defining "Type II error" in terms of "$H_0$ False" instead of [$H_1$ True] in the upper left panel of Table 12 helps maintain the distinction between inductive conclusion validity and statistical conclusion validity. Specifically, whereas NHSTP is used to decide between chance influences and nonchance influences (see Tables 2 and 7), inductive reasoning is employed to identify the nonchance factor involved (see Table 5). It is also important that $H_1$ is numerically nonspecific (see [P1.5] in Table 1).

To define *power* in power analysis, "Type II error" is defined as the error committed in the event that $H_1$ is true. That is, it is necessary to use the [$H_1$ True] heading in the upper left panel of Table 12. Moreover, $H_1$ is given a specific nonzero numerical value in power analysis. This effectively changes the conceptual meaning of $H_1$ from an implication of nonchance influences to the consequence of a specific efficient cause. This is reminiscent of the consequence of using "effect" in its vernacular sense discussed in section 22: "Effect: Vernacular and technical meanings." Consequently, $H_0$ and $H_1$ are no longer mutually exclusive and exhaustive in the power analytic account of NHSTP. More important, in making the meaning of $H_1$ numerically specific, power analysts may have eschewed the distinction between the two types of internal validity. NHSTP is given the additional role that should be played by inductive logic.

The power analytic practice of making $H_1$ numerically specific is consistent with the Multiple-$H_1$ Assumption view that there are, in fact, multiple numerical alternatives to $H_0$ (Neyman & Pearson 1928; Rozeboom 1960). However, this assumption should have no bearing on NHSTP, as may be recalled from the "$H_0$ and research manipulation efficacy" discussion in section 30. Why is there the emphasis on multiple numerically specific $H_1$'s? The answer may be the fact that the term "alternative hypothesis" is also used in another sense, albeit at a different level of discourse.

Given any to-be-explained phenomenon, there are alternative explanatory theories at the conceptual level (Popper



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

1968a; 1968b). This state of affairs may be characterized as the *Reality of Multiple Explanations* view in subsequent discussions. In fact, different psychologists often explain the same phenomenon with various substantive hypotheses. Moreover, diverse hypothetical structures or functions are postulated in these competing theories.

For example, some psychologists prefer Fillmore's (1968) case grammar or Yngve's (1960) "Depth" model to Chomsky's (1957) transformational grammar. These three substantive hypotheses lead to different research and experimental hypotheses (*à la* the schema depicted in Table 1). As these experimental hypotheses may implicate different independent and dependent variables in diverse experimental situations, they lead to *qualitatively* different $H_1$'s. The distinction between the *Multiple-$H_1$ Assumption* and the *Reality of Multiple Explanations* views depicted in Table 13 can be used to defend NHSTP against the *Multiple-$H_1$ Assumption* critique of NHSTP.

## 32. Alternative substantive hypothesis versus statistical alternative hypothesis

Consider first the *Multiple-$H_1$ assumption* column in Table 13. In terms of the number of extra words recalled, $H_1$ of the kernel-negative experiment is shown in Row [a]. Two additional alternatives to $H_0$ are shown in Rows [b] and [c] in the *Multiple-$H_1$ assumption* column. Each one of these statistical alternative hypotheses is a point prediction. However, if the Multiple-$H_1$ Assumption and the Reality of Multiple Explanations view were the same, it would be necessary to show something like what follows: Alternative [a] is an implication of the transformational grammar, Alternative [b] is derived from the case grammar, and Alternative [c] follows logically from Yngve's (1960) "Depth" model. Ironically, were this the case, the researcher should be very unhappy about the three theoretical alternatives for the following reason.

Such a state of affairs occurs when the three numerical alternatives are possible outcomes in the same experimental context (e.g., the same independent variable is manipulated, as indicated by the subscripts used in the *Multiple-$H_1$ assumption* column). This takes place when there is no *qualitative* difference among the theoretical structures or mechanisms envisaged in the three substantive hypotheses. Consequently, they do not differ in terms of how well they explain our linguistic competence at the conceptual level. This means that the three hypotheses give the same qualitative prescription in a well-defined task context. In other words, the three hypotheses are merely variations of the same genre under such situations. The choice among the three alternatives becomes a nontheoretical one. In what sense does the quantitative difference in question matter if it does not make any difference at the explanatory level?

Table 13. *The distinction between* statistical alternative hypothesis *and* alternative explanatory hypothesis

| Multiple-$H_1$ assumption | Reality of multiple explanations |
|---|---|
| [a] $H_1$: $(u_{negative} - u_{kernel}) < 0$ | [i] $H_1$: $(u_{negative} - u_{kernel}) < 0$ |
| $H_0$: $(u_{negative} - u_{kernel}) = 0$ | $H_0$: $(u_{negative} - u_{kernel}) = 0$ |
| [b] $H_1'$: $(u_{negative} - u_{kernel}) = -3$ | [ii] $H_1'$: $(u_{ca} - u_a) > 0$ |
| $H_0'$:$(u_{negative} - u_{kernel}) = 0$ | $H_0'$: $(u_{ca} - u_a) = 0$ |
| [c] $H_1''$: $(u_{negative} - u_{kernel}) = 5$ | [iii] $H_1''$: $(u_{n-p} - u_{n-n}) \neq 0$ |
| $H_0''$: $(u_{negative} - u_{kernel}) = 0$ | $H_0''$: $(u_{n-p} - u_{n-n}) = 0$ |

ca = counter-agent; a = agent
n−p = negative sentence with positive meaning; n−n = negative sentences with negative meaning

Consider now the "Reality of multiple explanation" column in Table 13. To begin with, additional independent variables generally are necessary for the experiment to test multiple explanatory hypotheses simultaneously. For example, it may become necessary to manipulate *Sentence modality* (e.g., *agent* versus *counter-agent*) to test the case grammar. The depth of the sentence structure would have to be manipulated if the "Depth" model is being tested. Specifically, it may be necessary to manipulate the type of negative sentences being used (e.g., a negative sentence with a positive meaning versus one with a negative meaning).

A prerequisite for a successful theory-corroboration experiment is that different experimental prescriptions are indicated by the qualitatively different theories. For example, the prescription of the transformational grammar is in Row [i]. The case grammar prescribes $H_1'$ in Row [ii]. The "Depth" model prescribes $H_1''$ in Row [iii]. As may be seen from the subscripts of the various means, the three different statistical alternative hypotheses are the implications of their respective experimental hypotheses at the statistical level.

## 33. A triad of hypotheses: The substantive hypothesis, $H_1$ and $H_0$

Two things should be emphasized. First, multiple conceptual alternative hypotheses give rise to their respective statistical alternative hypotheses. Second, the multiple statistical hypotheses (e.g., $H_1$, $H_1'$, and $H_1''$ in Table 13) are not alternatives to one single $H_0$. They have their own null hypotheses (viz., $H_0$, $H_0'$, and $H_0''$, respectively) even though these null hypotheses may be numerically equal to zero. They are zero under different conditions, however. This is indicated by the fact that three different conceptual hypotheses implicate different independent variables (viz., sentence-type, case, and negative-type, respectively, as may be seen from the subscripts in Table 13). In other words, each of these multiple null hypotheses describes what chance variations are like under its own unique set of conditions.

In short, the differences among the experimental expectations prescribed by diverse alternative substantive hypotheses at the conceptual level are *not* a matter of numerical differences such as $u_1 = 5$, $u_2 = 10$, $u_3 = 15$, and the like. Consequently, the exclusion of unwarranted alternative hypotheses at the conceptual level is also *not* a matter of choosing among numerically different $H_1'$s (Chow 1989). It involves testing different $H_1'$s defined by dissimilar data-collection conditions. Each of these $H_1'$s has its own $H_0$.

## 34. Statistical power and sample size

The best known use of statistical power is for disambiguating the difficulties brought about by the arbitrariness, ambiguity, or anomaly attributed to NHSTP. It is argued that, if the test is of sufficient power, one can be sure that



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

the statistically significant result is genuinely significant, and that a nonsignificant result is really nonsignificant. An important stipulation is that an appropriate sample size be determined with the help of the general purpose *Sample Size Tables* (Cohen 1987). Researchers can determine the appropriate sample size with reference to (a) the desired power, (b) the desired effect size, and (c) the $\alpha$ level to be adopted. An alternative set of tables may be found in Kraemer and Thiemann (1987). This important function of statistical power is best summarized as follows:

> From a power analysis at, say, $\alpha = .05$, with power set at, say, .*95*, so that $\beta = .05$, also, the sample size necessary to detect this negligible effect with .95 probability can be determined. Now if the research is carried out using that sample size, and the result is not significant, as there had been a .95 chance of detecting this negligible effect, and the effect was not detected, the conclusion is justified that no nontrivial effect exists, at the $\beta = .05$ level. (Cohen 1990, p. 1309) (Quote 3)

This mechanical approach to sample-size determination is inappropriate for experimental studies. To begin with, no reference is made in power analysis to the experimental design used. In general, fewer subjects are required when the repeated-measures design is used than when using the completely randomized design. There are other considerations when the matched-pair (for the 1-factor, 2-level design) or matched-group (for the multi-factor, multi-level design) is used. The stability of the data may be influenced by the success of the matching procedure.

Another unsatisfactory feature of the power analytic way of determining the sample size is its disregard for how well-trained the experimental subjects are. This, in turn, is dependent on the nature of the experimental task used. Given the same experimental task, data stability may be secured by using a few well-trained subjects when an unusual task like Sperling's (1960) partial-report task is used. Sometimes the nature of the investigation demands a large sample of naive subjects (e.g., Keppel & Underwood's 1962 study of proactive interference). Furthermore, the number of subjects required may be influenced by the number of experimental sessions (as well as the number of trials within an experimental session). These important procedural considerations have not been taken into account in power analysis.

In sum, the concern with sample size has a lot to do with data stability. This issue cannot be settled with a mechanical procedure or a general purpose tool for the simple reason that data stability is not determined by the sample size alone. It is also affected by the nature of the experimental task, the amount of practice the subjects have before data collection, and the experimental design used. These considerations are some of the a priori considerations recommended by Neyman and Pearson (1928) in Quote 1 (sect. 15).

## 35. A criticism of NHSTP with a Bayesian overtone

As may be recalled from the "Null-hypothesis significance-test procedure (NHSTP)" discussion in section 4, important to NHSTP is the associated probability, $p$. It is the conditional probability, $p(\text{Data}|\text{H}_0)$. Some critics find the reliance on $p$ unsatisfactory for various reasons. For example, $p$ is often misunderstood or knowing $p(\text{Data}|\text{H}_0)$ is not knowing $p(\text{H}_0|\text{Data})$. At the same time, the following assertion is made in power analysis:

> Now, what really is at issue, what is always the real issue, is the probability that $H_0$ is true, given the data, $P(H_0|D)$, the inverse probability. (Cohen 1994, p. 998) (Quote 4)

This concern with the inverse probability is like the Bayesian appeal to the posterior probability. To find NHSTP wanting for this reason goes beyond statistics, however. It raises instead questions about the nature or purpose of conducting empirical research. Of interest to the present discussion are methodological issues underlying the Bayesian, as well as the power analytic, theme that empirical data are collected to ascertain the posterior (or inverse) probability of the hypothesis of interest. The methodological issues are (a) the prototype of empirical research envisaged in the Bayesian approach, (b) the nature of the Bayesian hypothesis, and (c) the role of replication studies in empirical research.

## 36. The Bayesian hypothesis and sequential sampling procedure

It is assumed in the Bayesian approach that, before collecting data, the researcher attributes a prior probability (viz., the degree of belief) to the hypothesis. The research objective is not to ascertain the tenability of the hypothesis. Instead, data are collected to adjust the prior degree of belief with the Bayesian theorem. The new degree of belief in the hypothesis is the posterior probability. Given the Bayesian theorem, the evidential support for the hypothesis offered by the current data may not be sufficient to overcome the impact of the prior degree of belief. Hence, it is a Bayesian theme that researchers must take into account the prior probability of the hypothesis when they interpret the data.

That the Bayesian approach is of limited applicability to psychological research may be seen more readily after a discussion of the type of data collection exercise suitable for Bayesian analysis. More important, it may be seen that the Bayesian approach cannot be used for theory-corroboration purposes because it cannot be used to test explanatory hypotheses. The following example is adapted from Phillips's (1973) illustration. The Bayesian emphasis on the prior probability is antithetical to objectivity.

Suppose that a newspaper editor, E, would endorse the centre party in the coming election when it is preferred by 75% of the prospective voters polled. Editor E commissions polls about the impending election to determine if the criterion for endorsing the centre party is met. Columns 1 through 5 in Table 14 represent 5 successive polls conducted. Entries in the "Prior Probability" rows represent Editor E's prior degrees of belief in the three parties winning the election before the polling period specified by the column number. Specifically, Editor E assigns prior probabilities of .50, .60, and .40 to the left-wing, centre, and right-wing parties, respectively, before the first poll (see the "prior probability" entries in Column 1 in Table 14). The entry in the cell intersecting a column and the "evidence" row represents the poll result about the centre party in the poll in question. In other words, the percentages of people polled who chose the centre party are 30, 38, 45, 55, and 50 in Periods 1 through 5, respectively.

The "likelihood of evidence" is the probability of the evidence (e.g., 30% of those polled favor the centre party) if the centre party actually wins the election (viz., .35). As may be seen from the "prior probability $\times$ likelihood" and



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Table 14. *The accumulation of data and the conversion of the prior probability (Prior DOB) into its corresponding posterior probability at successive research stages*

| | | Poll-data inspection period | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Prior probability | $H_L$ | .50 | .36 | .31 | .27 | [.27] |
| | $H_C$ | .60 | .40 | .44 | .59 | [.59] |
| | $H_R$ | .40 | .32 | .24 | .14 | [.14] |
| Evidence | | 30% | 38% | 45% | 55% | [50%] |
| Likelihood[1] of evidence | $H_L$ | .38 | .38 | .40 | .30 | [.40] |
| | $H_C$ | .35 | .50 | .60 | .75 | [.60] |
| | $H_R$ | .32 | .35 | .25 | .35 | [.25] |
| Prior probability × likelihood | $H_L$ | .19 | .14 | .12 | .08 | [.11] |
| | $H_C$ | .21 | .20 | .26 | .44 | [.35] |
| | $H_R$ | .13 | .11 | .06 | .05 | [.04] |
| Posterior probability | $H_L$ | $\frac{.19}{.53} = .36$ | $\frac{.14}{.45} = .31$ | $\frac{.12}{.44} = .27$ | $\frac{.08}{.57} = .18$ | $\left[\frac{.11}{.50} = .22\right]$ |
| | $H_C$ | $\frac{.21}{.53} = .40$ | $\frac{.20}{.45} = .44$ | $\frac{.26}{.44} = .59$ | $\frac{.44}{.57} = .77$ | $\left[\frac{.35}{.50} = .70\right]$ |
| | $H_R$ | $\frac{.13}{.53} = .32$ | $\frac{.11}{.45} = .24$ | $\frac{.06}{.44} = .14$ | $\frac{.05}{.57} = .09$ | $\left[\frac{.04}{.50} = .08\right]$ |

$H_L$ = The left-wing party will form the next government. [H7-1]
$H_C$ = The centre party will form the next government. [H7-2]
$H_R$ = The right-wing party will form the next government. [H7-3]
Evidence = The percentage of voters polled indicated a preference for the centre party in the present example.

**1.** Likelihood of evidence = The probabilty of the evidence, given that $H_C$ ($H_L$ or $H_R$) is true.

"posterior probability" rows, the posterior probability is given by the Bayesian theorem represented by Equation 5:

$$\textit{Posterior Probability} = \frac{\textit{Prior Probability} \times \textit{Likelihood of Evidence}}{\textit{Sum of All (Prior Probability} \times \textit{Likelihood of Evidence)}} \quad \text{(Eq. 5)}$$

Editor E's mode of decision-making is called the "sequential sampling procedure" (Phillips 1973, p. 66) because of the following characteristics of the data-collection procedure. First, evidential information is gathered in stages. Second, the status of the evidence is examined at the end of every stage (e.g., a percentage in Table 7.1 in my book). Third, the evidence collected in successive stages is accumulated in the following way: the posterior probability of any stage serves as the prior probability of its immediately succeeding stage. Fourth, the data collection procedure is self-terminating in the sense that it stops when the posterior degree of belief assumes a certain value.

It is important to emphasize that these four sequential sampling features are not found in a typical experiment (e.g., the kernel-negative experiment described above). Moreover, Phillips's (1973) "sequential sampling procedure" characterization does not reflect four other important features of the Bayesian approach. For ease of exposition, these four additional features will be called the "reflexive" features of the Bayesian data collection procedure.

First, none of the hypotheses is proposed to explain a phenomenon that invites the investigation. Instead, they are hypotheses about an uncertain event in the future. This is unlike the explanatory substantive hypothesis depicted in Table 1. Second, the Bayesian analysis is not about the truth of hypotheses at all. Editor E does not collect data to accept or reject any of the hypotheses (e.g., $H_C$). Rather, Editor E is interested in the "probabilification" of the hypothesis (Earman 1992, p. 79). Hence, the reasoning shown in Table 3 or 4 cannot be carried out in the Bayesian approach.

Third, the procedure is *reflexive* in the sense that the termination of the data collection procedure depends fortuitously on when the periodic data inspection is carried out. Had the fourth inspection been delayed until Inspection 5, the evidence would be 50% instead of 55%. (Hence, the "prior probability" entries in Column 4 and the "likelihood of evidence" entries in Column 3 are duplicated in Column 5.) In such an event, the posterior probability for $H_C$ is only .70, which is not sufficient for Editor E to stop the polling. On the other hand, the Bayesian sequential sampling is also open-ended. Specifically, the poll does not stop after the third inspection because the posterior probability (viz., .59 in Table 14) is smaller than the one desired by Editor E (viz., .75). The Bayesian modus operandi is best summarized as follows:

> The scientist can design an experiment to enable him to collect data bearing on certain hypotheses which are in question, and as he gathers evidence he can stop from time to time to see if his current posterior opinions, determined by applying Bayes' theorem, are sufficiently extreme to justify stopping the experiment. (Phillips 1973, p. 66) (Quote 4)

A concomitant feature of the reflexivity and open-endedness found in the Bayesian methodology is that the size of the data set is ill defined. It is determined fortuitously by the data collection procedure.

Fourth, the decision to stop data collection is made on the basis of a criterion not related to what is said in the hypothesis. Note that $H_C$ is the hypothesis that the centre



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

party will form the next government. At the same time, Editor E's decision criterion is not the truth of $H_C$, but how certain Editor E is of $H_C$. That is, whether or not the centre party actually forms the next government has no bearing on the reason why the poll is conducted. It is also for this reason that Bayesians do not (and cannot) assess their data with reference to a well-defined criterion in a way independent of the prior probability. Consequently, Bayesians do not talk about *objectivity* because there is no objective entity or event against which Editor E's decision may be assessed.

In contrast to the third and fourth *reflexive* features of the Bayesian sequential sampling procedure, experimental psychologists do not treat their data in such a chance-dependent way. Instead, the size of the data set in experimental psychology is determined before data are collected. That is, what is said in Quote 4 is the exact opposite of what experimental psychologists would (or should) do. Specifically, experimenters adhere to their experimental plans, in which are stated, among other things, (a) the number of subjects, (b) the number of sessions a subject has to undergo, and (c) the number of trials per session. There is nothing chance dependent about the size of the data set. Moreover, the experimenter has to assess the consistency between the phenomenon and the substantive hypothesis, as well as that between the experimental prescription and the data. In other words, objectivity is not only possible, it is important in experimental psychology.

In short, the data collection procedure congenial to the Bayesian analysis is not appropriate for most psychological research. At the same time, the Bayesian hypothesis is not a prospective explanation of a phenomenon that exists before, as well as independently of, the data collection procedure. The reflexive dependence of the Bayesian hypothesis on the data collection procedure is responsible for the Bayesian disregard of objectivity. Consequently, the applicability of the Bayesian method to psychological research in general, and theory-corroboration experiments in particular, is questionable. The tenability of explanatory hypotheses cannot be ascertained by appealing to the researchers' subjective degrees of belief if the phenomenon to be explained exists prior to, as well as independently of, the data collection exercise.

## 37. Methodological criticisms in disguise

It has been mentioned that the role of the associated probability, *p*, in NHSTP is questioned by critics. Specifically, although *p* is not an index of the replicability of the result, the researcher tends to stop further investigation after a significant result (Bakan 1966). Critics find this wanting; they argue that researchers should conduct replication studies. This is important to Bayesians because their objective of empirical research is to revise the prior probability in light of new data. It is also necessary to conduct replication studies if one subscribes to the meta-analytic approach (Glass et al. 1981; Hunter & Schmidt 1990; Rosenthal 1984). Moreover, it has also been said that NHSTP results may be disambiguated with replication studies (Thompson 1996).

These arguments for conducting replication studies indicate that the real concern implicit in some criticisms of NHSTP has nothing to do with it being a statistical procedure. The criticisms are methodological critiques in disguise (inadvertently, perhaps). For this reason, another argument in support of NHSTP may be offered by showing that replication is not sufficient for theory corroboration. Worse still, successful replications may actually be misleading. Instead, the tenability of an explanatory substantive hypothesis is ascertained with a series of converging operations (Garner et al. 1956). NHSTP is used in every study in the series.

## 38. How important is replicability?

Several reasons may explain why replicability has captured critics' attention favorably. First, there are the ambiguity-anomaly criticisms of NHSTP, the collective point of which is that statistical significance may be reached fortuitously. There is the additional assumption that a fortuitous result is unlikely to be replicated. However, these criticisms can be answered, as may be recalled from "the sample size-significance dependence problem revisited" discussion in section 15. Be that as it may, some critics see another source of ambiguity.

They emphasize the arbitrariness of the choice of the $\alpha$ level. Specifically, setting $\alpha = .05$ is a convention. The question is raised as to why $\alpha$ is not set at the .10, .07, .01, or .005 level. More importantly, a result significant at the .05 level may not be significant at the .01 level (e.g., when the calculated *t* is 2.4 for a 1-tailed test with df = 18). By the same token, a result not significant at the .05 level may be significant at the .1 level (e.g., when the calculated *t* is 1.5 for a 1-tailed testwith df = 18). In short, statistical significance may simply be the fortuitous choice of the $\alpha$ level.

This "fortuitous choice of $\alpha$" criticism of NHSTP seems like a demand for an absolute proof for the substantive hypothesis. This demand cannot be met on logical grounds because it is impossible to prove any theory (i.e., substantiate any theoretical claim) with absolute certainty. It is not a limitation brought about by using NHSTP. It is the result of having to affirm the consequent of the major premises of the conditional syllogistic arguments implicated (see Table 3). It so happens that deductive logic does not allow drawing a definite conclusion about the antecedent of the major premise by affirming its consequence.

There is another reason why the "fortuitous choice of $\alpha$" criticism cannot be answered logically: the reality of multiple explanations for any phenomenon. To ask for absolute certainty under such circumstances is to require the elimination of all possible alternative explanations. Because this is impossible, the best one can do is to draw a tentative conclusion with the help of the inductive principle underlying the experimental design (see Table 5). The important point is that the criticism that the experimental result does not provide the conclusive evidential support should not be directed to NHSTP at all because it is not a difficulty brought into the research procedure because NHSTP is used.

At the same time, it is possible to show that the "fortuitous choice of $\alpha$" criticism is itself misguided. The criticism should be considered with reference to the fact that the $\alpha$ level is determined *before* data collection. That is, the decision is made *before* data collection that the level of strictness for rejecting $H_0$ stipulated by $\alpha$ is sufficient for the research in question. Consequently, subsequent decisions about the experimental task, experimental design, number of subjects, amount of training given to the sub-



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

jects, number of test sessions, number of trials per session, and the like are all made with the understanding that the chosen α level is sufficiently strict. Had a more stringent α level been deemed necessary, the other features of the experiment would have been different. Moreover, the "fortuitous choice of α" criticism is vacuous because critics can always stipulate a stricter criterion (viz., .001, .0005, etc.) after the completion of data collection.

Suppose that the statistical significance of the result is deemed not fortuitous and that the α level is accepted as adequate. Critics may still point out that replication studies are necessary because there is still the .05 probability of committing the Type I error. What does it mean to have rejected $H_0$ incorrectly? It means attributing the observed effect (in the technical sense of the word) to the experimental manipulation when, in fact, another variable may be used to explain the result. This is another way of saying that the experimental manipulation may have been confounded with an unknown variable. Seen in this light, the "fortuitous choice of α" critique is a design issue, not a difficulty inherit in NHSTP. It is important for the present discussion to note that this difficulty cannot be eliminated by conducting replication studies because it is a prerequisite of a replication study that data collection conditions must duplicate those of the original study. To the extent that this duplication is successful, the original confounding may still occur. In other words, absolute certainty about the substantive hypothesis cannot be established by successful replications. This state of affairs calls into question the necessity of conducting replication studies.

To recapitulate, critics' insistence on conducting replication studies seems to be motivated by (a) the wish to establish the tenability of the substantive hypothesis with absolute certainty and (b) the desire to disambiguate the NHSTP outcome. In defence of NHSTP, it is suggested that these concerns are methodological, not statistical. For example, some critics suggest that researchers should report the associated probabilities, *p*, of individual studies so that meta-analysis can be carried out.

It can be concluded that, although replicability may be the necessary condition, it is not the sufficient condition. A more positive defence of NHSTP consists of showing that (a) the "fortuitous choice of α" criticism cannot be used to justify the meta-analytic approach, (b) cognitive psychologists ascertain a substantive hypothesis by eliminating alternative substantive hypotheses and unknown confounding variables with converging operations, and (c) NHSTP is used in all such attempts. This positive defence may be presented in the context of studying the iconic store experimentally.

## 39. The "perceive more than can be recalled" phenomenon

Suppose that you take a quick glance at the rearview mirror while driving. You can report only a few things despite the feeling that you have seen more. This not uncommon experience is the phenomenon to be explained by the iconic store (Neisser 1967; Sperling 1960). The iconic store is said to have a relatively large storage capacity and a very short retention interval. Forgetting from the iconic store occurs because of information decay. Lastly, only sensory unprocessed information is available in the iconic store.

Suppose that there are 12 studies of the iconic store (see the "Study" column in Table 15). Although the result is significant in eight studies, it is nonsignificant in Studies 4, 5, 9, and 12 (see the "*p* of Test Statistic" column). Among the 8 studies with statistically significant results, the *p* values of Studies 6 and 8 are very close to .05, whereas that of Study 3 is exactly .05. Although the result is nonsignificant in Study 5, the *p* value does not differ by much from .05. This state of affairs may reinforce the "fortuitous choice of α" criticism.

## 40. Meta-analysis and its difficulties

Some critics suggest that the *p* values of the test statistic may be used to obtain a combined Z or that effect-size estimates may be used to obtain the combined effect size. A statement about the overall statistical significance is then made on the basis of the combined Z, called "combined significance level" (Harris & Rosenthal 1985). That is, the *p*, or effect-size, values from individual studies are treated as raw data (hence "raw data" in Table 15's title) and subjected to statistical analysis at a higher level of abstraction. This more abstract analysis is called meta-analysis or the "analysis of analysis" (Glass 1976; 1978; Glass & Kliegl 1983; Glass et al. 1981; Harris & Rosenthal 1985; Schmidt 1992).

Recently, meta-analysis has been promoted as a theory-corroboration tool, in addition to being an antidote for rectifying the harm done by using NHSTP (Cooper 1979; Cooper & Rosenthal 1980; Schmidt 1996). Specifically, it is an important meta-analytic assumption that knowledge grows with the accumulation of research results. The binary nature of the NHSTP outcome is deemed incompatible with the incremental growth of knowledge. This difficulty is amplified by the "fortuitous choice of α"criticism.

Some meta-theoretical issues must be settled before meta-analysis can be accepted as a valid theory-corroboration tool. They are: (a) the selection, (b) a lack of independence, (c) the unjustifiable disregard for research quality, and (d) problems brought about by a lack of commensurability among the to-be-aggregated studies (see Chow 1987b; 1987c; Cook & Leviton 1980; Eysenck 1978; Gallo 1978; Leviton & Cook 1981; Mintz 1983; Rachman & Wilson 1980; Sohn 1980; and Wilson & Rachman 1983). Of interest here is the problem of a lack of commensurability among the studies included in the meta-analysis.

To use the combined Z or effect size of the 12 studies in Table 15 to ascertain the tenability of the iconic store is to assume that it is legitimate to combine the data from the 12 experiments and to use them *in toto*. This assumption must be questioned in view of the fact that different independent variables are used (see the "Independent variable" column in Table 15). At the same time, different dependent variables are implicated. For example, the dependent variables may be the number of items available in one study; in another study they could be the correct reaction times, or the number or types of error made. How is it meaningful to take the average of the effect measured in terms of correct reaction times and that measured in terms of the frequencies of different kinds of errors? What would the average mean at the conceptual level? In other words, to combine the information from qualitatively different experiments is



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Table 15. *The incommensurability difficulty of meta-analysis illustrated with fictitious "raw data"*

| Study | *p* of test statistic | Effect size[1] | Independent variable | Property or function of the iconic store studied |
|---|---|---|---|---|
| 1 | .021* | 0.7 | ISI[2] | Rate of decay |
| 2 | .001* | 0.3 | Type of task | Relatively large storage capacity |
| 3 | .050* | 0.5 | Number of concurrent tasks | Independence from the short-term store |
| 4 | .110 | 0.11 | What to recall | Independence of location and identity information |
| 5 | .068 | 0.17 | Stimulus material | Nonassociative information |
| 6 | .049* | 0.4 | Type of task | Visible persistence |
| 7 | .02* | 1.5 | Type of material | Unprocessed information |
| 8 | .046* | 1 | Time of probe presentation | Select before processing |
| 9 | .070 | 0.06 | ISI | Visible persistence |
| 10 | .04* | 0.18 | Stimulus duration | Information registration rate |
| 11 | .038* | 0.2 | Stimulus duration within a fixed SOA[3] | Information registration rate |
| 12 | .066 | 0.29 | Type of material | No identity information |
| | Combined Z | Combined effect size | | |

*denoted significance at the 0.05 level

**1.** J. Cohen's (1987) $d = \frac{(u_E - u_c)}{\sigma_E}$

**2.** "ISI" refers to the interstimulus interval, the interval between the offset of the stimulus and the onset of the partial-report tone (see N. 2 in Ch. 5)

**3.** "SOA" refers to stimulus-onset asynchrony, the interval between the onset of the stimulus and the onset of the mask.

like the illegal practice of mixing apples and oranges (Cook & Leviton 1980; Mintz 1983; Presby 1978).

Glass et al. (1981) answer the "apples and oranges" reservation by saying that apples and oranges are both fruits. They argue that things that are incommensurable at one level of discourse become commensurable if they are subsumed in a higher-order category. However, this is not a good answer. Specifically, the outcome of Study A may be caused by the acidity of oranges, which is not a property common to all fruits. At the same time, the texture of apples may be the reason for the outcome of Study B, and the texture is also not a property common to all fruits. That is to say, meta-analysts have not provided a good theoretical justification for ignoring the qualitative differences among the diverse sets of research data.

## 41. Converging operations

It is an meta-analytic assumption that our understanding of an issue improves when more data are accumulated and used *in toto*. Moreover, merely knowing the significance or nonsignificance of individual studies is not suitable for such a data-accumulation exercise. However, the "apples and oranges" difficulty shows that the meta-analytic argument is debatable because "accumulate" is used in a quantitative and mechanical sense. The disregard for the qualitative differences among various studies prevents meta-analysts from seeing that knowledge evolves in a far from straightforward fashion. There are a lot of trials and errors at the conceptual level. NHSTP plays an important role in every one of these steps. This view (an alternative to the meta-analytic one) may also be used to illustrate (a) the rationale of conducting converging operations, (b) how the difficulty of perpetuating confounding variables in replication studies may be minimized, and (c) the difficulty caused by the Bayesian insistence on interpreting current data with reference to prior probability.

## 42. The rationale of converging operations

The relations among the quartet of hypotheses implicated in the theory-corroboration experiment depicted in Table 1 become more complicated as the investigation of the substantive hypothesis progresses. Successive rows of Table 16 represent successive stages of the theory-corroboration endeavor. The series of studies need not (and often is not) undertaken by the same experimenter.

The only requirement for the tenability of the iconic store before any data collection is that what is attributed to the iconic store (i.e., what is said in H) be consistent with P, the "perceive more than can be recalled" phenomenon (see the "Before experimentation" row in Table 16). Phenomenon P is the prior data in the sense that it exists before the substantive hypothesis. There is no evidential data for the iconic before the first experiment because Phenomenon P itself cannot be used as the evidence (hence, there is no entry in the cell defined by the intersection of the "Data" column and the "Before Experimentation" row). To use the original phenomenon for such a purpose is to commit the circularity error.

In the absence of any evidential data, the implication of the first study (viz., $I_1$) has to be consistent with P, in



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Table 16. *The phenomenon-hypothesis-implication-data (P-H-I-D) consistency at different research stages of the substantive hypothesis, H*

| Experiment | Phenomenon/prior data | Hypothesis | Implication | Data | P-H-I-D consistency |
|---|---|---|---|---|---|
| Before Experimentation | P | H | | | Yes |
| 1 | P | H | $I_1$ | $D_1$ | Yes |
| 2 | P + $D_1$ | H | $I_2$ | $D_2$ | Yes |
| 3 | P + $D_1$ + $D_2$ | H | $1_3$ | $D_3$ | Yes |
| . . . | | H | . . . | . . . | Yes |
| . . . | | H | . . . | . . . | Yes |
| n − 1 | P +$D_1$ . . . + $D_{n-2}$ | H | $I_{n-1}$ | $D_{n-1}$ | Yes |
| n | P + $D_1$ + . . . $D_{n-1}$ | H | $I_n$ | $D_n$ | Yes |
| . . . | | H | . . . | . . . | Yes |
| t | P + $D_1$ + . . . $D_{t-1}$ | H | $I_t$ | $D_t$ | Yes |

addition to being a theoretical derivation from H. The data from this study (i.e., $D_1$) are compared to what is said in $I_1$. Necessary for the tenability of H is that $D_1$ be consistent with $I_1$. Hence, the emphasis is on the phenomenon-hypothesis-implication-data (P-H-I-D) consistency in Table 16. This is the basis of objectivity, something ignored in the Bayesian and meta-analytic approaches.

The data-implication consistency (i.e., the consistency between $D_i$ and $I_i$, where i represents the row number in Table 16) is ascertained with NHSTP, as described in Tables 2 and 3 (or 4, as the case may be). Only a binary decision is required to initiate the chain of deductive reasoning depicted in Table 3 or 4. Hence, the binary nature of the NHSTP decision is adequate for this purpose. Why do critics consider the binary decision incompatible with the incremental growth of knowledge? One reason is that, as has been suggested earlier, critics identify NHSTP with the theory-corroboration process itself. The other reason is that there are two ways to look at the information accumulated, as well as how it is used.

Meta-analysts accumulate data via the test statistics of individual studies (see the "*p* of Test statistic" or "Effect size" column in Table 15). Bayesians accumulate raw data via the role played by the prior probability in the Bayesian theorem (see Table 14). In both cases, all data accumulated to date are used as the evidence. In contrast, only data collected in the current study are used to ascertain the tenability of the substantive hypothesis (e.g., as may be seen from Table 16, only $D_n$ is used to assess the tenability of $I_n$ in Study n; see also Tables 2, 3, and 4). More importantly, research data obtained in previous studies have no evidential role in the current study. Instead, conclusions drawn from earlier data are used in cognitive psychology as theoretical constraints on the "If H, then Implication" derivation. For example, Implication $I_n$ has to be consistent with P + $D_1$ + . . . $D_{n-1}$ (see Row [n - 1] of Table 16).

Each of the implications in Table 16 is a criterion of rejection for the substantive hypothesis, H, in the following sense. H has to be rejected if the data (e.g., $D_n$) are inconsistent with what is stipulated in any study (viz., $I_n$ in the case of Study *n*). In more concrete terms, studies of the iconic store are attempts to substantiate the theoretical properties that have been attributed to the iconic store (viz., those tabulated in the "Property or function of the iconic store studied" column of Table 15). For example, a fast rate of decay has been attributed to the information residing in the iconic store. An implication of this theoretical property is that performance on Sperling's (1960) partial-report task should decline rapidly within 250 to 500 msec when the interstimulus interval (ISI) is manipulated. The postulation of the iconic store would be untenable if this theoretical prescription were not demonstrated.

It follows that the "Independent variable" and "Property or function of the iconic store studied" columns in Table 15 jointly represent attempts to falsify the theory of the iconic store from various angles (viz., by testing the various theoretical properties of the iconic store with different experimental tasks in diverse settings by different experimenters). Suppose all these falsification attempts fail. That means that, as the research progresses, more and more of its theoretical properties are substantiated in qualitatively different situations. Hence, these various studies may be said to converge on the tenability of the iconic store. Researchers' confidence in the iconic store as a theoretical mechanism increases as more and more of these falsification attempts fail. In short, the series of experiments collectively form the converging operations used in validating the iconic store.

The studies depicted in Table 15 are not replication studies because they differ greatly among themselves. Suppose that critics question the statistically significant result of Study 8. As has been suggested earlier, this reservation is mostly a question about data interpretation. One source of ambiguity is the presence of an unknown variable that has varied systematically with the research manipulation. That is, the data may do something other than ascertaining the property of the iconic store. This ambiguity cannot be eliminated by replicating the experiment because every time the original study is repeated, the confounding variable would also occur.

The situation is very different in the case of converging operations. The theoretical property investigated in Study 4 has a different implication in another setting (hence, Study 8). Although Studies 4 and 8 are about the same theoretical property of the iconic store, radically different tasks and experimental manipulations are involved. Hence, in general terms, it is less likely that the same confounding variable is found in all of the studies when radically differ-



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

ent data-collection conditions are used in the series of converging operations. This is the reason why conducting converging operations is more satisfactory than conducting replication studies.

## 43. Objectivity and the Bayesian prior probability

Bayesians bemoan the fact that non-Bayesian researchers do not take into account the prior probability of the hypothesis when research conclusions are drawn. This Bayesian point may be illustrated with column 1 of Table 14. The evidence is that 30% of the voters polled indicate a preference for the centre party (i.e., $H_C$). The likelihoods are .38, .35, and .32 for $H_L$, $H_C$, and $H_R$, respectively. That is, one may argue that the evidence is more favorable to $H_L$ than to $H_C$ if one ignores the impact of the prior probabilities of the three hypotheses. Bayesians would point to the fact that the posterior probabilities are .36, .40, and .32, respectively for $H_L$, $H_C$, and $H_R$. In Phillips's (1973) words:

> Whether or not the [poll result] tells the [editor] something about the [election], but the information conveyed by the [poll] is far less than that shown in the prior probabilities. In this case the prior probabilities swamp out the information in the [poll], so that the posterior probabilities are determined more by the priors than by the likelihoods. The extra information given by the [poll] does not change the prior probabilities enough to warrant [changing the editor's editorial decision]. (Phillips 1973, p.74, explications in parentheses added) (Quote 5)

The Bayesian treatment of the prior probability of the hypothesis is the opposite of what experimental psychologists should accept. First, what does a higher posterior probability mean? It means simply that Editor E has the highest confidence in $H_C$. However, this is not the same as saying that $H_C$ is necessarily or inevitably true. Second, Bayesians find "probabilification" satisfactory because they treat their reflexive sequential sampling task as the prototype of all empirical research. Consequently, they do not find it necessary to assess the data with reference to the consistency between (a) the to-be-explained phenomenon and the substantive hypothesis and (b) the experimental prescription and the data.

There is a third reason why what is said in Quote 5 is debatable. Do Bayesians mean to suggest that research should be designed and data interpreted in ways determined by how the researcher feels about the hypothesis? Is this not an invitation to inject biases into the research processes? This is the opposite of what should be the case in view of the rationale described in Tables 1, 2, 3, and 5.

Experimenters following the research rationale represented in Tables 1 and 16 always give the to-be-corroborated hypothesis the benefit of the doubt. That is, the derivation of the experimental hypothesis (and hence, the design and execution of the experiment) is based on the assumption that the substantive hypothesis is true. This assumption is made even when the experimenter does not like the hypothesis. Moreover, the statistical, experimental, and theoretical conclusions are drawn with reference to the rules described in Tables 2 and 3 (or 4). How the researcher feels about the hypothesis or data has no role in the statistical, inductive, or deductive reasoning. It is necessary to consider the Bayesian methodology more judiciously.

## 44. Summary and conclusions

Not much is said in this defence about the criticism that various aspects of NHSTP are often misunderstood (e.g., the meaning of *p*, identifying $\alpha$ as an index of replicability, etc.) for the simple reason that they are not difficulties inherent in NHSTP. Instead, the emphases in this defence are on some meta-theoretical assertions about NHSTP. The most notable one is the commonly accepted view that the null hypothesis is never true. This view is problematic because the null hypothesis is not used in NHSTP as a categorical proposition descriptive of the world.

The null hypothesis appears in two different conditional propositions. First, it is the implication of the hypothesis that chance influences are responsible for the experimental result. What is said in the null hypothesis is (and should be) true if the control and experimental conditions are set up properly. Second, the null hypothesis is used to stipulate the lone sampling distribution to be used in making the statistical decision about chance influences. This utility of the null hypothesis shows that NHSTP is often misrepresented at the graphical level. More important, it shows that *statistical significance* and *effect size* (or *statistical power*) belong to different levels of abstraction (viz., the level of sampling distribution versus the level of raw scores).

The present defence of NHSTP is conducted in the context of theory corroboration. It is argued that NHSTP is not theory corroboration. However, NHSTP does provide the objective means to exclude chance influences as an explanation of research data. This statistical decision provides the minor premise for the first of three embedding syllogisms. The asymmetry between the *modus tollens* and *affirming the consequent* arguments can be tentatively resolved by appealing to an inductive rule underlying the experimental design. These considerations lead to the conclusion that many criticisms of NHSTP are actually questions about the inductive conclusion validity of the research.

The putative importance of the effect size is called into question by showing that the size of the effect is not an index of the evidential support for the substantive hypothesis offered by the data. Nor can the effect size, by itself, be the index of the practical importance of the research result in the case of the utilitarian experiment. It is made clear that statistics and practical validity belong to different domains.

Some difficulties with power analysis are illustrated by the affinity between NHSTP and TSD envisaged in power analysis. A notable example is that, being a conditional probability, statistical power cannot be the probability of obtaining statistical significance. Because there is a Bayesian overtone in power analysis, such analysis can be questioned to the extent that the Bayesian assumptions about research methodology are debatable. At best, the Bayesian approach has a very limited applicability in psychological research because it is applicable only to the sequential sampling procedure. It cannot be used to investigate explanatory hypotheses.

In short, what motivates some criticisms of NHSTP may be understood by the fact that although statistical significance provides a rational basis for rejecting chance influences as an explanation of data, it is not informative as to what the nonchance factor is. Moreover, statistical significance says nothing about the real-life importance of the data. It is argued in this defence that the real issue concerns why NHSTP is expected to furnish such information, given that NHSTP is a statistical procedure. What also has to be said is that the alternative numerical indices suggested by critics of NHSTP (viz., effect size and confidence interval) are also incapable of pinpointing the nonchance factor



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

responsible for the research result. The point is simply that statistics and practical importance belong to two different domains. Why should the tools from one domain be used to settle questions belonging to another domain?

Be that as it may, there is a positive way to look at the criticisms of NHSTP. The critiques are attempts to challenge NHSTP users to rationalize the research procedure in general, and the role of NHSTP in such a procedure in particular. The present defence of NHSTP is one such attempt. This account of NHSTP is not expected to be the final one. Nonetheless, it will fulfill its purpose if it serves as the basis for further exploration of the issues raised in the course of the present argument. It is hoped that a coherent view of NHSTP pertinent to empirical research will emerge from the ensuing discussion.

# Open Peer Commentary

*Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.*

## The null-hypothesis significance-test procedure: Can't live with it, can't live without it

Charles F. Blaich

*Department of Psychology, Wabash College, Crawfordsville, IN 47933*
**blaichc@wabash.edu**

**Abstract:** If the NHSTP procedure is essential for controlling for chance, why is there little, if any, discussion of the nature of chance by Chow and other advocates of the procedure. Also, many criticisms that Chow takes to be aimed against the NHSTP (null-hypothesis significance-test) procedure are actually directed against the kind of theory that is tested by the procedure.

Social scientists, especially psychologists, seem to be in the grip of a split personality, at least statistically speaking. On the one hand, there appears to be an uprising against statistical inference (e.g., Bakan 1966; Carver 1978; Cohen 1994; Hunter 1997; Loftus 1996; Morrison & Henkel 1970; Rozeboom 1960). On the other, despite token changes in the content of most texts, the null-hypothesis significance-test procedure (NHSTP) has been taught more or less the same way for the last 40 years.

Chow's book offers one of the few recent defenses of this procedure. Chow advances two basic arguments. First, that the NHSTP's limited but important function is to control for the possibility that chance factors alone could plausibly account for our data. Second, that many of the faults that have been attributed to the NHSTP, for example the fact that it does not provide information about effect size or the practical importance of a finding, are misdirected, because the procedure is not designed to accomplish the tasks the critics want done.

Chow's first argument repeats what a number of critics of the NHSTP have already stated: that too often researchers interpret the phrase "statistically significant" to imply that a study is of theoretical or practical "significance" (Guttman 1977; Shaver 1993; Thompson 1996). One reason, in my view, that many researchers have a hard time sticking to the strict meaning of the NHSTP is, despite the effort they expend learning different permutations of the NHSTP, they do not view chance as a plausible alternative explanation. Indeed, social scientists appear to have a curious love-hate relationship with chance: they love to reject the idea that chance explains their data (i.e., get a statistically significant result), but they are reluctant to discuss explicitly the chance processes they are so bent on rejecting. Apart from making vague references to sampling error, few textbooks describe the chance events modeled by the NHSTP in different research designs (See Ramsey & Schafer 1997 for a recent exception). Nor is there any discussion of whether it is appropriate to use the same inferential procedures (a) to analyze designs in which we randomly sample individuals from existing populations, (b) to randomly assign subjects to different conditions, or (c) when there is neither random assignment nor random sampling. Chow's book continues this tradition with virtually no discussion of chance processes.

One sometimes wonders how many researchers actually make the connection between the NHSTP and the real events that occur in their research (Tversky & Kahneman 1971). Too often the same researchers who apply the most stringent post hoc tests to control experiment-wise Type I error, turn around and repeatedly run replications of the same study "until it works" – that is, until the results are statistically significant. This is not a criticism of the NHSTP so much as it is a question of whether or not social scientists are philosophically committed to the idea that chance events affect our data. If they are not, then using the NHSTP is more of a ritual than an active effort to control for chance (Shaver 1993).

Chow also argues that the inability of the NHSTP to provide direct estimates of effect size is not important for theoretical research because the "exact magnitude of the effect plays no role in the rationale of theory corroboration" (Chow 1996, p. 96). Chow does allow that estimating effect size and the practical utility of a finding are useful for utilitarian research, but it is not clear that effect size and statistical significance can be this neatly separated. Effect size plays an important role in determining whether or not a statistical test is significant, and therefore whether or not a theory is corroborated. Nonetheless, there is a rationale to Chow's argument. According to Chow, theories lead to qualitative, "more, less, or not the same" hypotheses about the pattern of data. For example, the theoretical assertion that the linguistic competence of native speakers of English is an analog of the transformational grammar leads to the research hypothesis that it should be more difficult to process one kind of sentence than another (Chow 1996, p. 69). This in turn leads to a specific prediction about which sentences should take longer to process. Of course, the theory does not lead us to a more specific prediction of how much more time one sentence should take to process than another.

Chow's view of the qualitative prediction that a theory should produce (i.e., this group should be more or less than that group, or these groups should not be the same) is certainly consistent with the way that theory corroboration has been practiced in the social sciences. It is also clear that the NHSTP fits nicely with qualitative hypotheses. We want to predict that the mean of one group should be different from another, and we want to exclude the possibility that the "difference" is the effect of chance.

It might be useful, however, to view recent arguments in favor of emphasizing the use of confidence intervals, effect sizes, power analysis, and meta analyses not as critiques of the NHSTP but as tentative steps towards establishing a tradition for both theoretical and utilitarian research, in which we specifically predict the magnitude of the relationships among our variables. I don't want to fall prey to "physics envy," but the assertion "F = ma" seems both more satisfying and more risky than the prediction that there is a positive relationship between force, mass, and acceleration. Even if point predictions are out of reach, theories which generate predictions of "a little more" or "a lot less" would enrich our field (See Tukey 1969, p. 86 for discussion on this point).

Chow's book shows that the controversy surrounding the NHSTP will continue. Unfortunately, with a few exceptions, the



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

important points raised by critics of this procedure have not changed the way that editors and reviewers evaluate research. Although the NHSTP is, as Chow argues, just another control procedure, like random assignment or double-blind data collection, it differs in one critical respect. We cannot apply it until after data are collected. In a world in which only significant results are published, this makes researchers into gamblers, whose careers depend on the outcome of the chance events they are attempting to control for. It is not surprising, therefore, that the NHSTP continues to be practiced in a way that most editors and reviewers demand.

ACKNOWLEDGMENTS

I would like to thank Frank Howland and Louisa Blaich for reading previous drafts of this manuscript.

## On the position of statistical significance in the epistemology of experimental science

Charles E. Boklage

*Department of Pediatrics, East Carolina University, Laboratory of Behavioral and Developmental Genetics, Greenville, NC 27858-4354*
**boklage@brody.med.ecu.edu**

**Abstract:** Although various statistical measures may have other valid uses, the single purpose served by statistical significance testing in the epistemology of experimental science is as a peremptory rebuttal of one potential alternative interpretation of the data.

When I used bacterial viruses instead of humans to study developmental genetics, we routinely sampled five hundred million lives in a pipette. Most experimental answers came from differences among containers in the numbers of those lives. A two-fold difference required a retest against having added a limiting factor twice. A ten-fold difference raised prospects of miscounted titration dilutions. We repeated each experiment with appropriate variations to satisfy ourselves that we knew how the system came to behave as it did. When we were confident we could reproduce the answer, we did it again, for pretty results for publication. We knew every tool we used was imperfect: I made a personal experimental study of how best to handle pipettes in making dilutions, to appreciate that source of error intimately and minimize it. We could make the risk of refutation approach negligible, and our worst-case outcome was the prospect of someone else producing a reasonable alternative explanation for our results. One time, a reviewer asked for statistical confirmation of the difference between the lines in a graph showing how different viruses were affected by a particular manipulation. We made the calculations necessary to answer the silly question: the prospect that the lines differed only by chance in sampling was orders of magnitude below any reasonable preset criterion of negligibility we could have imagined.

In questions of human development, experimental control is of a much different character, and samples are orders of magnitude smaller. A reasonable alternative explanation is still my worst case, and the interpretation that the results may be due to chance alone can indeed be a reasonable alternative. In the biomedical literature, many editorial reviewers expect statistical tests in most submissions. Still, I have acted as if I thought everyone knows significance testing serves only to protect research results from that one alternative interpretation. The rest of what we really need from a good piece of research will always have to come from the experimental logic and the underlying chain of evidence.

These truths are not, I'm afraid, self-evident. There was a time when I did not yet know them, and a time before that when the people who taught me did not yet know them. Chow's answers are cogent and thorough, and his book should be a help in teaching statistical epistemology. It is not as easy as it might be to notice what he has to say about the valid practical uses of the alternative understandings he addresses. Perhaps that should be another book.

## Statistical significance testing was not meant for weak corroborations of weaker theories

Fred L. Bookstein

*Institute of Gerontology, University of Michigan, Ann Arbor, MI 48109.*
**fred@brainmap.med.umich.edu**

**Abstract:** Chow sets his version of statistical significance testing in an impoverished context of "theory corroboration" that explicitly excludes well-posed theories admitting of strong support by precise empirical evidence. He demonstrates no scientific usefulness for the problematic procedure he recommends instead. The important role played by significance testing in today's behavioral and brain sciences is wholly inconsistent with the rhetoric he would enforce.

Chow's book sets out a rhetoric of null-hypothesis significance-test procedures (NHSTP) in the context of "theory corroboration" as a style of behavioral research. The basic contention is set out clearly on page 67: "The limited, but important, role of NHSTP in empirical research is to supply the minor premiss required to start the chain of embedding conditional syllogisms implicated in theory corroboration."

Yet this volume never gives examples of the empirical corroboration of any interesting or plausible theory. In fact, it contains no empirical examples at all. Table by table, its "data" are pure simulations or formularies (Tables 2.5, 2.6, 7.2) or frankly fictitious (Tables 1.1, 5.1, 5.4, 5.5, 7.1, and 7.3). Had Chow access to examples demonstrating the wisdom of his suggested rationale, such as opposing theories (one of which ultimately came to dominate based on the accumulation of appropriately couched significance tests), surely he would have put them forward. From the fact of their complete absence from the volume, one may infer that there exists no such example: no empirical scientific practice to which the author's version of NHSTP is capable of making a meaningful contribution.

This predicament is a direct consequence of Chow's impoverished notion of "theory corroboration," the activity that he insists is the only proper context of significance-testing. For instance, this fundamental misconstruction pervades the extensive verbal chart (Table 5.2) contrasting "the agricultural (utilitarian) model" with these "theory-corroboration experiments." Entry after entry in the theory-corroboration column reiterates the same unfortunate reduction of empirical inquiry to some sort of rarefied logical exercise. The "impetus" is "to explain a phenomenon[,] independent of data collection"; the subject matter is an "unobservable hypothetical entity and its theoretical properties"; there is "no closure to the investigation"; the question is "why" rather than "how [much]"; the experimental hypothesis is "qualitatively different from the . . . substantive hypothesis"; the experimental manipulation and the dependent measure are each "different from the to-be-explained phenomenon"; the effect is "the difference between the means of two conditions." Theories of which the corroboration is hobbled by these rules could never be persuasively argued by empirical evidence at all – Chow has designed a logic of significance-testing perfectly specialized for theories that cannot ever be tested in stronger ways.

But why would anyone pursue such "corroborations" when so many superior forms of investigation come readily to hand? Most of today's most interesting theories of cognitive neuroscience can be stated clearly enough to justify strenuous attempts at compelling empirical support or refutation. Aspect by aspect, they contravene all the characterizations from Chow's table just quoted. Every attempt is made to tie explanation and hypothesis to data that are more and more explicitly measured and more and more effectively visualized, so that ultimately the theory and the reliable production of a strongly predictable empirical pattern amount to the same thing.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Let me illustrate using the theme of brain–behavior relations, my principal research interest at present. The "impetus" of work like mine is to discern and explain patterns in the empirical association between channels of data recorded using different instruments (e.g., an MR scanner and a questionnaire) on the same subjects, perhaps a sample of schizophrenics. The theory is that of "endophrenology": there exist structural features of the brain, perhaps including some malformations of the corpus callosum, that are associated tightly enough with psychiatric symptoms to underwrite etiological hypotheses. There are no "hypothetical entities" with "theoretical properties," but instead reliable measurements in each of two practical domains, neuroanatomy and psychiatric diagnosis. The issue is not "why" but *what* predictable associations there are between the two domains, in this case etiologic associations that hint equally at potential developmental understandings and treatments. There was no "experiment," if by that one means an intentional intervention (e.g., to induce a case of schizophrenia); in its place there is the excruciatingly careful accumulation of data and its defense against all the familiar threats to their validity – the validity of the data, mind you, not the theory. The "effect" of interest in a typical study is not the sign of a difference between two means but the demonstration of a distributed pattern, such as a dose–response relationship, a distinction of shape, or a system of peaks of blood flow localized in space and time. It is mainly from the values of regression coefficients like these that the modern endophrenologist extracts his most exciting aperçus into neurophysiology and neuropsychology. (See, for instance, Bookstein 1997 or Bookstein et al. 1996.)

From the unsuitability of this rhetorical structure to the conditions set out by Chow for the validity of NHSTP, one might infer that statistical-significance tests play no role in this domain; but of course they do. As in every other aspect of modern cognitive neuroscience, the exploitation of NHSTP is ubiquitous. Its role, however, is to *accept* a properly drawn null hypothesis, not to rule it out. For in this domain, to affirm a theory is to say something quantitative. One significance test might, for instance, assert the low rank of a cross-covariance matrix, as in the method of Partial Least Squares, and so sustain a claim of empirical simplicity for some pattern of bivariate associations. Another might be concerned with agreement between the spatial variation of the contrast between two averaged functional brain images and the distribution expected on a Markov random field hypothesis, so that one might be permitted to refer to that spatial field as "noise" rather than further regional signal requiring further investigation. A third and fourth test might sustain the claim that there is a shape difference between groups in one part of the brain, but nowhere else. This theory-laden area relies on NHSTP for rational scientific discourse in all these ways, none of which seem to find any resonance in either column ("agricultural model" and "theory corroboration") of Chow's Table 5.2.

Thus Chow's argument, to the extent that it can be taken seriously at all, quickly brings us to a *reductio.* Either brain–behavior studies are not empirical research, or they cannot use NHSTP properly, or they do not involve theories. All these possibilities are absurd, but Chow's approach leaves us no others. Better, of course, to conclude that his book is a description of obsolete academic exercises, not any sort of practical reason – that there is just no interesting scientific activity to which this characterization of NHSTP applies. Chow manages to "justify" the old-fashioned psychologist's use of NHSTP only by constraining its domain of application until its role becomes exiguous, perhaps even imaginary. He certainly shows no empirical contexts in which his version of NHSTP is of any relevance to modern behavioral or brain science. To the extent that those sciences make appropriate use of NHSTP in future work, those uses have no rhetorical relation to the logic of inference put forward here. We will need to turn to the discussions of others to learn a proper rhetoric of significance-testing in the behavioral and brain sciences.

## Null hypothesis tests and theory corroboration: Defending NHSTP out of context

Reuven Dar

*Department of Psychology, Tel Aviv University, Tel Aviv 69978, Israel.*
**ruvidar@freud.tau.ac.il**

**Abstract:** Chow's defense of NHSTP 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.

Systematic desensitization is a reliable procedure for treating specific phobias; but it is inappropriate, even dangerous, as a treatment for paranoid psychosis. Every formal procedure must be evaluated in the context of its actual use. NHSTP may be adquate for the modest task of ruling out chance as the explanation of the data. But in psychology, like it or not, NHSTP is the principal tool for testing substantive hypotheses in theory-corroborating studies. And in that capacity it is not only inadequate, but may be destructive to psychology as a scientific discipline.

Chow's recurrent tactic in contesting criticisms of NHSTP in terms of its actual use in psychological research is to reject them as representing nonstatistical issues. The most critical faults in NHSTP, however, are exactly those that "go beyond statistics" to the actual role of NHSTP in psychology. Chow labors to make a distinction between the statistical and substantive hypotheses; but in actuality, as any random study in psychology research journals will confirm, the move from the substantive to the statistical hypothesis is a swift one: researchers interpret statistical significance as confirming the substantive hypothesis and therefore as corroborating the theory from which the hypotheisis was derived. And what choice do they have? NHSTP is the only formal procedure currently available to the psychology researcher for deciding whether the substantive hypothesis was confirmed. In fact, without this direct link between NHSTP and the substantive hypothesis, NHSTP becomes an isolated actuarial procedure which has no role in theory development.

In stating that "the evidential support for the theory is secured first by excluding chance influences" (p. 88), Chow seems to presuppose that NHSTP is an essential component of sound scientific practice. But as opponents of this ritual have noted, some respected sciences, such as physics, seem to get away without this essential step. In fact, as Meehl (1978) observed, so did the finest research programs in psychology. When Skinner's pigeons learned to play ping-pong by operant conditioning, no one insisted that NHSTP be conducted to compare their playing ability to that of a wait-list control group.

Whereas NHSTP is not a universally endorsed criterion for good science, most scientists and philosophers of science would agree on this criterion: theories should make bold predictions and be tested severely (Chow, in fact, explicitly subscribes to this Popperian attitude). But in psychology, as argued above, NHSTP is the only test to which a theory is put in any experiment. And, as several opponents of NHSTP have noted, passing this test constitutes an extremely weak corroboration of the theory from which the experimental predictions were derived.

On this background, Chow's insistence that "the exact magnitude of the effect plays no role in the rationale of theory corroboration" (p. 96) is truly puzzling. Doesn't this assertion mean that a theory that successfully makes accurate predictions is in no better shape than a theory that predicts only that the magnitude of an effect will exceed zero? Indeed, this position makes sense only if one accepts Chow's premise that NHSTP is not a test of the theory – a premise which, as argued above, is patently false in the reality of research in psychology. This premise is presumably also Chow's rationale for discarding as "unnecessary" (p. 57) Serlin and



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Lapsley's (1985) Good-enough Principle, which represents efforts to make NHSTP more demanding and thus to increase its significance (in the real sense of the word) as a test of substantive hypotheses.

From this perspective, which ignores the actual role of NHSTP in psychology, it is hardly surprising that Chow disregards the possibility that when substantive hypotheses in theory-corroborating experiments are tested with NHSTP, this necessarily affects the way predictions are formulated. When researchers can only conclude, following a rejection of the null hypothesis, that a mean difference was larger than zero, their predictions (the substantive hypotheses!) would naturally be that this parameter will exceed zero. That is why, in Chow's own example, the researcher predicts that it would be "more difficult" to remember extra words after negative than after kernel sentences, rather than making a more accurate (and risky) prediction regarding the size of this difference.

I have argued (Dar 1987) that this state of affairs is detrimental to the development of theories in psychology and may be intimately related to Meehl's (1978) alleged slow progress in soft psychology. The ritual of NHSTP, as shown above, leads to trivial predictions. Trivial predictions are in turn likely to lead to weak and simplified theories: it takes much less thinking to formulate a theory that leads to predictions of nonzero differences than a theory that leads to more specific predictions. Consequently, as Meehl (1978) observed, theories in the area of soft psychology are typically short-lived: a theory cannot be compelling if all it is able (or willing) to predict is a nonzero difference between groups. Furthermore, when the only prediction in a theory-corroborating study is that the group difference will exceed 0, either success or failure in achieving this result can be easily accounted for by ad hoc explanations, which end up undermining the theory.

Lakatos, who elaborated Popper's view of science (which Chow claims to endorse), had this to say about the role of NHSTP in psychology: "after reading Meehl (1967) and Lykken (1968) one wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phony corroborations and thereby a semblance of 'scientific progress' where, in fact, there is nothing but an increase in pseudo-intellectual garbage" (Lakatos 1978, p. 88). This is a bit sharp, I admit; but if we wish to dull the edge of this criticism, we must find a better way to test our theories. Chow's defense of NHSTP does not lead in this direction.

## The logic of null hypothesis testing

Edward Erwin

*Department of Philosophy, University of Miami, Coral Gables, Fl 33124.*
**eerwin@umiami.ir.miami.edu**

**Abstract:** In this commentary, I agree with Chow's treatment of null hypothesis significance testing as a noninferential procedure. However, I dispute his reconstruction of the logic of theory corroboration. I also challenge recent criticisms of NHSTP based on power analysis and meta-analysis.

Chow's incisive (1996) book is timely, given that the American Psychological Association's Board of Scientific Affairs has recently been considering banning the reporting of significance tests in all APA journals (Shrout 1997). In this commentary, I will not address the practical question of the likely effects of such a ban, but will instead focus on an epistemological question: Is the null-hypothesis significance-test procedure (NHSTP) deeply flawed?

Is it? That depends on the procedure we are discussing. Different writers give different accounts of the form of the null hypothesis, and different definitions of such crucial concepts as "alpha," "beta," and "Type I and Type II errors." The result is that there is a family of related but different procedures all given the name "NHSTP."

In one very common version, NHSTP is an *inferential* procedure for indirectly testing a research hypothesis by assuming the truth of the null hypothesis and calculating the probability of getting the experimental results if chance alone were operating. Stated this abstractly, I suspect that many will agree, but differences will emerge when we specify the form of $H_0$ (the null hypothesis) and $H_1$ (the alternative hypothesis). To illustrate, sometimes $H_1$ is specified to be the hypothesis that the difference in the results between the conditions in an experiment is due to the independent variable, and $H_0$ is stipulated to be the negation of $H_1$. In other words, $H_0$ is the negation of the research hypothesis, such as "marijuana slows reaction time" (see Pagano 1994, p. 222).

If $H_0$ takes this form, then no mistake is made when a researcher infers $H_1$ from not-$H_0$ alone, but another rather obvious problem arises: the usual decision rules for rejecting the null hypothesis are incorrect (assuming that such rules are "correct" if and only if satisfying their conditions justifies the inference they specify).

For example, the following rule, which appears in an excellent statistics textbook, is incorrect even if alpha is set at 0.001, rather than the more usual 0.05 or 0.01:

> (R) If the obtained probability is less than or equal to alpha, reject $H_0$. (Pagano 1994, p. 223)

The key problem, of course, contrary to what the rule permits, is not merely that we are not entitled to infer from statistically significant results *alone* that the research hypothesis is true. If that were the sole problem, we could modify the rule to direct us to infer not the falsity of $H_0$ but the weaker conclusion that there is *some* evidence for its falsity. The deeper problem is that the finding of statistically significant results is not generally any evidence at all that $H_0$ is false *if* it is the negation of the research hypothesis. At the very least, other plausible competitors to the research hypothesis (besides a chance explanation) need to be ruled out.

It may be that those who are endorsing inference rules such as (R) are tacitly presupposing that the design of a particular experiment is adequate for ruling out all plausible competitors to the research hypothesis except a chance explanation. If so, some warning should be given that R is not to be invoked in quasi-experimental and epidemiological studies where the presupposition is not met. The warning is not needed, however, if we simply include the presupposition in a new version of the rule:

> ($R_1$) If the design of an experiment is sufficient for ruling out all rivals to $H_1$ that are of equal or greater credibility, then if the obtained probability is less than or equal to alpha, then reject $H_0$.

For a reason to be given shortly, $R_1$ is also incorrect, but even if we waive this objection, there is another problem. Whether correct or not, $R_1$ is not an informative inference rule. Without some specification of what constitutes meeting the first condition, the rule does not make clear when $H_0$ is to be rejected.

Chow avoids the above problems by, among other things, stipulating a different form for the null hypothesis, and by not endorsing any simple inductive rules for inferring the research hypothesis from experimental results. In his account, it is an oversimplification, if not a misrepresentation, to consider NHSTP an inductive procedure (p. 67). The sole function of NHSTP, in his view, is to exclude chance influences as an explanation (p. 88). This exclusion is not in itself confirmatory, but it does play an important role in theory corroboration as reconstructed by Chow.

I agree with Chow's strategy of defending NHSTP by treating it as a noninferential procedure, but I have some questions about his reconstruction of theory corroboration.

In his theory corroboration example, $H_1$ is that the mean number of extra-sentence words recalled after negative sentences is less than the mean number recalled after kernel sentences; and $H_0$ is that the mean number of extra-sentence words recalled after negative sentences is greater than or equal to the mean number recalled after kernel sentences (see p. 47). $H_0$ is to be rejected,



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

then, if the results are statistically significant. If $H_0$ is shown to be false, then we can use a disjunctive syllogism to show that $H_1$ is true (see p. 15). Given what $H_1$ says, however, how do we move from it to the conclusion that the results support the research hypothesis, namely that the linguistic competence of native speakers of English is an analogue of the transformational grammar? Chow's answer takes the form of three embedded syllogisms (see p. 70, Table 4.2).

$H_1$ serves as one of the premises of one of these syllogisms, the conclusion of which serves as a premise of a second syllogism, which in turn has a conclusion that serves as a premise for the final syllogism. Given the premises of this last syllogism, the argument would commit the fallacy of affirming the consequent if its conclusion were the research hypothesis (the same fallacy would be committed in the other two arguments if the conclusion were the antecedent of the conditional premise). Chow is well aware of the fallacy of affirming the consequent; to avoid it, he substitutes for the research hypothesis in the final syllogism the proposition that the research hypothesis "is true in the interim (by virtue of experimental controls)." He makes a similar substitution in the conclusion of the other two syllogisms. Chow later explains that "true in the interim" indicates that the conclusion qualified by this phrase is tentative (p. 77).

Inserting the phrase "true in the interim" renders each of the three arguments innocent of the charge of the fallacy of affirming the consequent, but it does nothing to guarantee validity. In fact, all three of Chow's syllogisms in table 4.2 (p. 70) are logically invalid. For that reason alone, his account does not explain how experimental data provide evidential support for a theory.

There are other problems as well. Chow intends to justify instances of the "in the interim" qualification by virtue of the stipulation "by virtue of experimental controls" (p. 72). This introduces the same kind of unclarity I referred to earlier in discussing Rule $R_1$: without some general specification of what constitutes adequate experimental controls, Chow's account does not give us a useful formal procedure for telling when experimental data are supportive.

There is a further problem to which I alluded earlier. Suppose that in any given case the experimental controls are adequate for discounting all known credible rivals to the research hypothesis, and that a chance hypothesis is then ruled out. Can we then infer that the research hypothesis has at least some tentative empirical support? To answer "yes" is to assume the validity of the following rule, which is one version of what philosophers refer to as the rule of "inference to the best explanation":

> ($R_2$) If H explains data D (i.e., it would explain the data if it were true), and H is more credible than any known rival explanation, and the data are statistically significant, then infer H (or at least that H has some empirical support).

$R_2$, however, is invalid. In some cases, we have evidence that the set consisting of H and its known credible rivals contains the true causal hypothesis that explains D. In other cases we do not; and consequently, satisfying the antecedent of the rule is no guarantee of empirical support (for further discussion, see Erwin 1996, pp. 62–73; and van Fraassen 1989).

It may be that Chow and I are not far apart, given that he expresses reservations about his attempt to lay out the logic of theory corroboration. This topic is important in its own right, but even if Chow has not succeeded here, this failure would not undermine his defense of NHSTP, nor would it affect his criticisms of power analysis, meta-analysis, or Bayesianism.

I will finish with a brief comment on arguments published after the publication of Dr. Chow's book.

John Hunter (1997) argues that NHSTP *as currently used* is a disaster, and that it should be banned. His key argument is that the error rate for the significance test is not the 5% that most psychologists believe it to be but is, rather, on average 60% in psychology. If the treatment in a study has an effect, he points out (1997, p. 4), then the only error possible is a Type II error: falsely concluding that the treatment has no effect. Studies show, Hunter continues, that the Type II error rate in some areas is 90%, and is 60% on average.

As part of his evidence, Hunter relies on a study that found low power rates for most of the 64 experiments that were reviewed (Sedlmeier & Gigerenzer 1989). Why is this finding relevant to the determination of Type II error rates? Hunter assumes that it is relevant because he equates low statistical power with a high Type II error rate (p. 5). This is a mistake, given his definition of a Type II error ("falsely concluding that the treatment has no effect when it actually does have an effect," p. 4).

Suppose that a drug treatment for clinical depression has no effect on depression, but clinical studies show improvement due to placebo factors, which are not controlled for. If some of these studies have low statistical power, the effects may not be statistically significant, but in failing to reject the null hypothesis, that is, that the observed effects are not caused by the treatment, no Type II error is committed. In general, showing that studies are not sensitive enough to detect a statistically significant result even if there is one cannot by itself show that Type II errors are being committed. An additional argument needs to be made that the independent variable really did cause the effects specified in the research hypothesis.

Of the 64 experiments reviewed by Sedlmeier and Gigerenzer, 7 had null hypotheses as research hypotheses (i.e., they were hypothesizing no difference between treatments). In all 7 cases, the experimenters took the lack of significance as a confirmation of their research hypothesis. Sedlmeier and Gigerenzer object (p. 312) that the power of the statistical tests was too low to warrant any of these inferences, but they provide no evidence that in any of these cases the null hypothesis was *false;* nor do they even claim to provide such evidence for any of the remaining 59 experiments.

Hunter does have a second argument. He points out that in Lipsey and Wilson's (1993) review of 302 meta-analyses, in only 3 cases is the effect size 0, thereby suggesting that the null hypothesis is true in a little less than 1% of the research domains considered (p. 5). How does Hunter get from the premise that the average effect size in the meta-analyses being reviewed is almost always above zero to the conclusion that the observed effects were caused by the treatment being studied? He does not say, but he may be relying on the following inference rule (which might as well be put on the table, given that it appears to be relied on in some meta-analytic reviews):

> ($R_3$) If the average effect size in studies of treatment T is above zero, infer that T caused the observed effects, or at least that this hypothesis has some degree of empirical support.

$R_3$ is incorrect. Given the standard ways of calculating effect sizes (e.g., see Smith et al. 1980), the finding of effect sizes above zero is neutral between competing hypotheses as to what caused the effects. Whether or not anyone would ever rely on $R_3$, we can still ask: Of the hundreds of meta-analytic reviews that have been published, how many provide solid evidence for their conclusions? To answer that question, the type of review done by Lipsey and Wilson (1993) is insufficient. Rather, what is called for is an epistemological review of the arguments of meta-analytic reviewers.

No one has published such a review for any large sample of meta-analytic reviews. There have, however, been epistemological reviews of individual meta-analyses, such as the one done by Smith et al. (1980), and the meta-analytic arguments have been found flawed (Erwin 1997, Ch. 8; see Shapiro 1997 for examples from cancer research). These examples are too few to support a general skeptical conclusion, but they do suggest that it is premature to conclude, on the basis of the meta-analyses done so far, that in studies of psychological treatments, the null hypothesis is false 99% of the time, or even most of the time.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## Chow's defense of null-hypothesis testing: Too traditional?

Robert W. Frick

*Department of Psychology, State University of New York at Stony Brook, Stony Brook, NY 11790.* **rfrick@sunysb.edu** **www.psy.sunysb.edu/rfrick/**

**Abstract:** I disagree with several of Chow's traditional descriptions and justifications of null hypothesis testing: (1) accepting the null hypothesis whenever $p > .05$; (2) random sampling from a population; (3) the frequentist interpretation of probability; (4) having the null hypothesis generate both a probability distribution and a complement of the desired conclusion; (5) assuming that researchers must fix their sample size before performing their study.

Critics of the null-hypothesis statistical-testing procedure (NHSTP) do not tend to criticize one another, despite differences in their positions. For example, NHSTP is criticized but power analyses are not, even though a power analysis assumes the existence of NHSTP. Researchers are advised to report effect size in statistical units such as Cohen's *d* (e.g., Schmidt 1996) or to report confidence intervals (e.g., Loftus & Masson 1994), but they are not told to report confidence intervals for effect size reported in statistical units. Cohen (1994) criticized the underlying logic of NHSTP but then suggested that researchers report confidence intervals because that accomplished NHSTP for all possible null hypotheses.

One might expect the defenders of NHSTP to ally, but this alliance too would be unnatural. I agree with Chow that NHSTP plays an essential and irreplaceable role in science (Frick 1996). I agree with many of his points, especially that effect size is not relevant in the theory-corroboration experiment. However, I disagree with many of the justifications Chow provides for NHSTP. In this commentary, I will focus on ways that Chow is in a sense too traditional. In assessing NHSTP, the actual practice of researchers must be distinguished from the way it is described in textbooks and the attempts to justify that practice logically. In each of the following criticisms, Chow has defended the traditional description or justification of NHSTP rather than the actual practice of researchers.

First, Chow implies that the null hypothesis is accepted whenever $p > .05$. Good researchers do sometimes argue that their evidence supports a hypothesis of no effect or no difference, but they use more evidence than just $p > .05$ (e.g., Frick 1995).

Second, Chow uses random sampling from a population to justify the construction of the requisite probability distribution. This implies that researchers should sample randomly from populations and that the business of statistical testing is making claims about populations. I disagree. To make a claim about a pattern in the data, such as that one treatment is more effective than another, the researcher must address the possibility that this observed pattern occurred just by chance. As Chow notes, statistical testing accomplishes this, with *p* being a measure of the strength of the evidence against the just-by-chance hypothesis. The outcome of statistical testing and a lack of artifacts – which I call the finding – is a conclusion about the subjects tested. No assumption of randomsampling is needed for this interpretation (Frick, in press b).

Third, Chow defends the frequentist interpretation of probability, in which probability is defined as the limiting ratio of an infinite sequence of trials. This definition confuses probabilities with the method of measuring probabilities. In other words, it is the operationalism Chow decries (p. 153). A propensity definition of probability better justifies the procedures of NHSTP (Frick, in press b).

Fourth, in the traditional justification of NHSTP, the null hypothesis plays two roles – it generates the probability distribution underlying the determination of *p*, and it is the complement of the researcher's desired conclusion. These two roles are incompatible. To generate the probability distribution, a point hypothesis, for example, $\mu_1 = \mu_2$ is needed. However, the complement of this is $\mu_1 \neq \mu_2$, which is not the claim researchers make and – as critics of NHSTP are fond of noting – not even a claim worth making. Researchers in practice make a directional claim, such as $\mu_1 < \mu_2$. To allow this claim, Chow describes the null hypothesis as being directional, for example, $\mu_1 \leq \mu_2$. However, this leads Chow to the awkward position of primarily defending the use of a one-tailed test, which researchers rarely use. This definition also does not support the definition of *p* as the probability of achieving the observed results or larger given the null hypothesis.

A solution is this: A point hypothesis is used to generate the probability distribution. Following the conventional rules of science, $p < .05$ allows rejection of this hypothesis, and it would also allow rejecting the hypotheses even more discrepant from the observed data. Therefore, a directional conclusion can be made. This is exactly the process Chow describes (and Fisher before him), but it cannot be described with a single null hypothesis serving two roles.

Fifth, Chow equates NHSTP with the fixed-sample stopping rule, in which the number of subjects is determined in advance. Do researchers actually use the fixed-sample stopping rule? Do researchers never (a) give up part way through a study because the results were discouraging, (b) test less than the planned number of subjects because *p* was already less than .001, or (c) test more subjects than planned when *p* was slightly greater than .05? These actions seem rational to me, but they violate the fixed-sample stopping rule. Fortunately, the alternatives to the fixed-sample stopping rule – sequential stopping rules in which the number of subjects is not fixed in advance – are compatible with NHSTP. Because of their increased efficiency and practicality, sequential stopping rules should usually be preferred to the fixed-sample stopping rule (Frick, in press a).

## We need statistical thinking, not statistical rituals

Gerd Gigerenzer

*Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, 14195 Berlin, Germany.* **gigerenzer@mpib-berlin.mpg.de**

**Abstract:** What Chow 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 chance as an explanation for data. I disagree. This ritual discourages theory development by providing researchers with no incentive to specify hypotheses.

Future historians of psychology will be puzzled by an odd ritual, camouflaged as the *sine qua non* of scientific method, that first appeared in the 1950s and was practiced in the field for the rest of the twentieth century. In psychology and education textbooks of this period they will find this ritual variously referred to as "statistical inference," null hypothesis testing, significance testing, and most recently, NHSTP. These historians will be surprised to learn that the ritual was quickly institutionalized, although (1) the eminent psychologists of the time – including Sir Frederick Bartlett, R. Duncan Luce, Herbert Simon, B. F. Skinner, and S. S. Stevens – explicitly wrote against its use (Gigerenzer & Murray 1987); (2) the statisticians Sir Ronald Fisher, Jerzy Neyman, and Egon S. Pearson would all have rejected NHSTP as an inconsistent mishmash of their ideas (Gigerenzer et al. 1989, Chs. 3 and 6); (3) hardly any eminent statistician of the time endorsed it; and (4) although it was presented to psychologists as the scientific method, it never caught on in the natural sciences.

Chow (1996) responds to a paper (Gigerenzer 1993) in which I used a Freudian analogy to capture how the conflicts between Neyman and Pearson's doctrine (the superego), Fisher's null hypothesis testing (the ego), and the Bayesians's approach (the id) have been projected into the psyches of textbook writers and researchers in psychology. The results are wishful thinking, sup-



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

pression of conflicts, and a statistical practice – null hypothesis testing – that resembles ritualistic handwashing. For instance, many textbook authors and the majority of experimenters do not understand what its final product – a *p*-value – actually means (see Acree 1978; Gigerenzer 1993; Oakes 1986; Sedlmeier & Gigerenzer 1989). Chow acknowledges this, but argues that if we can strip NHSTP (his term for an inconsistent hybrid of Fisherian and Neyman-Pearsonian ideas) of the mental confusion associated with it, something of limited but important use is left. According to Chow, NHSTP's usefulness is "restricted to deciding whether or not research data can be explained in terms of chance influences" (p. 188). This sounds like a reasonable and modest proposal, and Chow succeeds in pointing out many sources of confusion about significance testing. I do not, however, believe that even in this purified form NHSTP has much value for psychological research. Rather, this ritual undermines progress in our field by giving researchers no incentive to specify their hypotheses and by replacing statistical thinking with a mindless statistical procedure.

***Is testing unspecified hypothesis against "chance" a good research strategy?*** No. The single most important problem with null hypothesis testing is that it provides researchers with no incentive to develop precise hypotheses. To perform a significance test, one need not specify the predictions of either one's own research hypothesis or those of alternative hypotheses. All one has to do is test an unspecified hypothesis ($H_1$) against "chance" ($H_0$). In my experience, the routine of testing against chance using NHSTP promotes imprecise hypotheses.

To be sure, there are cases where testing against chance makes sense, such as in parapsychology.[1] But read John Arbuthnot's proof of God against chance in 1710 – the earliest null hypothesis test of which I know – and you see the flaws in this program (Gigerenzer & Murray 1987, pp. 4–5). In a science striving for precise process models, one needs methods that test the predictions of one model against those of alternative models, not a ritual that tests an unspecified hypothesis against chance.

Recall that statistical thinking involves making an informed choice among the various techniques available. Avoiding statistical thinking in the name of "objectivity," as Chow's implicitly advocates, has produced blind spots in research (Gigerenzer 1987). There is a toolbox of statistical methods for testing which of several predictions, if any, comes closest to the data. For certain problems least squares are useful, for others maximum likelihood, Neyman-Pearson analysis, Wald's sequential analysis, or Bayesian models. But even simple descriptive statistics can be better than null-hypothesis testing at discriminating between hypotheses. For instance, Anderson and Cuneo (1978) proposed two hypotheses about the processes underlying children's estimates of the area of rectangles ("adding" versus "multiplying" height and width). Following the null hypothesis-testing ritual, they identified one with chance ("adding") and did not specify the predictions of the other. Because the anova test was not significant, they took this as evidence for the "adding" process. However, had the authors specified the precise predictions of *both* hypotheses, they would have seen that the data pattern was in fact close to that predicted by the "multiplying" process and not by the null hypothesis (see Gigerenzer & Murray 1987, p. 100; Gigerenzer & Richter 1990). This example illustrates one blind spot that results from using NHSTP, which requires that the prediction of only one hypothesis be specified. Hypothesis testing should be symmetric, not asymmetric.

NHSTP allows researchers to get away with imprecise hypotheses and predictions. Testing an unspecified hypothesis against chance may be all we can do in situations where we know very little. But when used as a general ritual, this method ironically ensures that we continue to know very little.

***Compulsory rules.*** Chow proclaims that null hypothesis tests should be interpreted mechanically using the conventional 5% level of significance. This is what Fisher suggested in his 1935 book, a practice that was subsequently codified by many textbook writers into a religious doctrine of "objectivity." Later, this practice was rejected by both Fisher and Neyman and Pearson, as well as practically every other eminent statistician (Gigerenzer et al. 1989). The reason Fisher adopted a conventional level of significance of 5% (or 1%) in the first place seems to have been that he had no table, for other significance levels, partly because his professional enemy, Karl Pearson, refused to let him reprint the tables Pearson had. In the 1950s, Fisher rejected the idea of a conventional significance level: "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; rather he gives his mind to each particular case in the light of his evidence and his ideas" (Fisher 1956, p. 42). He then recommended reporting the exact level of significance instead (e.g., $p = .03$, but not $p < .05$).

In my opinion, statistical thinking is an art, not a mechanical procedure. Chow's view reminds me of a mechanical maxim regarding the critical ratio, the predecessor of the significance level: "A critical ratio of three, or no Ph.D."

What we need to teach our students is neither NHSTP nor any other statistical ritual. We need to teach them statistical *thinking:* how to generate bold hypotheses, derive precise alternative predictions, set up experiments to minimize real error (rather than just to measure and insert error into the F-ratio), analyze data for each individual separately if possible rather than automatically aggregating, and perform sound descriptive statistics and exploratory data analysis. And we need to teach them that there are several important statistical schools and tools, rather than pretending that *statistics is statistics is statistics is statistics.*[2] We should give students examples of situations where each tool works and where each does not work. Students should learn why Neyman believed that null hypothesis testing can be "worse than useless" in a mathematical sense (e.g., when the power is less than alpha), and why Fisher thought that Neyman's concept of Type II error reflects a "mental confusion" between technology (such as in Stalin's 5-year plans) and science (Fisher disdained the Russian-born Neyman; see Gigerenzer 1993). We can make statistics fun and interesting by scrapping the thoughtless ritual advocated by Chow and instead teaching students about the real statisticians and controversies behind the diverse array of statistical tools we have. Choosing among these tools requires statistical thinking, not rituals.

NOTES

**1.** Null-hypothesis testing (*t*-test and anova) was first applied in parapsychology and education, from which it spread to basic research. Danziger (1990) offers an interesting argument for why this happened in the United States and not in Germany.

**2.** Chow acknowledges that there exist different logics of statistical inference. But at the same time he falls into the it's-all-the-same illusion when he asserts: "To K. Pearson, R. Fisher, J. Neyman and E. S. Pearson, NHSTP was what the empirical research method was all about." (p. xi). This statement is incorrect. Neyman and Pearson spent their careers arguing against Fisher's null hypotheses testing and developing their own alternative, which rests on two precise hypotheses (rather than one null hypothesis) and the concept of Type-II error (which Chow declares not germane to NHSTP). Furthermore, for Fisher (1955; 1956), null-hypothesis testing was only one of several useful statistical methods, such as maximum likelihood and fiducial probability (Gigerenzer et al. 1989, Ch. 3; Hacking 1965).

## Stranded statistical paradigms: The last crusade

Judith Glück and Oliver Vitouch

*Institute of Psychology, University of Vienna, A-1010 Vienna, Austria.*
**judith.glueck@univie.ac.at; oliver.vitouch@univie.ac.at**

**Abstract:** Chow tries to show that for the case of hard-core experimentation, the criticisms of NHST are not valid. Even if one is willing to adopt his epistemological ideology, several shortcomings of NHST remain. We argue for a flexible and thoughtful application of statistical tools (including significance tests) instead of a ritualized statistical catechism that relies on the magic of $\alpha$.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Chow's (1996) book, a sequel to similar earlier publications (e.g., Chow 1988), is devoted to defending the holy grail of null-hypothesis significance testing (NHST) against an armada of increasingly influential critics. In his system of "theory-corroboration experimentation," Chow refers to idealistic experimental settings best served by *experimenta crucis.* In fact, he earnestly argues that ecological validity is detrimental to the quality of a study (pp. 102, 104).

***Chow's logic of scientific discovery.*** Implying that he practices Popperianism, Chow uses Popper's "modus tollens" system of falsification in a dubious way: "weak" NHST is promoted instead of strong testing of highly specific theoretical predictions (cf. Meehl 1978). In addition, Chow takes no notice of the critics of theory corroboration itself: Lakatos (e.g., 1970) is not mentioned at all; Kuhn (e.g., 1970) is referenced only en passant; not to mention modern constructivist perspectives (see Folger 1989 for an excellent discussion of Chow's insufficient logic).

Chow's seemingly modest argument is that "all we need" is a deterministic binary decision as to whether or not there is an effect. Unfortunately, statistics do not give us any instrument to test the existence of an effect independent of its size. Even if a researcher is not at all interested in effect size, it will be effect size, in combination with sample size, that determines significance.

Advocating binary decisions, Chow directs a film that is strictly black and white (accept/reject, pass/fail, heaven/hell). Unfortunately, it is even mostly black (a *film noir*). For 35 years, a long series of studies have shown that in vivo, power is usually much too small to reliably reject a false $H_0$ (e.g., Cohen 1962; Sedlmeier & Gigerenzer 1989; and most recently, Clark-Carter 1997). While Cohen & Co. typically focus on real-life research situations, Chow persistently labels every point that could threaten his system as "extra-statistical."

***Sometimes, size does count.*** The book does not give any indication of what sample size should be chosen (because this is impossible without reference to the concept of statistical power). Chow, usually devoted to strictness, relies on good fortune on this point: he seems to see researchers as a special type of *idiot savant* who generally lack judgment, but have an intriguingly good sense for appropriate sample sizes. If they do not get sample size right (whatever "right" means in Chow's view), they are either "cavalier" or "cynical." Sadly, Tversky and Kahnemann (1971) showed that a huge percentage of the audience at a mathematical psychology meeting must have been cynics, then. The cynical attitude also seems to be widespread in the field of brain imaging, where researchers draw inferential conclusions from median sample sizes as small as eight (Vitouch & Glück 1997).

***Taking rigidity for rigor: The α fetish.*** Chow defends α as a mathematically well-grounded and objective criterion (it makes good sense that there are giant α's on the book's cover). However, mere replication of this statement does not alter the fact that α is a normative and, to some degree, arbitrary convention (.05 and .01 are primarily "beautiful numbers"). Remarkably, Chow argues that a chosen α level is well-defined, whereas a proportion of explained variance is ill-defined. For a correlation coefficient, most researchers would agree that beyond significance, the size of the correlation is essential. But they often do not realize that a *t*-test can also be expressed as a correlation between group membership and the dependent variable, and the consideration of corresponding anova indices of association strength or explained variance ($\eta^2$, $\omega^2$) has not been generally established yet.

Whether *p* is .000007 or .048 does not make any difference in Chow's system; whether *p* is .048 or .051 is crucial. Chow believes in a hybrid system of pure statistics (Gigerenzer 1993), adopting the Neyman-Pearson concept of binary decisions based on a pre-set α level (the Fisherian *p* is interpreted numerically), but rejecting their related notion of pre-considering statistical power. That is why he keeps affirming that "A smaller α value indicates a stricter criterion" (p. 38), but never reflects that strictness against the Type I error is inversely related to strictness against the Type II error which may be equally crucial for the future of a theory. Chow argues that the probability of a Type II error cannot be computed because the true effect is unknown. However, since Type II errors do occur, it seems more reasonable to take power into account than to ignore the problem completely. Strict protection against the wrong error can regularly be observed in cases of inverse testing (like testing statistical assumptions or model fitting), where β should be rigorously controlled. Here, *p* values – denoting the probability of the data if the model were true – of .013 are often happily accepted.

***Style.*** The style of the text is often unfair and unobjective. On p. 110, Chow states that "meta-analysis was originally developed as a means to influence decision makers in some bureaucracy. This motivation renders it understandable (but not justifiable) why conceptual rigour or research quality is not deemed important in meta-analysis." Chow ignores the fact that introducing study variables as breakdown criteria enables the meta-analyst to test design influences on the outcome of studies – sorting apples from oranges, and even gaining information about the difference between them. In his own words, "that meta-analysis may sometimes be misused is an insufficient reason to abandon the method" (Chow 1988, p. 109, "meta-analysis" replacing "the significance test" – certainly the latter has far more often been misused).

Chow's terminology is notoriously connotative: whereas power analysis is "mechanical" and assessing effect sizes is "utilitarian," NHST is "qualitative," "objective," and "well-defined." Promoting a technical, thoughtless procedure (NHSTP) himself, Chow manages to accuse Cohen's power analysis of being technical and thoughtless, whereas Cohen (1994, p. 1001) clearly states: "First, don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist."

Another technique can be summarized as follows: present one good and one poor critical argument. Then, draw on why the poor argument is foolish for many pages, until everybody has forgotten the good argument. Subsequently, refer to the good argument using the phrase "As has been shown." For instance, intending to prove that confidence intervals are useless, Chow only demonstrates at length that they cannot resolve design flaws (Ch. 4.8).

***A case for statistical enlightenment.*** We agree that statistical conventions are necessary; but everybody should know that they are conventions, and they should be handled thoughtfully and flexibly, not as ritualized and mechanical procedures. Significance tests can be a useful tool, but they are not magical, definite, or self-contained – and they do not make decisions. It is always the researcher who decides, based on the detailed information he has at hand. Here, "Making sense is more important than making numbers" (K. G. Jöreskog, personal communication, June 29, 1997). We hope that future researchers will be able to explain and justify why they used a certain statistical approach instead of depending on any magic formulas. Catechisms are no longer needed – thinking is called for.

## Understanding Bayesian procedures

Robert A. M. Gregson

*Division of Psychology, School of Life Sciences, Australian National University, Canberra ACT 0200.* **robert.gregson@anu.edu.au**

**Abstract:** Chow's 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.

Arguments about the use of null-hypothesis significance testing by psychologists now abound (Gonzalez 1994; Townsend 1994) and seem to be of particular concern to the American Psychological Association. These disputes have taken a rather different character in other disciplines such as economics or physics, and in other



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

locations such as Britain (Howson & Urbach 1994; Sivia 1996) or Australia (Walley 1991). Chow starts with the well-worn disputes between Neyman/Pearson and Fisher and notes, as others have done, that what many psychologists actually do is a quaint hybrid of the two positions. My first amazement, as one who studied under Egon Pearson in the 1950s, is the assertion (p. 27) that inverse probabilities were admitted in the Neyman/Pearson-developed framework. It is true that Egon Pearson himself did modify his metatheoretical stance a bit over the years, but the Neyman/Pearson received doctrine was enshrined by F. N. David in Pearson's department; she admitted the use of Bayes' theorem, and hence inverse probability, only in some very restricted problems in genetics. She damned Bayesian inference and Sir Harold Jeffreys by consigning them to footnotes and not mentioning them supportively in lectures.

Chow would do well to learn a little of the sociology of science, for Lindley, who succeeded Pearson, advanced the resurgence of Bayesian methods and produced critical refutations of frequentist significance testing (Lindley 1965, p. 69) which are not covered by Chow. Bayesian inference is now back in the authoritative text with a volume to itself (O'Hagan 1994). In fact, one can argue that the objections to subjective probability and Bayes are a nineteenth century aberration (Dale 1991) which has been redressed in the last thirty years – except by psychologists. Chow's book reveals an extensive ignorance of what modern Bayesian inference is about, and its underlying rationale. It substitutes uninformed polemic (Chow, p. 144, et seq.) drawn from secondary derivative texts for a serious review of contemporary statistical source literature.

My second amazement was caused by the absence of any references to exchangeability (de Finetti 1974), sufficient statistics, sample spaces, stopping rules, and maximum entropy (Heidbreder 1996), all of which play important roles in Bayesian criticisms of frequentist inference. Perhaps it is as well to note that there are in fact Bayesian analogies of significance testing, and for some decisions in simple situations they will lead to the same end result as the frequentist cookbooks. The difficulties begin for the frequentists, who are forced into a lot of ad hoc constructions, when situations are complicated enough to stand a chance of resembling psychological reality (Smith & Roberts 1993; Sun et al. 1996).

The obscurity of Chow's criticism begins again on p. 145, where some axioms are mentioned but not given. In fact the axioms that support Bayes' theorem are of two sorts: those on probability spaces (Hartigan 1983, p. 30) and those on conjoint probability, where the symmetry of

$$p(X \cap Y) = p(X|Y) \cdot p(Y) = p(Y|X) \cdot p(X)$$

supports the substitution of $H$ for $X$ and $E$ for $Y$ and then some rearrangement. But this in practice requires that the $p(H)$ terms be coherent; they are not just any old subjective degrees of belief or willingness to bet. If they do not satisfy the condition of coherence then contradictions will eventuate, and if they are coherent then they will sufficiently satisfy the axioms of probability, an apparently little-known result (Cox 1946). I am aware that in econometrics one can have nonadditive probabilities in choice theory, but that is peripheral to the arguments as applied to decisions about experiments.

Chow completely fails to grasp the logical status of probabilities and wants to argue that Bayesian degrees of belief about the truth-value of hypotheses are in fact frequency statements; this seems to be a muddle between (1) $p(E|H)$ likelihoods which are expressible as relative frequencies if and only if we have a specifiable causal model, and (2) $p(H_i)$ for some series of $n$ mutually exclusive and exhaustive $H_i, i = 1, \ldots, n$, which are revisable degrees of belief as data accumulate. The sad thing about Chow's missed opportunities is that there are rigorous and comprehensible books about Bayes (Bernardo & Smith 1994; Robert 1994) and about using a Bayesian approach to the psychologists' much-loved anova (Box & Tiao 1973; Carlin & Loui 1996) with adequate worked examples. To quote a philosopher (Hanson 1958, p. 272) on reviewing a book of similar worth: "Some books deserve to fall stillborn from the press. A collection of confusions more comprehensive . . . could hardly be invented."

## "With friends like this . . .": Three flaws in Chow's defense of significance testing

Richard J. Harris

*Department of Psychology, Logan Hall, University of New Mexico, Albuquerque, NM 87131.* **rharris@unm.edu**

**Abstract:** Chow's book should be read only by those who already have a firm enough grasp of the logic of significance testing to separate the few valid, insightful points from the many incorrect statements and misrepresentations.

Chow (1996) makes a number of valid points about significance testing. Principal among them is that rejection of $H_0$ is only one element in the process of testing a conceptual hypothesis. Chow's distinctions among the various purposes of research (e.g., theory corroboration versus testing practical utility) in terms of the likely size of the differences among conceptual, research, and statistical hypotheses are also useful. However, the few positive contributions of Chow's small paperback are lost in a sea of invalid conclusions and misrepresentations. Examples follow:

***1. The folly of accepting null hypotheses.*** Chow (p. 71) explicitly claims that "Not to reject $H_0$ is to deny $H_1$"; and he makes the same point via symbolic logic in the accompanying Table 4.3 (and elsewhere). This leads him to reverse the usual assertion that a statistically significant result is less ambiguous than a nonsignificant result. Instead, Chow claims that a nonsignificant result proves that the alternative hypothesis (and the theory that generated it) is wrong (and $H_0$ is therefore true), whereas statistical significance has a host of interpretations based on the many alternative theories that could also have generated $H_1$. But accepting $H_0$ (rather than simply failing to reject it) leads us quickly into logical contradictions. For instance, it is not uncommon to find that $M_1$ is not significantly different from $M_2$ and that $M_2$ is not significantly different from $M_3$ but that $M_1$ is significantly different from $M_3$. Assuming two-tailed tests, accepting $H_0$ would lead us to conclude (from the first two tests) that $\mu_1 = \mu_2$ and that $\mu_2 = \mu_3$, which jointly imply that $\mu_1 = \mu_3$, and which directly contradicts our conclusion (from the third test) that $\mu_1$ differs from $\mu_3$. If, however, we "fail to reject" $H_0$ our conclusions become that we have insufficient evidence to be sure that $\mu_1$ differs from $\mu_2$ or that $\mu_2$ differs from $\mu_3$, but we *do* have sufficient evidence to be confident that $\mu_1$ differs from $\mu_3$ – a perfectly consistent set of conclusions. The consequences of accepting $H_0$ when using one-tailed tests (which Chow uses almost exclusively, about which more below) are even more perverse. Assume that we had predicted ($H_1$) that $\mu_1 > \mu_2 > \mu_3$ and assume further that the sample means had come out in exactly that order, but with the (one-tailed) tests of $M_1$ versus $M_2$ and of $M_2$ versus $M_3$ not quite reaching statistical significance, while $M_1$ versus $M_3$ is significant. Accepting $H_0$ would lead us to conclude that $\mu_1 \leqslant \mu_2$ (despite the fact that $M_1 > M_2$!) and that $\mu_2 \leqslant \mu_3$, which jointly imply that $\mu_1 \leqslant \mu_3$, thereby coming to two conclusions that are inconsistent with the direction of our sample differences among the means and which jointly contradict both the observed $M_1$ versus $M_3$ difference and the conclusion of our third significance test.

***2. Consequences of "buying into" two-valued hypothesis-testing logic.*** As Kaiser (1960) pointed out, insisting that there are only two possible conclusions from a test of the difference between two means ($\mu_1 = \mu_2$ vs. $\mu_1$ not $= \mu_2$ for a two-tailed test; $\mu_1 \leqslant \mu_2$ vs. $\mu_1 > \mu_2$ or $\mu_1 >= \mu_2$ vs. $\mu_1 < \mu_2$ for a one-tailed test) leaves us with no stipulation as to the direction of the population



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

difference if we achieve two-tailed significance, and leaves us unable ever to disconfirm our research hypothesis, no matter what the outcome of a one-tailed test. Over 3.5 decades later, I pointed out (Harris 1997a; 1997b) that "the researcher who takes traditional [significance-testing] logic seriously is thus faced with an unpalatable choice between (a) being unable to come to any conclusion about the sign of the effect or (b) violating the most basic tenet of the scientific method (empirical data as the arbiter of our conclusions)." Fortunately, most researchers ignore the dictates of traditional, two-valued logic and instead implicitly adopt Kaiser's suggestion that every two-mean test be interpreted as yielding three possible conclusions – that $\mu_1 > \mu_2$, that $\mu_1 < \mu_2$, or that we have insufficient evidence to be confident of the direction of the population difference. The one-tailed test simply eliminates one of the two directional alternatives (whichever one is opposite to the predicted direction) from consideration, and thus represents scientifically unacceptable research behavior.

Properly interpreted, one-tailed tests are so antithetical to the scientific method and two-valued logic leads so inexorably to a preference for one-tailed tests that any analysis of significance testing that endorses (as does Chow's book) both one-tailed tests and two-valued logic constitutes a threat to the proper socialization of researchers.

***3. Power misunderstood and misrepresented.*** Chow claims that the concept of power has nothing to do with significance testing (cf. all of Ch. 6). However, his examples and arguments in favor of this claim are seriously flawed. For instance, in each of his examples of "power analyses" he changes the mean about which *t* is centered to the actual population mean as the latter departs from zero, thus leading, for example, to two purportedly null distributions (Panels C and D of Fig. 6.2) centered about population mean differences of 1.0 and 3.0, respectively. He further argues that since power requires a consideration of two distributions while significance testing is based solely on the null distribution, the two cannot be related. This has the same logical status as arguing that factor A has nothing to do with the A × B interaction because a plot of A ignoring B requires a single line, while the A × B interaction requires a different line (set of means for different levels of A) for each level of B. Most peculiarly, Chow argues that power and Type II error (which other authors persist in seeing as complimentary) have nothing to do with each other, because power – Pr(reject $H_0$|not $H_0$) in Chow's notation – cannot be computed until we know the specific amount by which $\mu_1 - \mu_2$ departs from its null-hypothesized value, whereas the Type II error rate – $\beta$ = Pr(fail to reject $H_0|H_1$) – is based on the broad alternative hypothesis that incorporates all possible departures from $H_0$. This is nonsense, since we cannot compute $\beta$ either, until we have a specific value of $\mu_1 - \mu_2$ (or a standardized version thereof) with which to conjure.

## Reconnecting data analysis and research design: Who needs a confidence interval?

Andrew F. Hayes

*Department of Psychology, University of New England, Armidale, NSW 2351, Australia.* **ahayes@metz.une.edu.au www.une.edu.au/~ahayes/**

**Abstract:** Chow illustrates the important role played by significance testing in the evaluation of research findings. Statistics and the goals of research should be treated as both interrelated and separate parts of the research evaluation process – a message that will benefit all who read Chow's book. The arguments are especially pertinent to the debate over the relative merits of confidence intervals and significance tests.

In *Statistical significance*, Chow argues that the null hypothesis significance testing procedure (NHSTP) plays a "very limited, albeit important" role in the research enterprise (e.g., p. 65). The rejection of a null hypothesis allows one to rule out "chance" as a plausible explanation for a research finding, thereby initiating a sequence of logical deductions about the substantive hypothesis or theory under investigation. But the validity of those deductions depends on the quality of the research design and what the research was intended to accomplish. In this book, Chow convincingly relinks statistics with research design, two mutually dependent areas that often seem separated in both the science curriculum and the minds of researchers (including critics of NHSTP), while at the same time reminding the reader to keep the role played by statistics and that played by research design separate when they should be. NHSTP should not be criticized for a failure to provide answers to questions that are better answered through good research design (e.g., theory support) or that no statistical procedure can or should be used to answer (e.g., practical importance of a research finding).

Chow's arguments are especially relevant to the current debate over the comparative benefits of significance testing compared to the seemingly more informative effect-size and parametric confidence interval. Advocates of confidence intervals claim that (1) NHSTP retards the growth of scientific knowledge because it ignores the quantitative information contained in a research finding, that (2) statistically significant effects may be trivially small, and that (3) rejection of a null hypothesis tells us nothing about the truth of the substantive hypothesis. These criticisms of NHSTP have many implicit assumptions which, when evaluated, are found wanting.

The first assumption is that confidence intervals provide more information about the truth of the substantive hypothesis than do significance tests. But all inferential statistical procedures, including confidence intervals, have only limited meaning in the context of a study. Significance testing is simply a means of ruling out "chance" as an alternative explanation for a research finding. Critics of NHSTP are right that rejecting the null hypothesis of chance tells us little about whether or not the substantive question or theory the study was designed to test is true or supported (see Ch. 3), but neither does a confidence interval that does not include zero. Whether the "nonchance" mechanism producing a research result (i.e., one that yields a nonzero effect) is the one proposed by the investigator can only be determined by good research design – namely, the elimination of competing explanations through proper control of potential confounds and a convincing translation of the substantive question into an empirical hypothesis. Neither NHSTP nor confidence intervals can or should be used to decide whether or not a research result supports an empirical hypothesis or substantive theory.

It is also assumed by advocates of confidence intervals that they provide more useful information than the qualitative information a significance test yields – namely, the size of the effect and its "practical importance." But the information provided by a confidence interval can be quite ambiguous, as the size of an effect and the width of a confidence interval will depend on such factors as how the variables are operationalized and how the effect is represented statistically. Different operationalizations of the same conceptual variables may yield different effect sizes (cf. Prentice & Miller 1992). The relation of some measures of effect size, such as the correlation coefficient or measures of "variance accounted for" to the levels or values of the variables observed in a sample may be artifactual. And, as Chow argues, there are many research questions where the size of the effect is uninformative or irrelevant. In "theory corroboration" or "generality" studies, for example, what is of interest is not the size of an observed effect but whether chance can be ruled out as an alternative explanation for the obtained results (i.e., a qualitative decision). Measures of effect size are relevant only for the "utilitarian" experiment, where the goal is to determine how much of a change is caused by various levels of the experimental manipulation. A large effect does not directly imply that the effect is important, nor does a small effect mean that the intervention is not "worth it" in cost-benefit terms. Cost-benefit questions cannot be answered by relying solely on a statistical index or procedure (Chs. 3, 4, and 5).

However, Chow's exclusive reliance on classical statistical the-



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

ory based on the sampling distribution of a statistic continues the tradition followed by confidence-interval advocates of overlooking the mismatch between the predominant statistical inference method and the data collection procedures typically used. Rarely do researchers randomly sample from a specified population or populations; instead, they usually collect their data from willing and conveniently available participants. In the absence of random sampling from a specified population, there is neither a statistical basis for inferring values of "parameters" (as confidence-interval advocates seem to desire) nor a solid theoretical justification for using the sampling distribution as the reference distribution for evaluating whether "chance" should be discounted as a plausible explanation for the findings (cf. Edgington 1966; 1995; May & Hunter 1993). But Chow's arguments are still valid ones, as they apply to any NHSTP, not just one based on random sampling and parametric inference.

Finally, while the messages contained in *Statistical significance* are important, they are unfortunately not always easy or entertaining to find. The book reads much like a textbook on logic, filled with jargon words and phrases (e.g., "disjunctive syllogism," "affirming the consequent"), which I fear may limit its potential influence. In several chapters, the reader is carefully guided through a long series of logical arguments but must remember a similarly long series of conditional propositions, antecedents, consequents, and syllogisms that, after being initially explained, are presented only symbolically (e.g., [P3.1.1′]). Thus, while the writing is disciplined, the reader must be similarly disciplined to follow it. Nevertheless, the payoff is worth the effort. All users, advocates, and critics of significance testing will find at least one valuable lesson here if the book is given the attention it deserves.

## Testing significance testing: A flawed defense

John E. Hunter

*Department of Psychology, Michigan State University, East Lansing, MI 48824.* **hunterj@pilot.msu.ed.**

**Abstract:** Most psychometricians believe that the significance test is counterproductive. I have read Chow's book to see whether it addresses or rebuts any of the key facts brought out by the psychometricians. The book is empty on this score; it is entirely irrelevant to the current debate. It presents nothing new and is riddled with errors.

Over the last 10 years, more and more psychometric experts have recommended that the significance test be abandoned. Chow's book is widely touted as the "answer" to these criticisms. Chow is a true believer and his book presents the significance test as the greatest discovery in the history of science.

This is a curious book; written in a clear and articulate style, but riddled with mathematical and methodological errors. It is the worst statistics book that I have ever read, by a wide margin. A literature search shows that the author has written an undergraduate stat text and a series of articles on the teaching of statistics; hence the polished writing style. However, that literature search shows no instance of either empirical research or a substantive literature review. Thus, the author is totally insensitive to real issues of study quality and the devastating effect that sampling error has had on the research review process.

About one third of the book is either directly or indirectly devoted to an immensely wordy, repetitive, and convoluted presentation of one simple and well known fact. The evaluation of a substantive hypothesis usually requires the examination of a wide variety of different specific numerical findings. The author castigates techniques which do not accomplish "theory corroboration" but merely consider numerical findings. Thus, the author rejects those who want quantitative measurement of effect sizes, those who use or develop power analysis, and those who do meta-analysis.

But the significance test doesn't do "theory corroboration" either! The significance test is a purely mechanical procedure for judging an isolated numerical result. The significance test totally ignores study quality; it makes no use of information on construct validity, reliability, or biased sampling. In contrast, those who use quantitative effect size measurement have long been concerned with correcting values for study imperfections such as imperfect measurement, range restriction, and so on. See Schmidt et al. (1976) for an article showing exactly how statistical power is related to specific objectively measured aspects of study quality.

Chow's chapter on effect sizes reminds me of the "barefoot and pregnant" strategy for wife management. If we do not teach researchers to measure effects quantitatively, they will have only a very tenuous grasp of sampling theory and thus they will not understand statistical power: this has been proven by the teaching methods and results of the last 60 years. Hence researchers will continue to use significance tests blindly despite empirical studies showing that the significance test has an average error rate of about 60% (Hunter 1997).

The chapter on power is primitive and error-ridden. There are no references to empirical studies on power or the implications of meta-analysis findings for the determination of power in current research (Hunter 1977). The author falsely claims that you cannot determine power unless you already know the population value of the treatment effect, but you can easily use a baseline value from meta-analyses of similar studies. My personal reading suggests that Chow really does not understand that sampling error will occur in studies where the null hypothesis is false. He never shows any recognition of the massive problem produced by false indications of conflict in the literature stemming from Type II errors; a problem that is very severe when the significance test is wrong 60% of the time!

Chow's treatment of meta-analysis is extremely bad. He makes many errors in misstating other papers and ignores virtually all the work on meta-analysis done by experts on sampling error. He correctly indicates that the meta-analysis statistical techniques are not "theory corroborative," but meta-analysis was never intended to evaluate theory directly: the objective of meta-analysis is *fact finding;* the precursor operation to theory evaluation. Each separate analysis is applied to a set of results where all studies have the same independent and dependent variable (though this may be a hypothesis to be tested rather than an assumption to be justified). Meta-analysis evaluates specific numerical findings, not substantive theories as such. On the other hand, how can you have definitive theory testing if you do not have definitive determination of the specific numerical questions required for theory corroboration? The function of meta-analysis is precisely to evaluate the specific numeric questions that lie at the heart of theory corroboration.

Does the significance test aid in theory corroboration? The error rate for the significance test is 60%; how can it aid in anything? Suppose that evaluating a substantive hypothesis requires a correct answer to 5 questions. If you use significance tests, then you have a 60% error rate for each answer. The probability of getting a correct overall inference is thus $(.40)^5 = .01024$. That is, using significance tests, you have a 60% error rate for specific findings but a 99% error rate for the overall substantive evaluation.

Most psychometricians believe that the significance test is counterproductive (Cohen 1994; Hunter 1997; Kirk 1996; Loftus 1996; Schmidt 1996). Researchers need to think of effects quantitatively and to understand confidence intervals as a way of realizing the true extent of uncertainty in their findings. I have carefully read Chow's book to see whether it addresses or rebuts any of the key facts brought out by the psychometricians. The book is completely empty on this score; it is entirely irrelevant to the current debate. Alas, the book also says nothing new and makes no positive contributions to the field of methodology.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## If you've got an effect, test its significance; if you've got a weak effect, do a meta-analysis

John F. Kihlstrom

*Department of Psychology, University of California, Berkeley, Berkeley, CA 94720-1650.* **kihlstrm@cogsci.berkeley.edu socrates.berkeley.edu/~kihlstrom**

**Abstract:** Statistical significance testing has its problems, but so do the alternatives that are proposed; and the alternatives may be both more cumbersome and less informative. Significance tests remain legitimate aspects of the rhetoric of scientific persuasion.

I admit it: after more than 25 years of reading, writing, reviewing, and editing scientific research in psychology and related fields, I still cannot understand the fury that whirls around statistical significance testing. Yet the critics seem to be gaining ground: the *American Journal of Public Health* virtually banned tests of statistical significance from its pages, at least for a time, and the American Psychological Association (APA) has seriously contemplated doing the same. Whatever the outcome of the APA's deliberations, the pages of *Psychological Science,* the flagship journal of the American Psychological Society, will remain open to significance tests so long as I remain editor. The reasoning behind this policy is more pragmatic than mathematical, but I am glad to have my view bolstered by Chow's (1996) cogent, scholarly analysis of the debate.

Criticisms of significance testing, at least within psychology, take two broad forms (for representative samples of these criticisms, see Gonzalez 1994; Hunter 1997; Loftus 1996; Schmidt 1996; for responses to Hunter's paper, see Abelson 1997; Estes 1997; Harris 1997; Scarr 1997; and Shrout 1997). On the one hand, it is argued that when the sample size is large enough, even trivial effects can achieve statistical significance. Thus, effects can be touted as "significant" that are in fact utterly trivial from the standpoint of either theory or practice. On the other hand, it is argued that the failure to achieve statistical significance causes investigators (and other consumers of research) to discount effects that might well be of theoretical interest or practical importance. Thus, significance tests either deliver too much, by portraying negligible effects as consequential, or too little, by insinuating that genuine effects are nonexistent.

Rather than test for statistical significance, researchers are sometimes advised to report confidence intervals instead. But confidence intervals only make sense when the goal of the research is to make a point estimate – for example, of the mean family income for African Americans, or how many people will vote Republican in the next election. In such cases, it is ridiculous to test the null hypothesis, and researchers are well advised to calculate confidence intervals as an index of the precision of their estimates. But psychologists rarely wish to estimate population parameters; rather, we generally test hypotheses about the effects of particular treatments (e.g., two levels of distraction on memory), or about the relations between particular variables (e.g., two dimensions of personality), which have been manipulated or assessed because they are theoretically or practically interesting.

Suppose, for example, that a researcher publishes a study in which psychiatric patients who receive imipramine score, on average, 5 points lower on a depression scale than those who do not, whereas the difference averages 10 points for those who receive fluoxetine. Should a researcher simply report these point estimates? Certainly not, because point estimates cannot speak for themselves. In the first place, we're not interested in the point estimates, because they would be entirely different if the researcher had used a depression test with different scaling properties. What we really want to know is: do either of these effects differ from what would be observed in a placebo group? Do any of these effects differ from zero? And do any of these effects differ from each other?

These questions can be answered by calculating the confidence intervals around each mean, and then determining the extent to which these intervals overlap. But isn't it much easier on everyone if the researcher simply reports the results of an analysis of variance followed by planned comparisons, adopting a conventional level of statistical significance like $p < .05$ or .01? It is important to bear in mind, as Chow (1996) clearly demonstrates, that comparing confidence intervals and testing statistical significance are, for all intents and purposes, mathematically equivalent (remember the debate over analysis of variance versus multiple regression?). And significance tests give you a $p$ value to boot!

Of course, in this instance, significance testing might well indicate that neither of the drugs differs from placebo and that none of the means differ either from the others or from zero. Now suppose that a dozen more such studies are published, each yielding null results, but that a meta-analysis of the baker's dozen shows that, in fact, the effects of fluoxetine are greater than those of imipramine, which in turn are greater than those of placebo, which in turn are greater than zero. In this case, it is true that the failure of the first study to reject the null hypothesis is misleading: fluoxetine and imipramine are better than nothing. But the problem does not lie in statistical significance testing; rather, it lies in the researchers' failure to perform studies with enough power to reject the null hypothesis in the first place, the reviewers' failure to detect this flaw, the editor's willingness to accept the papers for publication, and the readers' willingness to take them seriously.

Even if the initial study had yielded significant results, of course, there might have been problems. With huge *N*s, even trivial differences can achieve statistical significance. So, investigators and consumers of research alike always have to ask themselves whether they should really care about a "statistically significant" result. How much variance is accounted for by the effect? Reporting effect sizes helps in this assessment, but in the final analysis the standards for small, medium, and large effects (Cohen 1992) are no less arbitrary (and no less context-specific) than the standards for statistical significance. In any event, it should be understood that none of these alternative techniques – statistical significance testing, comparison of confidence intervals, or meta-analysis – has any privileged status with respect to another important question: Are any of the treatment effects clinically significant (Jacobson & Christensen 1996; Jacobson & Truax 1991; Jacobson et al. 1984)? Clinical significance is sometimes assessed in terms of something like effect size, although it is not clear that the simple expedient of adopting stricter criteria for statistical significance would not yield the same conclusions. In the final analysis, however, the problem of clinical significance concerns the criteria by which treatment outcome is assessed rather than the statistical tools by which significance is documented.

I have dwelt on an example drawn from clinical research, but it should be clear that similar considerations apply to basic, theory-oriented research as well. Theories (formal or informal) generate hypotheses about the effects of certain manipulations, or the relations among certain variables, and statistical significance is often the most convenient way of testing these hypotheses. Chow (1996) does us a great service by pointing out that confidence intervals and effect sizes have little to offer when we wish to corroborate a scientific theory, where the hypotheses at stake are not at the same level of abstraction as "$H_0$ = P does not exist, $H_1$ = P does exist" – and I wish he had said more about Fisher's own role in the mistaken equation of significance testing with null hypothesis significance testing. Estes (1997) likewise reminds us that tests of statistical significance are the chief means of testing how well mathematical models or computer simulations of mental processes fit actual empirical data. Given that theory testing is the goal of science, and that formalisms such as operating computer simulations represent psychological theorizing at its best (Simon 1969), it would seem foolhardy to abandon statistical significance testing – even for those, like myself, whose theorizing never gets beyond the vague and verbal.

Significance tests are not our only means of analyzing and interpreting data, though, and we probably do rely too heavily on them. That statistical significance testing has become something



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

of a fetish is indicated by the reflexive way in which many researchers (and not just novices) report artificially precise values (e.g., $p < .0438$) ripped from their computer printouts, instead of adopting conventional (and more conservative) ranges like .05, .01, .005, and .001); by their persisting tendency to report one-tailed tests when two-tailed ones would do just fine; and by their inclination to conclude that $p < .01$ is "more significant" than $p < .05$). While I am grateful for Chow's (1966) mathematical exegesis, I wish that he had said more about these sorts of practical matters.

In the final analysis, the value of significance testing is practical, as a component of the rhetoric of science (Abelson 1995). Researchers can have their own subjective opinions about their own and others' results, but statistical significance tests are – how else to put it? – public, empirical, *tests of significance.* They constitute a principled way for researchers to claim that their experimental results are worth knowing about, and for consumers to evaluate researchers' claims. At least since the time of Neyman and Pearson (1928) and Fisher (1935), significance testing has kept the behavioral, cognitive, and social sciences from lapsing into solipsism, and they can continue to play this role, along with all the other procedures in our statistical repertoire.

## Statistical significance: A statistician's view

Helena Chmura Kraemer

*Department of Psychiatry and Behavioral Science, Stanford University, Stanford, CA 94305.* **hck@leland.stanford.edu**

**Abstract:** From a statistician's viewpoint, the concepts discussed by Chow relating to "statistical" significance bear little resemblance to the concept developed in statistics. Whether or not "statistical significance" has a place in psychological research is a decision for psychologists, not statisticians, to make, but the decision should be based on a less flawed version of what is being considered.

I generally agree with Chow's conclusion but disagree with much of his book. My objections would be allayed, however, were Chow to rename his book something like "Psychological significance" and to point out that his concepts had but a tangential relationship to statistical significance testing as developed in the field of statistics.

As Chow states, every research project begins with a substantive hypothesis. To establish its truth usually requires a convergence of evidence from many approaches, with null-hypothesis significance-testing procedures (NHSTP) but one of the many to be considered when the psychologists claim:

*Claim: If I prove, beyond reasonable doubt, that "such and so" is true, the credibility of the substantive hypothesis will increase.*

The key phrase here is "beyond reasonable doubt," encapsulated in NHST in the significance level, $\alpha$. Significance level is not a probability, conditional or otherwise. It is a number between 0 and 1 selected by the proponent as the *proposed upper limit* of the probability of any false claim that "such and so" is true. As such, it reflects (1) the scientific standards of the proponent and (2) what is acceptable to peer reviewers. What is so holy about $\alpha = 5\%$ or 1%? Nothing. Why can't it be 6% or 10% or 20%? However, $\alpha$ must be set *before* the evidence (data) is collected and analyzed and (2) peer reviewers must accept the levels as appropriate in the field of application.

Critical also is translating "such and so" into the "null hypothesis" $H_0$, which has two components: $H_0$: "such and so" is not true, and certain assumptions are true. Only if one proposes to demonstrate merely that "something nonrandom is going on" does the null hypothesis posit chance or randomness. As others have so eloquently pointed out, something nonrandom is almost always going on, and it seems a trivial exercise to redemonstrate that fact. Chow appears to believe that every $H_0$ posits randomness.

Moreover, certain design or mathematical assumptions are always incorporated into $H_0$ – assumptions with which the psychologists and statisticians (often tacitly) agree: normal distributions, equal variances, linear associations, and so forth, which play a key role in NHSTP.

*Problem 1: Chow's definitions of "significance level" and "null hypothesis" are either incomplete or imprecise.*

In NHSTP the psychologists propose a research design that is to produce certain data, and the statisticians propose that $H_0$ be rejected when a selected test statistic falls into a specified region. To show that this is a valid $\alpha$-level test, the statistician must show mathematically that *whenever* the null hypothesis is true, using this design and this proposed test, the probability of rejecting $H_0$ by the proposed rule never exceeds $\alpha$.

*Problem 2: Under the null, as well as under the non-null hypothesis, there are typically many distributions, one for each possibility. The appropriate graphic is the operating characteristic curve, a graph of the probability of rejecting $H_0$ when each such possibility is true, and not any one or two "bell-shaped" distributions.*

In any case it is uncommon that the distributions on which probabilities are based are exactly "bell-shaped" at all, but that's a quibble.

*Problem 3: The formulation of the proposed testing procedure specifically depends on the form of the alternative hypothesis.*

To take the simplest example, the difference between a proposal for a one-tailed versus two-tailed *t*-test depends strictly on the formulation of the alternative hypothesis. Generally, selecting a NHSTP sensitive to the researcher's specific claim is an essential part of the process of selecting an appropriate NHSTP.

*Problem 4: For any claim, there are many different valid NHSTPs, among which a choice must be made. Chow removes the primary basis for such a choice when he recommends against power analysis.*

Again, take the very simplest situation of a two-sample *t*-test: What is the proposed total sample size? What proportion of the total sample will be assigned to or selected from the two groups? Will the sample be stratified or matched? Will there be only an endpoint observation for each subject? If more, when and how? Is the two-sample *t*-test the best choice of test? Each of these decisions changes the NHSTP. Statisticians would base the choice on comparisons of power. How would Chow choose among them?

When researchers either reject or do not reject $H_0$ what is it that is rejected or not rejected? If one rejects $H_0$, what one logically accepts is Not-$H_0$: *Either the claim is true OR some of the assumptions are not true.* Any power calculations were done for the so-called alternative hypothesis: *The claim is true AND all the assumptions are true.* But the psychologists' claim was: *The claim is true.*

How well Not-$H_0$ or the alternative hypothesis corresponds to the claim depends on how well those assumptions correspond to reality. If crucial assumptions are not reasonably well satisfied, what does it matter whether or not results led to rejection of $H_0$? The results are likely invalid.

*Problem 6: If the NHSTP is valid, rejecting $H_0$ or not rejecting $H_0$ is a comment on the strength of the evidence to make a certain claim, not a comment on the truth or falsehood of either the null or alternative hypotheses, on the effect size, or on the future replicability or confirmability of the conclusion. Each of these interpretations is at some time indicated by Chow's presentation.*

When one validly rejects $H_0$, one in effect says, "The evidence is strong enough to risk making a claim that 'such and so' is indeed true." When the result is "non-significant," one says: "The evidence is not strong enough to risk making any claim with regard to



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

'such and so.'" The evidence may not be strong enough for a variety of reasons, all related to inadequate power (See Problem 5).

*Problem 7: The essence of a research paper is the evidence presented, not the conclusion the authors draw. Such evidence must include descriptive statistics, which would inevitably include or indicate effect sizes.*

Chow's opposition to effect sizes is inexplicable to me, since it amounts to saying "I have strong evidence supporting my claim, but I won't show you what it is!"

*Problem 8: The operating characteristic curve of a NHST has nothing to do with the receiver operating characteristic (ROC) in signal detection theory application.*

In signal detection theoretic application, there are a series of independent trials, on each of which one observes ($T_i$, $S_i$), where $T_i$ corresponds to what in the NHST is the test outcome, and $S_i$ indicates whether on that particular trial, the signal was delivered or not. The crucial fact is that the signal is present on some trials and absent on others.

In NHST the "signal" is presumably the truth or falsehood of $H_0$. But if one considers multiple independent tests (trials) of $H_0$, $H_0$ is always true, or always false, not true on some trials, false on others (even in the Bayesian approach). Without trial to trial variability in the truth/falsehood of $H_0$, there is no ROC. Perhaps Chow is confused by the fact that terms such as "operating characteristic," "false positive," and "false negative" are used in both contexts.

In the end, whether or not to use NHST in psychological applications is the choice of psychologists, not of statisticians. As a statistician, my concern is only that if NHSTs are to be used (or any other statistical approach), they be used correctly and to full advantage. Those who find NHST incomprehensible or unconvincing, or who feel there are better alternatives, should not use NHST in their own research, should not as members of peer reviewer groups, reviewers or editors of papers, or readers of research reports, accept results based solely on NHST. If there are enough such psychologists, NHST would quickly disappear from psychology without any need for a formal "Ban the *p*-value!" movement. That would neither please nor bother me.

What I find troublesome is that psychologists like Chow create a version of NHST ill-corresponding to that in the statistical literature and advocate its use, for they will likely misuse or abuse the approach. But the fact is that these same psychologists will likely misuse or abuse statistical methods *regardless* of what statistical approaches were to replace NHST. If banning is the only solution, "Ban the *p*-value!" may not be enough. We would have to ban use of all statistical methods in psychology. Surely the situation is not as bad as that?

## The Ego has landed! The .05 level of statistical significance is soft (Fisher) rather than hard (Neyman/Pearson)

Lester E. Krueger

*Department of Psychology, Ohio State University, Columbus, OH 43210-1222.* **krueger.2@osu.edu cog.ohio.state.edu/homepage/cfaculty/krueger.html**

**Abstract:** Chow pays lip service (but not much more!) to Type I errors and thus opts for a hard (all-or-none) .05 level of significance (Superego of Neyman/Pearson theory; Gigerenzer 1993). Most working scientists disregard Type I errors and thus utilize a soft .05 level (Ego of Fisher; Gigerenzer 1993), which lets them report gradations of significance (e.g., $p < .001$).

I agree with Chow (p. 117) that the null-hypothesis significance-test procedure (NHSTP) has a highly circumscribed yet vital role as the initial gatekeeper in scientific research. As he writes, "statistical significance means nothing more than the decision that chance influences may be ruled out as an explanation of the data with reference to a particular criterion of strictness" (p. 188). But is Chow's "criterion of strictness" (conventionally, the .05 level of significance) hard (the alpha of Neyman/Pearson theory) or soft (the *p* value of Fisher)?

For Fisherians, the .05 level is simply a convenient benchmark, with other levels being important as well, that is, there are gradations of significance. Chow (p. 39) faults this flexibility, equating it with ambiguity in the choice of the criterion. He prefers "the Neyman–Pearson recommendation that the $\alpha$ value be set before data analysis" (p. 39). However, the "criterion-choice ambiguity" that Chow attributes to the Fisherians is not really evident. Except for special cases (e.g., multiple comparisons; see Cowles 1989, pp. 171–74) and certain aberrations (e.g., the late Fisher; see Gigerenzer 1993, pp. 316–17), the significance level of .05 seems to be the one universally adopted (Cowles & Davis 1982; Hogben 1957, p. 320). Nor is ambiguity evident regarding what the gradations of significance mean. Conventional usage dictates that one term $p < .10$ "marginally significant," $p < .05$ simply (or barely) "significant," $p < .01$ "highly significant," and $p < .001$ "very highly significant" (see, e.g., Cohen 1990, p. 1309).

In the short-term Fisherian approach, a plausibility or *p* value is attached to each test statistic, that is, "the level of significance carries a meaning with respect to a single experiment" (Gigerenzer & Murray 1987, p. 26). Thus, "Fisher . . . would have preferred that the exact level of significance, say $p = .03$, be reported, not upper limits, such as $p < .05$, which look like probabilities of Type I errors but aren't" (Gigerenzer 1993, p. 329).

In the long-term Neyman/Pearson approach, all that matters is whether the outcome is significant or not; the particular *p* value is irrelevant (or worse!). *P* values that might seem quite close, such as .048 and .052, actually differ like night and day. Chow agrees that treating .048 and .052 differently is "not nonsensical" (p. 97) and that "the rigid adherence to this rule speaks ill of neither NHSTP nor its users" (p. 97). Thus, for Chow, ruling out chance influences "is an all-or-none issue" (p. 183); contrary to the Bayesians and others, "whether or not data from Study S warrant accepting the to-be-corroborated explanatory hypothesis is an all-or-none matter" (p. 187).

Chow's adherence to the hardline Neyman/Pearson decision-making approach (aptly termed Superego by Gigerenzer 1993) is also evident in his focus on a single test statistic. Because it deals with plausibility rather than errors, the softer Fisherian approach (termed Ego by Gigerenzer 1993) allows a multitude of test statistics. That flexibility has been exploited in medical studies on bioequivalence (Rogers et al. 1993), which utilize not only the traditional test statistic, based on zero difference between name-brand and generic versions of a drug, say, but also the test statistics for the two one-sided hypotheses representing upper and lower just-nontrivial differences. Such flexibility is taken to the extreme in the case of the (1 − alpha)% or 95% confidence interval, which encompasses (and thus identifies) all those null hypotheses for which the corresponding test statistic would produce a significant result at the alpha or .05 level.

Chow (p. 23) depicts Neyman and Pearson as Bayesian. However, Neyman and Pearson had only flirted with the Bayesian position (Oakes 1986, p. 110). In fact, Gigerenzer (1993) has argued that "Neyman and Pearson took the frequentist position more seriously than Fisher" (p. 316), citing "Fisher's quasi-Bayesian view that the exact level of significance somehow measures the confidence we should have that the null hypothesis is false" (p. 318). Likewise, Hogben (1957) contrasted the Backward Look or retrospective orientation of Fisher and the Bayesians, who relied on the sample outcome to infer the plausibility or likelihood of a hypothesis, with the Forward Look of Neyman/Pearson theory, which focused on predicting the proportion of correct decisions in future tests of a hypothesis.

Chow writes that "the choice of the $\alpha$ level is arbitrary" (p. 5),



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

owing to the lack of a priori and theoretical justification. In Neyman/Pearson theory, however, alpha is anything but arbitrary, because it represents the conditional probability of Type I errors. Alpha has bite (i.e., real consequences!) in Neyman/Pearson theory. Type I errors have no role for the Fisherians: "the significance level connotes [for Fisher] no probability of erroneous decisions due to rejecting or accepting the null hypothesis" (Inman 1994, p. 8). (In his Table 2.3, p. 21, Chow errs in listing Type I errors and alpha under Fisher as well as Neyman/Pearson.)

The extra bite of Neyman/Pearson theory comes at the price of strong assumptions (e.g., repeated random sampling; replications; see Hogben 1957), which Fisher was not shy about questioning. Chow does not make clear whether he accepts those assumptions, nor does he exhibit all that much concern about Type I errors, whose level, alpha, he terms "arbitrary." When he states that "statistical significance means simply that chance influences may be discounted" (p. 65), and that we ought to replace the term "significance" with "not chance," owing to "the unwarranted connotative meanings of the word 'significance'" (p. 128), Chow seems to be backsliding toward the simpler, but weaker, Fisherian position.

Chow very aptly extolls the many fine qualities of the theory-corroboration experiment as compared with the utilitarian one. Fisher did the same thing in railing against the Neyman/Pearson decision-making position, which he wrote was better suited for the factory floor than the scientific laboratory (Oakes 1986, p. 125). Chow might have credited (or criticized!) Fisher for that.

For Chow, NHSTP is concerned solely with ruling out chance. This is a bit odd, given that a nonsignificant result may actually have a logically firmer or more valid effect on the higher-order, nonstatistical (i.e., substantive, research, experimental) hypotheses than a significant result (Chow, pp. 71–72). If a well-designed, highly sensitive study yields a resoundingly nonsignificant test statistic, should the null hypothesis be simply, but firmly, retained (i.e., "safely ignored," Chow, p. 119), or should it be accepted? Chow recognizes the dilemma (p. 63), but is loath to sanction accepting the null hypothesis, even though that may in many cases be just as meaningful as rejecting the null hypothesis (Frick 1995; Greenwald 1975; Rogers et al. 1993). Chow (p. 56) does help pave the way for the acceptance of the null hypothesis, though, by quite effectively dismissing the view that the point-null hypothesis, as a categorical proposition, is necessarily always false.

Chow does a masterly job in providing an overview and in setting the stage for his defense of NHSTP. By making numerous valuable distinctions, such as that between statistical and nonstatistical (e.g., effect size) matters, he dispels many conceptual confusions. Alas, some key distinctions (hard vs. soft .05 criterion; accepting vs. rejecting the null hypothesis) were not drawn as sharply. Also, Chow errs in attributing to signal detection theory a reliance on a threshold (N. 2, p. 142), and in citing the central limit theorem (e.g., p. 35) when what he really meant was the law of large numbers.

## Logic and the foundations of statistical inference

Henry E. Kyburg, Jr.

*Department of Computer Science, University of Rochester, Rochester, NY 14627.* **kyburg@cs.rochester.edu www.cs.rochester.edu./u/kyburg**

**Abstract:** The rapprochement between methodology and statistics suggested by Chow's book is a much needed one. His examples suggest that the situation is even worse in psychology than in some other disciplines. It is suggested that both historical accuracy and attention to recent work on the foundations of statistics would be beneficial in achieving the goals that Chow seeks.

Chow has made a commendable effort to bring together the two worlds of statistical practise (in psychology) and scientific methodology. Unfortunately, the results of his efforts provide powerful evidence of the need for precisely the kind of rapprochement that is his worthy goal. There is more information about and understanding of statistics and methodology out there than one would suppose from what Chow says. He is quite right, though, in saying that it is not easy to get at, and that many people make no serious effort to get at it. His demonstration that some disputes about statistical evidence in the social sciences are uninformed is persuasive.

Consider the methodological insights of John Stuart Mill, for example, Mill's methods (1843) were for many years standard diet in courses in general logic. This has not been the case for roughly 50 or 60 years. The reasons are simple: Mill's methods are simplistic, and philosophical writers after Nagel (1961) have thought it more important to emphasize the richness and complexity of good empirical methodology rather than the "catch phrases" of Mill. One may indeed argue that the pendulum has swung too far, and that simplicity is a virtue, so long as its shortcomings are also realized. It may well be that Mill's methods will once again play a role in beginning courses in general logic. But a bit of Millian salt is not going to turn statistical hash into roast duckling.

Logic, however, is not Chow's strong point. Page 30 contains the following three statements: (1) "The truth value of a conditional proposition is determined by the relationship between its antecedent and consequent." (2) "a conditional proposition is false only if its antecedent is true while its consequent is false." (3) "The general point is that the truth value of a conditional proposition is not determined solely by the truth value, its consequent or its antecedent." And also on page 132: "It is incorrect to treat $H_1$ as synonymous with 'not-$H_0$', although $H_1$ and $H_0$ are mutually exclusive and exhaustive alternatives" (Chow 1996).

In Chapters 3, 4, and 5, Chow distinguishes among several kinds of experiments and applies ideas derived from Mill's methods (agreement, difference, concomitant variation) to each. Given the real examples that he analyzes, this exercise is not useless: some worthwhile criticisms emerge from his analyses. It is these critical studies of psychological methodology that provide the best argument for the explicit study of inductive methodology and constitute one of the most valuable features of the book. Chow is well equipped with common sense, and it is well applied.

Chow argues that statistical significance testing is an important ingredient in empirical research, but that its role is severely limited. One step in the investigation of a phenomenon is to rule out the possibility that the alleged phenomenon under investigation is an artifact – that an apparent difference is merely the result of chance. Of course, this is exactly Fisher's attitude. "[T]ests of significance are used as an aid to judgment, and should not be confused with automatic acceptance tests" (Fisher 1963, first edition 1925). Also, "In general, tests of significance are based on *hypothetical* probabilities calculated from their null hypotheses. They do not generally lead to any probability statements about the real world."; and "[t]he force with which such a conclusion [that the effect is not due to chance] is supported is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution is not true." (Fisher 1956, pp. 44, 39).

While Fisher thought of statistical decision theory as "statistics for shopkeepers," Neyman, perhaps the most influential of statisticians, thought of decision theory as lying at the heart of statistical methodology. Neyman was perfectly clear about inverse probability: it makes sense when you have a statistically justified or assumed prior distribution and this is, practically speaking, almost never. The joint paper cited by Chow (Neyman & Pearson 1928) contains a vaguely tolerant passage about inverse probability. After 1933 this tolerance, particularly on the part of Neyman, begins to disappear. But it is perfectly clear that from the beginning the idea of looking for inverse probabilities of the form $P(H|D)$ did not enter Neyman's mind as a practical possibility. To characterize the Neyman/Pearson approach as being focussed on inverse probability (p. 36) is the opposite of the truth.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Probability, for Neyman, was a strictly frequentist notion. He was explicit, for example, about the fact that probability does not apply to single cases (except for probabilities of 0 and 1), and thus did not apply to hypotheses or even to "false rejection" after experimentation or observation. It is a mistake (from the classical point of view) to say that the probability is .95 that a frequency lies in a specific .95 confidence interval. Inverse probability was anathema to Neyman (1942; Neyman & Pearson 1933).

Furthermore, it is often difficult to know what the author is trying to say about probability. For example, speaking of effect size and statistical power in the introduction to the chapter on Bayesianism, Chow claims that "a motivation underlying power analysis is to adopt a direct means of ascertaining the probability of the truth of the substantive hypotheses," and, in the very next sentence, claims that "the meaning of probability (viz., a relative frequency)" is not challenged. But it clearly makes no sense whatever to talk of the relative frequency of truth of the substantive hypothesis; a hypothesis is either true or false.

Bayesianism does not fare much better. The key source for Bayesianism is Jeffreys (1939), rather than Savage (1954) or the voluminous recent literature. When Chow gives an example of Bayesianism, the three alternative hypotheses, apparently to be considered exclusive, are given a total probability of 1.5. It is hard to understand this, because the author correctly characterizes a prior subjective probability of .60 as corresponding to the willingness of the agent to pay $60 for a return of $100 if the corresponding hypothesis turns out to be correct. Is the agent just contributing $50? That the probabilities do not add up to 1.0 is excused (in a footnote) on the grounds that "some evidence may be compatible with more than one hypothesis" (Chow, 1996, p. 174).

The interesting thing about Chow's book is not that he gets his history wrong, since that is not difficult to correct, or that he is unclear about the possibilities of interpreting probability; it is the attempt to tie together conventional statistical wisdom (or foolishness) and issues of methodology. Mill's methods have not been fashionable for many decades. Philosophers have mostly focussed on probability, if they have been concerned with quantifying scientific inference, or with "paradigms" and "revolutions" if they have been concerned with the qualitative historical shifts in scientific knowledge. The law of the pendulum suggests that we may well have gone too far in playing down methodological issues, and Chow provides us with both a welcome corrective for our neglect and a new slant on conventional statistical methods.

Nevertheless, it could be argued that his work suffers from its own pendulum excess. It could be said that this defect stems from Chow's own lack of understanding of the issues underlying the interpretation of probability and the foundations of statistical inference. The interested reader might want to take a look at Godambe and Sprott (1971), who present the results of a lively conference on the foundations of statistical inference, or Kyburg (1974), which contains, in addition to a particular thesis, an analysis of the main alternatives in the foundations of statistical inference.

## A defense of statistical power analysis

Brian R. Lashley

*Department of Psychology, University of Connecticut, Storrs, CT 06269.*
**lashley@NECA.com**

**Abstract:** Chow attacks statistical power analysis on theoretical grounds. I argue that if significance testing is defensible, so is power analysis. A number of Chow's criticisms seem to suggest that power analysts are confused about certain fundamental issues. I claim that few power analysts make the mistakes Chow describes. Finally, I address Chow's claim that power analysis is irrelevant to NHSTP because it deals with a different issue.

Although statistical power analysis is an integral component of methodological development in empirical research, Chow debates its validity. That is a puzzling stance for someone who defends statistical significance testing. I find it puzzling because researchers should be aware of the Type I and Type II error rates of the research designs and statistical techniques they use. These error rates can be assessed by certain types of power analyses.

An important assumption in the present discussion is that in disdaining power analysis Chow also disdains computer simulations which provide quantitative assessments of research designs and data analysis techniques via computer simulation.

To ensure the accuracy of their conclusions, researchers should use significance testing techniques that minimize both kinds of error. Some statistical significance testing procedures have both lower Type I error and lower Type II error rates than others (e.g., Lashley & Bond 1997; Lashley & Kenny 1997). This is important information for researchers who need to know which effects in their research are statistically significant and which ones are not. Monte Carlo Power analyses provide quantitative assessments of statistical significance tests and allow researchers to choose the most effective test.

The thesis of the present critique is that Chow criticizes power analysis inappropriately. Several of Chow's criticisms seem to suggest that power analysts are either unaware of certain issues, or unwilling to acknowledge them.

For starters, Chow (p. 119) says that the probability $(1 - \beta)$ is an unknown value. This is technically true as long as the population value of the parameter of interest is unknown. However, quantitative researchers often use computer simulations in which they conceive a population and sample from it repeatedly. These investigators know the population values of the parameters they investigate. Typically, these parameters include the mean and standard deviation, which are used to compute $(1 - \beta)$. In such cases, the power is known.

On the same page, Chow says that the conditional nature of statistical power is not acknowledged in assertions about statistical power. Quantitative writers may not always remind readers about the fact that statistical power is based on the condition that the null hypothesis is false, but some authors of computer simulation research (e.g., Lashley & Bond 1997) refer to power as the proportion of false null hypotheses rejected. That wording alone should suffice as evidence that the authors are aware of the conditional nature of power. In other cases, authors might safely assume that the reader is aware that only when the null hypothesis is false can a significance test have statistical power.

Chow (p. 124) says that "the statistical alternative hypothesis, $H_1$, is identified with the hypothesis about Phenomenon P in power analysis . . . it is asserted that if Phenomenon P exists, its effects must be detectable." In other words, power analysts treat statistical power as the probability that an empirical investigation will provide support for the substantive hypothesis. I cannot speak for all power analysts, but I think power analysts realize the difference between quantitative and substantive information. I think all power analysts would realize that the probability they are assessing is that of obtaining a test statistic greater than or equal to a critical value, not that of detecting a behavioral or psychological phenomenon.

On a similar note, Chow (p. 128) writes that "statistical power may be treated as an index of the a posteriori confidence of the researcher about the statistical decision, but not as an index of the a priori probability that $H_1$ is true." I know of no one who thinks that statistical power is an indication of the probability that $H_1$ is true. In certain situations, one may conduct a power analysis to assess the credibility of a statistical decision. But it should be clear to everyone that power is the probability of obtaining a significant test statistic given certain population values. For example, it is the probability of finding a significant *t* ratio given that the samples will come from populations with means $=\mu_1$ and $=\mu_2$ and variances $=\sigma_1$ and $=\sigma_2$.

Continuing that argument, Chow (p. 128) says that "[Quote 6-1] suggests that statistical significance is reached by virtue of the numerical index statistical power." Several of the words in [Quote



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

6-1] deserve a closer look. It says that "the sample size necessary to detect *this negligible effect* with .95 probability can be determined" (my emphasis). Cohen's (1990, p. 1309) original text was apparently dealing with a very small effect and $(1 - \beta) = .95$. Thus, the corresponding sample size must have been quite large. Chow seems to be criticizing Cohen's quotation for implying that a power analysis is being used to determine the sample size necessary to "yield" a desired result (i.e., a statistically significant result) as if the result were the product of a factory. However, the end of [Quote 6-1] says that "if the research is carried out using that sample size, and the result is *not* significant . . . the conclusion is justified that no nontrivial effect exists." Although I agree in part with Chow's criticism that Cohen seems to be attributing to power analysis an inferential capability that it does not have, I must ask: If one uses an unusually large sample and finds no significant effect, can one not feel safe in concluding that the population effect size is trivial or zero? I believe that this rhetorical question is the point being made by Cohen. Of course, one can use a large sample size and feel a certain amount of confidence without using power analysis; but I suspect that Cohen's quote may have been taken out of context and used as an example of misguided thinking by power analysts.

My final defense of power analysis against Chow's criticism pertains to Figure 6.2. To simplify my discussion, I would like to refer to panels A and B as "stage 1" and to panels C and D as "stage 2." Chow points out correctly that whereas power analysis is involved with stage 1, NHSTP is involved with stage 2; and the problem is that stage 1 has two distributions, stage 2 has a single distribution. Thus, power analysis "misrepresents" significance testing by implying that it is based on two separate distributions. I wish to point out that both graphically and conceptually, power analysis is meant to deal with a different stage in the inferential process than NHSTP. It attempts to predict, using the information in stage 1, the distribution of stage 2 and the outcome of the significance test. Stage 1 is in fact part of NHSTP; stage 2 cannot be derived without it. Thus, power analysis does pertain to a stage of NHSTP – the stage at which distributions for each of the competing hypotheses have been derived.

## The critics rebutted: A Pyrrhic victory

Stephan Lewandowsky and Murray Maybery

*Department of Psychology, University of Western Australia, Nedlands, W.A. 6907, Australia.* **lewan@psy.uwa.edu.au**
**www.psy.uwa.edu.au/user/lewan/**

**Abstract:** We take up two issues discussed by Chow: the claim by critics of hypothesis testing that the null hypothesis ($H_0$) is always false, and the claim that reporting effect sizes is more appropriate than relying on statistical significance. Concerning the former, we agree with Chow's sentiment despite noting serious shortcomings in his discussion. Concerning the latter, we agree with Chow that effect size need not translate into scientific relevance, and furthermore reiterate that with small samples effect size measures cannot substitute for significance.

Chow's response to the many recent criticisms of hypothesis testing is certainly timely. It is equally certain that it will fail to convince many of the critics; not because the critics are necessarily right, but because Chow repeatedly overshoots the mark in his rebuttal. For example, the commendation of those who rigidly treat $p = .052$ differently from $p = .048$ (p. 97), or some of his scathing criticisms of meta-analysis cannot be supported. We focus on two issues here; the critics' claim that the null hypothesis ($H_0$) is always false, and the claim that measuring effect size is more appropriate than hypothesis testing.

***$H_0$ is always false.*** Recall the criticism: "The null hypothesis, phrased as a statement of no treatment effect, is seldom if ever true in psychological research. . . . These deviations from a point null hypothesis, however trivial, will become significant with sufficient power." (Hammond 1996, p. 105). Chow deals with the criticism first by claiming that $H_0$ is a conditional proposition whose truth value does not depend on its consequent and antecedent being true (p. 30). At the same time, he presents the null hypothesis as both the consequent of one conditional proposition (p. 32) and the antecedent of another (p. 32). A further statement purports that the null hypothesis is not a "proposition about the to-be-studied phenomenon. Instead, it is about the data-collection procedure" (p. 32). A later clarification (p. 51) invokes chance variation, the existence of two populations "defined in terms of data collection procedure," and the fact that the "two populations . . . are identical in all aspects but one" for "a proper interpretation of $H_0$." Finally, Chow suggests that "it is not inconceivable that $H_0$ . . . is true because it may be possible to control properly the data-collection procedure in the experimental approach" (p. 56). Now, it is possible that all this adds up to a coherent, albeit complex view that the null hypothesis *can* be true, contrary to the critics' claim. Alas, Chow's unnecessarily convoluted treatment, earlier versions of which have attracted much criticism (Anthony & Mullen 1991; Bernieri 1991; Harris 1991; Rozeboom 1991), has obscured rather than further clarified this issue, thus perhaps lending inadvertent support to the critics of hypothesis testing. Lost in the mist is a much simpler and more compelling argument about the status of $H_0$ that, we believe, can be stated as follows:

1. For true experiments (as defined on p. 12), the null hypothesis of no difference is tantamount to postulating that the experimental manipulation was ineffective.

2. Therefore, anyone who claims that the null hypothesis is always (or nearly always) false would be committed to accepting that, given a sufficiently powerful experiment, any (or nearly any) experimental manipulation would yield a significant outcome.

3. Not only is this conclusion demonstrably false because even an experiment with 25,000 subjects may fail to reject $H_0$ (Oakes 1975, cited in Leventhal 1994), but

4. it is also questionable because the conclusion would entail the a priori acceptance of the fact that knocking on wood will prevent the occurrence of dreaded events, that black cats crossing the road are better predictors of future mishaps than white cats, or of any other superstition that can be put to an experimental test with sufficiently large sample sizes.

***Effect size over statistical significance.*** Hypothesis testing has been criticized on the grounds that a numerically large effect may escape detection because sample size is too small or, conversely, that a trivially small effect may turn out to be significant because sample size is sufficiently large. Some critics have therefore advocated the use of effect size in preference to relying on the significance criterion of hypothesis testing. On this issue, we disagree with Chow's radical slant towards (nearly) dismissing the utility of effect size altogether, but we share his emphasis on decoupling the statistical size of an effect from its scientific relevance. His case that no single effect-size criterion exists to evaluate research is convincing. In particular, different research domains and different designs would require different criteria (e.g., in a study of adult height, a 0.1 mm effect in a repeated-measures design is probably more newsworthy than a 10 cm effect in a between-subjects design). Moreover, even if it were possible to select a criterion effect size specific to a domain and design, difficulties would nonetheless arise because research often proceeds by unpacking an effect into its components. For example, one of us (M. M.) has recently identified a priming effect with spatial relations – accessing information pertinent to one direction (e.g., left) primes information pertinent to the opposite direction (right). Efforts are now being directed at dividing this effect into (a) a general semantic priming component, (b) a component concerning the alignment of a common spatial axis, and (c) a component reflecting the obligatory tagging of opposite locations or regions in space. Would our understanding in this field be advanced by imposing an effect-size criterion that may identify the general effect as newsworthy, but the lesser component effects as not newsworthy? We propose that it would be more appropriate to



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

note that all component effects, established by a criterion of statistical significance, need to be accounted for by a viable cognitive model.

Going beyond Chow's discussion, one critical aspect of effect size merits reiteration because it is often lost in debate, namely, the vulnerability of effect size measures to small sample sizes. The much-heralded property of effect size, that it is not affected by sample size, only applies to its expected (mean) value: as with any other sample statistic, the variability of effect size measures necessarily increases as sample size decreases. This can be readily demonstrated by Monte Carlo simulation (e.g., Carroll & Nordholm 1975). In consequence, although a statistically *non*significant effect can "look big," apparent size is of little informational value owing to the large uncertainty associated with the sample statistic.

So we endorse hypothesis testing but our reasoning differs from Chow's: on the one hand, it is meaningful to test null hypotheses because some are surely true; on the other hand, it would be meaningless to adopt an evaluative criterion based exclusively on effect size because, first, effect size is a sample statistic subject to error like any other, and second, no single criterion would seem viable even if set within a particular research domain.

## When the coefficient hits the clinic: Effect size and the size of the effect

Brendan Maher

*Department of Psychology, Harvard University, Cambridge, MA 02138.*
**bam@wjh.harvard.edu**

**Abstract:** The usefulness of effect-size differs in utilitarian experiments from its use in theory corroborations. Chow introduces the question of the relationship of effect-size to practical validity and the role of the assessment of "importance" in this. This review develops this question and suggests the actuarial table as a replacement for effect-size in practical decision-making.

The null-hypothesis significance-test procedure (NHSTP) and the basic concept of statistical significance has come under increasing criticism in recent years, mainly because it is seen as leading us to reject findings that are valid. Chow has systematically clarified the issues and reminds us of the basic reasons why our present procedures were developed. The focus of this review is the real world meaning of effect size and its usefulness in decision making. Before turning to this, we might reflect on the larger context in which these debates occur.

Psychologists (like other professionals) have a vested interest in demonstrating their claim to knowledge and technical competence in the field of behavior and psychological welfare of others. Statistically significant results improve publication possibilities, which helps academic career advancement; they appear to prove the efficacy of psychological services and so justify the fees that are charged for them, and so on. There is no evidence that psychologists are more or less altruistic than members of other professions, and we may assume that when faced with alternative ways of analyzing and presenting data there will be an understandable bias toward the method that is likely to be most favorable to the investigator's interests. Any proposal to adopt statistical procedures that increase the number of apparent discoveries, theory corroborations, patients improved, or other indices of successful application, must therefore be regarded with some caution.

Chapter 5, "Effect size and related issues" speaks to one of the most controversial of current statistical procedures, that is, the calculation of an "effect size." In this review I comment specifically on Chow's discussion of the utilitarian experiment in contradistinction to the theory-corroboration experiment. The utilitarian experiment investigates practical problems, mainly the effects of interventions designed to change a client's behavior. Here the magnitude of an effect size becomes informative only when translated into its practical validity, that is, the real-life consequences of the intervention.

Chow (p. 96) presents the well-known example cited in Rosnow and Rosenthal (1989) in which numerically small differences were found in the rate of myocardial infarction (MI) in a group taking aspirin and another taking placebo. Although the effect size of the difference was only .034, this translated into a survival rate in the aspirin-treated group that was 3.4% higher than that in the control group. Rosnow and Rosenthal comment that this has implications that are "far from unimpressive." Given that MI is frequently fatal, the comment is justified. The cost of aspirin is small and the possible benefit is life rather than death. Chow recognizes this, but points out that the criteria for what is "impressive" are not defined, and that it is not at all clear how such judgments are to be made objectively. He is right in this, but does not address the core issue, which is are there alternative statistical procedures that might help the client make practical decisions? Chow has given us the basis for developing possible answers; it would have been interesting to see him pursue it further.

Rosnow and Rosenthal's example is a double-edged sword. Small effect sizes may have large practical implications, and large effect sizes may have trivial practical implications. The relevance of the magnitude of an effect size cannot be assessed by looking at it. It is unfortunate that the word "effect" is used at all in labeling a computation that tells us nothing about the concrete effects produced by an intervention. Cost-benefit analysis offers an improved basis for the estimate, but the computation of both sides of the equation is plagued by the difficulties attendant on the creation of a quantitative measure of the "utility" of psychological costs and benefits. This is an assessment that the clients can make for themselves and statisticians cannot.

In the light of all this, perhaps we should consider the wisdom of translating any effect size in a utilitarian experiment into an actuarial table of real-world consequences. If, for example, we find that in a comparison of psychotherapy versus placebo in treating depression an effect size of 1.0 actually represents an average reduction of two points on a self-report depression scale, we need to know a great deal more about what that means in terms of genuinely important outcomes, such as changes in insomnia, appetite, days in or out of hospital, and so on. If these are negligible, the change in the scale score is basically irrelevant, no matter how large the computed effect size.

We also need to know how many individual patients in each group showed improvement in the direction implied by the difference between group means, and what was the distribution of given magnitudes of gain, including zero gain or losses. We need to know whether and how reliably these differences in individual outcome are related to such factors as current age, gender, education, age of onset, length of illness, and so forth. If they are to make rational decisions, people who are deciding to invest time and money in an intervention need such a table, not a single-quantity effect size.

Individuals also need to know the probability that an actuarial table, based on data already accumulated, is likely to continue to apply to new clients coming from the same general population. In the utilitarian experiment findings that fail to reject the null hypothesis means that any actual difference between the groups cannot provide a reliable basis for choosing one alternative over the other. Here the appropriate significance test provides the best estimate of the stability of the table. It cannot determine whether it is better for an individual to make a choice on an unreliable basis than on no basis at all. In practical terms, this is equivalent to saying that the alpha level for personal "significance" is set by the client but knowledge of the *p* values is a crucial element in the decision and there is no better substitute available.

All of these implications are inherent in Chow's remarkably clear exposition. Readers will be able to develop them quite readily.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## Some problems with Chow's problems with power

Deborah G. Mayo

*Department of Philosophy, Virginia Polytechnic Institute, Blacksburg, VA 24061.* **mayod@vt.edu.**

**Abstract:** Chow correctly pinpoints several confusions in the criticisms of statistical hypothesis testing but his book is considerably weakened by its own confusions about concepts of testing (perhaps owing to an often very confusing literature). My focus is on his critique of power analysis (Ch. 6). Having denied that NHSTP considers alternative statistical hypotheses, and having been misled by a quotation from Cohen, Chow finds power analysis conceptually suspect.

Standard statistical testing (null-hypothesis significance tests and Neyman–Pearson methods), while widely used in diverse sciences, have been the subject of considerable criticism and controversy among philosophers (especially Bayesians) and others. There is no doubt that these methods are in need of defense from someone who can clarify the complex issues and disentangle the disagreements and confusions involved, especially in the psychological literature.

Chow's book (1996) raises a number of important and correct points against critics: many of the criticisms, Chow rightly notes, are based on confusing statistical inference with substantive inductive and scientific inference, on poorly designed or misinterpreted tests, and on a misplaced desire for a probability that these methods are not designed to supply: a posterior probability of a hypothesis. Chow is at his best when emphasizing what critics tend to overlook: that statistical tests concern hypotheses about a sampling distribution (e.g., of a test statistic), that such hypotheses must be distinguished from what he calls experimental and research hypotheses, and that the result of statistical testing must be distinguished from corroborating a scientific hypothesis, although progress toward theory corroboration may be afforded by combining sufficiently numerous and probative statistical tests. Correct too is Chow's distinction between the context of theory corroboration in science and the Bayesian context. These ideas warrant further attention. But first Chow should rethink the version of standard statistical testing theory worthy of being defended.

***Null-hypothesis statistical-testing procedure, NHSTP.*** NHSTP is the hybrid of Fisherian and Neyman-Pearson (NP) tests that Chow imagines practitioners use, and it is the one he is defending. Although the essential contribution of NP theory was the introduction of alternatives to the null hypothesis and the corresponding power function – Chow discards this from NHSTP (e.g., "Fisher was correct not to consider Type II error because it plays no role in NHSTP," p. 43). What Chow keeps from NP theory is the conception of a test as a decision procedure to reject or accept a null hypothesis (of chance) according to whether data reach a preset critical value of a test statistic. To some, it might seem as if Chow's NHSTP ejects the best parts of each approach.[1] Given this view of tests, it is not surprising that Chow finds the notion of *power* problematic.

***Chow's critique of power analysis.*** Chow faults the field of power analysis for two reasons: (1) "The a priori probability of obtaining statistical significance is said [by power analysts] to be given by the power of the test" (p. 131) but this is false; and (2) NHSTP is restricted to only the null hypothesis $H_0$, but power analysis depends upon alternative statistical hypotheses. Chow's charges against power analysts accordingly boil down to arguing first, that a power calculation does not give an unconditional probability (of a statistically significant result) and second, that the calculation of power is impossible for an account that excludes alternative statistical hypothesis. Both charges are correct, yet they are not damaging to a correct use and interpretation of power in standard Neyman-Pearson testing. I will take these up in turn:

1. Chow is misled throughout by an unfortunate quote from Cohen (1987, p. 1) that "The power of a statistical test is the probability that it will *yield* statistically significant results" (Chow's emphasis; quote 6-2 on p. 120). Thus, Chow charges that in power analysis, "the power of the statistical test is treated as the probability of $H_1$ being true (by virtue of the fact that it represents, to power analysts, the probability of obtaining statistical significance)" (p. 124). Chow seems also to be confusing (or alleging that the power analyst confuses) the probability a test correctly rejects $H_0$ and accepts $H_1$, with the probability that $H_1$ is true. Anyone who treats power as either the probability of a significant result or the probability of $H_1$ is justly castigated by Chow – but does anyone commit such egregious errors?

2. According to Chow (p. 132), the probability of a Type II error should be defined as (i) p(Accept Chance|not-$H_0$) while in power analysis it is defined as (ii) p(Accept Chance|$H_1$). But (i) is not defined in NP theory unless not-$H_0$ is a point hypothesis, and since Chow's NHSTP excludes such alternatives it is not surprising Chow concludes that "it is impossible to represent statistical power graphically in the sense envisaged in power analysis without misrepresenting NHSTP" (p. 137). But if so, then it is NHSTP that forces a nonstandard interpretation of the probability of a Type II error. For Chow, (i) refers to the probability that the test Accepts $H_0$ when some (substantive) nonchance factor is really responsible – a calculation which he admits a statistical method cannot supply. We are not told why such a nonstandard notion should be preferred to the standard statistical one, nor why we should oust alternative statistical hypotheses from our methodology of testing.

Moreover, since power analysts are working within NP testing theory where it is entirely appropriate to consider the power of a test to reject $H_0$ for various different point alternatives – that is, power curves – Chow's criticism misses its target.

By restricting himself to the single hypothesis of the Fisherian test, Chow's defense of NHSTP is forced to accept an overly limited role for statistical analysis: "NHSTP answers the question as to whether or not there is an effect. However, it is not informative about the magnitude of the effect." (p. 7). In fact, considering a test's ability to detect alternatives can provide information about the magnitude of the effect that is or is not indicated by a statistical result. For example, if a test had a high [low] power to detect an effect of a magnitude specified in $H_1$ then failure to reject the null hypothesis (of 0 effect) would be a good [poor] indication that the magnitude of the effect was less than $H_1$ asserts. Thus, power considerations offer a good way to scrutinize the meaning of statistical results,[2] and Chow has given us no reason to abandon them.

Chow overlooks the fact that, although his one-sided tests may be articulated with reference to the null hypothesis alone, their justification as good or best tests had to be derived by considering alternative statistical hypotheses (e.g., as in deriving uniformly most powerful tests). Chow's NHSTP tests are cut off from their logical foundation in NP theory.

***An error regarding the goal of NP statistics.*** On pp. 21 (Table 2.3), 23, 42, and elsewhere, Chow asserts – quite erroneously – that the probability of interest in NP statistics is "the inverse probability, p(H|D)" (p. 21) and declares that "the Neyman–Pearson preference for the inverse probability" is not consistent with the mathematical foundation of NHSTP (p. 42). Indeed, such an NP preference would also be inconsistent with NP theory (which was designed for cases where no such inverse probability was even meaningful)! Chow's mistake, if uncorrected, will only supply further grist for the Bayesian mills which regularly accuse NP theory of unsoundness (alleging it to be interested in posterior probabilities while supplying only error probabilities).

NOTES

**1.** My own preference would be to reverse what gets ejected: to retain the NP use of alternative hypotheses while replacing the decision-theoretic interpretation of tests with an inferential one. Power calculations are needed to specify tests, but to infer what is and is not indicated by a specific result may be achieved by calculating error probabilities using that result (rather than a preset cut-off). Suppose, for example, that the *p*-value observed is not small and so $H_0$ is "accepted." To interpret this one might calculate, not the usual power of the test against an alternative $H_1$, but



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

rather, the probability of a result more significant statistically than the one obtained given $H_1$. How this relates to using tests inferentially is discussed in Mayo 1996.

**2.** An analysis sensitive to the specific value of the result is also possible (see N. 1).

## Significance tests cannot be justified in theory-corroboration experiments

Marks R. Nester

*Queensland Forestry Research Institute, MS 483, Fraser Road, Gympie, 4570, Australia.* **nesterm@qfril.se2.dpi.qld.gov.au**

**Abstract:** Chow's one-tailed null-hypothesis significance-test procedure, with its rationale based on the elimination of chance influences, is not appropriate for theory-corroboration experiments. Estimated effect sizes and their associated standard errors or confidence limits will always suffice.

I have read 160 pages of Chow's book, including the preface but excluding Chapter 7. On those pages I have annotated 194 comments, queries, and criticisms and so this brief commentary will outline only some of my thoughts.

1. Yes, Chow's book does contain a few factual errors and careless remarks, though I regard them as irritating rather than disastrous for Chow's thesis. For example, contrary to p. 8, statistical significance is considered by many to be extremely likely, rather than "assured," if a large enough sample is used. If the first sentence of section 2.7.2, p. 35, is taken literally, then it is nonsense. If I am correctly guessing Chow's intention here then the notion of approximation should be associated with the central limit theorem. The complements of the research and experimental hypotheses presented in Table 3.1 on p. 47 are not true complements. On p. 181 Chow insists that asking about the magnitude of an effect is a "pragmatic consideration" rather than a "statistical concern." This is an extraordinary claim when one considers how much effort statisticians have devoted to obtaining estimates of parameters.

2. On p. 11, Chow reports that "submission to statistical authority" may be a contributing factor in "the continual reliance on NHSTP [null-hypothesis significance-test procedure]." I am compelled to dismiss this idea by quoting one of the early statistical authorities, Yates (1951, p. 32), who not only condemned some of his fellow statisticians because of their obsession with significance tests but also stated: "scientific research workers . . . pay undue attention to the results of the tests of significance . . . and too little to the estimates of the magnitude of the effects they are investigating." The argument that utilitarian experiments were foremost in Yates's mind will be rendered obsolete below.

3. Table 6.1 and the associated discussion on pp. 140–41 represent a flawed approach to testing a theory (substantive hypothesis). Chow states that NHSTP is used at each stage of a falsification process. It seems to me that the more stages there are then the more likely it is that one of the NHSTPs will lead to the rejection of an experimental hypothesis and indirectly to the rejection of the theory. This is fine for those who believe all theories are wrong. On the other hand these people do not need to test any theories. In practice, it may be necessary to accept a theory tentatively, irrespective of the outcome of any NHSTP. Chow might agree with this! Chow expects a falsification process to be based on the NHSTP; he demands that alpha levels be strictly enforced (e.g., p. 97); but states (on p. 92) that "research conclusions . . . are not accepted or rejected on the *sole* basis of statistical significance."

4. I support some of Chow's condemnations of the occasional sloppy thinking used by critics of NHSTP. It is true that effect size per se is not a measure of evidential support for a statistical hypothesis, let alone for a substantive hypothesis. However, I do believe that the effect size and its corresponding standard error (or confidence limits) together may provide some measure of support for a statistical hypothesis – thus the larger the effect size and the smaller its standard error, the stronger the evidence.

5. One fact Chow missed is that there is a utilitarian experiment inside every theory-corroboration experiment. With regard to the phenomenon of linguistic competence described in Table 3.1, p. 47, surely it would be interesting to know the number of words which can be recalled after negative and kernel sentences. This may, for instance, have application in communications between pilots and control towers. No harm can come from reporting these magnitudes and their associated standard errors or confidence limits. It is even possible that one day there will be psychological theories which will yield quantitative predictions, and then the publishing of estimated magnitudes will have been justified.

6. There are several different experimental hypotheses which can be associated with a particular theory. If the true magnitudes of the effects associated with each experimental hypothesis are known or accurately estimated, then it is possible for each of us to make a subjective assessment of the importance of the theory. This is different from evidential support. Thus, with regard to linguistic competence and Chow's particular experimental hypothesis, suppose that the intrinsic utilitarian experiment reveals that 4.1 words can be remembered, on average, after a kernel sentence and only 3.9 words after a negative sentence. I personally would not care which theory explained linguistic competence because I regard the number of words which can be remembered as being quite similar. Furthermore, I would allocate low priority to the search for such a theory.

7. The effects estimated in the intrinsic utilitarian experiment can still be used in Chow's three embedding syllogisms, Table 4.2, p. 70, with no loss in "objectivity" by substituting 95% confidence limits, say, for the 5% NHSTP. This renders the NHSTP unnecessary.

8. Consider the linguistic competence phenomena and develop an experimental hypothesis from some hypothesized theory. In Chow's case, the experimental hypothesis deals with numbers of words which can be remembered. Now reflect that there may be many alternative theories which can explain the results of the proposed experiment. By merely considering the direction (sign) of an effect we can simultaneously confirm or reject many theories. Thus, if the experiment indicates that more words can be remembered after kernel sentences, then all theories which imply otherwise can be rejected. Similarly, if we discover that more words can be remembered after negative sentences, then the rejected theories in the first case become the accepted ones. Thus in order to discount a whole range of theories all we need to know is the direction of an effect. Careful study of Chow's one-tailed NHSTP with its rationale based on the elimination of chance influences (e.g., pp. 31–32, 36, 37, and 175) reveals that the NHSTP is not designed to estimate the direction of an effect! Why not just estimate and report the effect itself?

On a personal level I must congratulate Chow on a valiant, sincere, generally well-written and sometimes clever attempt at justifying the continued use of the NHSTP in special circumstances. On a professional level I hope no one else reads this book. I fear that the scientifically or statistically naive will be overwhelmed by Chow's confident and occasionally superficially convincing arguments.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

# Significance testing – does it need this defence?

Günther Palm

*Department of Neural Information Processing, University of Ulm, D-89069 Ulm, Germany.* **palm@neuro.informatik.uni-ulm.de**

**Abstract:** Chow's (1996) *Statistical significance* is a defence of null-hypothesis significance testing (NHSTP). The most common and straightforward use of significance testing is for the statistical corroboration of general hypotheses. In this case, criticisms of NHSTP, at least those mentioned in the book, are unfounded or misdirected. This point is driven home by the author a bit too forcefully and meticulously. The awkward and cumbersome organisation and argumentation of the book makes it even harder to read.

*Statistical significance* by Chow (1996) is a defence of null-hypothesis significance testing. When I started reading the book, I was surprised not only that this ubiquitous procedure had to be defended at all, but also about the strangely tense tone of this defence. Having worked through the book, I fully agree with (most of) the opinions of the author, but my impression is that the case for significance testing could have found a better advocate. The problem is mainly the style of the book. The wording is sometimes strange; for example, the colloquial use of the adjective "wanting" or the use of "efficacious" instead of "effective." The sentences are often clumsy and complicated. Most of the various tables in the book are more confusing than helpful. And the book is much too long: the main arguments are repeated too often against the various slightly different angles of putative or real criticism. Also the abbreviation of "NHSTP" for the central subject of the book is not very attractive.

The book begins by presenting the main points of criticism directed against NHSTP. The most important ones are:

(1) One should report more data that entered the statistical analysis instead of just reporting the significance of the outcome. Here I basically agree that it cannot hurt to display these data (means and standard-derivations of measurements, obtained *p*-value).

(2) The null-hypothesis $H_0$ is never exactly true. I don't quite understand this argument. It may simply be the objection that the experimenters have chosen an easy-to-falsify null-hypothesis to simplify their task.

(3) The choice of the significance criterion (e.g., 5% or 1%) is arbitrary. This is obviously true, but in many disciplines there is a general agreement about the proper significance value.

(4) The significance probability *p* obtained from the statistical analysis is often misinterpreted.

(5) The so-called Type II error is not considered, and sometimes erroneous conclusions are drawn from insignificant results. Like the last point, this is a criticism of some incorrect applications of significance testing rather than of the procedure itself. In a typical application of NHSTP one simply cannot determine the probability of the Type II error and one cannot draw any conclusions from an insignificant result. Hence the choice of the word "significant."

Chow's general line of argument against these criticisms is simply that they are criticisms against some incorrect applications of NHSTP rather than against NHSTP itself. In exemplifying and clarifying this point, the author could give some useful advice on the proper use of significance testing at least in the process of theory corroboration. But instead he produces a meticulous and tiring defence against criticisms of NHSTP.

The core of Chow's argumentation is the prototypical use of significance testing in theory corroboration. This argument bears some meta-scientific overtones which I do not share with the author, so I will redescribe it in my own words. In the post-Popperian spirit that has been generally adopted in the scientific community, it goes without saying that a theory cannot be verified, it can only be falsified experimentally. Since the goal of a research group obviously cannot be to falsify their own theory, they try to "corroborate" it, essentially by falsifying a different theory. Sometimes (but rarely) there are indeed two conflicting theories. More often, one has to make a reasonable prediction based on generally accepted background knowledge and common sense which is at variance with the to-be-corroborated new theory, and one falsifies this prediction by an appropriate experiment. If this falsification is only probabilistic, it is normally demonstrated by significance testing.

Logically this means that background knowledge B implies the null hypothesis $H_0$, which provides a probability distribution of some outcome (or test-statistic) T with expectation 0. The new hypothesis N, however, implies that $T > 0$. After the experimental result that $T = t$, one calculates the probability $p = \text{prob}[T \geq t|H_0]$, and if $p$ is sufficiently small one consideres $H_0$ to be probabilistically falsified, therefore $H_1$ and consequently N is corroborated.

The amount of evidence obtained for the theory N depends on the tightness of the implication $N \rightarrow H_1$ and on the smallness of $p$. In most actual cases it depends more strongly on the tightness of the implications and on the design of the experiment than on the smallness of $p$. It appears that criticisms of NHSTP are sometimes really directed against the inductive method of theory corroboration or against the tightness, or even the correctness of the implications involved in specific experiments.

With respect to Bayesian criticism of NHSTP, Chow starts with a strange example (p. 146), which I did not quite understand. Instead, I find it useful to distinguish four kinds of problems (this distinction is perhaps made implicitly but unfortunately not explicitly in the book): the problems of *corroboration* of versus *decision* between *general* versus *specific* hypotheses. These four cases differ mainly in the knowledge that one may assume about the probabilities involved: $p(H_0)$, $p(H_1)$, the a priori probabilities of the two hypotheses, $p(T|H_0)$, $p(T|H_1)$, and the probability distribution of the test-value T under the two hypotheses, $H_0$ and $H_1$.

The prototypical case of theory corroboration treats the problem of corroboration of a general hypotheses (i.e., a "theory"). In this case one knows only $p(T|H_0)$, so NHSTP is the only reasonable statistical procedure. In the much rarer case of a decision between two theories, one may know $p(T|H_0)$ and $p(T|H_1)$, but usually there is no intersubjective way of determining $p(H_0)$ and $p(H_1)$. In this context it may be doubted whether the concept of probability of scientific theories makes sense at all. Thus, the use of Bayes's formula to get from $p(\text{Data}|H_1)$ to $p(H_1|\text{Data})$ may give unreasonable results.

On the other hand, in the case of a decision between two specific hypotheses – for example, in some medical applications where one has to decide based on statistical tests whether or not a specific person has a specific illness – one often knows all the four probabilities. In this case one should certainly make use of them and try to determine $P(H_1|\text{Data})$.

Finally, in the case of corroboration of a specific hypothesis, as in some legal cases or medical applications, the problem is to corroborate the well-founded opinion that a specific person has a certain illness or has committed a certain crime. In this case, different people may have different subjective probabilities, $p(H_0)$ and $p(H_1)$; one typically has agreement on $p(T|H_0)$, but one often does not really know $p(T|H_1)$. In most actual situations of this type, it is probably best to resort to the argumentation of theory corroboration, that is, NHSTP.

Chow devotes a substantial part of the book to making a further distinction in the case of corroboration of a general hypothesis. This has to do with the question of whether the general hypothesis is a consequence of a theory (as in $N \rightarrow H_1$ above) or a more practical statement (which the author calls "utilitarian") for example, that a certain treatment has an effect on T. This provides a different context and motivation for doing the statistical test, but in my view it does not change the interpretation of the result. This distinction introduces several additional controversial problems, such as the author's distinction between experiments, quasi-experiments and non-experiments (p. 91), which in my view only confuses the main issue.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

In summary, the book is essentially an elaboration on the point that there is really nothing to defend NHSTP against. If one agrees with this, one should not read the book.

## Some statistical misconceptions in Chow's *Statistical significance*

Jacques Poitevineau[1] and Bruno Lecoutre[2]

[1]*LCPE, C.N.R.S., 92120 Montrouge, France.* **jacques.poitevineau@ens.fr.**
[2]*C.N.R.S. et Université de Rouen, 76821 Mont-Saint-Aignan Cedex, France.* **bruno.lecoutre@univ-rouen.fr http://epeire.univ-rouen.fr/labos/eris**

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

**Abstract:** Chow's book 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.

***Frequentist theories.*** In presenting current practices (NHSTP), Chow rightly refers to the "hybridism" of Fisher and Neyman/Pearson's theories; this was identified long before Gigerenzer (1993), although this particular term was not used: see, for example, Morrison and Henkel 1970, p. 7. But to say that these authors may be responsible for the hybridism because they "sometimes changed positions" and "incorporated their opponent's ideas" (p. xi) is far from the truth, since their basic theories evolved little.

Chow rightly emphasizes that $\alpha$, $\beta$, and $p$ are conditional probabilities in the frequentist framework. This follows from the fundamental mathematical notion of sampling distribution (conditional on the parameters). But, with regard to this distribution, the assertion that "the standard error of the mean . . . is . . . $s/\sqrt{n}$ if $\sigma$ is not known" (p. 27) is nonsense, since $s$ is in this case a random variable that varies from one sample to another.

***Fisher.*** Nothing is said about Fisher's concepts of probability, which are of direct importance for the objectives Fisher assigned to statistical methods. If, in early writings, he effectively mentioned the idea of a "conventional $\alpha$" (Chow's Table 2.3, p. 21, line 9), he later came to repudiate any systematic predetermined level of significance (Fisher 1956/1990, p. 45). Moreover, he argued against the interpretation of $p$ as the relative frequency of error when sampling repeatedly in a same population (Fisher 1956/1990, pp. 81–82).

Last, contrary to what Chow implies, Fisher was evidently interested in p(H|D) (although this probability cannot, of course, be identified with $p$) as it emerges from his work, not only on the fiducial theory (e.g., Fisher 1935b; 1956/1990), but also on the Bayesian method (a fact often ignored) in his last years (Fisher 1962).

***Neyman/Pearson.*** It is first rather surprising that the single reference is the (Neyman & Pearson) 1928 paper that presents only the premises of their theory. In Neyman's own words: "Our first paper on the subject was published in 1928, over twenty years ago. However, it took another five years for the basic idea of a rationale theory to become clear in our minds." (Neyman 1952, p. 58). Actually, some verbal formulations in the introduction of this paper could be misleading, and above all the concept of statistical power did not appear. Relevant references are the 1933 papers (Neyman & Pearson 1933a; 1933b).

Contrary to what is indicated in [Table 2.3, line 5], the probability of interest is not "the inverse probability, p(H|D)." Neyman and Pearson devised a precise method which is independent of the prior probabilities p(H) and hence cannot say anything about p(H|D) (Neyman & Pearson 1933b). Moreover, as a radical frequentist, Neyman later came to discard $p$(H), and consequently $p$(H|D), when it could not be assigned a frequentist meaning (Neyman 1950; 1952).

Also $H_0$ is not "the hypothesis of zero difference" (Table 2.3, p. 21, line 6); on the contrary, it should be the hypothesis of interest, the one for which it is more important to avoid error. This fact is clear and more than implicit, even in the 1928 paper; and it was firmly stressed by Neyman (Neyman 1950, p. 263). Consequently, $H_1$ is not "the substantive hypothesis itself" (Table 2.3, line 7) but the challenging one.

Moreover, "Multiple $H_1$'s" (Table 2.3, line 8) are not necessarily assumed. The alternative hypothesis (and also $H_0$) may be simple as well as multiple. More important again is the fact that $H_0$ and $H_1$ are always assumed to be mutually exclusive and exhaustive (what Chow considers to be only a Fisherian feature), so that "not-$H_0$" is really equivalent to $H_1$ (or $H_1$'s), contrary to Chow's claim (p. 132). Thus it cannot be said that "the identity of the non-null hypothesis (viz., not-$H_0$) is not known" (p. 186). Even in the case of a multiple hypothesis (e.g. $\mu > 0$), the possible values of the parameter under the alternative hypothesis are perfectly known (e.g., the strictly positive part of the real line). What is indeterminate is only the knowledge of the *true* value.

It is basically nonsense to say that "the meaning of 'Type II Error' is changed when statistical power is introduced" (p. 131). The concept of statistical power was introduced by Neyman and Pearson as the complement of type II error (Neyman & Pearson 1933b), and plays a basic and explicit role in their theory (a Neyman-Pearson's test consists of building a critical region that minimizes $\beta$, or equivalently that maximizes power, for a fixed $\alpha$). Power, with regard to any simple hypothesis $H_i$, is the probability that the test statistic falls into the critical region (given that $H_i$ is true), and is consequently equivalent to $1 - \beta_i$ ($\beta_i$ denoting the probability of committing a type II error given $H_i$). In the case of multiple $H_i$'s, power can be calculated for each of the simple hypotheses (which is perfectly determined, as indicated earlier), thereby leading to the well-known *power function.*

Concerning the implicit appeal to TSD by power analysts (Ch. 6), Chow seems to be unaware that it is signal detection theory that drew upon Neyman-Pearson's theory (Green & Swets 1966, p. 1), and not the reverse.

Last, Chow forgets to report the fact that the sample size $N$ must be fixed *a priori* (before experiment); and, curiously enough, he does not mention the criticisms raised about the "decision characterization" of the Neyman-Pearson's theory (e.g., Rozeboom 1960).

***Statistical Bayesian theory.*** Chapter 7 is only Chow's personal views about "Bayesianism." This has nothing to do with the rational statistical Bayesian theory, and is consequently completely misleading. Chow ignores the foundations of Bayesian statistics and appreciable theoretical developments that are ever more strongly challenging the frequentist approach in all domains (for recent accounts, see in particular Bernardo & Smith 1994; Robert 1995), including the related methodological debates peculiar to medicine (e.g., see Ashby 1993). Moreover, realistic uses of Bayesian methods for analyzing experimental data have been proposed (e.g., Spiegelhalter et al. 1994; Lecoutre et al. 1995; Rouanet 1996).

## Null-hypothesis tests are not completely stupid, but Bayesian statistics are better

David Rindskopf

*Educational Psychology, City University of New York Graduate Center, New York, NY 10036.* **drindsko@email.gc.cuny.edu**

**Abstract:** Unfortunately, reading Chow's work is likely to leave the reader more confused than enlightened. My preferred solutions to the "controversy" about null- hypothesis testing are: (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.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

As I read Chow's précis, I felt confused: Did I not understand the author, or did the discussion contain major errors and misunderstandings? Unfortunately, after re-reading the précis and reading the book, I concluded that the latter was the case. I wish I could say more positive things about the book, but more is wrong or misleading than is right. Although I agree with some of the author's statements (e.g., that null-hypothesis tests can be defended), I disagree with most of his conceptualization of the problem and most of his reasoning. A full critique would take too much space, and would be unnecessary. Instead, I will give a few examples, and then briefly summarize my position about hypothesis testing.

For a book that is mainly concerned with a discussion of null-hypothesis tests, it is surprising that it does not begin with a definition of "null hypothesis." To most statisticians, the null hypothesis is a statement that some parameter is equal to zero. That is why the term "null" applies: null means zero, as in zero effect, or zero relationship. But Chow's main example concerns a directional hypothesis, so most of his discussion is irrelevant to the usual arguments about problems with null-hypothesis tests.

Why should null (i.e., zero) hypotheses be of such importance? Occam's Razor provides the answer. A scientific theory should be no more complex than is necessary to explain the observed facts. Therefore, until there is clear evidence that a treatment is effective, or that a relationship between two variables exists (as two examples), one should act as if there were no effect. The simplest theory, that nothing has an effect, is only rejected after evidence to the contrary.

Now let me turn to Bayesian statistics, where Chow's discussion and example are tortuous. The example he uses, insofar as it is comprehensible at all, is clearly wrong: he has prior probabilities for three alternatives as .5, .6, and .4; the probability that one of these will occur is therefore .5 + .6 + .4 = 1.5. Even beginners in probability can see that there is a problem here.

As final examples, consider some statements in Chow's summary and conclusions section of the Précis:

1. "[B]eing a conditional probability, statistical power cannot be the probability of obtaining statistical significance." No sensible person has ever claimed that anyway.

2. "As there is a Bayesian overtone in power analysis, power analysis can be questioned to the extent that the Bayesian assumptions about research methodology are debatable." No one else sees a Bayesian overtone in power analysis, so the conclusion is unwarranted.

3. "The Bayesian approach has a very limited applicability in psychological research because it is applicable only to the sequential sampling procedure." Another incorrect statement, and therefore an invalid conclusion.

So what is the correct attitude about null hypothesis tests? First, most people do not believe that any effect is exactly zero. Instead, they are trying to choose between two alternatives: (1) The effect is so small that it is unimportant, or (2) the effect is large enough to be important. Therefore, researchers really want to test the "small," not null, hypothesis. As long as they follow the usual rules of applied statistics, they will not go wrong: an effect should be declared significant only if it is both statistically significant and large enough to be of practical importance. (Further details on my views about this are in Rindskopf 1997.)

But even though null hypothesis tests are not so bad if properly done, there is a better approach for most problems in statistics: Bayesian methods. To see how natural Bayesian methods are, consider the following interpretation of a confidence interval: "There is a 95% chance that the parameter is in this interval." In classical statistics, this is an erroneous statement, but all teachers hear this interpretation from most of their pupils every semester. In Bayesian statistics, this is the correct interpretation of a 95% credible interval, which can often be calculated in the same way as a classical statistician's 95% confidence interval. Bayesian statistics is about using information (from data) to modify beliefs about unknown quantities (parameters). We start with prior beliefs, which can be very vague (uninformative) if we have little information about a situation before gathering data. After gathering data, we modify our beliefs to arrive at posterior beliefs (technically, a posterior probability density).

Once we have our posterior beliefs, we can either summarize them in various ways, or use them to make decisions, along with information about the costs and benefits (utilities) of various actions. One simple summary that would be informative is to decide how big an effect would have to be in order to declare it of practical importance, and then calculate the probability that the effect is at least that large. For example, one would be able to make a statement such as: "There is a 5% probability of a large negative effect, a 20% probability that the effect is small, and a 75% probability of a large positive effect." These calculations and interpretation can only be done in Bayesian statistics; such an interpretation does not make sense in classical statistics. (Again, further details are in Rindskopf 1997.)

Finally, some suggested reading for those interested in comparative statistics (i.e., a comparison among different approaches to statistical inference). Two reasonably simple books are those of Barnett (1982) and Oakes (1986). At a higher level of abstraction, Silvey (1975) provides an overview of the major approaches to statistical inference.

## Meta-analysis, power analysis, and the null-hypothesis significance-test procedure

Joseph S. Rossi

*Cancer Prevention Research Center and Department of Psychology, University of Rhode Island, Kingston, RI 02881.*
**kzp101@uriacc.uri.edu www.uri.edu/research/cprc/**

**Abstract:** Chow's (1996) defense of the null-hypothesis significance-test procedure (NHSTP) is thoughtful and compelling in many respects. Nevertheless, techniques such as meta-analysis, power analysis, effect size estimation, and confidence intervals can be useful supplements to NHSTP in furthering the cumulative nature of behavioral research, as illustrated by the history of research on the spontaneous recovery of verbal learning.

Chow (1996) raises many compelling points in defense of the null-hypothesis significance-test procedure (NHSTP) and I am sympathetic to much of what he has to say. In particular, some of the arguments made by critics of NHSTP are long overdue for critical examination. For example, as I am frequently engaged in the conduct of large randomized clinical trials of behavioral interventions for health promotion and disease prevention, I could only wish that the null hypothesis was, in fact, never true! Chow reminds us that NHSTP is but one step in the research process, and not necessarily the most important one. The sequence of deductive and inductive logic and the establishment of proper experimental controls provides the crucial foundation on which inference and knowledge are based within the context of research design. No degree of reliance on meta-analysis, power analysis, effect sizes, confidence intervals, or NHSTP itself for that matter will alter the tentative nature of the logic on which the conclusions of an experiment are based.

And yet the consensus of methodological expertise for the past four decades is that reliance on NHSTP has failed to produce a cumulative science, a position well-articulated by Meehl (1978) and others. Chow's position seems to be that if we would only get back to rigorous adherence to the underlying logic of experimental design, all would be well. Although I certainly endorse such adherence, I do not believe it will be sufficient. Effect size indices and confidence intervals along with techniques such as meta-analysis and power analysis can surely be used to enhance NHSTP, and not only for utilitarian experiments, which Chow occasionally seems to admit, but for theory-corroboration experiments as well.

An interesting case is provided by the history of theory-corroboration research on the spontaneous recovery of verbal



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

associations (Rossi 1990; 1997). Researchers in this field were no less meticulous than Sperling (1960) and others in the work on iconic storage described by Chow as an illustration of well-controlled experimental research. Yet 20 years of NHSTP-prescribed research (c. 1948–1968) did not result in clear consensus or rejection of the existence of spontaneous recovery. Instead, there was continued controversy concerning the phenomenon, with some studies showing the effect and others not, until researchers eventually lost interest and moved on to other topics. Meta-analysis and power analysis can be helpful in understanding what happened.

The interference theory of learning predicted the existence of spontaneous recovery but gave researchers no indication of the size of the effect. In any event, it was not the custom to select sample sizes based on expected effect size or on statistical power considerations. Instead, laboratory tradition dictated the number of subjects that would be run, much as in the iconic storage research. Traditional sample sizes served well in the investigation of strong phenomena easily brought under experimental control, such as proactive and retroactive inhibition. But spontaneous recovery was a more subtle phenomenon. Meta-analysis of spontaneous recovery studies indicated an average effect size (d) of 0.39 (95% confidence interval = 0.27–0.48). Using this as the expected effect size, the average power of spontaneous recovery studies was .38 (alpha = .05). Not surprisingly, this figure agrees well with the proportion of spontaneous recovery studies that were statistically significant ($p < .05$): .43. Assuming the effect to be of a magnitude similar to that of other laboratory phenomena in the study of human learning resulted in the selection of sample sizes too small to show the effect consistently.

Given the extent to which researchers in this field attended to issues of experimental control, it is not surprising that there was a great deal of confusion and dismay over these results. The issue is still considered "one of the unresolved issues of interference theory" (Zechmeister & Nyberg 1982, p. 112). I cannot help but think that attention to issues of effect size and statistical power would have greatly benefited investigators. Rather than conclude that spontaneous recovery did not exist, researchers might have used statistical power to set an upper limit on the magnitude of the effect as a guide for future investigators, much as is done in the physical sciences (Rossi 1990; 1997). Based on the average sample size of 80 typical of work in this area, type II error rates can be determined across a range of possible effect sizes. For example, the type II error rate is .05 for an effect size of $d = 0.82$ and .10 for an effect size of 0.74. There is only a 25% chance that the effect size is greater than 0.59. This procedure provides a sensible method of evaluating null results and may be useful in designing further research.

Researchers might also use meta-analysis to aid in the design of future studies by using the average effect size as the basis for selecting a sample. Researchers could easily determine that a sample size of 208 subjects would be necessary to achieve power of .80. Alternatively, the lower bound of the 95% confidence interval could be used as the estimate of effect size. In this case, a sample size of 432 would be necessary. Although large for laboratory research in human learning, sample sizes this large and larger were sometimes used in spontaneous recovery research. If such sample sizes were too large for an investigator, additional avenues of exerting experimental control over the situation would be necessary to increase the expected size of the effect. Such additional approaches would be desirable in any case as a means of further delineating and understanding the conditions under which the phenomenon occurred.

NHSTP should not be abandoned, but it does need help. As Chow indicates, more rigorous adherence to the underlying logic of experimental design along with increased attention to issues of experimental control are important parts of that help. But so too are techniques such as meta-analysis and power analysis, which can be essential for understanding the results of experiments and for aiding in the accumulation of results across studies. By such means the behavioral sciences may yet set out on the path the more developed sciences have already discovered.

## Significance testing in a Bayesian framework: Assessing direction of effects

Henry Rouanet

*UFR Math-Info, Université René-Descartes, 75270 Paris Cedex 06, France.*
**rouanet@math-info.univ-paris5.fr**

**Abstract:** Chow's efforts toward 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.

My first reaction to Chow's book was positive. I like the project of discussing theory corroboration (as opposed to "utilitarian" experimentation) in connection with a well-chosen example (the kernel-negative experiment), considering (1) the various sorts of hypotheses – substantive, research, experimental, and statistical – that are involved in experimental research as well as the role of significance testing as a minor premise in a series of implicit syllogisms in theory corroboration. Above all, I fully concur that the criticisms against significance testing have definitely gone too far, and that it would be foolish to abandon significance testing just because there are common "false beliefs" about it. The message of methodologists should be education, not eradication. Significance testing has a useful role to play in data analysis.

Unfortunately, as I read further, I became perplexed about Chow's peremptory treatment of many topics, such as effect size. For example, isn't it poor judgment to claim that the question "How effective is the treatment?" is not even meaningful in theory-corroboration experiments? What I find more than questionable in Chow's book, however, is the Bayesian chapter, and I will now concentrate my comments on that.

Chow's critique of the Bayesian approach is confined to the most narrow-minded personalistic "Bayesianism," a view that no contemporary Bayesian statistician would hold (Bayesian statisticians have learned to be tolerant). Current developments in the "noninformative" approach to Bayesian statistics are simply ignored by Chow. (In this connection, his citations of Jeffreys are misleading.) In the noninformative (as opposed to the "personalistic") approach, the prior distribution is chosen to express a "state of ignorance" about parameters, with the following motivation: if the prior probability expresses ignorance about parameters, the posterior probability expresses the evidence provided by the data. Using noninformative priors provides a link between Bayesian and frequentist procedures, which can be reinterpreted in terms of probabilities about parameters. Such a link is surely not coincidental, and deserved consideration in a book devoted to the rationale of statistical significance.

Chow's example of a fictitious poll is extraordinarily contorted, and certainly not congenial to actual Bayesian data analysis.[1] Why did Chow not take up a condition dealt with earlier in the frequentist framework? Let us, for example, take the negative kernel experiment, with a matched-paired design and $n = 20$ subjects (hence 19 d.f.). If the observed t-ratio is +1.727, the following frequentist statement holds (under the usual normal sampling model): $P(t > +1.729|\delta = 0) = 0.05$ (where $\delta$ denotes the true mean effect); that is, the probability of obtaining a result more extreme (on the upper side) than the data, if $\delta = 0$, is 0.05. Now in this situation, a standard Bayesian analysis consists in inducing a noninformative prior distribution with the parameters $\delta$ and $\sigma$ (SD of individual differences). Based on a well-known result in Bayesian statistics (e.g., Box & Tiao 1973, p. 102), under these prior assumptions the following statement holds: $P(\delta > 0|\text{data}) = 0.95$ (notice that this statement differs, of course, from



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

the notoriously false belief P(δ = 0|data) = 0.05). That is, the posterior probability that the true effect δ has the same sign (here positive) as the observed effect is 0.95. Intuitively, for the negative-kernel experiment, an observer in a prior state of ignorance will, after the experiment, become 95% sure that it is more difficult to remember additional words, after a negative sentence.

It seems to me that the foregoing Bayesian reinterpretation of the significance level in terms of *assessing the direction of effects* is a highly meaningful interpretation of significance testing, and as such provides a strong argument in favor of it. Of course, this interpretation is "bought" at the price of the additional assumptions about prior distributions; those who are not willing to make those assumptions may stick to the frequentist interpretation. But the two interpretations are not in conflict. Contrary to Chow's prejudice, turning to the Bayesian framework does not mean giving up significance testing. My claim is that Chow's plea for significance testing would have been reinforced by the reinterpretation of the significance level in the Bayesian framework. In this view, Bayesian procedures can be used to complement, rather than replace, the familiar frequentist procedures: this approach is presented in Rouanet (1996) and developed in Rouanet et al. (1991).

In conclusion, I have serious reservations about Chow's book. Whereas I appreciate the author's efforts to elaborate a methodology of theory-corroboration and I concur with the plea for significance testing, the book contains too many ill-considered claims. The discussion of Bayesian inference is misguided. A major omission is the failure to present significance testing in the Bayesian framework, with a reinterpretation of the observed significance level for assessing the direction of effects.

NOTE

**1.** The calculations of Table 7.1 (p. 146) are obscure. How do the data (i.e., the row labeled "evidence") enter the computations? Where do the likelihoods come from? Since the hypotheses are obviously not mutually exclusive, how can the posterior probabilities be correct? In addition, there is a significant discrepancy between the book and the Précis. In the book (p. 147), the fictitious editor E will endorse the Center Party if he is 75% sure about the hypothesis Hc; in the Précis (sect. 36, para. 3), when the party is preferred by 75% of the prospective voters.

## Costs and benefits of statistical significance tests

Michael G. Shafto

*Human-Automation Interaction Research Branch, NASA-Ames Research Center, Moffett Field, CA 94035-1000.* **shafto@simon.arc.nasa.gov olias.arc.nasa.gov/**

**Abstract:** Chow's book provides a thorough analysis of the confusing array of issues surrounding conventional tests of statistical significance. This book should be required reading for behavioral and social scientists. Chow concludes that the null-hypothesis significance-testing procedure (NHSTP) plays a limited, but necessary, role in the experimental sciences. Another possibility is that – owing in part to its metaphorical underpinnings and convoluted logic – the NHSTP is declining in importance in those few sciences in which it ever played a role.

I think the late Amos Tversky said that even experts should not try to reason intuitively about probability. As a case in point, the "null-hypothesis significance-test procedure (NHSTP)" has claimed a long list of victims, some of them well documented by Chow, ranging from the hapless introductory student to the would-be expert. Few of us who have inflicted the NHSTP on our students could measure up to the intellectual standard that Chow implies. Many of us who commit statistical analyses as a side-effect of our work will be reduced to befuddled self-doubt. It seems that every nuance of the NHSTP is plagued with conceptual difficulties even more paralyzing than we suspected.

What should we conclude from this widespread confusion, which, according to Chow, manifests itself too often in misconstruals of data, misinterpretations of experimental results, and possibly serious errors in the evaluation of competing theories? One conclusion might be that practitioners need a more sophisticated understanding of the logic and application of the NHSTP. Chow has done a thorough and systematic job of trying to defend that position against the arguments reviewed in Chapter 1, "A Litany of Criticisms."

Chow provides strong responses to these criticisms. In later chapters he also provides careful discussions of the difficulties and errors surrounding statistical power and effect-size estimation. These discussions should be required reading for basic and applied researchers in behavioral and social sciences, as well as for instructors in quantitative methods. The general strategy behind his arguments, however, is to restrict the scope of applicability of the NHSTP. How restricted can the applicability become before the benefit is negligible compared with the costs?

Chow's thorough discussion of the misconceptions surrounding the NHSTP tends to weaken his principal conclusion – that the NHSTP has a crucial role to play in empirical research, especially in theory corroboration. I would suggest that the NHSTP adds little value and much confusion, as Chow's own analysis shows, to the research process.

Rather than being used in model testing, the NHSTP is often used in lieu of model testing (Simon 1979, Ch. 5.4). The dominant concerns in theory-development tend to be testable speculations about "laws of qualitative structure" (Langley et al. 1987, p. 21) or explorations of simple systems (Ohlsson & Jewett 1997). As more powerful theories and more specific models are developed, research efforts focus on finer-grained analyses of denser data sets, for example, psychophysiological data, sequential behavior, learning, and individual differences. When theory is strong, and specific models are at issue, we would expect hypothesis-testing to focus on more interesting questions than Theory A versus random chance.

In utilitarian research, the NHSTP is often (mis-)used as a filter. Faced with weak theory and many possible sources of variance, we monitor available data for "significant" correlations or trends. In contexts such as aviation safety or public health, we need a systematic method to help us interpret large, noisy datasets and to call our attention to apparent changes which may require intervention. The logic of the NHSTP is not followed in any clear sense, although components of it play a role in balancing sample size, variability, and intuitive assessment of possible costs. There is usually no real population and no real sample, but there is reasoning about data and about the risks and costs associated with any course of action, including inaction.

It is not clear that the NHSTP adds value in these situations, compared with other theory-based methods such as decision-theoretic analyses or Bayesian methods, or compared with rules of thumb such as those used in statistical process control. As Chow points out, statistical significance can easily be misinterpreted in large samples. The NHSTP may be most clearly and directly useful, in both utilitarian research and theory corroboration, as a safeguard against over-interpretation of subjectively large effects in small samples. This benefit, however, must be weighed against the considerable difficulties that Chow documents. Either he is wrong in his analysis, or the NHSTP entails profound conceptual difficulties even for many experts; and this is not a strong recommendation for its use as a fundamental method of analysis.

A few years ago I taught a course in quantitative methods for business majors. I was surprised to find that, despite covering some challenging applied mathematics, textbooks in this field made no mention whatever of the NHSTP. Related topics, such as decision theory, Bayesian inference, and stochastic modeling were covered. Methods were included that required solution by computer: linear, integer, and goal programming; Monte Carlo simulation. Yet, in three representative textbooks (Lapin 1994; Stevenson 1992; Winston 1991) comprising about 3,000 pages, there was only one brief mention of the NHSTP (Lapin 1994, pp. 997–1002). This experience, as well as interactions with



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

colleagues in chemistry, physics, biology, and engineering, has convinced me that the central place of the NHSTP in some behavioral sciences owes more to historical accident than to a favorable cost-benefit ratio.

## Inductive strategy and statistical tactics

Paul Snow

*P.O. Box 6134, Concord, NH 03303-6134.* **paulsnow@delphi.com**

**Abstract:** Chow ably defends classical significance testing by relating this method to venerable principles for inductive reasoning. Chow's success does not preclude the use of other approaches to statistical reasoning, which is fortunate not only for Bayesian rivals, but even for some fellow classicists.

Chow offers a fresh, but appropriately conservative defense of classical null-hypothesis significance testing. For Chow, statistical significance operates within a larger theory of inductive reasoning. "Inductive reasoning" is meant in the broad sense in which philosophers (e.g., Black 1969) sometimes use it to describe any nondemonstrative inference, rather than the narrower sense of generalization based on specific observations.

Chow builds his theory from the precepts of John Stuart Mill and Karl Popper's falsificationism. The mission of significance testing is to exclude chance as a plausible explanation for whatever evidence is being evaluated if the data warrant exclusion. Whatever else one might conclude from the evidence is attributed to other identifiable aspects of the inductive program, such as the design of experimental investigations.

The resulting clear division of labor between statistics and inductive logic deflects criticism that is sometimes directed at classical statistics onto those other aspects of the inductive program. It also provides a principled framework for the evaluation of empirical work which uses significance testing. Traps for the unwary are revealed (e.g., just because a result is not attributable to chance it does not follow that the investigator's favorite nonchance explanation is correct) along with the means of avoiding them (principally the correct use and interpretation of controls in experiments, analyzed here with much subtlety using apt examples from the psychological literature along with thoughtful hypothetical illustrations).

Some specific features of significance testing which trouble critics are deftly dispatched. For instance, significance has an "either-or, accept-reject" quality, in contrast to the more graduated conclusions of Bayesians. In Chow's account, dichotomy appears quite natural. Either chance is tenable as an explanation of what was observed, or else it is not. The issue can be treated somewhat like a methodological concern, such as a suspicion that a test tube might have been dirty. If contamination is a likely possibility, then that is cause to disbelieve and dismiss the research results in question; likewise if chance may be operating unchecked. That the line between the tenable and the untenable is arbitrary is admitted frankly and without apology.

Chow's argument on this point compares well with other classical accounts of the import of rejecting the null hypothesis (for a review of these, see Howson & Urbach 1993, Ch. 9). But consider the suggestion that one might report $p$ values numerically rather than as inequalities (e.g., $p = 0.0123$, rather than $p < .05$). Chow discusses this option almost as if it were an alternative to classical significance theory, rather than an aspect of classicism. "Either-or" ($p < .05$, say no more) is defensible, but surely $p$ as a number provides additional information of legitimate interest to some readers, and not just those readers who might misinterpret it in various ways discussed by Chow. Similarly, confidence intervals and power analysis appear here almost as rivals rather than as complements to significance. The defense offered, then, is highly specific to significance testing, as opposed to the full range of classical methods.

Some criticisms that arise outside the classical camp are not fully treated in Chow's book. The inflexibility of classical stopping rules, for instance, is of both practical and theoretical concern, sharply different from Bayesian methods, and a perennial subject of discussion (Berger & Berry 1988). Chow touches briefly on stopping rules as an aspect of Bayesian sequential sampling, but the force of the general Bayesian or "likelihoodist" position is neither acknowledged nor addressed.

A deeper problem may be that the quasi-syllogistic approach presented by Chow, following the lead of Mill and Popper, may not provide a sufficiently rich account of inductive reasoning. Few would dispute that (1) controlled experimentation can be an effective way to explore the world or deny that (2) modus tollens has a place in theory testing. The similarity between these techniques and the formalisms of deductive logic imparts a reassuring ring of truth to the inferential enterprise.

On the other hand, any appearance of soundness is illusory, since inductive reasoning is not demonstrative. More specifically, the "eliminative" pattern of inference advanced by Chow is simply not a syllogism. Chance explanations are not really eliminated by significance results; they are only made relatively implausible in the opinion of a particular observer.

That does not render a quasi-syllogistic approach useless, but it does suggest that other approaches might also be worth looking into. Those other approaches need not be hostile to classical insights. For example, the principles behind significance testing fare quite well in the personalist and otherwise "confirmation" oriented inductive theories of George Polya (1954).

Chow's argument does not succeed in excluding different approaches to the analysis of chance influences. Nor does it rule it out that within those other approaches some sort of graduated discounting, rather than outright rejection, of questionable research might be plausible. The failure to rule out alternative approaches and rationales is not devastating to a purely defensive work like Chow's. Significance testing deserves to be practiced if it has any foundation. Chow has furnished an entirely satisfactory "existence proof" for such a foundation, along with a tough-minded guide to the proper uses of significance testing in empirical research.

## The historical case against null-hypothesis significance testing

Henderikus J. Stam and Grant A. Pasay

*Department of Psychology, University of Calgary, Calgary, Alberta, Canada T2N 1N4.* **stam@acs.ucalgary.ca**

**www.psych.ucalgary.ca/people/faculty/stam/**

**Abstract:** We argue that Chow's defense of hypothesis-testing procedures attempts to restore an aura of objectivity to the core procedures, allowing these to take on the role of judgment that should be reserved for the researcher. We provide a brief overview of what we call the historical case against hypothesis testing and argue that the latter has led to a constrained and simplified conception of what passes for theory in psychology.

Like Truman Kelley (1923) before him, Siu Chow (1996) seeks to defend the canons of statistical methods, in particular null-hypothesis significance testing, from its multiple critics. Kelley's concern was to defend the notion that objectivity was to be found in methods, not in the judgment of the scientist (see Stout 1987, for an excellent discussion of the controversy). Similarly, Chow, both here and in his previous writings (e.g., 1991), takes us through an updated version of this argument and does us a great service by collating in one volume most of the criticisms and a great deal of material relevant to the recent debates concerning hypothesis testing. Our view, however, is that Chow has misinterpreted what we will call the historical critique, which focuses on the historical emergence of hypothesis testing and its use by psychologists, particularly its institutionalization (e.g., Danziger



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

1987; 1990; Falk & Greenbaum 1995; Gigerenzer 1987; 1993; Gigerenzer & Murray 1987; Gigerenzer et al. 1989). In addition, we argue that this institutionalization has stifled, and continues to stifle, theory development in psychology.

First, we recognize that Chow considers a more limited role for hypothesis testing procedures than most textbooks allow. He does this by giving us a third version of hypothesis testing, one that is different from both the Fisherian and Neyman-Pearson versions and one that constitutes a refined version of the hybrid model widely adopted in psychology and elsewhere. To his credit, Chow's hybrid version is explicitly so. His version is rhetorically structured within a logical framework, creating the impression of a potentially automated procedure, one that requires the researcher to follow certain rules of logic that inevitably lead to proper conclusions with proper use. This rhetorical reconstruction also allows Chow to argue that most criticisms of hypothesis testing "are based on critics' diverse assumptions about the prototype, goal and nature of empirical research" (p. 176) rather than on the details of hypothesis testing procedures themselves.

As Danziger (1987; 1990) has argued, one striking development in the history of twentieth-century North American psychology was the gradual emergence of the aggregate as the unit of analysis. As a response to the pressures for applied knowledge, psychologists began constituting research groups whose purpose was to serve as a vehicle for comparison with other groups, for example, those differing in intelligence. Gradually, individual scores came to be reported in the aggregate with departure from the aggregate as "error." But until their conjunction with inferential statistics, aggregate scores created difficult theoretical problems for mid-century psychology. In particular, statistical aggregates that referred to groups of individuals made it difficult to elaborate the theoretical intra-individual processes that were purportedly of theoretical interest to psychology. Danziger's historical analysis points out that the adoption of inferential statistics allowed psychologists to argue that "the hypothetical distributions of the statistical analysis could be identified with characteristics of real psychological systems, thus permitting a bridging of the gap between data that referred to groups and theoretical constructs that referred to individuals" (1987, p. 46). From this emerged the notion that constructs such as "memory" could be studied not by investigations of individual acts of remembering but by comparing experimental group performances on some restricted and controlled task. The aggregated numbers referred back to a stylized and idealized (or functional) conception of memory that was not representative of any single participant in the experiments.

The role of hypothesis-testing in this history changed the very nature of what counted as theory in psychology. Chow bypasses the question of how we come to have theories by invoking Popperian indifference – anything can count as a conjecture so long as it is "consistent" with the phenomenon. He uses "theory" and "hypothesis" interchangeably, assuming that both are mere "speculative accounts" (p. 46). Theories of any complexity are dispensable in hypothesis-testing scenarios, since these require only binary hypotheses and Chow's logic of the true theory-testing situation (Ch. 4) requires several translations from what he calls "substantive hypothesis" to "statistical hypotheses." At each step an unknown number of ceteris paribus conditions must hold. The final test of a hypothesis is only indirectly related to any theoretical account and must be a test of some aggregated set of numbers to fulfill the requirements of the hypothesis-testing mechanisms. These mechanisms allow for the testing of binary decisions, and theory "corroboration" is never truly that; rather, it is the repeated posing of binary questions that require little forethought. But as the historical case illustrates, what was alluring about the new procedures was their ability to appear objective. It is this feature of the tests that Chow wants most to defend. Years of sustained criticism have made hypothesis testing out to be the ambiguous procedure it is; Chow denies this by arguing that the tests have merely been badly used. He states that the "random sampling distribution of the test statistic determines . . . its meaning" (p. 13) and that the alpha-level has a meaning "independent of the researcher's theoretical preference." We don't know what it means for meaning to reside outside the researcher's beliefs or theories and outside the community to which these beliefs and theories make sense. Chow's rhetorical reconstruction goes so far as to deny meaning and intent to the researcher. He claims that the correct use of hypothesis-testing guarantees objectivity whereas the long-standing debate over its use demonstrates, instead, the highly negotiated nature of its meaning.

Significance testing has become widespread in psychology (e.g., Hubbard et al. 1997; Sterling 1959). At the same time, our theoretical vocabulary has changed significantly. The developments in theory seem curiously unrelated to developments in, or the use of method (the exception is the reverse, namely, the use of tools turned into theory, see Gigerenzer et al. 1989). Chow might argue that since it doesn't matter where theories come from, this is irrelevant. When a discipline's theoretical developments are largely unrelated to its empirical pursuits, however, that discipline's scientific credentials are surely suspect. Chow unwittingly gives an example of this in Chapter 4. Savin and Perchonock's (1965) study concerns a linguistic analogue of Chomsky's transformational grammar. Chomsky's (e.g., 1965) theory of transformational-generative grammar has developed and altered in important respects since its major 1957 formulation. Chomsky does bring a variety of observations and arguments to bear that have an empirical grounding but his theory has never depended on the kind of test proposed by Savin and Perchonok.

In Chapter 4, Chow invokes Boring's discussions of control in experimentation as a way of clarifying the logic of the experiment. This is interesting because Kelley's (1923) strongest defense of the objectivity of statistical methods came in reply to critiques by Boring (1919; 1920). Among other things, it was Boring's claim that it was not the "mathematical result" that determined the usefulness of obtained data, but the "scientific intuition of the experimenter and his public" (1919, p. 337). Like Chow, Kelley defended the statistical procedures of his day as a way to guarantee an objectivity of method and to remove Boring's "intuition" from science. After a half a century of indiscriminate use, we are not suggesting that hypothesis testing should be abandoned. We do argue, however, that it is time to accept that, in psychology at least, the reflexive nature of the subject matter requires that we recognize the researchers' role in the processes of investigation. This might liberalize the methods available to our researchers and students and allow theory to determine methodology, rather than vice versa.

## A plea for Popperian significance testing

Zeno G. Swijtink

*Department of Philosophy, Sonoma State University, Rohnert Park, CA 94928.* **swijtink@sonoma.edu**

**Abstract:** Even in a theory corroboration context, attention to effect size is called for if significance testing is to be of any value. I sketch a Popperian construal of significance tests that better fits into scientific inference as a whole. Because of its many errors Chow's book cannot be recommended to the novice.

Of all the human sciences, psychology must have the closest interest in statistics. This is no doubt because psychology has enough experimental control to stabilize variability, but not enough control to eliminate variability altogether. Hence, years ago, psychologists interacted with statisticians, like R. A. Fisher, Jerzy Neyman, and Egon Pearson.

Unfortunately, to judge from the references in the back of Chow's book, these two groups have lost contact since. I think this is a shame, because they can still learn from each other. Statisticians have overcome the conceptual confusion that dominated the acid disputes between Fisher and Neyman in the 1930–50s. They



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

now have a better understanding of the differences between Fisherian significance testing and Neyman-Pearson hypothesis testing (Cox & Hinkley 1974). Even the contacts between Bayesian and non-Bayesian statisticians, in my estimation, have become more productive, and if it has not all become ecumenical, there is a greater interest in comparing different types of analysis of the same material. Some of this story is told in our book *The empire of chance* (Gigerenzer et al. 1989).

But statisticians can also learn from scientists, including psychologists. An important point that Paul Meehl has argued for years is that statistical inference does not exhaust all of inductive inference (Meehl 1990), and Chow elaborates upon this in the most valuable part of this book. Chow's analysis, however, is flawed, and it remains a challenge to identify the place of statistical analysis in scientific inference in general.

Reading this book was a frustrating experience for me, not only because it is badly written and makes unreasonable demands on the reader, such as to remember the difference between [A7-1] and [A7-2] long after these propositions have been introduced. I also felt frustrated because I agree with Chow that there remains an important role for significance testing and I am in complete sympathy with his Popperian leanings. Yet, at the same time, I found his book full of historical errors, logical silliness, misrepresentations, and lack of understanding.

The book is not to be trusted as a serious discussion of the historical record. A most glaring mistake is made in Table 2.3 on page 21, and repeated in a number of other places. In this table Chow tries to trace back the NHSTP smörgåsbord to its sources in either Neyman-Pearson tests of hypotheses or Fisherian significance testing. He claims that the "probability of interest" in the Neyman-Pearson approach – not accepted in NHSTP – is the inverse probability p(H|D), the probability of the hypothesis given the data. Chow's reference that would support this is to an early paper by Neyman and Pearson (1928). Was Neyman indeed a closet Bayesian before he turned a frequentist (Neyman 1957; Neyman & Pearson 1933)? Not at all; the identification is erroneous. In their 1928 paper, Neyman and Pearson had not yet formulated their 1933 solution to statistical inference as inductive behavior based on the frequency concept, but their problem was the same: what makes all the statistical tests that are around, based on the *t*-statistic, on chi-square, on F, or what have you, good tests? By way of a solution, they show that some of the standard tests follow from a likelihood criterion that rejects the hypothesis when the likelihood of the observed sample under an alternative is sufficiently large compared to its likelihood under the hypothesis. Reducing the data to its value for the statistic leads to the same decision. But the likelihood ratio is defined in terms of P(D|H), the probability of the data given a hypothesis, or the likelihood of the hypothesis given the data, a frequency concept. Only in the 1933 paper did the classical NP point of view based on performance characteristics emerge, but the 1928 paper does not subscribe to inverse probability.

We can find an example of logical silliness on page 15. Rozeboom's refusal to accept the null in the absence of a significant result is countered by a "conditional syllogism" that is meant to back the acceptance of the null in this situation. In this manner any fallacious inference can be backed by a valid argument! Just add the additional premise: "If the premises, then the conclusion." Such a premise is unwarranted, and often false. I thought that Chow's Popperian leanings would have stopped him at this point: not proven false is not the same as proven true. As a Popperian at most he could have found corroboration for a whole disjunction of hypotheses around the null as "not (yet) falsified," if the test had been a severe one. This is in fact the use of significance testing that I myself favor.

I agree with Chow that attention to effect size is often because of a utilitarian interest, but I am not convinced with one of his central arguments for leaving classical NHSTP in place for theory-corroboration experiments. We may be willing to reject the statistical null just on the basis of having a significant result. But if we agree with Meehl's "crud" factor, that in the causal rush of things everything is, however minimally, correlated with anything, we need evidence of some effect size to conclude that there is evidence for the experimental hypothesis. A mere denial of the null is exceedingly weak, consistent with the tiniest shift in mean that may well be there because experimental controls are never mathematically perfect. This is to me the crux of the argument that the statistical null hypothesis is never true, that is, never experimentally realizable, even if the experimental null hypothesis is true.

Examples of misrepresentation abound in the chapter on Bayesianism. The probabilities of an exhaustive set of exclusive hypotheses, as in Table 7.1, should add up to one, especially if one uses them in the denominator of a posterior probability calculation; – contra note 1 on page 174 – evidence may be compatible with mutually exclusive hypotheses, and so on. But these are just silly mistakes that make the book unfit for a novice. What irritated me most about this chapter was the straw man Bayesian that is set up. This is not a serious engagement with the rich Bayesian literature on experimental design and analysis (Lad & Deely 1994). A basic misrepresentation is that in Bayesianism "empirical data are collected [in order] to ascertain the inverse probability of the hypothesis of interest" (p. 144). This is not the goal of collecting empirical data; conditionalization only comes up when the observations have been made. The goal may vary: an experiment may be chosen because it is likely to diminish one's own uncertainty about the explanation of a phenomenon, or because it is likely to diminish one's colleagues uncertainty. That is, choice of observational study or experiment is a matter of Bayesian decision theory, and not solely, as Chow assumes, of Bayesian confirmation theory.

What about significance testing? Chow's starting point, that one should evaluate its role within the whole scheme of inferences made in scientific research, is essential but, as I indicated above, I found his execution flawed. Statistical analysis has two parts: modeling (finding a [family of ] model[s] that may fit the data), and estimation (ranking the members of this family as to how well they fit the data). Significance tests come in to test whether the model is plausible at all (Box 1980). If the family of models is rejected, we are left with little, because the complement of the family is unwieldy. If there is no significant result, we are in business, not because we have shown the model to be correct but because we follow the Popperian methodological rule that we may use a model as long as we have not falsified it even though we tried hard. Admittedly, this turns Fisher on his Popperian head, but somewhere I still savor the thought that my construal is in a Fisherian spirit.

## Significance tests: Necessary but not sufficient

Louis G. Tassinary

*Environmental Psychophysiology Laboratory, Texas A&M University, College Station, TX 77843-3137.* **lou@archone.tamu.edu**
**red.www.nsf.gov/EHR/GERD/pff/fellows/tassinary-louis.html**

**Abstract:** Chow (1996) offers a reconceptualization of statistical significance that is reasoned and comprehensive. Despite a somewhat rough presentation, his arguments are compelling and deserve to be taken seriously by the scientific community. It is argued that his characterization of literal replication, types of research, effect size, and experimental control are in need of revision.

Over the past 10 years, Chow has been nearly alone his quixotic fight against the power-analysis juggernaut. *Statistical significance* (Chow 1996) is the culmination of his effort. In this book he confronts directly the blunderbuss attack on the null-hypothesis significance-testing procedure (NHTSP) and provides a conceptual framework for understanding why both the NHTSP and the proposed alternatives (i.e., statistical power, confidence intervals, and meta-analysis) continue to be misunderstood.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Before commenting on a few specific issues, let me make one general comment: Chow is to be commended for writing such a book. He is clearly fighting an uphill battle, and I hope the bulk of his arguments ultimately prevail in the midst of such silliness as the recent calls to ban the significance test (e.g., Hunter 1997). Chow's actual text, however, seems to distract from and even occasionally undermine the force of his arguments. The chapters are very repetitive, the figures of poor quality, and the tables confusing; there are also noticeable misspellings. The arguments warrant a better showing.

The crux of Chow's argument is that the NHSTP centers around the assessment of a particular conditional probability; that is, the probability of the data, given that the null hypothesis is true. By embedding this assessment within the context of a nested series of deductive syllogisms, it is possible to exploit the outcome of this assessment to test a particular theory-based prediction. Much of what Chow argues is not new. It was clearly articulated nearly two decades ago for both undergraduate (Walizer & Wiener 1978) and graduate audiences (Cook & Campbell 1979). What I believe is new is his novel suggestion that by taking the notion of conditional probability seriously we are driven to the conclusion that "no distribution based on $H_1$ is implicated in NHSTP" (p. 137). As Chow correctly points out, this simple conclusion has profound implications for our understanding of statistical inference generally, and places the burden of validity back squarely on the logic of the experimental design rather than on the elegance of formalisms (cf. Platt 1964).

Despite my agreement with his arguments against the use of either power analysis or Bayesian statistics to supplant the role of converging operations in the establishment of internal validity, I found Chow's rigid categorization of the different kinds of research (i.e., theory corroborative, utilitarian, clinic, and generality) somewhat anachronistic. Comparative research on the conditions and effects of various forms of action, and research leading to actual real world actions (Lewin 1946) was originally and is still currently the most common form of scientific research. The theoretical scientist sitting in his ivory tower advancing knowledge and the applied researcher out in the trenches building a better mouse trap have always been caricatures, but they are quickly becoming unrecognizable. Because of the confluence of theoretical and applied research it is counterproductive to preclude alternative indices such as effect size from having any role in theory corroboration. Indices such as this have clear implications for the sufficiency testing of quantitative models, whether neural or ecological, and it would be unfortunate if Chow's important insights about statistical power and conditional probability were discarded because of a failure to see the usefulness of quantitative indices in the process of model building and testing.

Given the well-argued position for the necessary (albeit limited) role of significance tests in the theory corroboration process, I was also struck by the offhand dismissal of literal replications as misleading (p. 141). If the goal of the NHSTP is to establish whether or not there is a difference worth explaining, then the literal replication must be as important, if not more so, in establishing the existence of an explanandum. The astronomical observation of a new comet, a report of a cure for AIDS, or the announcement of the successful achievement of cold fusion are all intriguing and have clear theoretical implications. Yet the inferential specificity of subsequent attempts to refine and test the theoretical implications of such reports hinges upon the ability of independent investigators in diverse laboratories to successfully replicate these results. Converging operations and/or constructive replications build upon but do not supplant the need for literal replications. This is especially true in fields where the experimentation is heavily dependent upon sophisticated procedures and elaborate instrumentation (cf. Church et al. 1996).

One final note. Chow uses Boring's (1954/1963) tripartite definition of experimental control as check, restraint, and guidance to explicate the methodological distinctions made by Cook and Campbell (1979) between a true experiment, a quasi-experiment, and a non-experiment. I found this discussion confusing and believe it stems from a misunderstanding of both Boring and Cook and Campbell. Boring defines guidance as the "alteration of the independent variable in accordance with precise known predetermination" (p. 113). This is clearly different from Chow's definition as "provisions to exclude procedural artifacts" (p. 76), something that applies equally well to all three senses of the experiment control. Cook and Campbell distinguish between the three types of inquiry in the following manner: in any inquiry there are elements that can be labeled treatments, outcome measures, and experimental units. The critical differences, however, are that investigators are unable to randomly assign experimental units to treatments in quasi-experiments; and in non-experiments, they are also unable to decide which outcome measures to acquire. Thus, the type of control is orthogonal to the type of investigation, although there is presumably a rough correlation between the overall degree of control and the degree to which an investigation approximates a randomized experiment.

## On various methods of reporting variance

Bruce A. Thyer

*School of Social Work and Department of Psychology, The University of Georgia, Department of Psychiatry and Health Behavior, Medical College of Georgia, Athens, GA 30602.* **bthyer@uga.cc.uga.edu**

**Abstract:** Chow's defense of NHSTP is masterful. His dismissal of including effect sizes (ES) is misplaced, and his failure to discuss the additional practice of reporting proportions of variance explained (PVE) is an important omission. Reporting the results of inferential statistics will be greatly enhanced by including ES and PVE when results are first determined to be statistically significant.

Chow (1996) can be read on at least two levels. On the first it is a persuasive response to the critics of null-hypothesis significance-test procedure (NHSTP), showing that many of their remarks are misplaced. In the second, it should be required reading for all students after completion of one or two courses in inferential statistics, as it forms an admirable foundation for understanding what statistics can and cannot do for the researcher. It will be an effective inoculum against a lifetime of erroneous understanding and application. Moreover it is a fairly simple text: complex statistical formulae are absent and examples of experiments used to illustrate the use of statistics are easy to follow.

Chow deals with some of the purported remedies for the abuse of statistics, conducting *a priori* analyses of statistical power before conducting a study, reporting effect sizes, and meta-analyses in a generally competent manner. His criticisms of meta-analytic procedures are particularly well-taken. Curiously, he fails to discuss another suggestion for augmenting conventional statistical reporting practices: that of routinely including the proportions of variance potentially explained, whenever one has statistically significant difference (see Good & Fletcher 1981; Hudson et al. 1985; Stocks 1987).

This is a common practice, or at least well understood, when reporting the Pearson *r* correlation coefficient. Calculating *r*2 yields the maximum potential predictive power that may be inferred from one variable to its correlate. A Pearson *r* of .30 between variables X and Y allows one to determine that a maximum of 9% of the variance of Y may be attributable to ("explained" by) its linear regression on X. An *r* of .80 yields an *r* of .64, and so forth.

Less common is reporting proportions of variance potentially explained (PVE) by independent variables subjected to inferential statistical analyses of differences among the dependent variables. Were this routinely done, "statistically significant differences" would never be erroneously assumed to be important or meaningful, or to imply anything of clinical significance. Take the instance of an anova where [$F(3, 984) = 14.8$; $p < .0001$]. Some less



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

sophisticated readers may infer powerful effects from this. Calculating the PVE (see Hudson et al. 1985) finds that only 4% of the variance in outcomes can be potentially attributable to the independent variable. This is a "potential" effect, for as Chow notes repeatedly, inferential statistics alone are not necessarily useful in determining the "causes" of any differences. Such determinations can only be meaningfully undertaken in the context of a given experimental design. Reporting [$F(3, 984) = 14.8$; $p < .0001$, PVE = .0432] aids immensely in the interpretation of this anova, compared to the more common practice. A further useful practice is to provide confidence intervals around a PVE.

Chow is quite correct in dismissing arguments that conventional means of reporting the results of statistical tests (e.g., Ns, *df*, *t* or *F* coefficient, alpha) should be *replaced* with ES (effect size) estimates. He misses the mark in arguing against *supplementing* conventional reporting with additional information about ES and PVE, *providing* conventional levels of statistical significance (e.g., $p < .05$) are attained. Nothing is lost, and there is much to be gained in reporting ES and PVE. Both are useful, ES in terms of standard deviation units, PVE in terms of explained variance; these are related but conceptually different concepts.

The *practical* importance of a given effect size or PVE is not solely a matter of statistical interpretation. Small effects can be quite important in some fields of practice (e.g., a nation-wide 5% reduction in automobile fatalities), less so in others (e.g., a 5% reduction in Beck Depression Inventory Scores for a treatment group of 40 clients). ES and PVEs obtained in the context of a poorly designed study do not alter its fundamental shortcomings. ES and PVEs added to a sound investigation can be very helpful.

Chow's discussion of statistical power is very well done. He omits a problem commonly encountered by applied researchers and those investigating new areas. To complete an *a priori* power analysis, one must establish the required alpha level, the statistical power desired, and *the effect size which is anticipated*. Researchers very often have little guidance in anticipating the magnitude of the effects (minimum expected or desired difference) from a given independent variable. This renders problematic the advocacy of using *a priori* power analysis to help establish the desired sample size of a prospective experiment. My experience is that this is the major reason that my students (and myself for that matter) fail to conduct *a priori* power analyses. Arbitrarily selecting a figure (a.k.a., guessing) does not seem like a sound practice.

The above minor points aside, Chow's *Statistical significance* is a very sound work which admirably clarifies many conceptual confusions on the role of inferential statistics in experimental research.

## Statistical inference: Why wheels spin

William S. Verplanck

*Professor of Psychology, Emeritus, University of Tennessee, Knoxville, Knoxville, TN 37916.* **wverplan@utk.edu**
**funnelweb.utcc.utk.edu/~wverplan**

**Abstract:** NHSTP is embedded in the research of "cognitive science." Its use is based on unstated assumptions about the practices of sampling, "operationalizing," and using group data. NHSTP has facilitated both research and theorizing – research findings of limited interest – diverse theories that seldom complement one another. Alternative methods are available for data acquisition and analysis, and for assessing the "truth-value" of generalizations.

Since 1955, the cumulative number of papers cited by Chow is a linear function of the year. A short fall-off in rate of citations per year in the early seventies is made up by an increased rate in 1990, when "questions began to be asked." The few references prior to 1955 are dated over a period of years: one in 1763, the rest from 1927 to 1954. These are "basic" works on the philosophy of science in books and at work.

George Miller's 1956 paper, entitled "The magical number seven, plus or minus two: Some limits on our capacity for processing information," often taken as the beginning of the "cognitive revolution," introduced "information processing" into psychology. "Information processing" became the subject for theorizing in "cognitive science."

In due course, first Popper and then Kuhn confirmed and endorsed, at least by implication, the "cognitive revolution," the new paradigm, ensuring that the theory of the "scientific empiricists" about theories and theory-testing would be wedded by still further theories (e.g., about "truth" and "falsification") to NHSTP.

Chow's citations may reflect only his decisions of what to cite. They nevertheless clarify the historical development of the wedding of "cognitive science" to inferential statistics.

Back in the forties, somebody measured the Hullian "e-bar-dot" by dint of 26 or so assumptions made about data from rats. Ever since, this reviewer has been suspicious of "assumptions" and is inclined to hunt them out, and to sound an alarm when they remain hidden.

Chow states the assumptions upon which NHSTP is based. In a complex series of arguments and presentations, he enables the reader to identify the kinds of data – and guesses (hypotheses, theories) – based on them that are suited to treatment by NHSTP.

Chow does not consider (a) the samples of individuals whose behaviors provide the data used with NHSTP for the development and testing of theories on the cognitive functions of the "mind" or "brain" of the human (and some other) species, (b) an evaluation of the concept of "operationalizing," (c) the relationship of statistical measures derived from group data to the behavior of any one individual in the group, or (d) the implications for both research and theory in psychology of NHSTP methodology for rejecting "untruth."

Sets of data used in the cognitive sciences are most often derived from the behavior of experimental groups made up of students at college/university A in the year X. Most often, these are students in psychology courses who were required to serve as subjects, or who served as subjects to make up for an exam they missed, or who were paid to serve, or who volunteered. Many of these may have liked, disliked, or not known the individual who "ran" them. Can findings on such a sample be replicated using a sample of students at college B, C, or D, in years X + 10, X + 20, or X + 30? Are these appropriate samples of the human species – even of young Americans? (Time for a *t*-test!) Are such samples appropriate for generating theories purporting to find out about "cognitive structures" of the brain or "mind"?

In "operationalizing" theoretical terms and statements, Boring and Pratt stood "operationalism" on its head. One published research – reprinted in a book of readings – "operationalized" the Freudian identification with the father-figure (or some such). He measured this by a count of spools packed in boxes by Stanford undergraduates in 195x, following the single instruction "pack the spools in the boxes" (with no further instruction or "feedback") until they stopped. How many other theoretical entities can this procedure "operationalize?"

"Operationalization" produces garbage; most psychologists have failed to note that Bridgman's "operationism" developed from the methods used in measuring "time" and "space" – before "Big Bang" theory.

Theorists use group data from samples using "operationalized concepts" to construct falsifiable (seldom falsified) theories about the structures of the human mind or brain that "process information." What structures? Where? In the mind or brain of that .56 infant of the 1.56 infants that statistics tell us is/are born to the N female graduates of Z University in the 25 years following graduation? What are the assumptions underlying the application of probability theory about the distribution of errors to such group data?

In analyzing group data, one adds a datum (information; 0, 1) on each identifiable thing a single subject does, first with another such datum-data, then with other things that this individual does;



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

these new "data" are then added to the equivalent new "data" of every other subject, producing newer "data." Such a procedure seldom fails to produce normally distributed "data," suitable for NHSTP. That the occurrences of each specific action (response) of each subject might show orderliness – "lawfulness" – not suitable for NHSTP methodology is ignored, even though this is easily demonstrated by research in both "psychophysics" and "learning."

The wedding of NHSTP with cognitive science, with the blessing of "theory-construction," has been successful: count the number, since 1955, of papers given at meetings and published in refereed journals, then duly summarized in "secondary sources." Count the number of kinds of memory discovered by "operationalizing."

NHSTP has enabled research to be carried out easily; computer programs can both produce and analyze data, all but untouched by human hands – or thought. Doing such research is easier than observing, counting, and classifying. That most findings are trivial, that the theories are all but irreconcilable, that answers to most questions lie buried under ten to the *n*th bytes of "information" is becoming evident. A cognitive scientist now wonders publicly whether they've been "spinning (our) wheels" for the past thirty years or so.

Behavioral science needs data on the individual behaviors of individual organisms, each finding "verified" – replicated – by data taken from a number of other individuals, one by one. The visual methods introduced by Tufte, non-parametric "quick and dirty," and descriptive statistics (excluding means and standard deviations), suffice in testing generalizations, confirming or disconfirming them.

## Four reasons why the science of psychology is still in trouble

Kim J. Vicente

*Cognitive Engineering Laboratory, Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario M5S 3G8, Canada.*
**benfica@mie.utoronto.ca www.ie.utoronto.ca/IE/HF/kim/home.html**

**Abstract:** Chow's monograph exhibits four prototypical symptoms of psychology's enduring scientific crisis: (a) it equates empirical science with statistical analysis; (b) it settles for qualitative rather than quantitative theories; (c) it ignores the role of ecological validity in the generalizability of theories; and (d) it puts rigid adherence to arbitrary but documentable rules over critical thinking about the meaning of results.

Chow's exceptionally well-written monograph shows why the science of psychology is in trouble, an opinion that has been consistently expressed by prominent psychologists over a disturbingly long period (e.g., Allport 1975; de Groot 1990; Gibson 1967/1982; Hammond et al. 1986; Loftus 1996; Meehl 1967; 1978; Neisser 1976; Newell 1973). Specifically, Chow's monograph exhibits four typical symptoms of psychology's enduring crisis.

***1. Empirical science = Statistical analysis.*** Chow equates empirical science with statistical analysis of data from highly controlled experiments. This view is partially conveyed in the very first sentence of the book: "To conduct empirical research is to engage in an exercise which requires conceptual, theoretical and *statistical* skills" (p. ix, emphasis added; see also p. 168). This view is re-emphasized in one of the last sentences of the book: "At a minimum an empirical research is good if it has statistical conclusion validity and inductive conclusion validity" (p. 187). Both of these opinions would come as a surprise to Nobel Laureate Konrad Lorenz (1973), who not only did not exhibit statistical skill in his research, but never even published a paper with a graph in it! Instead, Lorenz devoted his life to describing and identifying phenomena as they occurred in nature. His rationale was that naturalistic observation is a legitimate form of empirical science which should precede formalization, quantification, and controlled experimentation. Before one can meaningfully formalize, quantify, or experiment, one should identify a natural phenomenon that is worthwhile investigating in more detail, and categorize the dimensions of that phenomenon to know what should be manipulated experimentally. Lorenz's rich, multi-faceted view of empirical science contrasts with the impoverished, unidimensional view that Chow refers to as "*the* scientific procedure" (p. 42, emphasis added).

Perhaps because of a preoccupation with statistical analysis, Chow undervalues the role of theory and intuition in science. For example, he states that psychology is an "empirical discipline" (p. 164), that theories are speculative (pp. 46, 61), and that the term "theory" has grandiose connotations (p. 46). Holton (1988), a well-known historian of science, has observed and criticized this narrow view of science: "The younger sciences . . . are now (erroneously, in my opinion) trying to emulate the older physical sciences by restricting their area of investigation, even if artificially, to . . . phenomenic (empirical) and analytical statements" (p. 3). But as Holton's rich case studies illustrate, science is much more than this, encompassing naturalistic observation, qualitative description and categorization, inductive leaps of faith, and axioms that can never be empirically tested. Again, Chow's view of science is comparatively narrow, a fact that has important implications (see below).

***2. Settling for low-hanging fruit.*** Chow also claims that "the exact magnitude of the effect plays no role in the rationale of theory corroboration" (p. 96). This is certainly true, if the intent is to capture most existing theorizing in psychology. However, it overlooks the fact that there are a continuum of theories, allowing us to make predictions about the impact of the independent variables on dependent variables with increasing specificity. Areas on this continuum include: categorical, ordinal, interval, and point predictions. As the preceding quote makes clear, Chow believes that these last two categories play no role in theory corroboration. But surely a theory that makes more specific predictions (e.g., Simonton 1997) is a more mature theory, and one that we should be seeking (Meehl 1978)? By overlooking this point, Chow is settling for "low-hanging fruit" rather than striving to develop more sophisticated and powerful theories.

***3. A science of the laboratory.*** On the one hand, Chow claims that psychologists take the external validity of experiments very seriously (p. 92). On the other hand, he follows Ebbinghaus's (1885/1964) legacy, stating that ecological validity is detrimental to the validity of theory-corroboration experiments (pp. 102, 171). It is difficult to reconcile these two statements. The fact is that external validity has *not* been taken very seriously by many psychologists because of an effort to keep the experimental setting "pure" and cleansed of the rich details that characterize ecologically valid settings (e.g., Banaji & Crowder 1991). As a result, many psychological results do not generalize beyond the experimental laboratory. As Neisser (1976) puts it, "the artificial situation created for an experiment may differ from the everyday world in crucial ways. When this is so, the results may be irrelevant to the phenomena that one would really like to explain" (p. 33). And as Gibson (1967/1982) observed, "when a science does not usefully apply to practical problems there is something seriously wrong with the theory of the science" (p. 18).

***4. Cargo-cult science.*** Finally, Chow also holds "objectivity," "rigor," and "integrity" as the holy grails of scientific criteria. Perhaps the most extreme example of this attitude is the claim that treating $p = 0.048$ and $p = 0.052$ differently is "doing the right thing" (p. 97). The fact that there is no ontological basis for doing so (Rosnow & Rosenthal 1989) is given secondary importance (see also n. 3 on p. 118). Chow prefers "rigid adherence" (p. 97) to arbitrary but documentable rules over thinking critically about the meaning of results if the latter involves criteria that are not completely objective. In doing so, Chow (p. 168) overlooks the fact that science inevitably involves making decisions that involve intuition, aesthetics, and subjective preferences (Holton 1988).

The cause of this attitude may lie in the "physics-envy" that has plagued psychology since its inception. The result is cargo-cult



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

science – research that seems to follow the norms of scientific investigation, but that nevertheless misses something essential (Feynman 1985). The shackles imposed by such a restricted view of rigorous science may explain, for example, why the first 100 years of memory research largely served to confirm what the average middle-class third-grader already knows about human memory (Kreutzer et al. 1975).

***Conclusion.*** If we follow Chow's logical arguments, we will continue to have a science of psychology that holds a narrow view of science, that only seeks weak qualitative theories, that has little to say about activities outside of the laboratory, and that strives so hard to look like "real science" that it puts itself into an intellectual straitjacket. After over a 100 years of experience, we should know better than to repeat the errors of our predecessors.

ACKNOWLEDGMENT

The writing of this commentary was sponsored by research grants from the Natural Sciences and Engineering Research Council of Canada.

## Statistics without probability: Significance testing as typicality and exchangeability in data analysis

John R. Vokey

*Department of Psychology and Neuroscience, University of Lethbridge, Lethbridge, Alberta, Canada T1K 3M4.* **vokey@uleth.ca**
**www.uleth.ca/~vokey**

**Abstract:** Statistical significance is almost universally equated with the attribution to some population of nonchance influences as the source of structure in the data. But statistical significance can be divorced from both parameter estimation and probability as, instead, a statement about the atypicality or lack of exchangeability over some distinction of the data relative to some set. From this perspective, the criticisms of significance tests evaporate.

Chow (1996) equates statistical significance with the rejection of "chance influences" as an explanation for patterning or structure in the data, such a rejection then serving a very limited but important role as inductive evidence (in the form of corroborating an experimental implication) in a hierarchical, logical argument in support of a to-be-corroborated theory. He argues cogently that by correctly recognising the different levels of this logical argument, and the position of statistical significance within it, many of the criticisms of significance tests and equally the proposed remedies offered can be seen to be either irrelevant or misplaced. This laudable, point-by-point deconstruction and refutation of critics' arguments is in line with, but exceeds in depth those of other recent defenses of null hypothesis testing (Frick 1996; Greenwald et al. 1996; Hagen 1997; Macdonald 1997), as well as earlier attempts of his own (e.g., Chow 1988). As far as these arguments go, I am in thorough agreement. But they don't go far enough.

With very few exceptions (e.g., May et al. 1990), significance tests are routinely presented in textbooks as probability-based, binary statements about population parameters. Null hypotheses are stated in terms of population parameters (e.g., $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$), and the principal result of a significance test is the conditional probability of the data, given the null hypothesis, $p(\text{data}|\text{null})$ (which includes random sampling from a specified population distribution, typically normal). Critics and defenders alike, Chow included, appear to accept this representation as canonical, generally in the example form of a *t*-test for independent samples. It would appear that this conflation of significance testing with random sampling from populations and, hence, parameter estimation and probabilistic inference is responsible for much of the debate about significance testing.

*If* significance testing is accepted as synonymous with this one representation, then the criticisms of significance testing outlined by Chow, such as that the null can never be true, that what is really desired is the inverse probability of the null given the data rather than the data given the null, that confidence intervals around estimates of population parameters are preferable to a simple binary decision about them, and that the Bayesian or subjective probabilistic approach is preferable to the frequentist can appear reasonable, even conclusive. But the logic of significance testing, especially for theory-corroborative research, does not *require* parameter estimation and random sampling, as attested to, inter alia, by randomisation testing and other nonparametric (e.g., rank based) permutation tests (e.g., Edgington 1966; 1995; Fisher 1935; Hunter & May 1993; Kempthorne 1955; Pitman 1937a; 1937b; 1937c), and by the fact that for theory-corroborative research – the focus of Chow's exposition – random-sampling and inferences about extant populations are not necessary, may often be undesirable (e.g., Eysenck 1975; Mook 1983), and are frequently unobtainable (or at least intentionally unobtained) in behavioural research (e.g., Hahn & Meeker 1993). These approaches are not merely approximations to parametric tests, even though they are often presented as such; they represent a fundamentally different conceptualisation of statistical inference (e.g., Camilli 1990). Clearly, if there is no random sampling then there can be no estimates of and inferences about population parameters; the null hypotheses in these cases refer to *effects* in particular samples with the statistical validity provided by an assumption (or act) of random assignment rather than random sampling.

The idea can be taken further to eliminate also the dependency on probabilistic (random) considerations, eliminating the probability-based criticisms of statistical significance in the process. Rouanet et al. (1986) discuss significance testing as the assessment of what they refer to as the *typicality* of the data relative to some specified *set* with regard to some aspect or measurement. Although the usual probability calculus is used, no probabilistic considerations are involved. Instead, the 0–1 range of probabilities is used as a scale of typicality, and the resulting "*p*-values" are not taken as probabilities, but as the proportion of the different groups of observations from the specified set that are at least as extreme as the obtained one.

For example, with the canonical *t*-test in mind, consider two groups of 4 observations each. The result that the four highest scores all fall in one group is "significant" in that of the set of all possible ways of dividing 8 scores into two groups of 4, less than 5% (1 out of 70) would be this extreme. That is, the result is *atypical* of the specified set, and remains so regardless of the basis of forming the groups in the first place (i.e., whether or not any random process was involved). If, for example, the 8 scores were the heights of 4 adult males and 4 adult females, the result of the four tallest falling in the male group is still atypical of the permutations of the set, however highly probable or expected the result is. The same would be true if the result were coded as, say, a mean difference, a *t*-statistic, or an *F*-ratio; and of the 70 permutations of the scores, less than 5% of the statistics of the set were as extreme as the observed one.

The reference set need not be restricted to values directly observed. The atypicality or "significance" of a result or collection of scores can be computed relative to any set, including infinite sets. In these cases, the distribution of the set statistics would be obtained from the more traditional sampling distributions (e.g., normal, $\chi$-square, etc.) for the determination of the *p*-values. With the appropriate assumptions (e.g., random-sampling or randomisation), any of these *p*-values could be converted to probabilities, but it is not obvious what would be gained by such a move, unless the assumptions were true.

Similar arguments have been advanced by Draper et al. (1993), for whom significance testing is seen as an assessment of exchangeability (i.e., can the scores be seen to be exchangeable or equivalent over some distinction, e.g., sex, with respect to some aspect, e.g., height, relative to some set, e.g., the set observed, or some larger set?). From these perspectives, statistical significance is not about the rejection of "chance influences," but rather simply a statement about the presence or absence of structure (lack of



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

exchangeability) in the data. Ascertaining the putative *source* of the structure and its extension or generality is, as always, a function of other aspects of the research design and the theory-corroborative argument. It would appear to be this meaning of statistical significance or something like it that is implicit in the vast amount of behavioural and other research for which neither random-sampling nor assignment is claimed, but for which a presumably meaningful "statistically-significant" result is reported nonetheless.

## The *non*-significance of straw man arguments

Niels G. Waller[1] and Wesley O. Johnson[2]

*[1]Department of Psychology, University of California, Davis; [2]Division of Statistics, University of California, Davis, Davis CA, 95616.*
**ngwaller@ucdavis.edu psychology.ucdavis.edu/waller**

**Abstract:** We demonstrate that *Statistical significance* (Chow 1996) includes straw man arguments against (1) effect size, (2) meta-analysis, and (3) Bayesianism. We agree with the author that in experimental designs, $H_0$ "is the effect of chance influences on the data-collection procedure . . . it says nothing about the substantive hypothesis or its logical complement" (Chow 1996, p. 41).

Chow wisely reminds us, as have others before him, that (1) we must carefully distinguish substantive theory from statistical hypothesis (cf. Meehl 1978; see also Bolles 1962); that (2) as a decision procedure, NHSTP plays only a very limited role in empirical research; and that (3) "statistical conclusion validity and practical validity belong to two unrelated or independent domains (Chow 1996, p. 108). Few readers will quibble with these claims. Considered in globo, however, we suspect that Chow's book will invite much criticism because it is considerably more catholic and, like the fields at Rothamsted Experimental Station, it is filled with straw men.

Even the book's title is misleading because it suggests a thorough treatment of null hypothesis testing. Yet Chow focuses exclusively on NHSTP "in the context of experimentation because two prominent critics of NHSTP [apparently Cohen and Meehl] have made the point that their criticisms of NHSTP are directed to non-experimental studies" (Chow 1996, p. xi). We wonder whether Chow agrees with these critics – in the context of quasi- or non-experimental studies – that statistical hypothesis inference testing (like Cohen, we will avoid the acronym) "has not only failed to support the advance of psychology as a science but also has seriously impeded it" (Cohen 1994, p. 997) or that significance testing is "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" (Meehl 1967, p. 265)?

***Straw man arguments against effect size.*** In Chapter 5, Table 5.1, Chow offers four studies that putatively illustrate the "ambiguity arising from the dependence of statistical significance on the sample size" (Chow 1996, p. 90). For illustrative purposes, these data are reproduced in Table 1. Notice that there are several features of this table (which we have faithfully reproduced) that invite confusion. For example, notice that Chow uses lowercase Greek letters ($\mu_E$, $\mu_C$) to denote statistics – that is, sample estimates of population parameters – whereas in earlier chapters he used (the more traditional) Roman uppercase letters (e.g., $\bar{X}_E$, $\bar{X}_C$). Notice also that the Yes/No entries in the table have no apparent relation to the other columns. For instance, if we assume that the experimental and control conditions have equal variances and sample sizes then none of the test statistics surpasses conventional threshold values.

Nonetheless, Chow asks us to consider studies C and D because they supposedly illustrate the "incommensurate significance-size problem critique of NHSTP." According to this critique it is paradoxical to consider the results of Study C nonsignificant when the effect size in Study C is almost twice as large as that in Study D. Chow contends that the "incommensurate significance-size problem critique of NHSTP seems to be predicated on the assumption that the size of the effect is indicative of the degree of evidential support for the hypothesis offered by the data" (Chow 1996, p. 91) and he suggests that "it seems intuitively reasonable to assume that the magnitude of the effect size is indicative of the degree of evidential support" (Chow 1996, p. 94). The implication of this straw man argument is that, ceteris paribus, larger effect sizes provide greater probative weight than smaller effect sizes (Chow is arguing against this position).

Table 1 (Waller & Johnson). Chow's Table 5.1: *The putative ambiguity and anomaly of significance tests illustrated with four fictitious studies*

| Study | $u_E$ | $u_C$ | Effective size* $d = (u_E - u_C)/\sigma_E$ | Statistical test (e.g., *t*) significant? | df |
|---|---|---|---|---|---|
| A | 6 | 5 | .1 | Yes | 22 |
| B | 25 | 24 | .1 | No | 8 |
| C | 17 | 8 | .9 | No | 8 |
| D | 8 | 2 | .5 | Yes | 22 |

*J. Cohen (1987)

We believe that strong theories generate point predications (or predictions of functional forms, e.g., exponential growth) and weak theories generate range predictions (e.g., the correlation will be between .3 and .7, see Lykken 1991; Meehl 1990), although we realize that most social scientists eschew anything but ordinal predictions. Point and range predictions are easier to make when measuring instruments are linked to a common metric (Waller et al. 1996). We also believe that when investigators make theoretically informed predictions of effect sizes it is "reasonable to assume that the magnitude of the effect is indicative of the degree of evidential support." Observations that are close to their theoretically predicted values provide more support for a theory than those that are relatively distant (see Meehl 1990 for further discussion of this point). For instance, if scales E and F putatively measure constructs G and H, the latter being theoretically distinct, then $r_{E,F} = .9$ does not support the discriminant validity of E or F or the theoretical distinctiveness of G and H. Moral: *effect sizes can be too large.* Chow does not consider this issue because he contends that "only *qualitative* information is required in theory corroboration" (1996, p. 112, emphasis added).

***Straw man arguments against meta-analysis.*** Chow also believes that meta-analysis is seriously flawed as a research tool because "research quality is not deemed important" (p. 110). For instance, he claims "[m]any more meta-theoretical issues arise if meta-analysis is used as a theory-corroboration tool . . . [such as] the unjustifiable disregard for the quality of the research" (p. 112). We find these sweeping statements unfounded and puzzling because they suggest that Chow is unfamiliar with current meta-analytic practice.

Detailed procedures for weighting research quality have been discussed in the meta-analytic literature for more than a decade and a half (e.g., see Rosenthal 1984, pp. 54–62). In physics, for example, Hedges (1987) notes that differential weighting of parameter estimates is standard practice in meta-analytic summaries of physical constants and that physicists routinely disregard up to 40% of available studies. Psychologists have also paid attention to research quality and, interestingly, they find that the incorporation of questionable research (however defined) generally does not bias meta-analytic conclusions because biases which are introduced by poor studies typically average out with aggregation



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

(Lipsey & Wilson 1993; see Utts 1996, Ch. 25, for an excellent introduction to the topic of meta-analysis).

*Statistical significance* is brimming with syllogisms, minor premises, and other tools of formal logic and thus it is doubly surprising to hear Chow claim that "conceptual rigour is not an issue in meta-analysis, as may be seen from the fact that *some* meta-analysts deny the distinction between good and poor research (Chow 1996, p. 110, emphasis added). The illogical nature of this reasoning seems obvious to us and reminds us of a passage from Fisher's *Statistical methods and scientific inference.* Recounting Venn's diatribe against the rule of induction, Fisher noted that "it seems that in this chapter Venn was to such an extent carried away by his confidence . . . that he became uncritical of the quality of the arguments he used" (Fisher 1973, p. 28).

***Straw man arguments against "Bayesianism."*** Chow appears to treat "frequentistism" as if there were no subjective aspects (cf. Berger & Berry 1988). On the other hand, he also appears to be making the assumption that virtually all data of interest to him would be collected as random samples from normal populations with equal variances between samples. Nowhere in his treatise do we find mention of the importance of checking these assumptions in order to be reasonably confident in final inferences. His invocation of the central limit theorem would of course only apply to large sample situations in which case there would be no need to mention Student's *t* distribution (Hogg & Tanis 1993, p. 432).

Although sequential analysis, hypothesis testing, and decision theory are substantial areas of emphasis that are studied in both Bayesian and frequentist modes (Berger 1985; Berry & Lindgren 1996; DeGroot 1970; Ferguson 1967; Kass & Raftery 1995; Siegmund 1985), there is no requirement in Bayesianism or frequentistism that data be looked at sequentially, that hypotheses either be tested or not, or that decisions of any kind be made. If Bayesianism had caught on before frequentistism, standard statistical practice might well have required that the posterior probability for the null be less than or equal to .05 in order to decide in favor of the alternative. This would not be a property of the Bayesian approach; it would merely be a property of common statistical practice which utilized the Bayesian mode of inference. And in our opinion, it would be just as silly as the current practice of using a .05 cut off for *p*-values.

We would further argue that there are few single experiments for which the ultimate conclusion, reject or not, would convince the scientific community at large of a real effect. We expect that it will only be after repeated experimentation and an accumulation of evidence that the scientific community will reach some form of consensus as to the presence of a real statistical effect. This observation, it seems to us, puts the .05 criterion in its proper place. For example, if study after study, with moderate or small sample sizes, resulted in either posterior probabilities or *p*-values in a neighborhood of .06 and if the "effect" was both consistent across studies and of real importance (cf. Utts 1996, Ch. 24), it would seem foolish indeed to deny the existence of a real phenomenon.

We would finally argue that the personalistic vs. frequentist dichotomy alluded to by Chow is not correct. Both types of probability exist and there need be no resultant inconsistency. Frequentists and Bayesians would probably agree that probability associated with the sampling distribution of the data is frequentist. It is also likely that both groups would agree that the probability of rain tomorrow would have to be based on a subjective determination and could only be given a frequentist interpretation by thinking about multiple tomorrows, a possibly dubious construct. The only real issue here is that "diehard frequentists" would simply be unwilling to specify such a probability: if they were willing to do so, there would be no disagreement as to how to cope with the mix of subjective and so called objective probabilities from the prior probability and likelihood, which would be merged via Bayes Theorem, regardless of one's religious persuasion.

In summary, we agree with what we perceive to be Chow's cardinal thesis in *Statistical significance,* namely, that in experimental designs, $H_0$ is "the hypothesis about the effect of chance influences on the data-collection procedure . . . it says nothing about the substantive hypothesis or its logical complement (Chow 1996, p. 41). We only wish that Chow had made this point without senselessly attacking other important and useful statistical concepts, such as effect sizes and confidence intervals, meta-analysis, Bayesian posterior distributions, and point predictions.

## A viable alternative to null-hypothesis testing

Bruno D. Zumbo

*Departments of Psychology and of Mathematics, University of Northern British Columbia, Prince George, B.C. V2N 4Z9, Canada.*
**zumbob@unbc.ca quarles.unbc.ca./psyc.edgeworth2.html**

**Abstract:** This commentary advocates an alternative to null-hypothesis testing that was originally represented by Rozeboom over three decades ago yet is not considered by Chow (1996). The central distinguishing feature of this approach is that it allows the scientist to conclude that the data are much better fit by those hypotheses whose values fall inside the interval than by those outside.

Of all the concepts in scientific methodology, statistical significance testing (and in particular null hypothesis testing, NHT) has the noble distinction of simultaneously being: (a) criticized fiercely and disavowed, and yet (b) guarded as one of our sacred cows. This distinction can be clearly seen in Chow's (1996) book.

Even though they have been vast in number, the criticisms of NHT will continue to have little impact until something better is offered to the practising scientist. Chow discusses many of the commonly suggested alternatives to NHT (i.e., effect sizes and conventional confidence intervals), however he is clearly less than sanguine about them (for example see pp. 5, 87, and 98 of the book). The primary purpose of this commentary is to remind the readers of an alternative to NHT that I believe was not presented by Chow (or, for that matter, in most other discussions of statistical significance) and is a contender as a viable alternative.

***Workable alternative to NHT.*** It is important to remember that the alternative presented herein was first offered over 35 years ago (Rozeboom 1960) but it appears we have ignored it until recently. In fact, as in Chow (1996), Rozeboom's paper is often cited as being one of the first and most penetrating criticisms of NHT, but what is not often mentioned is that Rozeboom also offers a workable alternative to NHT in the form of confidence intervals (although it is essential to note that he does have a particular interpretation of them).

Briefly, as a metascientific foundation, following Rozeboom (1960), it can be shown that instead of using hypothesis testing as the ultimate mechanism through which we make decisions and commit ourselves to action on the basis of experimental data, we should use a methodology that reflects the scientific process as a cognitive activity. That is, the scientific process is a cognitive activity in which we make an appropriate adjustment in the degree to which we accept or believe knowledge claims given the empirical finding(s) at hand.

Conventional confidence intervals are encountered by most students of the social and behavioral sciences. However, to be useful as an alternative to NHT, confidence intervals should not be interpreted simply in terms of probability of coverage, as is often taught in introductory courses, but rather as an impartial simultaneous evaluation of all the alternatives under consideration for the current data. Quite simply, by using confidence intervals a scientist can conclude that the data are much better fit by those hypotheses whose values fall inside the interval than by those outside. Note that in Rozeboom's suggestion the commonly used 5% chance of error refers to the incorrect simultaneous dismissal of a large part of the total set of alternative hypotheses and is the total likelihood of error, not just of Type I.

Rozeboom's interpretation of confidence intervals is not the current norm of what is taught in statistical science textbooks or



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

used in scientific practise. A semblance of Rozeboom's suggestion can be seen, however, in Mosteller and Tukey's (1977) classic textbook wherein they introduce the one-sample t-test and end their discussion of testing against various hypothesized population means by stating, "It [a hypothesized population mean] might take on, in turn, all possible values, as when we seek a confidence interval" (p. 3).

Two points are noteworthy at this juncture. First, the Rozeboom interpretation of confidence intervals appears to me, at least on a surface level, to be akin to Bayesian credibility intervals (or highest density regions). In this same light, it should be noted that it is inherent in the Bayesian (and also the fiducial) approaches to confidence intervals that the parameter in question is, in some sense, random. It is important to note that this is not the case for more commonly used probability of coverage perspectives on confidence intervals, where only the interval is random and may or may not cover the fixed but unknown value of the quantity we are trying to estimate. Second, Rozeboom's description of confidence intervals could also be achieved in the framework of likelihood theory (Edwards 1972).

We take Rozeboom's suggestion one step further by systematizing the process of combining new results with existing knowledge. That is, rather than an unspecified reference to an adjustment in the degree to which we accept or believe knowledge claims, in some situations it is better handled more systematically by computational aids such as Bayes' theorem or empirical Bayesian methods.

Rozeboom's suggestion could accordingly be interpreted in our current climate of anti-NHT as suggesting that rather than salvage the sacred cow of hypothesis testing, we butcher the beast and move to a more nutritious feast offered to us over three decades ago. The main course at this feast would be an ample serving of easily used techniques that are useful to scientists, because these techniques would allow them to conclude that given the: (a) design for obtaining the data, (b) measurement error, and (c) stochastic and deterministic assumptions of the specified family of statistical models, the current data are much better fit by the parameters (and hence hypotheses) inside the confidence interval than by those outside.

# Author's Response

## The null-hypothesis significance-test procedure is still warranted

Siu L. Chow

*Department of Psychology, University of Regina, Regina, Saskatchewan, Canada S4S 0A2.* **chowsl@leroy.cc.uregina.ca**

**Abstract:** Entertaining diverse assumptions about empirical research, commentators give a wide range of verdicts on the NHSTP defence in *Statistical significance.* The null-hypothesis significance-test procedure (NHSTP) is defended in a framework in which deductive and inductive rules are deployed in theory corroboration in the spirit of Popper's *Conjectures and refutations* (1968b). The defensible hypothetico-deductive structure of the framework is used to make explicit the distinctions between (1) substantive and statistical hypotheses, (2) statistical alternative and conceptual alternative hypotheses, and (3) making statistical decisions and drawing theoretical conclusions. These distinctions make it easier to show that (1) $H_0$ can be true, (2) the effect size is irrelevant to theory corroboration, and (3) "strong" hypotheses make no difference to NHSTP. Reservations about statistical power, meta-analysis, and the Bayesian approach are still warranted.

### R1. Introduction

For ease of exposition, "NHSTP defence" is used to refer to the defence of the null-hypothesis significance-test procedure (NHSTP) presented in *Statistical Significance.* "Synopsis" refers to the synopsis of the NHSTP defence, and "rejoinder" refers to the present response to the commentaries. There are four main reasons for the wide range of verdicts on the NHSTP defence. First, commentators entertain disparate ideas about various aspects of empirical research. Second, logistic concerns of empirical research sometimes make it impossible to observe the nuances important to philosophers, logicians, or statisticians (**Kraemer, Mayo**). Such departures may be justified when no logical or mathematical rule is broken.

The third reason is the use of the collective terms "power analysts," "meta-analysis," and "Bayesian" to discuss critically particular versions of these techniques in order not to sound personal and to indicate that the criticisms are about the ideas, not their proponents. The fourth reason is the need to distinguish between a technique and the assumptions about empirical research held by experts who use it. If the NHSTP defence were a critique of Bayesian statistics, meta-analysis, and power analysis at the technical level, it would be necessary to direct the criticisms at their more recent, mature versions. The critique, however, is about what power analysts, meta-analysts, and Bayesians think about empirical research.

It is argued in the NHSTP defence that the validity of the theory-corroborative experiment is assessed in terms of conceptual, theoretical, and methodological criteria, not a numerical index, an account accepted by some commentators in general terms (**Boklage, Hayes, Tassinary, Thyer, Vokey**). Specifically, it is essential that the substantive explanatory (hence, not necessarily quantitative) hypothesis to be tested be consistent with the phenomenon to be understood in a nontautological way. The experimental hypothesis should be a valid deduction from the substantive hypothesis in the context of the specific experimental task. The exclusion of recognized alternative explanations is made possible when data are collected and analyzed according to the experimental design that satisfies the formal requirements of an inductive rule (e.g., Mill's [1973] method of difference). NHSTP is used to exclude chance influences as an alternative explanation of the data (i.e., to choose between chance and nonchance). Using the NHSTP outcome, the experimenter interprets the data with reference to the implicative relationships among the substantive, research, and experimental hypotheses (i.e., to isolate what the nonchance factor is). What the experimental data mean at the conceptual level is determined by the theoretical foundation of the experiment, not by statistics or any other nonconceptual concerns (e.g., the practical importance of the result).

The issues that will be discussed in the present response are (1) the propriety of using statistics in psychological research, (2) the formal structure of scientific investigation versus the sociology of science, (3) the differences between substantive and statistical hypotheses, (4) the validity of the hypothetico-deductive framework, and (5) some reservations about effect size, statistical power, meta-analysis, and Bayesian statistics.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

## R2. The historical perspective

The NHSTP defence is a rationalization of cognitive psychologists' modus operandi, namely, conducting theory-corroboration experiments in Popper's (1968a; 1968b) "conjectures and refutations" framework. The outcome of NHSTP is used to start the chain of syllogistic arguments that leads to the theoretical conclusion. The issue is whether or not it is wrong from the historical perspective.

**R2.1. Neyman and Pearson and the inverse probability.** Reading too much into a statement made by Neyman and Pearson (1928), I suggested that they subscribed to using the inverse probability. This was a mistake (**Gregson, Krueger, Kyburg, Mayo, Poitevineau & Lecoutre**).

**R2.2. Statistics and nomothetic research.** Statistics is used when the concern is about what can be said about a well-defined group of individuals (i.e., nomothetic) rather than of individuals as unique beings (i.e., idiographic). **Stam & Pasay** as well as **Verplanck** raise the possibility that the NHSTP defence perpetuates a historical error, namely, the error psychologists made when they were diverted from doing idiographic research as a result of using statistics. However, there are qualities that are found in all members of the group. Statistics is used to describe how these properties are distributed among the members of the group. For example, the mean of the group is sensitive to the magnitude of individual scores. To use the mean to represent the group is to describe how the individuals are distributed along the dimension in question, not to suggest that individual differences are not important. Statistics is used in nomothetic research to ascertain what is true of the group, despite individual differences.

The uniqueness of X is described in terms of how X differs from others in terms of demeanor, dress code, tastes, social skills, and the like. This description is meaningful only when there is information about the demeanor, dress code, tastes, social skills, and the like, common to the group. That is to say, statements in idiographic research are meaningful only against the backdrop of nomothetic research.

There is another reason why using statistics is not incompatible with idiographic research. Suppose that it is necessary to ascertain individual X's memory capacity. Not only is X different from other individuals, there are also intra-individual differences in different situations or on different occasions. It is legitimate to ask what X is really like, despite such intra-individual differences. Statistics (particularly NHSTP) can, and should, be used for such a purpose.

**R2.3. The hybrid nature of NHSTP.** The hybrid NHSTP is faithful to neither the Fisherian account nor that of Neyman and Pearson (**Gigerenzer** 1993). For example, from Neyman and Pearson's approach is adopted the practice of fixing the critical associated probability ($p$) before data collection and of making a binary decision about $H_0$. The features from Neyman and Pearson render NHSTP rigid and mechanical. From Fisher what is adopted is his appeal to only one numerically nonspecific $H_1$ as a complement of $H_0$. To critics, this Fisherian feature is responsible for discouraging numerically specific hypotheses in psychology when NHSTP is used. Hence, using NHSTP impedes theory development (**Gigerenzer**) and prevents researchers from engaging in two types of "statistical thinking."

The "rigid" and "mechanical" characterizations of NHSTP are correct. Nonetheless, "strict" is a better characterization than "rigid." The rigidity is necessary for inter-researcher agreement. The issue is really not that the decision is rigid, but whether it is well defined and appropriate. The meaning of the associated probability, $p$, of the test statistic is well defined in terms of the sampling distribution in question. It is appropriate as an index of the decision maker's strictness.

Three issues seem relevant when one asks why $p = .048$ is treated differently from $p = .052$ (**Krueger, Vicente**). First, the choice of $\alpha = .05$ is called into question. The answer is simply that the rationale of NHSTP is not affected if any other value is chosen for $\alpha$. Second, the importance of using $\alpha$ in a strict manner may be seen by considering why it is important for a teacher to maintain a consistent passing grade. The third issue concerns why it is necessary to fix the $\alpha$ level before data collection (**Lewandowsky & Maybery**). The reason is that all design and procedural decisions about the experiment are made with reference to $\alpha$ as the criterion of strictness used to reject chance. Had a different $\alpha$ value been used, concomitant changes in the design or procedure would have to be made.

It bears reiterating that "mechanical" is not a derogatory term. A mechanical procedure is one that guarantees the same outcome if it is carried out properly. Hence, using NHSTP does not render an experiment a "mindless" exercise. Cognitive psychologists do engage in **Gigerenzer**'s Type I "statistical reasoning," namely, choosing among alternative statistical procedures in an informed way. Nor does using the mechanical NHSTP release the researcher from the need to consider other conceptual, theoretical, and methodological factors.

## R3. The Popperian structure

NHSTP is defended by illustrating its important, though very restricted, role in the theory-corroboration experiment in Popper's (1968a; 1968b) "conjectures and refutation" perspective. It is not clear why **Shafto** denies NHSTP's contribution. The defence can be strengthened by settling the following issues suggested by **Erwin, Glück & Vitouch, Gregson, Nester,** and **Waller & Johnson:** (1) the effect of auxiliary assumptions on using modus tollens, (2) the sociology of science, (3) the invalidity of affirming the consequent, and (4) the neglect of theory discovery.

**R3.1. The implication of auxiliary assumptions on modus tollens. Glück & Vitouch** refer to Folger's (1989) point that the experimental expectation is the implication of the conjunction of the substantive hypothesis and additional auxiliary assumptions. Hence, modus tollens does not guarantee the rejection of the experimental hypothesis because the experimenter may blame the auxiliary assumptions. Chow's (1989) reply was that a responsible, well-informed, and noncynical experimenter should have good reasons not to blame the auxiliary assumptions in the face of an unexpected result. The reasons are (1) the specificity of the substantive hypothesis, (2) the methodological assumptions commonly held by workers in the same area of research, and (3) the well-established theoretical ideas in cognate areas. If there are good reasons to suspect any of the auxiliary assumptions, the experimenter is obliged to con-



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

duct additional experiments to substantiate the suspicion. Using auxiliary assumptions does not mean being cynical or cavalier toward research.

**R3.2. The formal structure versus the sociology of science.** The issue of the sociology of science is raised (**Gregson**) because scientists do not behave in the way envisaged in the Popperian perspective. This sentiment is reinforced by **Glück & Vitouch**'s reference to Lakatos's rendition of Popper's framework. Note that the distinction is not made in the sociology argument about the formal requirement of a particular ideal of knowledge and the activities or psychology (**Blaich**) of the professionals in some disciplines. Popper's is a defensible account of the former if the ideal consists of the following features:

1. Knowledge evolves by reducing ambiguity.
2. Conclusions must be justified with valid empirical evidence.
3. There are well-defined criteria for settling inter-researcher disagreement.
4. Objectivity is achieved when the critical criteria are independent of theoretical preferences of the disputants.

**R3.3. The invalidity of affirming the consequent.** Given a conditional syllogism, it is not possible to draw a definite conclusion about its antecedent when its consequent is affirmed. This is the case when $H_0$ is rejected. The interim solution in the NHSTP defence is that recognized alternative explanations are excluded by virtue of the experimental controls. How does the researcher know what constitutes adequate experimental controls (**Erwin**)? The researcher is guided by the formal requirement of Mill's canons of induction.

**Erwin** may legitimately push the point further and ask what the researcher should do if a confounding variable is discovered despite having dealt with the "auxiliary assumption" issue (see sect. 3.1). The solution is not just to appeal to another statistical index in the study in question. It is to conduct another experiment designed specifically to examine the confounding variable. The three embedding syllogistic arguments, whose sufficiency as a means to establish "warranted assertibility" is called into question by **Erwin,** are also involved in the new experiment.

**R3.4. The mechanical syllogistic arguments.** Drawing a theoretical conclusion from the data with a series of three embedding arguments is a mechanical exercise (**Vicente**). However, it does not follow that the hypothetico-deductive framework is antithetical to critical thinking. The experimenter's critical thinking, intuition, and subjective preferences have important roles to play in proposing the substantive hypothesis, devising the experimental task, choosing the appropriate experimental design, and the like. At the same time, it is important that the influences of these subjective factors are neutralized in data interpretation. It is for this reason that experimental controls, as a means of excluding alternative explanations, are necessary. It is also for this reason that the NHSTP defence is based on the theory-corroboration experiment (**Waller & Johnson**).

**R3.5. Converging operations and replication.** The various theoretical properties of a hypothetical structure envisaged in a cognitive theory have to be substantiated. The series of experiments designed for such purposes constitute converging theory-corroboration operations. It is emphasized in the NHSTP defence that these theory-corroboration experiments are not literal replications of the original study, because they differ from the original study as well as among themselves. What should also be said is that literal replication has a different role, a point made by **Tassinary.** Specifically, replication studies are essential for ensuring that the new discovery is not a fluke. NHSTP is used in each of the replication studies in the same way it is used in theory-corroboration studies, namely, to exclude chance influences as an explanation.

## R4. The nature of the substantive hypothesis

Substantive hypotheses are speculative explanations proposed to explain phenomena. Research data are collected to substantiate these hypotheses. There arises the distinction between the instigating phenomenon and the evidential data of the hypothesis as well as the following issues: (1) alternative conceptual hypotheses versus a statistical alternative hypothesis, (2) the nature and role of the effect size, (3) the nature of a good explanatory hypothesis, and (4) the numerically nonspecific versus numerically specific statistical hypotheses.

**R4.1. Instigating phenomenon versus evidential data.** Phenomenon P cannot provide the substantiating evidence for the hypothesis that explains P itself in a nontautological way. For example, the hypothesis "snake phobia" is proposed to account for the instigating phenomenon of an individual's irrational fear of snakes. [See Davey: "Preparedness and Phobias" *BBS* 18(2) 1995.] Consequently, this reaction to snakes cannot be used as evidence for the "snake phobia" explanation without rendering the argument circular. It is necessary to distinguish between the phenomenon of which the substantive hypothesis is an explanation (the instigating phenomenon) and the data that are used to substantiate the hypothesis (the evidential data).

**R4.2. Alternative conceptual hypotheses versus alternative statistical hypotheses.** While the substantive hypothesis explains the instigating phenomenon, the experimental hypothesis describes what the experimental data should be like. When the experimental hypothesis is expressed in statistical terms, it is the statistical alternative hypothesis, $H_1$. This is very different from **Zumbo**'s recounting of Rozeboom's (1960) view of what $H_1$ is. A successful substantive hypothesis (T) is one that makes the to-be-explained phenomenon understandable. The experimental hypothesis serves as the criterion of rejection of T in the sense that T is deemed untenable if data do not match what is said in the experimental hypothesis. The statistical null hypothesis is used to ascertain whether it is possible to exclude the explanation in terms of chance influences. **Dar** however, finds the distinctions among the substantive, research, experimental, and statistical hypotheses unnecessary.

Of interest is the fact that the substantive hypothesis (e.g., the view that the short-term store retains acoustic-verbal-linguistic information) takes the form of delineating the nature of the psychological structures or mechanisms underlying the phenomenon of interest. These are qualitative specifications, not quantitative stipulations. Moreover, the experimental hypothesis is an implication of the hypothetical structure in a particular experimental context. This



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

context is not quantitative (e.g., acoustically similar and acoustically dissimilar materials are used in the experimental and control conditions, respectively). Seen in this light, the insistence that a "strong" hypothesis is one that gives a specific numerical value to a parameter (**Vicente, Waller & Johnson, Zumbo**) raised several issues.

**R4.3. The nature of a good theory.** First, what is the criterion of a "strong" hypothesis apart from the fact that it give a specific nonzero parameter value? The characterization is not informative if the criterion for being "strong" is not independent of its being numerically specific. Second, a hypothesis that is "strong" in this sense is not necessarily a testable or a more informative hypothesis, as can be seen from Propositions [P1], [P2], [P3], and [P4]:

> There will be 1 inch of snow. [P1]
> There will be snow on Christmas day. [P2]
> There will be 1 inch of snow on Christmas day. [P3]
> There will be snow on Christmas day between 2 and 3 p.m. [P4]

Although [P1] is numerically specific, it is not testable because it is not clear under what condition [P1] can be shown to be wrong. [P2] is a testable hypothesis even though it is not numerically specific in the way [P1] is specific. It is true that [P3] is testable and more informative than [P2]. However, the superiority of [P3] over [P2] is a matter of specificity, not of being quantitative for the following reason. [P4] is similarly superior to [P2] because it is more specific. [P3] is "stronger" than [P4] in the sense envisaged by critics of NHSTP. However, it is not possible to say which of [P3] and [P4] is more informative. What can be said is that they are informative in different senses. The moral of the story is that even if it were possible to ignore the circularity problem (see sect. R4.1), the "strong" characterization is not an appropriate, let alone the only, criterion for theory assessment (**Waller & Johnson**). A more serious objection to the "strong" characterization is that it is not clear that a "strong" hypothesis is a necessarily explanatory hypothesis, as may be seen from [P5]:

> Every six hours of practice will improve performance by 3 points. [P5]

Suppose that [P5] is an empirical generalization of practical importance established after numerous meticulous replications, and it gives rise to a specific numeric parameter (**Bookstein**). Does it explain why such a functional relationship exists between practice and performance? [P5] itself invites an explanation. To explain functional relationships such as [P5], appeals to hypothetical mechanisms are inevitable. Specific theoretical properties are attributed to these mechanisms. Theory-corroboration experiments are conducted to substantiate these theoretical properties. This approach is different from, as well as superior to, the operationalization suggested by **Verplanck.** An insistence on using the "strong" criterion or indifference to hypothetical mechanisms (**Bookstein**) may actually impede theory development if psychologists stop asking the "Why" questions about statements like [P5].

Recognizing that the substantive hypothesis is more than a functional relationship between variables, the researcher would have to consider a number of issues in a light different from that envisaged by critics of NHSTP. For example, it becomes necessary to distinguish between an efficient cause and a material (or formal) cause. Hence, there are important differences among experimental, quasi-experimental, and nonexperimental research (**Palm**), as well as differences between utilitarian and theory-corroboration experiments. Specifically, while the outcome of the experimental manipulation is the phenomenon of interest in the utilitarian experiment, it is not so in the theory-corroboration experiment. Although the efficient cause is the concern of the utilitarian experiment, only the material (or formal) cause is important to the theory-corroboration experiment. These considerations change the complexion of the issues related to the nature of the statistical hypotheses, effect size, and statistical power.

## R5. More ado about the null hypothesis

It is necessary to belabour the point that $H_1$ is not the substantive hypothesis because it has not been taken up in the commentaries in the discussion of (1) the nature of $H_0$, (2) the testing of a numerically specific difference between test conditions, and (3) how to ascertain that the parameters are the same in two different conditions.

**R5.1. The nature of $H_0$.** Some commentators reiterate the criticism that, as a result of using NHSTP, it is easy to support weak hypotheses. This criticism was predicated on the assumption that the null hypothesis is a straw man to be rejected because it is always false (**Rindskopf, Swijtink, Waller & Johnson**). Critics of NHSTP seem to have in mind a hypothesis that explains or describes a phenomenon that is a complement of the to-be-explained phenomenon. They are satisfied that $H_0$ can never be true when it is shown that such a complementary phenomenon is not possible.

$H_0$ is neither an explanation nor a description of a phenomenon that is complementary to the phenomenon to be explained. Rather, it is derived from the complement of the substantive hypothesis in exactly the same way that $H_1$ is derived from the substantive hypothesis. $H_0$ can be true (and should be true in a properly designed and conducted experiment; see **Lewandowsky & Maybery**) because it is a prescription of what the data should be like if what is said in the substantive hypothesis is not true and chance influences alone determine the pattern in the data.

Apart from the fact that $H_0$ is not a straw man, it is never used as a categorical proposition. Instead, it appears as the consequent in [P6] and the antecedent in [P7]:

> If chance factors alone influence the data, then $H_0$ is true. [P6]
> If $H_0$ is true, then the sampling distribution of differences has a mean difference of zero. [P7]

The cogency of the commentaries is unclear when no attempt has been made to deal with [P6] and [P7]. For example, it is neither the case that [P6] is silly (**Swijtink**) nor that [P7] is inappropriate (**Frick**). [P6] is a statement about what should follow solely from chance influences in the case of the completely randomized one-factor, two-level experiment.

**R5.2. The case of expecting a numerically specific parameter.** An objection to NHSTP voiced in the commentaries is that in resting satisfied with "$H_1$: $\mu$ 0" or "$H_1$: $\mu > 0$" or "$H_1$: $\mu < 0$," the researcher is distracted from developing a "stronger" hypothesis that makes it possible to say "$H_1$: $\mu_E$



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

$- \mu_C = 5$," instead of "$H_1$: $\mu_E = \mu_C$." **Zumbo,** as well as **Harris,** subscribes to Rozeboom's (1960) view that there are multiple $H_1$'s with specific nonzero values. A difficulty with this position (over and above the one discussed in sect. R4.3) may be seen by considering the situation that suggests "$\mu_E - \mu_C = 5$."

The experimenter is justified in expecting a difference of 5 between the experimental and control conditions when it is an implication of a functional relationship like [P5] or the result of a computer simulation (**Lashley**). However, the decision about statistical significance is made on the basis of one sampling distribution of differences. Hence, this expectation of "$\mu_E - \mu_C = 5$" is not represented by "$H_1$: $\mu_E - \mu_C = 5$," but by "$H_0$: $(\mu_E - \mu_C) - 5 = 0$" because the numerator of the *t* statistic is "$(\mu_1 - \mu_2) - 5 = 0$" (Kirk 1984), and the denominator is the standard error of the difference.

**R5.3. Using $H_0$ to ascertain the equivalence between two conditions.** Some commentators suggest that it may be too negative to characterize a nonsignificant result as a failure to reject chance influences (**Frick**). Instead, NHSTP can be used in a positive way to ascertain or accept a properly drawn null hypothesis (**Bookstein**). The suggestion to use NHSTP to accept a null hypothesis is reminiscent of Rogers et al.'s (1993) "nonequivalence null-hypothesis" approach, the purpose of which is to ascertain statistically that the parameters (e.g., the means) from two conditions are equivalent.

The "non-equivalence null-hypothesis" approach is debatable for the following reasons. First, in view of the role played by the sampling distribution of the test statistic in tests of significance, a significant result is one that is deemed too unlikely to be the result of chance influences. What can it mean to say that a test statistic is "significant by chance" (Rogers et al. 1993, p. 554)? Rogers et al. (1993) seem to have conceptualized their equivalence test at a level of abstraction different from that of tests of significance.

Second, a nonsignificant result is made unambiguous if statistical equivalence is achieved when the confidence interval is included in the equivalence interval (Rogers et al. 1993). At the same time, the equivalence between two conditions is deemed established when the confidence interval falls within the equivalence interval. The difficulty of this position is that the equivalence interval is determined in terms of practical or clinical criteria, not statistical ones. The width of the equivalence interval is sensitive to the context, which includes, among other things, the researcher's vested interests. Objectivity becomes a concern, especially if the equivalence interval is determined after the significance test is carried out.

In short, questions about the ambiguity of the result of NHSTP are questions about data stability, for example, whether or not (1) the measurements are made properly, (2) subjects are selected or assigned properly, and (3) subjects are given sufficient training and the like. These are not statistical concerns. Nor can they be quantified. Hence, the equivalence interval cannot disambiguate the ambiguity of the statistical decision.

**R5.4. Experimental expectation and $H_0$.** An implication of Schneider and Shiffrin's (1977) model of automatic detection is that the subject's reaction time to a target is the same regardless of the set size (i.e., the number of items in the briefly shown visual display). In other words, the implication of the automaticity hypothesis is that there is no effect of set size (viz., $\mu_1 = \mu_2 = .. \mu_k$), and it is indistinguishable from the null hypothesis. Consequently, it seems that accepting the null hypothesis is more than accepting chance explanations (**Bookstein, Frick**).

There are two reservations. First, in view of the fact that NHSTP is based on the sampling distribution that is predicated by chance influences, it is inherently impossible to decide whether the absence of the set-size effect in Schneider and Shiffrin's (1977) study is the result of automatic, parallel detection or of chance influences. Findings of this kind become less ambiguous, however, when the $H_0$-like experimental expectation is placed in a mutually exclusive and exhaustive relationship with the expectation of a competing hypothesis. For example, the expectation of the serial controlled search model of target identification is unlike $H_0$. In such an event, the emphasis is on rejecting the serial controlled search, not on accepting the $H_0$-like automaticity hypothesis. The second reason is that it should be possible to derive an experimental expectation that is unlike $H_0$. The experimenter's inability to do so indicates that the substantive hypothesis is not as well defined as it should be.

## R6. The statistical alternative hypothesis, effect size, and statistical power

Apart from the issue of whether or not $H_1$ is the substantive hypothesis, there is also the question of its exact role in NHSTP. It is not possible to talk about effect size or statistical power if $H_1$ has no role in NHSTP. There are also two intertwining issues, namely, the graphical representation of effect-size or statistical power and the level of abstraction involved.

It is customary to discuss the effect size and statistical power in the context of the *t* test. Moreover, the discussion is carried out in the context of a distribution for the control condition and another one for the experimental condition. These two are distributions at the level of raw scores, as witnessed by the fact that the effect size is defined as the difference between the experimental and control means in units of the standard deviation of the control condition. They are labeled the $H_0$ and $H_1$ distributions, respectively. The effect size is represented by the distance between the means of the two distributions. Although the $H_1$ distribution is not used in making the decision about statistical significance, it is essential in defining the effect size and statistical power. This account of the *t* test will be called the "customary account" henceforth.

What actually takes place in the *t* test is not what is described in the customary account. The $\alpha$ level is defined in terms of the sampling distribution of differences, not the distribution of population scores. Nor is this sampling distribution about, or based on, the control condition. This is true not because psychological theories are not "strong" or because psychologists are not "bold" when they propose their hypotheses (**Gigerenzer**'s Type II "statistical reasoning"). Hence, any appeal to computer simulation for insights about the expected effect size becomes moot (**Lashley**). That is, even if the "strong" theory argument were not problematic for the reasons given in section R4.3, it is still not possible to represent the $H_1$ distribution in a way that reflects properly the probability basis of the *t* test. In other words, "effect size" and "statistical power" cannot be de-



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

fined at the level at which the statistical decision is carried out.

In short, to accept the customary account, it is necessary to show why the probability basis of the *t* test is not the theoretical sampling distribution of differences between two means. In the event that it is not possible for critics to do so, they have to provide a valid reconciliation between the customary and NHSTP accounts.

To the extent that the NHSTP account of the probabilistic basis of the *t* test is not refuted, the questions about the customary account raised in the NHSTP defence remain, particularly those about the dependence of effect size and statistical power on the $H_1$ distribution. **Lashley**'s commentary seems to be an attempt to reconcile the customary and NHSTP accounts by suggesting that the power analysis is meant to deal with a different stage in the inferential process. The power analytic argument is about stage 1 (at which two distributions are involved), whereas the test of significance is an exercise in stage 2 that utilizes only one distribution. Moreover, power analysis predicts the outcome of NHSTP.

**Lashley**'s effort raises the following questions: Do the two stages belong to the same level of abstraction? What is the basis of the researcher's ability to predict the lone distribution in stage 2 from the two distributions from stage 1? How is the prediction possible when nonstatistical influences on data stability are not taken into account (e.g., the difficulty of the experimental task, the amount of practice available to the subjects, etc.)? What additional numerical information is provided by the effect size that is not provided by the test statistic? How can the sample size be determined in the mechanical way suggested in power analysis? How is the magnitude of the effect related to the validity of the theoretical statement about a material or formal cause?

## R7. Some issues about the effect size

Effect size is of interest because it seems to indicate the amount of evidential support provided by the data for the hypothesis (the "evidential status of the data") or of the practical importance of the data (the issue of "practical validity"). The question of evidential status takes on a different complexion when distinctions are made between (1) the substantive and statistical hypotheses, and (2) the formal (or material) and efficient causes.

Consider Sternberg's (1969) study of short-term memory storage. He manipulated the memory set size (viz., 1, 2, or 4 digits) and found that subjects' correct reaction times increased linearly with increases in the size of the memory set. One can (and very often does) say that the manipulation of the set size was the cause of the increase in reaction times. Nonetheless, there is a less misleading way to describe the functional relationship between set size and correct reaction times.

In manipulating the memory set size, Sternberg (1969) provided the memory system with different contexts to operate. The increase in reaction times when given larger memory set sizes is a reflection of a property of the short-term store (viz., its inability to handle multiple items simultaneously). In other words, the observed functional relationship reveals what Aristotle would call a "material cause" or a "formal cause." This is different from an efficient cause (e.g., exerting force to move a stationary object), which is what critics of NHSTP have in mind when they talk about effect size. The material or formal cause of interest to cognitive psychologists is not ascertained by statistical significance or effect size (**Lewandowsky & Maybery**). Instead, it is determined by the validity of the series of embedding syllogisms. Furthermore, the effect size gives no information that is not available in the test statistic that is used to decide whether chance influences can be excluded. The conclusion is that effect size, (i.e., the magnitude of an efficient cause) is irrelevant to theory corroboration.

**Maher**'s suggestion to prepare and use actuarial tables for utilitarian research is appropriate for reasons of practical validity. The success of such an approach depends on having some valid, well-defined, nonstatistical criteria developed independently of the effect size itself (see also **Kihlstrom**). The more immediate lesson is that regardless of the index used, the effect size on its own (or any other statistical index) is informative of neither the practical validity nor the evidential status of the data (**Nester**).

## R8. Statistical power

The validity of power analysis can still be questioned because it has not been shown why the NHSTP account is incorrect or how the NHSTP and customary accounts may be reconciled. For the sake of argument, assume that the customary account had not been problematic. Power analysis is meant to be used to disambiguate the decision about statistical significance. The ambiguity arises because statistical significance depends on sample size, effect size, and $\alpha$ level. The power of the test is used to determine the correct sample size for a required effect size at the chosen $\alpha$ level. The decision about statistical significance is said to be unambiguous under such circumstances.

Suppose that the sample size stipulated in the statistical power table is 25, given that the expected effect size is 0.85 and the power of the test is .95 with a set at .05. How can one be sure that the result is unambiguous when it is not known whether the 25 subjects have been given the proper training on the task? Are they given enough trials in the experimental session? These questions become more important, regardless of the decision about statistical decision, when fewer than 10 well-trained subjects are typically tested in multiple 300-trial sessions in the area of research in question (**Lewandowsky & Maybery**). It is simply not clear how the numerical index, statistical power, can confer validity on matters that are not quantitative in nature.

## R9. A recapitulation of some reservations about the Bayesian approach

Psychologists often propose hypotheses to explain phenomena (see **Bookstein** for an exception). The minimal criterion for accepting an explanatory hypothesis is that it should be consistent with the phenomenon to be explained. Given the distinction between the instigating phenomenon and the evidential data in section R4.1, it can be seen that the data collection procedure is neither the criterion used to assess the validity of the hypothesis nor the unexplained phenomenon itself. Such a scenario is characterized as the "phenomenon → hypothesis → evidential-data" sequence in the NHSTP defence.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Also important in the defence of the NHSTP is the fact that, regardless of the experimenter's theoretical biases or preferences, the hypothesis to be corroborated is given the benefit of the doubt when experimenter derives the research and experimental hypothesis, designs the experiment, and tabulates the data. That is, what the data indicate is not affected by what the experimenter thinks (or feels) about the hypothesis before the experiment. This is ensured partly by stipulating the size of the data set (viz., the number of subjects, the number of trials per session, the number of sessions, etc.). Nor is the data collection procedure adjusted as a result of periodic examinations of the data accumulated. This account of psychological research is different from the scenario Bayes had in mind.

Bayes's concern was with a situation in which the hypothesis was about the outcome of the data collection itself. The size of the data set was ill defined. There was nothing to explain, and there was no criterion for assessing whether the subjective degree of belief in the outcome of data collection exercise is correct. Would a data analysis procedure based on such a scenario applicable to the "phenomenon → hypothesis → evidential-data" sequence? Given the Bayesian overtone of Rozeboom's (1960) alternative, this issue also applies to **Zumbo**'s suggestion.

The derivation of the posterior probability from the prior probability in the context of new data is not questioned in the NHSTP defence at the mathematical level. Instead, the issue raised concerns whether the Bayesian theorem is appropriate for analyzing data about the validity of the "phenomenon → hypothesis → evidential data" sequence. How can this nonmathematical issue be dealt with in the new Bayesian developments? The more serious consideration is an implication of the Bayesian theorem about data interpretation.

It was Bayes's practice to obtain the posterior probability by adjusting the prior probability. The extent to which the data change the prior degree of belief in the hypothesis depends on the prior belief itself. The data have less weight the higher the prior degree of belief, a point not taken into account by **Snow.** In non-Bayesian terms, this practice amounts to saying that the theoretical importance of the data depends on the researcher's degree of belief in the hypothesis before the experiment. This is antithetical to objectivity, and it is shown in the NHSTP defence why there is no reason to do so.

This objection would not apply to new Bayesian approaches if they no longer used the Bayesian theorem for such a purpose. **Rouanet** seems to suggest another possibility. The Bayesian exercise is still the derivation of the posterior probability from the prior probability. However, a "noninformative" Bayesian would assume a "state of ignorance" about parameters when choosing the "prior distribution." The posterior distribution then expresses the evidence provided by the data, presumably not contaminated by any nonzero prior probability.

There is a close relationship between the experimental design and the test of significance. For example, the *t* test and anova are used for experiments that use the one-factor, two-level and the one-factor, multilevel design (or factorial designs), respectively. Are these issues important in the "noninformative" Bayesian approach? How are the various aforementioned meta-theoretical and methodological considerations met in the "noninformative" Bayesian approach?

## R10. Further ado with meta-analysis – psychometric meta-analysis

**Snow** may be referring to Glass et al.'s (1981) meta-analytic approach when he suggests that specific information about the associated probability, *p*, may be useful. The issue of incommensurability was one of the difficulties of Glass et al.'s (1981) approach. Even though a group of studies is all about the same phenomenon, it is inappropriate to combine them in meta-analysis for an overall test of significance because it is inappropriate to mix apples and oranges.

Representing the psychometric meta-analytic orientation (Hunter & Schmidt 1990; Schmidt 1996), **Hunter** points out that studies dealing with the same independent and dependent variables enter into the meta-analysis only to obtain a better estimate for the parameter, not to do a test of significance. This does not overcome the "mixing apples and oranges" difficulty. For example, set size was the independent variable and correct reaction time was the dependent variable in both Schneider and Shiffrin's (1977) and Sternberg's (1969) studies. Be that as it may, it is not meaningful to include them in the same meta-analysis as Hunter recommends, because there are other important differences between the two studies.

Researchers are advised by psychometric meta-analysts not to draw conclusions about substantive issues on the basis of data from single studies, because the psychometric meta-analysis is more accurate and less ambiguous than individual tests of significance. An examination of the justification offered for this assertion is instructive. With 1,455 participants from 15 geographic sites, Schmidt et al. (1985) found a correlation coefficient of .22 between the performance on the test being validated and the ability to operate a special keyboard. The correlation coefficient was statistically significant. They then formed 21 random samples (without replacement) of 68 members each from the 1,455 participants. Each of these 68-member samples was treated as a ministudy. The correlation between task performance and keyboard operation was obtained for each of the 21 "ministudies." Statistical significance was found in only 8 of the 21 ministudies. The means of the 8 ministudies was .33, which differed from the "true" correlation coefficient of .22. This is the reason psychometric meta-analysts find individual tests of significance misleading. They also conclude that the meta-analytic result is more accurate.

Four things to note before accepting the argument for meta-analysis: First, the effective size of the population shrank as the number of ministudies increased because the samples were selected without replacement. This feature renders the independence among the ministudies suspect. The second point is that Schmidt et al. (1985) should have used cluster sampling, not simple random sampling, to form their ministudies in order to reflect the local characteristics of the 15 sites. That is, samples in their ministudies were not representative. Third, their "true" parameter ($r$ = .22) was not theoretically informed. It was the measurement obtained from a complete enumeration of all the participants, and they assumed that a complete enumeration necessarily gives an accurate result. There is no clear reason why this should be the case; the opposite is more likely to be true. Given the same extent of the resources for conducting the research, the chance of making mistakes is higher if there are more units or participants to be measured (see Slonium 1960). Fourth, LeLorier et al. (1997)



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

have reported that "the outcome of the 12 large randomized, controlled trials . . . were not predicted accurately 35 percent of the time by the meta-analyses published previously on the same topics" (p. 536).

Is meta-analysis a valid means of corroborating explanatory hypotheses (**Rossi**)? **Hunter** makes it clear that psychometric meta-analysts are not interested in explanatory hypotheses. Chow (1987c) has shown that Glass et al.'s (1981) approach was invalid as a theory-corroboration procedure. The objection to using meta-analysis is not that it may been misused (**Glück & Vitouch**), but that some of its underlying meta-theoretical assumptions are debatable. The difficulty with resolving the discrepancies in studies of spontaneous recovery may partly be due to the fact that there is insufficient theoretical insight (**Rossi**). Alternatively, why should there be theoretical unanimity when the phenomenon may have multiple underlying material or formal causes?

## R11. Summary and conclusions

The NHSTP defence is an attempt to rationalize the role of tests of significance in the theory-corroboration experiment. This approach was adopted because it has been acknowledged that the criticisms of NHSTP were not applicable to experimental studies in which all recognizable controls are properly instituted. There is no reason to believe that using NHSTP hinders theory development. There are difficulties with the characterization of the "strong hypothesis." The effect size has no evidential status. The more serious reservation about the effect size and statistical power is based on the fact that they are defined at a level of abstraction different from the level at which the decision about statistical significance is made. Without disputing the mathematics in the Bayesian or meta-analytic approaches, their role in theory corroboration may be questioned on methodological or conceptual grounds. In sum, there is as yet no reason to revise the NHSTP defence in any substantive way.

# References

**Letters "a" and "r" appearing before author's initials refer to target article and response, respectively.**


Abelson, R. P. (1995) *Statistics as principled argument.* Erlbaum. [JFK]

(1997) On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science* 8:12–15. [JFK]

Acree, M. C. (1978) Theories of statistical inference in psychological research: A historicocritical study. University Microfilms International (University Microfilms No. H790 H7000). [GG]

Allport, D. A. (1975) Critical notice: The state of cognitive psychology. *Quarterly Journal of Experimental Psychology* 27:141–52. [KJV]

Anderson, N. H. & Cuneo, D. O. (1978) The height + width rule in children's judgements of quantity. *Journal of Experimental Psychology: General* 107:335–78. [GG]

Anthony, T. & Mullen, B. (1991) Conceptual rigor mortis: Or, why people are supporting pornography when they don't send money to televangelists. *Theory and Psychology* 1:361–67. [SL]

Ashby, D., ed. (1993) Papers from the Conference on Methodological and Ethical Issues in Clinical Trials, 27–28 June 1991, special issue. *Statistics in Medicine* 12:1373–534. [JP]

Bakan, D. (1966) The test of significance in psychological research. *Psychological Bulletin* 66:423–37. [CFB, aSLC]

Banaji, M. R. & Crowder, R. G. (1991) Some everyday thoughts on ecologically valid methods. *American Psychologist* 46:78–79. [KJV]

Barnett, V. (1982) *Comparative statistical inference* (2nd edition). Wiley. [DR]

Berger, J. O. (1985) *Statistical decision theory and Bayesian analysis,* second edition. Springer-Verlag. [NGW]

Berger, J. O. & Berry, D. A. (1988) Statistical analysis and the illusion of objectivity. *American Scientist* 76:159–65. [PS, NGW]

Bernardo, J. M. & Smith, A. F. M. (1994) *Bayesian theory.* Wiley. [RAMG, JP]

Bernieri, F. J. (1991) Rigor is rigor: But rigor is not necessarily science. *Theory and Psychology* 1:369–73. [SL]

Berry, D. A. & Lindgren, B. W. (1996) *Statistics: Theory and methods,* second edition. Duxbury Press. [NGW]

Black, M. (1969) Some half-baked thoughts about induction. In: *Philosophy, science, and method,* ed. S. Morgenbesser, P. Suppes & M. White. St. Martin's Press. [PS]

Bolles, R. C. (1962) The difference between statistical hypotheses. *Psychological Reports* 11:639–45. [NGW]

Bookstein, F. L. (1997) Biometrics and brain maps: The promise of the morphometric synthesis. In: *Neuroinformatics: An overview of the human brain project. Progress in neuroinformatics, vol. 1.,* ed. S. Koslow & M. Huerta. Erlbaum. [FLB]

Bookstein, F. L., Streissguth, A., Sampson, P. & Barr, H. (1996) Exploiting redundant measurement of dose and behavioral outcome: New methods from the teratology of alcohol. *Developmental Psychology* 32:404–15. [FLB]

Boring, E. G. (1919) Mathematical vs. scientific significance. *Psychological Bulletin* 16:335–38. [HJS]

(1920) The logic of the normal law of error in mental measurement. *American Journal of Psychology* 31:1–33. [HJS]

(1954) The nature and history of experimental control. *American Journal of Psychology* 67:573–89. [aSLC]

(1963) The nature and history of experimental control (reprinted). In: *History, psychology, and science: Selected papers by Edwin G. Boring,* ed. R. I. Watson & D. T. Campbell. Wiley. [LGT]

(1969) Perspective: Artifact and control. In: *Artifacts in behavioral research,* ed. R. Rosenthal & R. L. Rosnow. Academic Press. [aSLC]

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

Box, G. E. P. & Tiao, G. C. (1973) *Bayesian inference in statistical analysis.* Addison-Wesley. [RAMG, HR]

Camilli, G. (1990) The test of homogeneity for 2X2 contingency tables: A review of and some opinion on the controversy. *Psychological Bulletin* 108:135–45. [JRV]

Campbell, D. T. (1969) Prospective: Artifact and control. In: *Artifacts in behavioral research,* ed. R. Rosenthal & R. L. Rosnow. Academic Press. [aSLC]

Campbell, D. T. & Stanley, J. C. (1963) *Experimental and quasi- experimental designs for research.* Rand McNally. [aSLC]

Carlin, B. P. & Loui, T. A. (1996) *Bayes and empirical Bayes methods for data analysis.* Chapman and Hall. [RAMG]

Carroll, R. M. & Nordholm, L. A. (1975) Sampling characteristics of Kelley's $\epsilon^2$ and Hay's $\omega^2$. *Educational and Psychological Measurements* 35:541–54. [SL]

Carver, R. P. (1978) The case against statistical significance testing. *Harvard Educational Review* 48:378–99. [CFB]

Chomsky, N. (1957) *Syntactic structures.* Mouton. [aSLC]

(1965) *Aspects of the theory of syntax.* MIT Press. [HJS]

Chow, S. L. (1987a) *Experimental Psychology: Rationale, procedures and issues.* Detselig. [aSLC]

(1987b) Some reflections on Harris and Rosenthal's thirty-one meta-analyses. *Journal of Psychology* 121:95–100. [aSLC]

(1987c) Meta-analysis of pragmatic and theoretical research: A critique. *Journal of Psychology* 121:259–71. [arSLC]

(1988) Significance test or effect size? *Psychological Bulletin* 103:105–10. [aSLC, JG, JRV]]

(1989) Significance tests and deduction: Reply to Folger (1989). *Psychological Bulletin* 106:161–65. [arSLC]

(1991a) Conceptual rigor versus practical impact. *Theory and Psychology* 1:337–60. [aSLC, HJS]

(1991b) Rigor and logic: A response to comments on "Conceptual Rigor." *Theory and Psychology* 1:389–400. [aSLC]

(1991c) Some reservations about statistical power. *American Psychologist* 46:1088–89. [aSLC]

(1992) *Research methods in psychology: A primer.* Detselig. [aSLC]

(1996) *Statistical significance. Rationale, validity, and utility.* Sage Publications. [CFB, JFK, HEK, DM, JSR, LGT, JRV, BDZ]

Church, R. M., Crystal, J. D. & Collyer, C. E. (1996) Correction of errors in scientific research. *Behavior Research Methods, Instrumentation, and Computers* 28:305–10. [LGT]

Clark-Carter, D. (1997) The account taken of statistical power in research




https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

published in the British Journal of Psychology. *British Journal of Psychology* 88:71–83.
Cohen, J. (1962) The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* 65:145–53. [GG]
(1965) Some statistical issues in psychological research. In: *Handbook of clinical psychology*, ed. B. B. Wolman. McGraw-Hill. [aSLC]
(1987) *Statistical power analysis for the behavioral sciences* (Revised edition). Academic Press. [aSLC, DM]
(1990) Things I have learned (so far). *American Psychologist* 45: 1304–12. [aSLC, LEK, BRL]
(1992a) Statistical power analysis. *Current Directions in Psychological Science* 1:98–105. [aSLC]
(1992b) A power primer. *Psychological Bulletin* 112:155–59. [aSLC, JFK]
(1994) The earth is round ($p < .05$). *American Psychologist* 49: 997–1003. [CFB, aSLC, RWF, JEH, NGW]
Cohen, M. R. & Nagel, E. (1934) *An introduction to logic and scientific method.* Routledge & Kegan Paul. [aSLC]
Coltheart, M. (1980) Iconic memory and visible persistence. *Perception and Psychophysics* 27:183–228. [aSLC]
Cook, T. D. & Campbell, D. T. (1979) *Quasi-experimentation: Design and analysis issues for field settings.* Rand McNally. [aSLC, LGT]
Cook, T. D. & Leviton, L. C. (1980) Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality* 48:449–72. [aSLC]
Cooper, H. M. (1979) Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology* 37:131–46. [aSLC]
Cooper, H. M. & Rosenthal, R. (1980) Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin* 87:442–49. [aSLC]
Copi, I. (1982). *Symbolic logic (6th edition).* Macmillan. [aSLC]
Cowles, M. (1989) *Statistics in psychology: An historical perspective.* Erlbaum. [LEK]
Cowles, M. & Davis, C. (1982) On the origins of the .05 level of statistical significance. *American Psychologist* 37:553–58. [JFK]
Cox, D. R. & Hinkley, D. V. (!974) *Theoretical statistics.* Chapman and Hall. [ZGS]
Cox, R. T. (1946) Probability, frequency and reasonable expectation. *American Journal of Physics* 14:1–13. [RAMG]
Dale, A. I. (1991) *A history of inverse probability.* Springer-Verlag. [RAMG]
Danziger, K. (1987) Statistical method and the historical development of research practice in American psychology. In: *The probabilistic revolution, vol. 2.: Ideas in the sciences,* ed. L. Krüger, G. Gigerenzer & M. Morgan. MIT Press. [HJS]
(1990) *Constructing the subject: Historical origins of psychological research.* Cambridge University Press. [aSLC, GG, HJS]
Dar, R. (1987) Another look at the Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist* 42:145–51. [RD]
De Finetti, B. (1974) *Theory of probability.* Wiley. [RAMG]
De Groot, A. D. (1990) Unifying psychology: A European view. *New Ideas in Psychology* 3:309–20. [KJV]
DeGroot, M. H. (1970) *Optimal statistical decisions.* McGraw-Hill. [NGW]
Draper, D., Hodges, J. S., Mallows, C. L. & Pregibon, D. (1993) Exchangeability and data analysis. *Journal of the Royal Statistical Society A* 156:9–37. [JRV]
Earman, J. (1992) *Bayes or bust? A critical examination of Bayesian confirmation theory.* The MIT Press. [aSLC]
Ebbinghaus, H. (1885/1964) *Memory: A contribution to experimental psychology.* Dover. [KJV]
Edgington, E. S. (1966) Statistical inference and nonrandom samples. *Psychological Bulletin* 66:485–87. [AFH, JRV]
(1995) *Randomization tests* (3rd edition). Dekker. [AFH, JRV]
Edwards, A. W. F. (1972) *Likelihood.* Cambridge University Press. [BDZ]
Erwin, E. (1996) *A final acounting: Philosophical and empirical issues in Freudian psychology.* MIT Press. [EE]
(1997) *Philosophy and psychotherapy: Razing the troubles of the brain.* Sage. [EE]
Estes, W. K. (1997) Significance testing in psychological research: Some persisting issues. *Psychological Science* 8:18–20. [JFK]
Eysenck, H. J. (1975) Who needs a random sample? *Bulletin of the British Psychological Society* 28:195–98. [JRV]
(1978) An exercise in mega-silliness. *American Psychologist* 33:517. [aSLC]
Falk, R. & Greenbaum, C. W. (1995) Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology* 5:75–98. [aSLC, HJS]
Ferguson, T. (1967) *Mathematical statistics: A decision theoretic approach.* Academic Press. [NGW]
Feynman, R. P. (1985) *"Surely you're joking, Mr. Feynman!": Adventures of a curious character.* Bantam. [KJV]
Fillmore, C. J. (1968) The case for case. In: *Universals in linguistic theory,* ed. E. Bach & R. T. Harms. Holt, Rinehart, and Winston. [aSLC]
Fisher, R. A. (1925, 1963) *Statistical methods for research workers.* Hafner. [HEK]
(1935a) *The design and analysis of experiments.* Oliver and Boyd. [GG, JFK, JRV]
(1935b) The fiducial argument in statistical inference. *Annals of Eugenics* 6:391–98. [JP]
(1955) Statistical methods and scientific induction. *Journal of the Royal Statistical Society (B)* 17:69–77. [GG]
(1956/1990) *Statistical methods and scientific inference.* Oliver and Boyd. (3rd edition 1973 reprinted, Oxford University Press, 1990). [GG, HEK, JP]
(1959/1973) *Statistical methods and scientific inference (2nd edition).* Hafner Publishing Co. [aSLC, NGW]
(1962) Some examples of Bayes's method of the experimental determination of probabilities *a priori. Journal of the Royal Statistical Society B* 24:118–24. [JP]
Folger, R. (1989) Significance tests and the duplicity of binary decisions. *Psychological Bulletin* 106:155–60. [rSLC, JG]
Frick, R. W. (1995) Accepting the null hypothesis. *Memory and Cognition* 23:132–38. [RWF, LEK]
(1996) The appropriate use of null hypothesis testing. *Psychological Methods* 1:379–90. [RWF, JRV]
(in press a) A better stopping rule of conventional statistical tests. *Behavior Research Methods, Instruments, and Computers.* [RWF]
(in press b) Interpreting statistical testing: Processes, not populations and random sampling. *Behavior Research Methods, Instruments, and Computers.* [RWF]
Gallo, P. S., Jr. (1978) Meta-analysis – A mixed meta-phor? *American Psychologist* 33:515–17. [aSLC]
Garner, W. R., Hake, H. W. & Eriksen, C. W. (1956) Operationalism and the concept of perception. *Psychological Review* 63:149–59. [aSLC]
Gergen, K. J. (1991) Emerging challenges for theory and psychology. *Theory and Psychology* 1:13–35. [aSLC]
Gibson, J. J. (1967/1982) James J. Gibson: Autobiography. In: *Reasons for realism: Selected essays of James J. Gibson,* ed. E. Reed & R. Jones. Erlbaum. [KJV]
Gigerenzer, G. (1987) Probabilistic thinking and the fight against subjectivity. In: *The probabilistic revolution, Vol.2. Ideas in the sciences,* ed. L. Krüger, G. Gigerenzer & M. S. Morgan. MIT Press. [GG, HJS]
(1993) The superego, the ego, and the id in statistical reasoning. In: *A handbook for data analysis in the behavioral sciences: Methodological issues,* ed. G. Keren & C. Lewis. Erlbaum. [arSLC, GG, JG, LEK, JP, HJS]
Gigerenzer, G. & Murray, D. J. (1987) *Cognition as intuitive statistics.* Erlbaum. [GG, LEK, HJS]
Gigerenzer, G. & Richter, H. R. (1990) Context effects and their interaction with development: Area judgements. *Cognitive Developments* 5:235–64. [GG]
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 University Press. [GG, HJS, ZGS]
Glass, G. V. (1976) Primary, secondary and meta-analysis of research. *Educational Researcher* 5:3–8. [aSLC]
(1978) Integrating findings: The meta-analysis of research. *Review of Research in Education* 5:351–79. [aSLC]
Glass, G. V. & Kliegl, R. M. (1983) An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology* 51:28–41. [aSLC]
Glass, G. V., McGaw, B. & Smith, M. L. (1981) *Meta-analysis in social research.* Sage. [arSLC]
Godambe, V. P. & Sprott, D. A., eds. (1971) *Foundation of statistical inference.* Holt Rinehart and Winston. [HEK]
Gonzalez, R. (1994) The statistics ritual in psychological research. *Psychological Science* 5:325–28. [RAMG, JFK]
Good, R. & Fletcher, H. J. (1981) Reporting explained variance. *Journal of Research on Science Teaching* 18:1–7. [BAT]
Green, D. M. & Swets, J. A. (1966) *Signal detection: Theory and psychophysics.* John Wiley. [JP]
Greenwald, A. G. (1975) Consequences of prejudice against the null hypothesis. *Psychological Bulletin* 82:1–20. [LEK]
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* 33:175–83. [JRV]
Guttman, L. (1977) What is not what in statistics. *Statistician* 26:81–107. [CFB]
Haber, R. N. (1983) The impending demise of the icon: A critique of the



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

concept of iconic storage in visual information processing. *Behavioral and Brain Sciences* 6:1–11. [aSLC]
Hacking, I. (1965) *Logic of statistical inference.* Cambridge University Press. [GG]
Hagen, R. L. (1997) In praise of the null hypothesis statistical test. *American Psychologist* 52:15–24. [aSLC, JRV]
Hahn, G. J. & Meeker, W. O. (1993) Assumptions for statistical inference. *The American Statistician* 47:1–11. [JRV]
Hammond, G. (1996) The objections to null hypothesis testing as a means of analysing psychological data. *Australian Journal of Psychology* 48:104–06. [SL]
Hammond, K. R., Hamm, R. M. & Grassia, J. (1986) Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychological Bulletin* 100:257–69. [KJV]
Hanson, N. R. (1958) New books. *Mind* LXVII:272–75. [RAMG]
Harris, M. J. (1991) Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory and Psychology* 1:375–82. [SL]
Harris, M. J. & Rosenthal, R. (1985) Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin* 97:363–86. [aSLC]
Harris, R. J. (1997a) Reforming significance testing via three-valued logic. In: *What if there were no significance tests?,* ed. L. Harlow, S. Mulaik & J. Steiger. Erlbaum. [RJH]
(1997b) Significance tests have their place. *Psychological Science* 8:8–11. [RJH, JFK]
Hartigan, J. A. (1983) *Bayes theory.* Springer-Verlag. [RAMG]
Hedges, L. V. (1987) How hard is hard science, how soft is soft science?: The empirical cumulativeness of research. *American Psychologist* 42:443–55. [NGW]
Heidbreder, G. R. (1996) *Maximum entropy and Bayesian methods.* Kluwer Academic. [RAMG]
Hogben, L. (1957) *Statistical theory: The relationship of probability, credibility and error.* W. W. Norton. [aSLC, LEK]
Hogg, R. V. & Tanis, E. A. (1993) *Probability and statistical inference,* fourth edition. Macmillan. [NGW]
Holton, G. (1988) *Thematic origins of scientific thought: Kepler to Einstein* (revised edition). Harvard University Press. [KJV]
Howson, C. & Urbach, P. (1993) *Scientific reasoning: The Bayesian approach,* second edition. Open Court. [PS]
(1994) Probability, uncertainty and the practice of statistics. In: *Subjective probability,* ed. G. Wright & P. Ayton. Wiley. [RAMG]
Hubbard, R., Parsa, R. A. & Luthy, M. R. (1997) The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917–1994. *Theory and Psychology* 7:545–54. [HJS]
Hudson, W. W., Thyer, B. A. & Stocks, J. T. (1985) Assessing the importance of experimental outcomes. *Journal of Social Service Research* 8(4):87–98. [BAT]
Hunter, J. E. (1997) Needed: A ban on the significance test. *Psychological Science* 8:3–7. [CFB, EE, JEH, JFK, LGT]
Hunter, J. E. & Schmidt, F. L. (1990) *Methods of meta-analysis: Correcting error and bias in research findings.* Sage. [arSLC]
Hunter, M. A. & May, R. B. (1993) Some myths concerning parametric and nonparametric tests. *Canadian Psychology* 34:384–89. [JRV]
Inman, H. F. (1994) Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from *Nature. The American Statistician* 48:2–11. [LEK]
Jacobson, N. S. & Christensen, A. (1996) Studying the effectiveness of psychotherapy: How well can clinical trials do the job? *American Psychologist* 51:1031–39. [JFK]
Jacobson, N. S., Follette, W. C. & Revenstorf, D. (1984) Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy* 15:336–52. [JFK]
Jacobson, N. S. & Truax, P. (1991) Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* 59:12–19. [JFK]
Jeffreys, H. (1939) *Theory of probability.* Oxford University Press. [HEK]
Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review* 67:160–67. [RJH]
Kass, R. E. & Raftery, A. E. (1995) Bayes factors and model uncertainty. *Journal of the American Statistical Association* 90:773–95. [NGW]
Kelley, T. L. (1923) The principles and techniques of mental measurement. *American Journal of Psychology* 34:408–32. [HJS]
Kempthorne, O. (1955) The randomization theory of experimental inference. *Journal of the American Statistical Association* 50:946–67. [JRV]
Keppel, G. & Underwood, B. J. (1962) Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior* 1:153–61. [aSLC]
Kirk, R. E. (1984) *Basic statistics* (2nd edition). Brooks/Cole. [aSLC]
(1996) Practical significance: A concept whose time has come. *Educational and Psychological Measurement* 56:746–59. [JEH]
Kraemer, H. C. & Thiemann, S. (1987) *How many subjects? Statistical power analysis in research.* Sage. [aSLC]
Kreutzer, M. A., Leonard, C. & Flavell, J. H. (1975) An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development* 40, Serial No. 159. [KJV]
Kuhn, T. S. (1970) *The structure of scientific revolutions,* second edition. University of Chicago Press. [JG]
Kyburg, H. E., Jr. (1974) *The logical foundations of statistical inference.* Reidel. [HEK]
Lad, F. & Deely, J. (1994) Experimental design from a subjective utilitarian viewpoint. In: *Aspects of uncertainty,* ed. P. R. Freeman & A. F. M. Smith. Wiley. [ZGS]
Lakatos, I. (1978) Falsification and the methodology of scientific research programmes. In: *The methodology of scientific research programs: Imre Lakatos' philosophical papers, vol. 1,* ed. J. Worral & G. Currie. Cambridge University Press. [RD]
Also in: *Criticism and the growth of knowledge,* ed. I. Lakatos & A. Musgrave. Cambridge University Press. [JG]
Langley, P. W., Simon, H. A., Bradshaw, G. L. & Zytkow, J. M. (1987) *Scientific discovery: An account of the creative processes.* MIT Press. [MGS]
Lapin, L. L. (1994) *Quantitative methods for business decisions* (sixth edition). The Dryden Press. [MGS]
Lashley, B. R. & Bond, C. F., Jr. (1997) Significance testing for round robin data. *Psychological Methods* 2:278–91. [BRL]
Lashley, B. R. & Kenny, D. A. (submitted) *Power estimation in social relations analyses.* University of Connecticut Press. [BRL]
Lecoutre, B., Derzko, G. & Grouin, J.-M. (1995) Bayesian predictive approach for inference about proportions. *Statistics in Medicine* 14:1057–63. [JP]
LeLorier, J., Grégoire, G., Benhaddad, A., Lapierre, J. & Derderian, F. (1997) Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine* 337:536–42. [rSLC]
Leventhal, L. (1994) Nudging aside Meehl's paradox. *Canadian Psychology* 35:283–98. [SL]
Leviton, L. C. & Cook, T. D. (1981) What differentiates meta-analysis from other forms of review. *Journal of Personality* 49:231–36. [aSLC]
Lewin, K. (1946) Action research and minority problems. *The Journal of Social Issues* 2:34–46. [LGT]
Lindley, D. V. (1965) *Introduction to probability and statistics. Part 2: Inference.* Cambridge University Press. [RAMG]
(1972) *Bayesian statistics: A review.* Siam, Regional Conference Series in Applied Mathematics, Philadelphia. [JP]
Lipsey, M. W. & Wilson, D. B. (1993) The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist* 48:1181–1209. [EE, NGW]
Loftus, G. R. (1996) Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science* 5:161–71. [CFB, JEH, JFK, KJV]
Loftus, G. R. & Masson, M. J. (1994) Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review* 1:476–90. [RWF]
Lorenz, K. Z. (1973) The fashionable fallacy of dispensing with description. *Die Naturwissenschaften* 60:1–9. [KJV]
Lykken, D. T. (1968) Statistical significance in psychological research. *Psychological Bulletin* 70:151–59. [RD]
(1991) What's wrong with psychology anyway? In: *Thinking clearly about psychology, vol. 1: Matters of public interest,* ed. D. Cicchetti & W. M. Grove. University of Minnesota Press. [NGW]
Macdonald, R. R. (1997) On statistical testing in psychology. *British Journal of Psychology* 88:333–47. [JRV]
Manicas, P. T. & Secord, P. F. (1983) Implications for psychology of the new philosophy of science. *American Psychologist* 38:399–413. [aSLC]
May, R. B. & Hunter, M. A. (1993) Some advantages of permutation tests. *Canadian Psychology* 34:401–07. [AFH]
May, R. B., Masson, M. E. J. & Hunter, M. A. (1990) *Application of statistics in behavioral research.* Harper & Row. [JRV]
Mayo, D. (1996) *Error and the growth of experimental knowledge.* The University of Chicago Press. [DM]
Meehl, P. E. (1967) Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34:103–15. [aSLC, RD, KJV, NGW]
(1978) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 46:806–34. [aSLC, RD, JG, JSR, KJV, NGW]
(1990) Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry* 1:108–41. [aSLC, ZGS, NGW]
Mill, J. S. (1973) *A system of logic: Ratiocinative and inductive.* University of Toronto Press. [arSLC, HEK]



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63:81–97. [aSLC, WSV]
(1962) Some psychological studies of grammar. *American Psychologist* 17:748–62. [aSLC]
Mintz, J. (1983) Integrating research evidence: A commentary on meta-analysis. *Journal of Consulting and Clinical Psychology* 51:71–75.
Mook, D. G. (1983) In defense of external invalidity. *American Psychologist* 38:379–87. [aSLC, JRV]
Morrison, D. E. & Henkel, R. E., eds. (1970) *The significance test controversy: A reader.* Aldine. [CFB, aSLC, JP]
Mosteller, F. & Bush, R. R. (1954) Selected quantitative techniques. In: *Handbook of social psychology: Vol. 1.-Theory and method*, ed. G. Lindzey. Addison-Wesley. [aSLC]
Mosteller, F. & Tukey, J. W. (1977) *Data analysis and regression: A second course in statistics.* Addison-Wesley. [BDZ]
Nagel, E. (1961) *The structure of science.* Harcourt, Brace and World. [HEK]
Neisser, U. (1967) *Cognitive psychology*. Appleton-Century-Croft. [aSLC]
(1976) *Cognition and reality: Principles and implications of cognitive psychology*. Freeman. [KJV]
Newell, A. (1973) You can't play 20 questions with Nature and win: Projective comments on the papers of this symposium. In: *Visual information processing*, ed. W. G. Chase. Academic Press. [KJV]
Neyman, J. (1942) Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society* 105:292–327. [HEK]
(1950) *First course in probability and statistics.* Holt. [JP]
(1952) *Lectures and conferences on mathematical statistics and probability*. (2nd edition) Graduate School, U.S. Department of Agriculture. [JP]
(1957) "Inductive behavior" as a basic concept of philosophy of science. *International Statistical Review* 25:7–22. [ZGS]
Neyman, J. & Pearson, E. S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inferences (Part I). *Biometrika* 20A:175–240. [arSLC, JFK, HEK, JP, ZGS]
(1933a) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A* 231:289–337. [HEK, JP, ZGS]
(1933b) The testing of statistical hypotheses in relation to probabilities *a priori. Proceedings of the Cambridge Philosophical Society* 29:492–510. [JP]
Oakes, M. (1986) *Statistical inference: A commentary for the social and behavioral sciences.* Wiley. [aSLC, GG, LEK, DR]
Oakes, W. F. (1975) On the alleged falsity of the null hypothesis. *Psychological Record* 25:265–72. [SL]
O'Hagan, A. (1994) *The advanced theory of statistics: Vol. 2B, Bayesian inference.* Edward Arnold. [RAMG]
Ohlsson, S. & Jewett, J. J. (1997) Simulation models and the power law of learning. In: *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, ed. M. G. Shafto & P. Langley. Erlbaum. [MGS]
Pagano, R. (1994) *Understanding statistics in the behavioral sciences*, 4th edition. West Publishing Company. [EE]
Phillips, L. D. (1973) *Bayesian statistics for social scientists.* Nelson. [aSLC]
Pitman, E. J. G. (1937a) Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society (Series B)* 4:119–30. [JRV]
(1937b) Significance tests which may be applied to samples from any population II. The correlation coefficient test. *Journal of the Royal Statistical Society (Series B)* 4:225–32. [JRV]
(1937c) Significance tests which may be applied to samples from any populations III. The analysis of variance test. *Biometrika* 29:322–35. [JRV]
Platt, J. C. (1964) Strong inference. *Science* 146:347–53. [LGT]
Polya, G. (1954) *Patterns of plausible reasoning.* Princeton University Press. [PS]
Popper, K. R. (1968a) *The logic of scientific discovery* (originally published in 1959). Harper & Row. [arSLC]
(1968b) *Conjectures and refutations* (originally published in 1962). Harper & Row. [arSLC]
Prentice, D. A. & Miller, D. T. (1992) When small effects are impressive. *Psychological Bulletin* 112:160–64. [AFH]
Presby, S. (1978) Overly broad categories obscure important differences between therapies. *American Psychologist* 33:524–615. [aSLC]
Rachman, S. & Wilson, G. T. (1980) *The effects of psychological therapy*. Pergamon Press. [aSLC]
Ramsey, F. L. & Schafer, D. W. (1997) *The statistical sleuth. A course in methods of data analysis.* Duxbury Press. [CFB]
Rindskopf, D. (1997) Testing "small," not null, hypotheses: Classical and Bayesian approaches. In: *What if there were no significance tests?*, ed. L. L. Harlow, S. A. Mulaik & J. H. Steiger. Erlbaum. [DR]
Robert, C. P. (1994) *The Bayesian choice: A decision–theoretic motivation.* Springer-Verlag. [RAMG, JP]
Rogers, J. I., Howard, K. I. & Vessey, J. T. (1993) Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin* 113:553–65. [rSLC, LEK]
Rosenthal, R. (1983) Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology* 51:4–13. [aSLC]
(1984) *Meta-analytic procedures for social research.* Sage. [aSLC, NGW]
Rosenthal, R. & Rubin, D. B. (1979) A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology* 9:395–96. [aSLC]
(1982) A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology* 74:166–69. [aSLC]
Rosnow, R. L. & Rosenthal, R. (1989) Statistical procedures and the justification of knowledge in psychological science. *American Psychologist* 44:1276–84. [aSLC, BM, KJV]
Rossi, J. S. (1990) Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology* 58:646–56. [JSR]
(1997) A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In: *What if there were no significance tests?*, ed. L. L. Harlow, S. A. Mulaik & J. Steiger. Erlbaum. [JSR]
Rouanet, H. (1996) Bayesian methods for assessing importance of effects. *Psychological Bulletin* 119:149–58. [JP, HR]
Rouanet, H., Bernard, J. M., Bert, M. C., Lecoutre, B. & Lecoutre, M. P. (1991) *L'Inference statistique dans la démarche du chercheur (Statistical inference in the strategy of the researcher).* Peter Lang. (English version in preparation). [HR]
Rouanet, H., Bernard, J. M. & Lecoutre, B. (1986) Nonprobabilistic statistical inference: A set-theoretic approach. *The American Statistician* 40:60–65. [JRV]
Rozeboom, W. W. (1960) The fallacy of the null-hypothesis significance-test. *Psychological Bulletin* 57:416–28. [CFB, aSLC, JP, BDZ]
(1991) Conceptual rigor: Where is it? *Theory and Psychology* 1:383–88. [SL]
Savage, L. J. (1954) *Foundations of statistics.* Wiley. [HEK]
Savin, H. B. & Perchonock, E. (1965) Grammatical structure and the immediate recall of English sentences. *Journal of Verbal Learning and Verbal Behavior* 4:348–53. [aSLC, HJS]
Scarr, S. (1997) Rules of evidence: A larger context for the statistical debate. *Psychological Science* 8:16–17. [JFK]
Schmidt, F. L. (1992) What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist* 47:1173–81. [aSLC]
(1996) Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods* 1:115–29. [arSLC, RWF, JEH, JFK]
Schmidt, F. L., Hunter, J. E. & Urry, V. E. (1976) Statistical power in criterion-related validation studies. *Journal of Applied Psychology* 61:473–85. [JEH]
Schmidt, F. L., Ocasio, B. P., Hillery, J. M. & Hunter, J. E. (1985) Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology* 105:309–16. [rSLC]
Schneider, W. & Shiffrin, R. M. (1977) Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* 84:1–66. [arSLC]
Sedlmeier, P. & Gigerenzer, G. (1989) Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105:309–16. [EE, GG, JG]
Shapiro, S. (1997) Is meta-analysis a valid approach to the evaluation of small effects in observational studies? *Journal of Clinical Epidemiology* 50:223–29. [EE]
Shaver, J. P. (1993) What statistical significance testing is, and what it is not. *Journal of Experimental Education* 61(4):293–316. [CFB]
Shrout, P. E. (1997) Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science* 8:1–2. [EE, JFK]
Siegel, S. (1956) *Non-parametric statistics for the behavioral sciences.* McGraw-Hill. [aSLC]
Siegmund, D. (1985) *Sequential analysis.* Springer-Verlag. [NGW]
Silvey, S. D. (1975) *Statistical inference.* Chapman and Hall. [DR]
Simon, H. A. (1969) *The sciences of the artificial.* MIT Press. [JFK]
(1979) *Models of thought*. Yale University Press. [MGS]
Simonton, D. K. (1997) Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review* 104:66–89. [KJV]



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press

Sivia, D. S. (1996) *Data analysis.* Clarendon Press. [RAMG]
Slonim, M. J. (1960) *Sampling.* Simon and Schuster. [rSLC]
Smith, A. F. M. & Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 55:3–23. [RAMG]
Smith, M., Glass, G. & Miller, T. (1980) *The benefits of psychotherapy.* Johns Hopkins University Press. [EE]
Sohn, D. (1980) Critique of Cooper's meta-analytic assessment of the findings of sex differences in conformity behavior. *Journal of Personality and Social Psychology* 39:1215–21. [aSLC]
Sperling, G. (1960) The information available in brief visual presentations. *Psychological Monographs* 74:11 (Whole No. 498). [aSLC, JSR]
Spiegelhalter, D. J., Freedman, L. S. & Parmar, M. K. B. (1994) Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society A* 157:357–416. [JP]
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. [HJS]
Sternberg, S. (1969) Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist* 57:421–57. [rSLC]
Stevenson, W. J. (1992) *Introduction to management science* (second edition). Irwin. [MGS]
Stocks, J. T. (1987) Estimating proportion of explained variance for selected analysis of variance designs. *Journal of Social Service Research* 11(1):77–91. [BAT]
Stout, D. A. (1987) Statistics in American Psychology: The social construction of experimental and correlational psychology, 1900–1930. Unpublished Doctoral Dissertation, University of Edinburgh. [HJS]
Sun Li Hsu, J. S. J., Guttman, I. & Leonard, T. (1996) Bayesian methods for variance component models. *Journal of the American Statistical Association* 91:743–52. [RAMG]
Thompson, B. (1996) AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher* 25(2):26–30. [CFB, aSLC]
Townsend, J. T. (1994) Methodology and statistics in the behavioural sciences. *Psychological Science* 5:321–25. [RAMG]
Tukey, J. W. (1960) Conclusions vs. decisions. *Technometrics* 2:1–11. [aSLC]
(1969) Analyzing data: Sanctification or detective work. *American Psychologist* 24:83–91. [CFB]
Tversky, A. & Kahneman, D. (1971) Belief in the law of small numbers. *Psychological Bulletin* 76:105–10. [CFB, JG]
Utts, J. M. (1996) *Seeing through statistics.* Duxbury Press. [NGW]
Van Fraassen, B. (1989) *Laws and symmetry.* Oxford University Press. [EE]
Vitouch, O. & Glück, J. (1997) "Small group PETting:" Sample sizes in brain mapping research. *Human Brain Mapping* 5:74–77. [JG]
Walizer, M. H. & Wiener, P. L. (1978) *Research methods and analysis: Searching for relationships.* Harper and Row. [LGT]
Waller, N. G., Tellegen, A., McDonald, R. & Lykken, D. T. (1996) Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality* 64:545–76. [NGW]
Walley, P. (1991) *Statistical reasoning with imprecise probabilities.* Chapman and Hall. [RAMG]
Wilson, G. T. & Rachman, S. J. (1983) Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology* 51:54–64. [aSLC]
Winston, W. L. (1991) *Operations research: Applications and algorithms* (second edition). PWS-Kent. [MGS]
Yates, F. (1951) The influence of *Statistical methods for research workers* on the development of the science of statistics. *Journal of the American Statistical Association* 46:19–34. [MRN]
Yngve, V. (1960) A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104:444–66. [aSLC]
Zechmeister, E. B. & Nyberg, S. E. (1982) *Human memory: An introduction to research and theory.* Brooks/Cole. [JSR]

## New BBS Web Site Service

BBS is instituting a new policy of temporarily mounting online manuscripts submitted for refereeing (with the authors' permission). The manuscripts will be mounted on the BBS Web Site as "BBS Submitted Manuscripts Under Review," accompanied by a flag indicating that the author holds the copyright. The purpose of this is two-fold: (1) to accelerate and facilitate the refereeing process and (2) to establish priority publicly for submitted manuscripts while they undergo refereeing. Please indicate with your submission whether you authorize BBS to archive your submitted manuscript on the Web during refereeing (it is not compulsory). After refereeing is completed, your manuscript will be withdrawn, and if it is accepted, the final draft will be archived for potential commentators in the BBS Preprint Archive.



https://doi.org/10.1017/S0140525X98501161 Published online by Cambridge University Press