

## We Need Statistical Thinking, Not Statistical Rituals

**Gerd Gigerenzer**

Center for Adaptive Behavior and Cognition,  
Max Planck Institute for Human Development, Berlin, Germany,

*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, Chaps. 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’ approach (the id) have been projected into the psyches of textbook writers and researchers in psychology. The results are wishful thinking, suppression 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.<sup>1</sup> 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  $p < .05$ ).

<sup>1</sup> 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.

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*.<sup>2</sup> 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.

---

<sup>2</sup> 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-hypothesis 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, Chap. 3; Hacking, 1965).