2

PHILOSOPHICAL ORIENTATION—SIGNIFICANT DIFFERENCE

> As you know, the notion of statistical significance underpins marketing research. (Burns & Bush, 2010, p. 504)
>
> My thesis, however, is that the hypothetico-deductive method . . . actually retards the progress of science. (Locke, 2007, p. 868)

2.1 Introduction

This chapter describes some key philosophical maxims—conception of knowledge, model of science, and the role of "negative" results—characteristic of the significant difference research paradigm. Thus, Section 2.2 illustrates that the conception of knowledge development shared by those in this paradigm is too simplistic, namely, rejecting the null hypothesis at the .05 level or better. This viewpoint continues unabated as the epitome of methodological rigor among management and social science researchers, something captured when examining the growth of empirical research and statistical significance testing in these areas from the end of World War II through 2007. It also accounts for the unfortunate habit of overgeneralizing the results of single studies with $p \leq .05$ results to other contexts and time periods.

The model of science adopted by members of the significant difference school, *hypothetico-deductivism*, is presented and critiqued in Section 2.3. In the main it is a framework shown to be philosophically untenable. It also is an inaccurate description of how real scientists actually behave.

The role that "negative" or null results (i.e., findings which do not conform with predictions, but more usually are thought to be statistically insignificant, or $p > .05$, results) play in the significant difference paradigm is recounted in Section 2.4 and constitutes the final third of this chapter. Such results mostly are viewed with suspicion and/or hostility, which has led to a bias against their publication. The consequences of this publication bias are explored. They include dire threats to the objectivity, integrity, and self-correcting nature of science itself through the production of empirical literatures contaminated with errors and outright falsehoods. Concluding comments are made in Section 2.5.

2.2 Conception of Knowledge

2.2.1 Unproblematic: Rejection of the Null Hypothesis, H_0

The conception of knowledge in this paradigm is taken to be largely unproblematic; a knowledge claim is made when a researcher can reject the null hypothesis (H_0) at $p \leq .05$. This methodological imprimatur was

handed down by arguably the greatest statistician of all time, Sir Ronald A. Fisher. The origin of the *p*-value is not due to Fisher, but rather to Karl Pearson (1900), who introduced it in his χ^2 test. Yet there is no doubt that it was Fisher who was responsible for popularizing statistical significance testing and *p*-values. He did this via multiple editions of his books *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935)—books that Savage (1954, p. 275) called the two most influential in the development of statistics in the 20th century.

Fisher used discrepancies in the data to reject what he dubbed the null hypothesis, that is, he calculated the probability of the data (x) on a true null hypothesis, or $\Pr(x \mid H_0)$. Formally, $p = \Pr(T(X) \geq T(x) \mid H_0)$. *P* is the probability of getting a test statistic T(X) greater than or equal to the observed result, T(x), in addition to more extreme ones, conditional on a true null hypothesis, H_0, of no effect or relationship. So the *p*-value is a measure of the (im)plausibility of the actual observations (as well as more extreme and unobserved ones) obtained in an experiment or other study, assuming a true null hypothesis. The rationale is that if the data are seen as being rare or highly discrepant under the null hypothesis, this constitutes *inductive evidence* against H_0.[1]

As for the sanctity of the .05 level, Fisher (1966, p. 13) simply commented: "It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance." With respect to his famous disjunction (Fisher, 1959, p. 39), a $p \leq .05$ indicates that "*either* an exceptionally rare chance has occurred, *or* the theory of random distribution [null hypothesis] is not true." And rejections at lower levels, such as $p < .01$, $p < .001$, and so on, are said to furnish even stronger evidence against H_0.

For Fisher (1926, p. 504) the *p*-value from a statistical test plays an important epistemic role by helping to certify scientific knowledge: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this [$p \leq .05$] level of significance." He cemented this conviction when proclaiming: "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (Fisher, 1966, p. 16). Moreover, Fisher (1973, p. 46) held that the *p*-value is an "objective" measure of evidence against the null hypothesis:

> The feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to and verifiable by other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief [in the null hypothesis] it engenders.

Fisher thought that statistics could play a central role in promoting inductive inference, that is, drawing inferences from the particular to the general, from samples to populations. For him, "inductive inference is the only process known to us by which essentially new knowledge comes into the world" (Fisher, 1966, p. 7), and he saw *p*-values from significance tests as being evidential (Hubbard & Bayarri, 2003, p. 172).

Finally, Fisher (1970, p. 2) championed the use of statistical analysis as an important vehicle for elevating disciplinary status: "Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences." So the use of statistical methods, and significance tests in particular, is presented as a key route to scientific progress.

In light of the above endorsements, researchers nervous about their scholarly credentials enthusiastically adopted Fisher's promise of a seemingly "objective" and "scientific" criterion for justifying knowledge claims. In fairness, Fisher (e.g., 1955, p. 74, 1966, p. 13) warned that the conclusions drawn from a test of significance are *provisional*, especially in the early stages of a research program. He saw the test of significance as one more piece of evidence, to be used with other relevant pieces, for assessing the merits of a hypothesis (Cochran, 1974, p. 1461). In addition, his daughter and biographer, Joan Fisher Box (1978, pp. 135–136), confides that her father despaired over the thoughtless application of his methods, though I am unaware of his making any public criticisms of such carelessness. Having said this, it is easy to see from the above how Fisher's words of caution were completely undone by the force of his own rhetoric. So much so that, in reflecting more than 50 years ago on the then influence of Fisher's (1925) *Statistical Methods for Research Workers* on applied research, Yates (1951, p. 33) already was lamenting that investigators often regarded the results of a test of significance to be the ultimate objective of an experiment. Today, this tendency is well nigh universal. Nelder (1985, p. 238), whose views are shared by Guttman (1985), Nester (1996), and Vickers (2010, p. 54), among other statisticians, makes this point bluntly: "The grotesque emphasis on significance tests in statistics courses of all kinds . . . is taught to people, who if they come away with no other notion, will remember that statistics is about tests for significant differences." Below I supply evidence confirming Nelder's misgivings. But first I wish to address the increase over time in empirical research published in the social and management sciences, something which would seem to augur well for the discovery of linchpins to scientific advance, namely, empirical generalizations.

2.2.2 The Growth of Empirical Research: 1945–2007

In this section I track over time the amount of empirical research published in the social and management sciences, beginning with the former.[2] Table 2-1 charts the proportion of empirical work in four social sciences: geography, political science, psychology, and sociology.[3] The data were obtained by inspecting all articles and research notes in a number of leading journals from each of these areas to determine whether they were empirical. Specifically, I content-analyzed a randomly selected single issue of each journal for every year from 1945 through 2007. Leading journals were targeted because they might be expected to mirror best research practices within their disciplines. The parentheses contain the journals included in each social science: geography (*Annals of the Association of American Geographers, Economic Geography, Professional Geographer*), political science (*American Journal of Political Science, American Political Science Review, Public Administration Review*), psychology (*Journal of Applied Psychology, Journal of Comparative Psychology, Journal of Consulting and Clinical Psychology, Journal of Educational Psychology, Journal of Experimental Psychology: General, Psychological Bulletin, Psychological Review*), and sociology (*American Journal of Sociology, American Sociological Review, Social Forces*).[4]

Table 2-1 is based on the examination of 10,874 papers, of which 7,928 (72.9%) are empirical. It reveals interesting patterns in the growth of empirical research. For example, with some minor exceptions, all four social sciences show an inexorable rise in the percentage of articles devoted to data-based research. Comparing 1945–1949 with 2000–2007, these figures grew as follows: geography (34.5%–69.6%), political science (2.4%–64.9%), psychology (70.5%–91.0%), and sociology (34.1%–90.8%).

Table 2-2 shows the increase over time in the percentage of empirical work published in the five business fields of accounting, economics, finance, management, and marketing.[5] They rest on the investigation of 15,505 papers, some 9,710 (62.6%) being empirical. As with the social sciences, prestigious journals were selected from the management areas. Unlike the social sciences, however, management disciplines are of more recent vintage; only the economics journals and two others (*The Accounting Review* and *Journal of Marketing*) spanned the entire time period 1945–2007. The journals representing each field, together with their initial dates of publication, are given in parentheses: accounting (*The Accounting Review, Journal of Accounting and Economics,* 1979; *Journal of Accounting Research*, 1963), economics (*American Economic Review, Economic Journal, Journal of Political Economy, Quarterly Journal of Economics, Review of Economics and Statistics*), finance (*Journal of*

Table 2-1 The Growth of Empirical Research in the Social Sciences: 1945-2007

	Geography			*Political Science*		
Years	*Total*[a]	*Empirical*[b]	*%*[c]	*Total*	*Empirical*	*%*
1945-2007	1,262	689	54.6	1,722	857	49.8
1945-1949	58	20	34.5	82	2	2.4
1950-1959	161	48	29.8	170	28	16.5
1960-1969	212	105	49.5	218	80	36.7
1970-1979	240	132	55.0	329	171	52.0
1980-1989	189	127	67.2	319	201	63.0
1990-1999	211	124	58.8	339	203	59.9
2000-2007	191	133	69.6	265	172	64.9
	Psychology			*Sociology*		
Years	*Total*	*Empirical*	*%*	*Total*	*Empirical*	*%*
1945-2007	6,004	5,058	84.2	1,886	1,324	70.2
1945-1949	261	184	70.5	164	56	34.1
1950-1959	809	688	85.0	330	172	52.1
1960-1969	1,129	1,005	89.0	300	210	70.0
1970-1979	1,216	1,045	85.9	295	224	75.9
1980-1989	988	774	78.3	295	228	77.3
1990-1999	888	713	80.3	274	227	82.8
2000-2007	713	649	91.0	228	207	90.8

[a]Total refers to the total number of articles and research notes.

[b]Empirical refers to the number of empirical articles and research notes.

[c]% refers to the percentage of empirical articles and research notes.

Finance, 1946; *Journal of Financial Economics,* 1974; *Journal of Financial and Quantitative Analysis,* 1966; *Journal of Money, Credit and Banking,* 1969), management (*Academy of Management Journal,* 1958; *Administrative Science Quarterly,* 1956; *Human Relations,* 1947; *Journal of Management,* 1975; *Journal of Management Studies,* 1964; *Organizational*

Table 2-2 The Growth of Empirical Research in the Management Sciences: 1945-2007

	Accounting			*Economics*			*Finance*		
Years	*Total*[a]	*Empirical*[b]	*%*[c]	*Total*	*Empirical*	*%*	*Total*	*Empirical*	*%*
1945-2007	1,481	717	48.4	4,594	2,315	50.4	1,768	1,116	63.1
1945-1949	60	5	8.3	202	43	21.3	24	7	29.2
1950-1959	146	17	11.6	531	160	30.1	60	30	50.0
1960-1969	248	42	16.9	671	262	39.0	125	64	51.2
1970-1979	298	116	38.9	927	425	45.8	431	203	47.1
1980-1989	311	218	70.1	928	514	55.4	450	283	62.9
1990-1999	229	173	75.5	793	526	66.3	381	289	75.9
2000-2007	189	146	77.2	542	385	71.0	297	240	80.8

	Management			Marketing		
Years	*Total*	*Empirical*	*%*	*Total*	*Empirical*	*%*
1945-2007	4,502	3,351	74.4	3,122	2,234	71.5
1945-1949	38	13	34.2	125	48	38.4
1950-1959	189	99	52.4	221	81	36.7
1960-1969	523	265	50.7	424	205	48.3
1970-1979	981	741	75.5	659	502	76.2
1980-1989	1,082	856	79.1	659	502	76.2
1990-1999	945	759	80.3	524	433	82.6
2000-2007	744	618	83.1	510	463	90.8

[a]Total refers to the total number of articles and research notes.

[b]Empirical refers to the number of empirical articles and research notes.

[c]% refers to the percentage of empirical articles and research notes.

Behavior and Human Decision Processes, 1966; *Strategic Management Journal,* 1980), and marketing *(Journal of Consumer Research,* 1974; *Journal of Marketing, Journal of Marketing Research,* 1964).

A picture similar to that encountered in the social sciences is apparent in the management areas. Here, too, an almost monotonic increase in the percentage of empirical work is observed in all five domains. Between 1945–1949 and 2000–2007, the frequency of published research that is empirically based rose in the following manner: accounting (8.3%–77.2%), economics (21.3%–71.0%), finance (29.2%–80.8%), management (34.2%–83.1%), and marketing (38.4%–90.8%).

In both the social and management sciences, empirical research is systematically displacing nonempirical/conceptual papers. Pointedly, in some disciplines this research has attained almost monopolistic status.[6] Yet despite this explosion in data-based research, few empirical regularities are discernible (Bamber, Christensen, & Gaver, 2000; Hubbard & Lindsay, 2002; Mick, 2001; Rossi, 1997).

Take marketing as an example of this state of affairs. Andrew Ehrenberg and his colleagues (e.g., Ehrenberg & Bound, 1993; Ehrenberg & England, 1990) have produced fine examples of empirical generalizations in the areas of buyer behavior and price elasticities. Bass (1993, 1995) also has had success with regard to new product diffusion. Yet it is highly doubtful that these stubborn facts would have emerged were it not for the research orientation—they were actively seeking regularities—and tenacity shown by the authors themselves. Sadly, these are the exceptions and not the rule. The lack of repeatable empirical facts, Barwise (1995, p. G30) reminds us, means that much marketing practice and teaching is based on only anecdotal evidence. In this sense, academic marketing research may have generated a collective output that is little or no better than Ries and Trout's (1993) popular press offering *The 22 Immutable Laws of Marketing,* which is entirely anecdotal in content. It is difficult to build a science relevant to the practice of marketing, or any other discipline—see Dawes's (1994) *House of Cards: Psychology and Psychotherapy Built on Myth* as a conspicuous example—that consists primarily of hearsay.

Lest the above account come across as excessively grim, we would do well to consult Armstrong and Schultz's (1993) attempt to find useful, empirical marketing principles. Specifically, they examined some 566 normative statements about products, pricing, place, or promotions (the four Ps, or marketing mix, and quintessentially core elements of the marketing curriculum) selected from nine marketing textbooks. Armstrong and Shultz reported that none of these statements had empirical support. They also found, based on the agreements among four raters, that only

20 of the 566 statements qualified as meaningful principles. When 20 marketing professors were asked how many of the 20 meaningful principles were correct, useful, surprising, and had empirical support, none met all four criteria. Disturbingly, Armstrong and Schultz discovered that 9 of the 20 principles were judged to be nearly as correct when their wording was reversed.

2.2.3 The Hegemony of Statistical Significance Testing: 1945–2007

The Social Sciences

The social sciences are awash with tests of statistical significance. It is almost as if, absent such tests, a research paper is somehow "unscientific." In this section I present evidence on the spread of statistical significance testing in the social sciences. I begin with psychology, a field that contained the earliest consumers of such methods.

Psychology. Given the crucial role that experimentation plays in the discipline, it is not surprising that psychologists became Fisherian converts long before those in the other social sciences. Joan Fisher Box (1978, p. 130) writes that her father's *Statistical Methods for Research Workers* (Fisher, 1925) did not receive a single favorable review. This may have been true in Great Britain. It was not true in the United States, where Harold Hotelling (1927, p. 412) pronounced it to be of "revolutionary importance."

Indeed, at Columbia University, Hotelling was one of a triumvirate of players in Rucci and Tweney's (1980) phylogeny of the spread of Fisher's methods in psychology in America. The other two were George Snedecor at Iowa State College, whom Fisher Box (1978, p. 313) describes as the "midwife in delivering the new statistics to the United States," and Palmer Johnson at the University of Minnesota. Hotelling studied with Fisher at Rothamsted Experimental Station in 1929, and Johnson learned from Fisher when the latter was a professor of eugenics at University College, London (his appointment there was made in 1933). Snedecor became a disciple and arranged for Fisher to lecture at Iowa State College in the summers of 1931 and 1936. Interestingly, because Fisher is not easy to read (Savage, 1976, p. 443), Snedecor's (1934, 1937) own expositions of the great man's work were well received among psychologists (Lovie, 1979, p. 169).

Each of these three important promoters of Fisherian methods in the United States passed on the message to their students in the 1930s and

beyond. For example, while at Stanford, Hotelling instructed the psychologist Quinn McNemar, who, in turn, taught others in the discipline, such as Lloyd Humphreys (Northwestern), David Grant (Wisconsin), Allen Edwards (Washington), and so on. Likewise, Snedecor's influence is evident in Edward Lindquist's (1940) popular book preaching the Fisherian gospel in education and psychology.[7]

After World War II, the Neyman–Pearson conception of significance testing claimed ascendancy over Fisher's approach among mathematical statisticians. But this was not the case in social science statistics textbooks, where an anonymous hybridization of Fisher's and Neyman–Pearson's methods typically are showcased (Gigerenzer, 1993; Gigerenzer et al., 1989; Goodman, 1993; Hubbard & Bayarri, 2003, 2005). This important issue is visited in some detail in Chapter 7.

I examined the publication incidence of empirical research using tests of statistical significance in the seven leading American Psychological Association journals listed in the previous section. This required the inspection of 5,058 empirical papers and research notes, of which 4,477 (88.5%) used significance tests.

By way of background, Gigerenzer and Murray (1987, ch. 1) allege that between about 1940 and 1955 an "inference revolution" took place in psychology. In this time inferential statistics, and especially *p*-values, were broadly adopted and eventually institutionalized as the single method of inductive inference. It was during this approximate period that rejection of the null hypothesis at $p \leq .05$ gradually infiltrated psychology textbooks on statistical analysis (Halpin & Stam, 2006; Huberty, 1993; Huberty & Pike, 1999).

The content of psychology journals strongly supports the notion of a 1940–1955 inference revolution. For example, two colleagues and I (Hubbard, Parsa, & Luthy, 1997), with data gathered from the *Journal of Applied Psychology,* showed that whereas 25.0% of empirical work published in this journal between 1940 and 1945 employed *p*-values, this number more than doubled to 59.7% during 1950–1954. Further confirmation of an inference revolution in psychology is provided by Patricia Ryan and me in our analysis of 12 American Psychological Association journals (Hubbard & Ryan, 2000). We found that while only 4.0% of empirical work in these journals reported *p*-values for 1935–1939, the corresponding numbers for 1940–1944 and 1950–1954 were 22.7% and 71.7%, respectively.

Beginning as they do in 1945, my data do not permit an additional empirical check of Gigerenzer and Murray's (1987) inference revolution hypothesis. What they do clearly reinforce for 1945–1949, however, is the

rapidity with which *p*-values were deified by psychologists; some 62.5% of empirical studies featured this index in their accounts (see Figure 2-1 and Table 2-3). This figure jumped to 81.3% during 1950–1959 and 86.9% in the following decade. A relentless—though by now subdued because the upper limit is approaching—increase in this number is observed thereafter: 1970–1979 (90.4%), 1980–1989 (91.9%), 1990–1999 (91.9%), and 2000–2007 (95.4%). These figures are largely corroborated by Parker's (1990) investigation of the growth of statistical significance testing in *Perception & Psychophysics* over the period 1966–1990.

Gigerenzer and Murray (1987) write that prior to the inference revolution, psychologists relied mostly on assorted, nonstandardized ways for making inductive inferences. These included the presentation of copious descriptive data for individual subjects or small groups. Occasionally, mechanistic criteria for gauging significance, such as three times the probable error and critical ratios, were utilized (see Hubbard, Parsa, & Luthy, 1997, for details). L. D. Smith, Best, Cylke, and Stubbs (2000), in an article

Figure 2-1 The Growth of Statistical Significance Testing in Empirical Work in Psychology and Sociology: 1945–2007

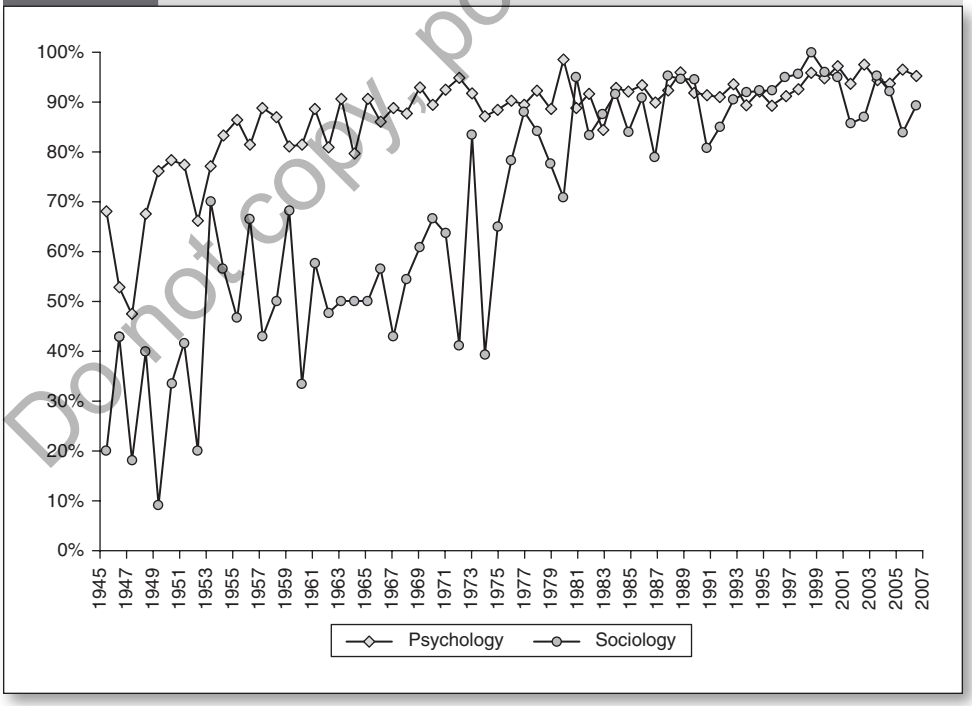


Table 2-3 The Growth of Statistical Significance Testing in the Social Sciences: 1945-2007

	Geography		*Political Science*	
Years	*Number*[a]	*%*[b]	*Number*	*%*
1945-2007	270	39.2	620	72.3
1945-1949	0	0.0	0	0.0
1950-1959	2	4.2	4	14.3
1960-1969	17	16.2	24	30.0
1970-1979	52	39.4	99	57.9
1980-1989	74	58.3	146	72.6
1990-1999	68	54.8	186	91.6
2000-2007	57	42.9	161	93.6
	Psychology		*Sociology*	
Years	*Number*	*%*	*Number*	*%*
1945-2007	4,477	88.5	956	72.2
1945-1949	115	62.5	15	26.8
1950-1959	559	81.3	87	50.6
1960-1969	873	86.9	107	51.0
1970-1979	945	90.4	154	68.8
1980-1989	711	91.9	198	86.8
1990-1999	655	91.9	208	91.6
2000-2007	619	95.4	187	90.3

[a]Number refers to the number of empirical articles and research notes using tests of statistical significance.

[b]% refers to the percentage of empirical articles and research notes using tests of statistical significance.

tellingly called "Psychology Without *p* Values," note that graphical techniques were a common means of conveying information in psychology's early days. Of particular interest, no consensus existed about which were the appropriate ways for drawing inferences from data, and, importantly, this was not seen to be a problem among members of the profession

(Gigerenzer & Murray, 1987). Finally, it is sobering to observe that some of the most respected theoreticians and experimentalists in psychology—H. Ebbinghaus, Wolfgang Köhler, Jean Piaget, B. F. Skinner, S. S. Stevens, Edward L. Thorndike, and Wilhem Wundt—had no use for inferential statistics, preferring instead to exercise their own judgment (Gigerenzer & Murray, 1987, p. 26; Smith et al., 2000, p. 260). It would be beneficial to resurrect such thinking.

Sociology, Political Science, and Geography. Data on the frequency of statistical significance testing were collected for the period 1945–2007 for sociology, political science, and geography using the same journals listed earlier. For sociology, this involved a total of 1,886 articles, 1,324 (70.2%) of which are empirical, with 956 (72.2%) of the latter using tests of significance. Corresponding numbers for political science are 1,722 total articles, with 857 (49.8%) empirical, and 620 (72.3%) of the empirical papers using significance tests. Finally, for geography, 1,262 articles in total, 689 (54.6%) empirical, and 270 (39.2%) invoking significance tests.

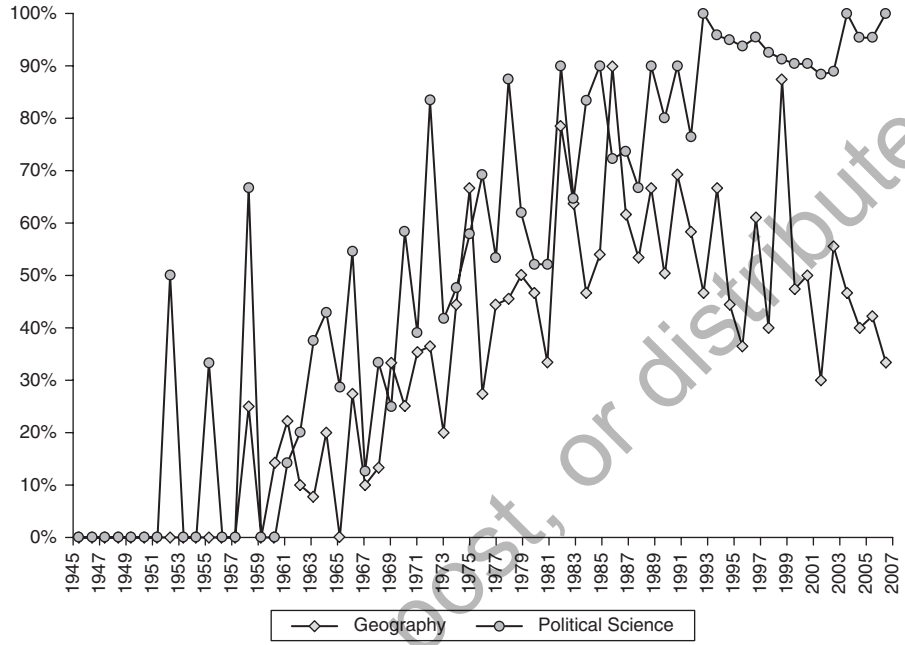
Over time, sociologists' zeal for *p*-values is exceeded only by the psychologists' (see Figure 2-1 and Table 2-3). Whereas none of the data-based works in political science and geography during 1945–1949 involved statistical tests, 26.8% already did so in sociology. This figure almost doubled in sociology during the 1950s (50.6%), while on average it remained low in political science (14.3%) and geography (4.2%). Nonetheless, a steady increase in statistical significance testing is seen in Figure 2-2 and Table 2-3 for political science, such that for the period 2000–2007 this discipline's count (93.6%) just eclipses that for sociology (90.3%).[8] For this same time period, geography (42.9%) yields a comparatively modest number, perhaps because the area of spatial statistics offers fewer testing options than its aspatial counterpart.

The Management Sciences

Figures 2-3 and 2-4 and Table 2-4 depict the almost uniform rise of empirical articles and research notes employing statistical significance tests in the business sciences for 1945–2007. I begin the discussion with economics, by far the oldest business discipline.

Economics. Keuzenkamp and Magnus (1995, p. 16) quip that if economists have natural constants, then the best known is .05. But it was not always this way. Morgan (1990, p. 235), for example, remarks that econometric work in the 1920s and 1930s was centered chiefly on the measurement of phenomena and that the role of inference was not viewed as

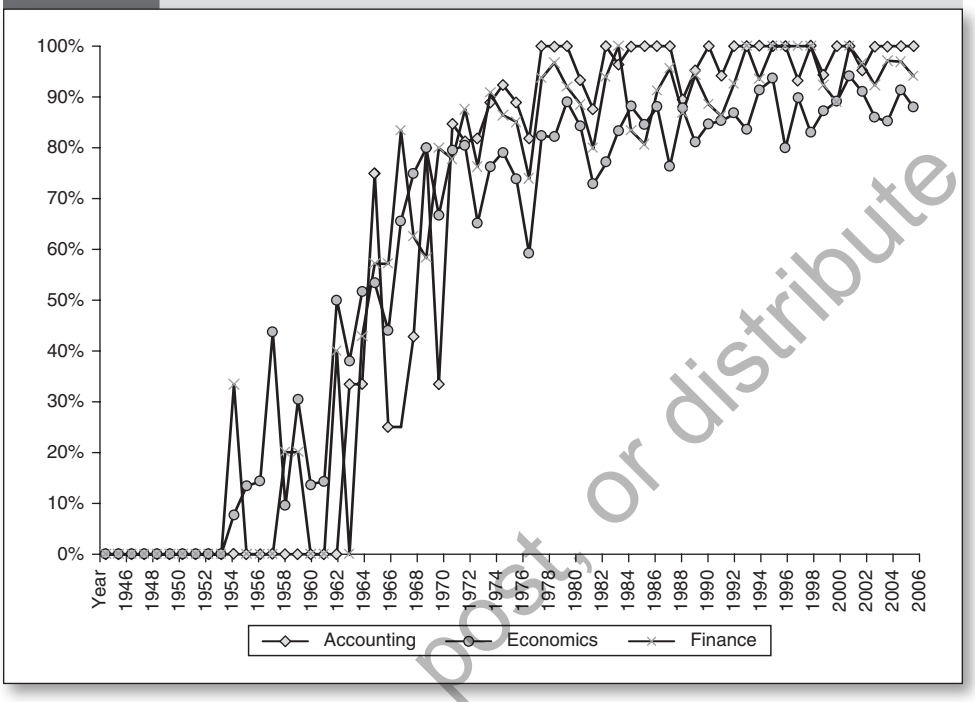
Figure 2-2 The Growth of Statistical Significance Testing in Empirical Work in Geography and Political Science: 1945–2007



important. This philosophy preceded, and subsequently was echoed in, the motto of the Cowles Commission for Research in Economics, created in the United States in 1932, that "Science is Measurement." Things changed, however, with the publication of Trygve Haavelmo's (1944) 115-page paper "The Probability Approach in Econometrics." In this paper Haavelmo gave a lengthy explanation of the Neyman–Pearson ideas on hypothesis testing, and urged that economic theories be cast as formal statistical hypotheses. Haavelmo's article greatly influenced the work of the Cowles Commission. It led to a new motto for the Commission in 1952: "Theory and Measurement." Now, the gathering of data would be dictated by neoclassical economic theory.

Haavelmo's article also marked a reorientation in empirical economics from theory *development* to theory *testing* (Keuzenkamp, 2000, pp. viii, 160–162; Morgan, 1990, pp. 257, 263–264; Ziliak & McCloskey, 2008, p. 113). This shift in emphasis was aided by another Cowles Commission member (and director from 1948 to 1954), Tjalling Koopmans, who

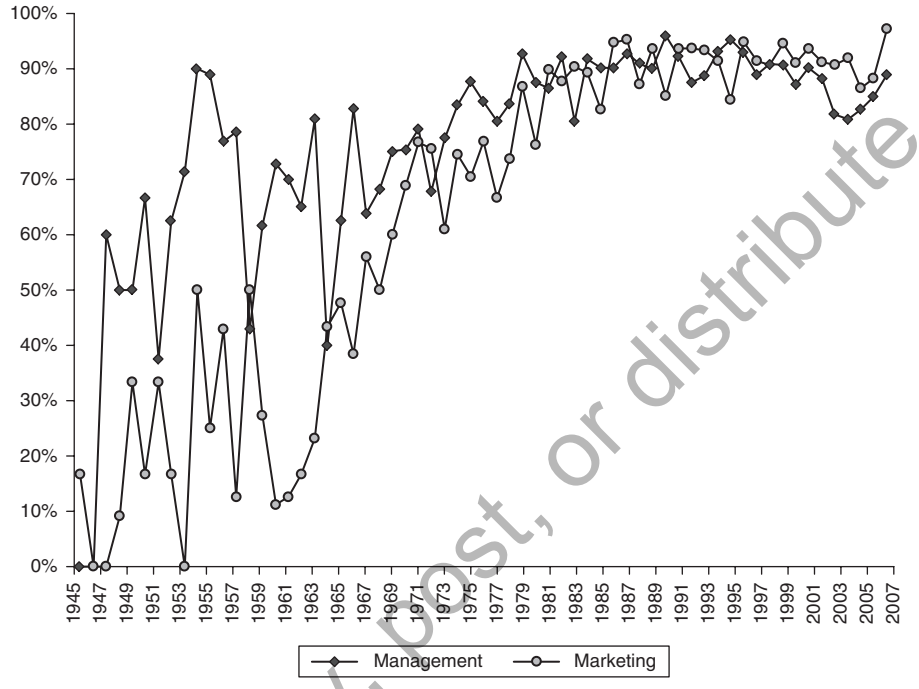
Figure 2-3 The Growth of Statistical Significance Testing in Empirical Work in Accounting, Economics, and Finance: 1945–2007



argued that the focus on measurement error had resulted in the neglect of sampling error. In what I regard as a critical mistake, over the years economists, through their obsession with significance testing, have de facto downplayed measurement error and elevated sampling error to dominance. Today this is personified in Hendry's (1980, p. 403) exhortation that "the three golden rules of econometrics are test, test, test."

Employing the same procedures used with regard to the social science literatures, I traced the growth of significance testing in economics from 1945 through 2007. The data are based on a content analysis of the five prestigious economics journals listed previously. While econometrics textbooks almost exclusively present the Neyman–Pearson model of hypothesis testing, these journals instead are saturated with Fisherian *p*-values, as shown in Figure 2-3 and Table 2-4. All told, some 4,594 articles were reviewed, with 2,315 (50.4%) being empirical. Of the data-based articles, 1,697 (73.3%) employed Fisherian significance testing.

Figure 2-4 The Growth of Statistical Significance Testing in Empirical Work in Management and Marketing: 1945–2007



In my sample, no statistical significance tests were performed during the period 1940–1945. Thereafter, a steady increase in their visibility is apparent. For instance, the decade 1950–1959 saw 13.8% of empirical articles featuring tests of significance, probably in response to the Cowles Commission's call for theory testing. This figure rose abruptly to 52.3% during the 1960s. Even during the 1970s (74.6%), when the contribution of econometrics was being questioned (Keuzenkamp, 2000; Morgan, 1990), growth in the use of such testing nevertheless occurred. And it continued to do so in the 1980s (83.1%), the 1990s (85.7%), and for 2000–2007 (89.1%). Economists take Hendry's (1980) mantra to heart.

As with psychology, in the earlier days in economics there was variation in the tests carried out by researchers, and no ground rules for presenting results were enforced (Morgan, 1990, p. 157). This altered, of course, when conventional wisdom made the reporting of *p*-values for all intents and purposes mandatory.

Table 2-4 The Growth of Statistical Significance Testing in the Management Sciences: 1945-2007

	Accounting		*Economics*		*Finance*		*Management*		*Marketing*	
Years	*Number*[a]	*%*[b]	*Number*	*%*	*Number*	*%*	*Number*	*%*	*Number*	*%*
1945-2007	643	89.7	1,697	73.3	967	86.6	2,854	85.2	1,792	78.2
1945-1949	0	0.0	0	0.0	0	0.0	7	53.8	6	12.5
1950-1959	0	0.0	22	13.8	3	10.0	67	67.7	23	28.4
1960-1969	18	42.9	137	52.3	35	54.7	185	69.8	88	42.9
1970-1979	100	86.2	317	74.6	173	85.2	606	81.8	351	73.3
1980-1989	211	96.8	427	83.1	253	89.4	764	89.3	443	88.2
1990-1999	170	98.3	451	85.7	275	95.2	696	91.7	395	91.2
2000-2007	144	98.6	343	89.1	228	95.0	529	85.6	423	91.4

[a]Number refers to the number of empirical articles and research notes using tests of statistical significance.

[b]% refers to the percentage of empirical articles and research notes using tests of statistical significance.

Accounting, Finance, Management, and Marketing. With the exception of economics, many of the other business journals originated during or after the hardware and software needed to make the computation of statistical tests effortless were routinely available. So indoctrination into the significant difference paradigm had already taken place.

"Historically," however, there are some counterintuitive empirical results in the spread of statistical significance testing in the business fields. Granted, the absolute numbers are small, nevertheless it is the "softer" disciplines of management (53.8%) and marketing (12.5%) that used significance tests during 1945–1949, whereas accounting, economics (as noted earlier), and finance saw zero usage. Management's high percentage for this time is due to the fact that only one journal, *Human Relations,* is in the database. And *Human Relations* has a pronounced psychology bent.

By the 1960s statistical significance testing in empirical research was firmly ensconced in accounting (42.9%), finance (54.7%), management (69.8%), and marketing (42.9%). Yet the 1970s witnessed a dramatic upsurge in such work: accounting (86.2%), finance (85.2%), management (81.8%), and marketing (73.3%). For 2000–2007, the numbers grew again: accounting (98.6%), finance (95.0%), management (85.6%), and marketing (91.4%). There is little room left for the further expansion of statistical significance testing in management science empirical studies; the practice has long since had a stranglehold in these fields.

2.2.4 Summarizing the Dominance of the Statistical Significance Test

Fisher (1959, p. 76) reminisced that "the common tests of significance . . . have come to play a rather central part in statistical analysis." As has been shown, even this strong assertion is, in fact, an understatement; these tests are considered to be almost the *only* legitimate way of making inductive inferences from numerical data. The test of significance is no mere statistical technique, but the glue that holds together the entire research process. It largely dictates how we formulate hypotheses; design questionnaires; organize experiments; and analyze, report, and summarize results (Hubbard & Armstrong, 2006, p. 114). To illustrate this, the test of statistical significance has come to be seen as a mark of scientific rigor (Lindsay, 1995, p. 35), as the sine qua non of scientific respectability (Cowles, 2001, p. 179), as the centerpiece of inductive inference (Hubbard & Ryan, 2000, p. 678), as an objective and universally defined standard of scientific demonstration (Gigerenzer et al., 1989, p. 108), as an automaticity of inference (Goodman, 2001, p. 295), and simply "as an end, in and of itself" (Cicchetti, 1998, p. 293).

In the significant difference paradigm statistical inference and scientific inference are interchangeable; asterisks outweigh substance.[9] To judge the validity of this account it is necessary only to raise the question: "How would one analyze data if the significance test was outlawed?"[10] I suspect that many researchers in the management and social sciences would be hard pressed to respond.

2.2.5 Overgeneralizing the Results of Single Studies

Another facet of the unproblematic conception of knowledge justification embedded in the significant difference paradigm needs to be made clear. This is the tendency for researchers to place far more confidence in the conclusions of single works—predominantly empirical in composition and sporting $p \leq .05$ outcomes, although not confined to this genre—than is warranted. Such studies are perceived as having an aura of finality about them, as if they have largely settled the matter of the research topic at hand (see, e.g., De Long & Lang, 1992, p. 1258). Perhaps because of this there is an accompanying proneness to overgeneralize the scope of the findings of single works (Gauch, 2003, pp. 254–255; Hubbard & Lindsay, 1995, p. 52; Starbuck, 2006, pp. 34–35), something Wells (2001, p. 494) denounces as "The Perils of $N = 1$." More broadly, Tversky and Kahneman (1974, p. 1124) call this phenomenon of generalizing on the basis of insufficient evidence the "representativeness heuristic" (see also Cassidy, 2009, p. 195).

Incredibly, this penchant for overgeneralizing the applicability of initial results from one-off studies persists even in those all-too-rare situations where later works somehow manage to circumvent the barriers against publishing replications and reexaminations (see Chapter 5) and rebut, often decisively, earlier findings. Five examples from the marketing, and one from the psychology, literatures presented below epitomize this inclination.

Chronologically, the first example is the study on subliminal advertising carried out at a New Jersey drive-in movie theater in 1956 whose results have been referred to many times in marketing and consumer behavior textbooks (cf. Wilkie, 1986). This study, publicized in the press, claimed to show how the subliminal messages "Hungry? Eat Popcorn" and "Drink Coca-Cola" flashed repeatedly on the screen for 1/3,000 of a second (far below the level of the limen, i.e., our ability to consciously perceive stimuli) boosted the sales of popcorn and Coca-Cola by 58% and 19%, respectively. These dramatic findings from an investigation notable for its absence of scientific controls have found no subsequent support in research addressing the connection between subliminal advertising and buyer behavior (see, e.g., Moore, 1982; Rosen & Singh, 1992; Theus, 1994).

Regardless, Kerin, Hartley, and Rudelius (2013, p. 118) report that customers spend $50 million a year for audiotapes containing subliminal messages to help them stop smoking, lose weight, and improve their self-esteem. Further, about two thirds of U.S. consumers say that subliminal messages are hidden in commercials, and roughly half believe that this technology can cause them to buy things against their will.

A second example is Julian Simon's (1979) reanalysis of the experimental data employed in Zielske's (1959) authoritative contribution on the remembering and forgetting of advertisements. Zielske's determination as to which method of scheduling advertisements is more effective—spaced or pulsed—is equivocal, and his published results pictured only idealized representations of the raw data. Simon's reanalysis of the latter points to a definitive conclusion: A spaced advertising schedule is more dollar-effective than one that is pulsed. Frustratingly, despite Simon's reworking of the raw data, it is still Zielske's idealized findings which are more likely to be discussed in the classroom. For example, in a convenience sampling of 15 consumer behavior textbooks published either for the first time or in subsequent editions between 1982 and 1988 (thereby allowing enough time for Simon's results to be included), only two cite both his work and Zielske's, and one of these fails to mention their different conclusions (Hubbard, Brodie, & Armstrong, 1992, p. 5). Thirteen of the 15 textbooks do not cite Simon's article, while 10 of them feature Zielske's experiment.[11] Reflecting this disparity, Google Scholar shows that Zielske (n = 196) has attracted more than six times as many citations as Simon (n = 30).

Levitt's (1960) hugely influential publication on "marketing myopia" is the third example. This is when executives are said to become overly enamored with their products, thereby losing sight of the underlying needs of their customers and the poor decisions which inevitably stem from this. Specifically, Levitt was concerned with what he considered to be the short-sighted thinking of those in charge of the U.S. railroad and film industries. As Morris (1990, p. 279) explains, Levitt's argument, "which has become familiar to almost every teacher and probably most students of marketing," was that railroad managers saw their business as running trains and did not anticipate a market shift toward freeway and air transport. Likewise, Levitt opined, movie moguls were focused on the cinema and were unprepared for the competition posed by television.

After painstaking research 30 years later, Morris (1990) was able to demonstrate that it was government regulations, not deficient marketing practices nor unwise management, which prevented the railroads from moving into alternative modes of transportation. Similarly, Morris showed

that regulations stymied the movie studios from encroaching on the television market. In the final analysis, Levitt's (1960) anecdotal offering that the film and railroad industries "failed to be marketing oriented is unsubstantiated by the historical evidence" (Morris, 1990 p. 282). Facts notwithstanding, it is still Levitt's conjecture which holds sway in the textbooks and classrooms. As a bizarrely lopsided indicator of this, Levitt's article has managed an astonishing total of 2,201 citations (Google Scholar), while Morris's, with a mere 3, is long since forgotten.

The fourth example is from Gorn (1982), who published an article stating that product preferences can be classically conditioned through a single pairing with background music. Kellaris and Cox (1989) believed that Gorn's results may have been due to demand artifacts. This concerned them because Gorn's article was drawing notice, being cited in the *Social Sciences Citation Index* 34 times between 1982 and 1988. However, the work was also being referred to in consumer behavior textbooks, at least one of which used it as a basis for declaring that classical conditioning of product preferences is "well established and widely used." In three well-designed experiments, Kellaris and Cox failed to replicate Gorn's results. Despite this, according to Google Scholar, Gorn (n = 603) continues to outperform Kellaris and Cox (n = 158) by a factor of almost 4 to 1 when it comes to garnering citations.[12]

The fifth example is provided by Bottomley and Holden (2001). In circumstances they labeled "unique," these authors conducted a study to determine the empirical generalizability of Aaker and Keller's (1990) model stating that consumers evaluate brand extensions on the basis of the quality of the original brand, the degree of fit observed in the parent and extension categories, and the interaction between the two. Like Gorn's (1982), Aaker and Keller's paper made an impact in intellectual circles, being cited some 92 times from 1990 to 2001, and their findings similarly were appearing in marketing textbooks (Bottomley & Holden, 2001, p. 494).

What makes Bottomley and Holden's (2001) paper unique is that they were able to gain access to Aaker and Keller's (1990) original data set as well as seven others from studies attempting to replicate the latter's work.[13] This is noteworthy because, to my knowledge, it represents the most comprehensive example of independent replication research in academic marketing. It also is anomalous because, as shown in Section 5.2, replication research seldom is seen in marketing.

In their reanalyses of the eight data sets, Bottomley and Holden (2001) were able to resolve many of the conflicting results between them. They also cautioned that a key lesson from their work is the danger of drawing "firm conclusions about theory on the basis of only one study" (p. 494).

An example from psychology completes the discussion of the risks involved in putting too much faith in the results of one-off studies. Kelley and Blashfield (2009) trace the history of Broverman, Broverman, Clarkson, Rosenkrantz, and Vogel's (1970) article on sex bias in the mental health profession. After analyzing questionnaire returns, Broverman and her colleagues wrote that a double standard was discernible among mental health clinicians in how they viewed men and women. In particular, they found that assessments of mental health were biased in favor of men.

Kelley and Blashfield (2009, p. 123) note that Broverman et al.'s (1970) study has had an enormous impact on the thinking of a generation of psychologists and mental health experts, along the way becoming one of the most highly cited papers in psychology. Using *Science Citation Index* and *Social Science Citation Index* databases, Kelley and Blashfield discovered that Broverman et al. had attracted an amazing 934 citations since publication of their article. To lend perspective on this result, Kelley and Blashfield also checked on the citations gathered by Stephen and Christine Abramowitz, the two most prolific authors on sex bias during the 1970s. Their best-known paper (Abramowitz, Abramowitz, Jackson, & Gomes, 1973) acquired 58 citations. Impressively, Kelley and Blashfield reveal, Broverman et al. have collected more citations than the top-cited articles from some of the field's 20th century luminaries, including Hans Eysenck ($n = 866$), B. F. Skinner ($n = 590$), and Robert Sternberg ($n = 226$).

And yet a critique by Stricker (1977) as well as replications by Phillips and Gilroy (1985) and Widiger and Settle (1987), each published in well-known psychology journals, have brought to light "fatal methodological flaws" in Broverman et al.'s (1970) work (Kelley & Blashfield, 2009, p. 126). Baffling to Kelley and Blashfield (2009, p. 128) are how Broverman et al.'s findings continue to be accepted in the field:

> It is almost as if these critical articles [i.e., Stricker, 1977; Phillips & Gilroy, 1985; Widiger & Settle, 1987] were never published. Researchers citing Broverman et al. seem to be unaware that the central conclusions from that article are erroneous.

Pursuing this lead, Kelley and Blashfield add that Broverman et al. were cited some 53 times between 2000 and 2008 alone, while plaudits earned for works over their entire lifespans by scholars attempting to correct the written record on this topic suffer in comparison: Stricker ($n = 82$), Phillips and Gilroy ($n = 26$), and Widiger and Settle ($n = 33$). Baffling indeed.

Given their criticisms over the apparent neglect among psychologists about issues relating to safeguarding the literature, it will be instructive to see how well Kelley and Blashfield's (2009) paper itself is received. While still early days, Google Scholar informs us that as of February 2013 they have been cited only 3 times.

Findings based on single works can be highly misleading. They serve to underscore Kendall's (1961, p. 5) warning that "the pathway of knowledge is littered with the wreckage of premature generalization." But the seriousness of the lessons for knowledge development in the examples presented above go well beyond this reproof. By continuing to dwarf the influence of the very studies exposing their fallacious results, these works offer compelling evidence that science in the significant difference paradigm is not self-correcting. Like urban legends, erroneous results maintain a life of their own. This chilling message about whether we can trust what we read in the literature is reinforced explicitly in Sections 2.4, 4.4.1, 4.4.2, 5.4, 5.5.2, 7.4.1, and 8.3.2, and implicitly elsewhere throughout this book. We need many more replications, and far greater recognition (e.g., citations) and rewards (e.g., promotion, tenure) for those performing such essential tasks.

2.3 Model of Science—Hypothetico-Deductivism

This paradigm is linked inextricably with the hypothetico-deductive (H–D) model of scientific explanation. According to the philosopher Hausman (1992, p. 304), the essence of this method is captured in four steps:

1. *Formulate* a model or theory, *T*.
2. *Deduce* a prediction, hypothesis, or some other empirical proposition, *P*, from *T* in conjunction with a number of other auxiliary propositions. The latter would include descriptions of initial conditions, other relevant information (theories), and *ceteris paribus* ("other things being equal") qualifiers.
3. *Test P*, because *T* can be evaluated only indirectly in the H–D model.
4. *Judge* whether *T* is confirmed or disconfirmed on the basis of whether *P* turned out to be "true" or "false."

The H–D conception of the scientific method has dominated a large part of 20th century philosophical thinking. It is anointed in a number of preeminent works. These include Braithwaite's (1953) *Scientific Explanation*,

Popper's (1959) *The Logic of Scientific Discovery,* Nagel's (1961) *The Structure of Science,* and Hempel's (1965) *Aspects of Scientific Explanation.* Recall how Fisher's declaration that the use of statistical analysis could raise the status of a discipline found a receptive audience among applied researchers. This same promise of scientific respectability also was offered by proponents of the H–D model. Consider Braithwaite's (1953, p. 9) assurance: "It is this hypothetico-deductive method applied to empirical material which is the essential feature of a science; and if psychology or economics can produce empirically testable hypotheses, *ipso facto* they are sciences." Small wonder that researchers in these areas rallied to the cause.

As chronicled in Section 2.2.3, the pages of the leading management and social science journals are flooded with the empirical (statistical significance) testing of hypotheses, to the virtual exclusion of other means of data analysis. Economists typically are in the vanguard of compliance with the H–D method, and more apt to use formal/mathematical analyses to deduce propositions from some axiomatic system. Researchers in finance and accounting strive to emulate their colleagues in economics. In the "softer" areas, such as management, marketing, anthropology, geography, political science, social psychology, and sociology, the "deduction" of hypotheses tends to be more ad hoc—what Meehl (1990, p. 199) terms a "loose derivation chain." Mostly this follows from a review of pertinent literatures that would seem to make the hypotheses offered plausible.

Over time, the primacy reserved for theory and theory *testing* has resulted in the following commonly held opinions and/or behaviors:

- Hypothesis generation (discovery) and testing (justification) are viewed as quite separate and distinct activities, with philosophers of science focusing on the latter (Suppe, 1977). This same situation holds in the social science and business disciplines where the majority of scholars, as noted above, continue to be almost totally absorbed with justification rather than with discovery. Depending on one's philosophical orientation, the objective is to either confirm (logical positivism/empiricism) or falsify (Popper, 1959, 1963) theories. In contradistinction, the process of theory development has received much less attention (Haig, 2005; Hubbard & Lindsay, 2002; Hunter, 2001; Wells, 2001).
- It is the application of the "scientific method" which distinguishes science from non-science. And the H-D model, with its proclivity for theory testing, is seen widely as the very embodiment of *the* scientific method (Hubbard & Lindsay, 2013b, p. 1394). Therefore, if the correct methodological recipe is adhered to, this makes the output "science." In keeping with this outlook, science progresses by following an orderly, mechanistic, sanitized protocol. Apostles of the significant difference paradigm reject Tukey's

(1980) claims that science does not begin with tidy questions, nor end with tidy answers. For them, each research question (hypothesis) is framed in meticulous detail—H_1, H_2, H_3, H_{3a}, H_{3b}, H_{3c}, H_4, and so on. And the findings are reported, via *p*-values, with a seemingly impressive degree of precision, as if addressing decisively the topic(s) at hand in *that* moment. All research endeavors involve exact questions and answers.

- A single study (e.g., Broverman et al., 1970) can yield immediate and lasting knowledge. Such beliefs are revealed in the idea of the "crucial experiment," implying that theories can be conceived, evaluated, and finalized over a very short period of time (Bauer, 1994; Box, 1994; Haig, 2014; Lakatos, 1970).[14] This is seen in Hausman's (1992) point 1 above, where a theory is simply "formulated." It is why Locke (2007, p. 867) bemoans that the H–D approach demands premature theorizing, a consequence of which is that "theories tend to be grounded in myths and superstitions" (Van de Ven, 2007, p. 17) rather than facts. One manifestation of this view is the lack of replications that are conducted across the management and social sciences (see Chapter 5). Another is that traditional statistics courses bypass the development of theories and concentrate only on testing models in the "one-shot" context (Box, 1994; Chatfield, 1995; Ehrenberg, 1968). Still another, T. Clark, Floyd, and Wright (2006, p. 655), T. G. Gill (2010, p. 295), Hambrick (2007, p. 1346), Helfat (2007, p. 185), and McGrath (2007, p. 1373) admonish, is the requirement by editors of high-quality journals that papers must make a *theoretical* contribution, even when this comes at the expense of those that report interesting and well-documented facts. On this account, facts are subordinate to theory.
- A good theory produces (or should produce) universal generalizations (Aram & Salipante, 2003; S. R. Clegg & Ross–Smith, 2003; Flyvbjerg, 2001; Krebs, 2001; Starbuck, 2006; Teigen, 2002).[15] This idea of laws being universal finds modern expression in Hempel's (1965) deductive-nomological, or covering law, model of scientific explanation. Here, causation is interpreted in Humean terms as one of the constant conjunction of events yielding universal or lawlike empirical regularities. Bishop (2007, pp. 318–320, 401; see also Easton, 2002; Manicas & Secord, 1983; Tadajewski, 2008; Yu, 2006) argues that this discredited positivist philosophy and its quest for apodictic knowledge endures as the received view in the social sciences. Views like this explain why the conception of generalization, articulated in Chapter 6, follows the "representative model" of inferential statistics (Cook & Campbell, 1979). They also help to account for the propensity to overgeneralize results (Bamber et al., 2000; Wells, 2001) noted earlier, and to display a bias against publishing negative findings (Hubbard & Armstrong, 1992, 1997; Lindsay, 1994; Sterling, Rosenbaum, & Weinkam, 1995).
- Theory testing in this paradigm is carried out predominantly on single data sets. As such, considerations of *internal validity* are paramount (D. T. Campbell & Stanley, 1966)—can I reasonably conclude from this

data set that changes in the dependent variable are, in fact, attributable to the manipulation of the independent variable(s)? *External validity*, or generalizability, issues usually are ignored or downplayed (Laurent, 2000; McQuarrie, 2004; Rogers & Soopramanien, 2009; Rozin, 2009; Steckler & McLeroy, 2008; Wells, 1993, 2001; Winer, 1999) or else consigned to the "conclusions" section, where an appeal is made for future research to address this issue.

- It is imperative, from the standpoint of publishing the manuscript, that these one-shot studies using single data sets find statistically significant results that buttress the "originality" or "novelty" claims made in the work. Once this creative idea is accepted for publication, the researcher switches to another novel topic. In other words, what should be seen as the first step in possibly establishing the tenability of a result all too often is accepted as the last word on the research problem. In this manner, the worshiping of original research with statistically significant outcomes guarantees the prevention of a cumulative body of knowledge against which to judge the credibility of future results. Nelder (1986, p. 112), in his presidential address to the Royal Statistical Society, disparaged this practice, calling it the "cult of the isolated study." This cult is ubiquitous in the management and social sciences. Its legacy is highly damaging, namely, a literature composed chiefly of fragmented, one-off results whose contributions to knowledge are of the most speculative kind. As an example, why should anyone care about the fact that intrinsic motivation was found ($p \leq .05$) to improve salesforce satisfaction if this has been found only once, for 56 salespeople in two companies in the United Kingdom in 1995 (Hubbard & Lindsay, 2002, p. 386).
- Good science preaches the virtues of *descriptive* correlational analyses. Fair enough. The downside is that insufficient attention is directed at the search for causal *explanations* (Bunge, 1997, 2004), the latter defined as "clarifying the mechanisms through which and the conditions under which the causal relationship holds" (Shadish, Cook, & Campbell, 2002, p. 9). While both description and explanation are vital for knowledge accumulation, it is the discovery of these mechanisms that is scientists' only protection against the "problem of confounders," something which is always a threat in observational (nonrandomized) studies (Hubbard & Lindsay, 2013b, p. 1395).

Given the above beliefs, it is understandable why tests of statistical significance have enjoyed such a privileged status in conventional social and business research methodology.

Unfortunately, the H–D model of science on which the behavioral and business disciplines rest squarely is deeply flawed. On the one hand it is susceptible to the Duhem–Quine argument, which states that one never tests scientific hypotheses by themselves (Hausman's, 1992, point 2 at the beginning of Section 2.3). When deducing a hypothesis from some

theory, numerous auxiliary propositions and initial and boundary conditions are conjoined with the deduced hypothesis. Accordingly, if a hypothetical prediction "fails" (e.g., the null is not rejected), this could be due to the falsity of one or more of the auxiliary propositions, and so on. In this manner we can continue to defend the failed hypothesis by blaming endlessly the support cast, thus leading to a state of "infinite regress" over the truth/falsity of the deduced hypothesis. On the other hand, the H–D model also is vulnerable to the fallacy of affirming the consequent. If a hypothetical prediction "succeeds" (e.g., the null is rejected), there can be other mutually incompatible hypotheses that entail identical predictions. The H–D approach is unable to discriminate between these competing alternative hypotheses, even if some of them are theoretically absurd (Ketokivi & Mantere, 2010, p. 318). The upshot, econometricians Darnell and Evans (1990, p. 37) observe, is that no attempts to test theories—whether by confirmation or falsification—can ever be conclusive.

Because of this, philosophers of science (e.g., Hausman, 1992; Nola & Sankey, 2007) have largely forsaken the epistemological warrant proffered by the H–D model; Psillos (1999, p. 174) labels it crude, Glymour (1980a, p. 322, 1980b, p. 36) calls it hopeless, Gorski (2004, p. 28) says it doesn't work, and Gower (1997, p. 15) states that it cannot ratify the acceptance of any scientific belief. Of course, no path to knowledge procurement is infallible. What is jarring, however, is that while the H–D method faces severe philosophical difficulties, one finds little evidence of such concerns in the behavioral and business disciplines. At the same time the steadfast hewing to the dictates of hypothetico-deductivism places the conduct of research in these same areas at odds with practicing investigators in all the natural sciences (Barwise, 1995, p. G32). In other words, members of the significant difference paradigm accede to a conception of knowledge development that neither philosophers nor practitioners of science endorse. No matter; to them the H–D model remains sacrosanct.

2.4 The Role of "Negative" ($p > .05$) Results

2.4.1 Publication Bias and the Credibility of Empirical Findings From Individual Studies

Many researchers in this paradigm are convinced that an editorial-reviewer bias exists in the management and social (and biomedical) sciences against publishing so-called negative or null results (see, e.g., Bakan, 1966; Bakker, van Dijk, & Wicherts, 2012; Doucouliagos & Stanley,

2013; Fanelli, 2009; Feige, 1975; C. J. Ferguson & Brannick, 2012; Gerber & Malhotra, 2008; Greenwald, 1975; Hubbard, 1995b; Hubbard & Armstrong 1992, 1997; Ioannidis, 2005b; Ioannidis & Trikalinos, 2007; Kruskal, 1978; Lindsay, 1994; Masicampo & Lalande, 2012; Mayer, 1993; Nosek, Spies, & Motyl, 2012; Salsburg, 1985; Simmons, Nelson, & Simonsohn, 2011). By the latter are meant results incompatible with predictions, but more broadly regarded as statistically insignificant, or $p > .05$, outcomes.[16] For instance, almost 50 years ago, when commenting on a former editor (Arthur W. Melton) of the *Journal of Experimental Psychology,* Bakan (1966, p. 427) had this to say: "His clearly expressed opinion that non-significant results should not take up the space of the journals is shared by most editors of psychological journals." Such views have persisted, indeed intensified, over the years, allowing Scott Maxwell (2004, p. 148) to broadcast without fear of contradiction that "typical editorial practices virtually [mandate] statistical significance as a prerequisite for publication," meaning that $p \leq .05$ findings must be secured "at all costs" (Fanelli, 2010, p. 1; Ioannidis, 2012, p. 647).

Fisher (1966) had an early hand in fanning this publication bias in two ways. First, he wrote that "the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation" (p. 16). Given the asymmetry in this proposition, it is easy to see why editors, reviewers, and researchers could interpret "insignificant" results as being, at best, inconclusive and thus less worthy of publication. In comparison, the rejection of H_0 comes across as definitive. Second, when selecting, arbitrarily, the .05 level to demarcate statistical significance, Fisher instructed experimenters to be "prepared to ignore all results which fail to reach this [.05] standard" (p. 13). So researchers are left with the distinct impression that $p > .05$ results are irrelevant.[17]

At issue is that this publication bias against null findings may seriously distort the content of an empirical literature, even to the degree where its probity cannot be taken for granted. To see this, consider the three options available to an investigator who obtains initially $p > .05$ results. First, null results are not as likely to be submitted for publication. Second, if submitted, null results are more likely to be rejected for publication. Third, confronted with null results the researcher opts to continue with the topic by searching for $p \leq .05$ outcomes. These three scenarios, all of which compromise trust in the scientific enterprise, are addressed in turn.

Negative Results Are Less Likely to Be Submitted for Publication

For example, Greenwald (1975), using intentions data from 36 *Journal of Personality and Social Psychology* authors, estimated the probability of

null results being submitted for publication to be .06, compared with .59 for those with "significant" results. Coursol and Wagner's (1986) inspection of 609 responses from a survey of counseling psychologists found that while 82% of the articles reporting positive outcomes were submitted for publication, this figure was only 43% for those with neutral or negative findings. The fact that scholars are less willing to offer manuscripts with $p > .05$ findings for publication misrepresents the volume of empirical work carried out in any given research area (C. J. Ferguson & Heene, 2012, p. 556; Gerber & Malhotra, 2008, p. 5).

Negative Results Are Less Likely to Be Published

Part of the explanation for this is that manuscripts with null findings often are considered to reflect poorly on the researcher's skills rather than on nature, and so once again are thought to be undeserving of publication (see, e.g., Hubbard & Lindsay, 2013b, p. 1394). There is some justification for this. For example, research in both the social (Bakker et al., 2012; J. Cohen, 1988; S. E. Maxwell, 2004; Ottenbacher, 1996) and management (Cashen & Geiger, 2004; T. D. Ferguson & Ketchen, 1999) sciences reveals that many studies have inadequate statistical power to reject a false H_0, a topic of sufficient concern as to necessitate further discussion in the appendix to this chapter. Alternatively, a null finding may simply indicate the absence of any substantive effect in the population (Nickerson, 2000, p. 261).

There is, however, weighty evidence to show that null results are less likely to be published than their non-null peers. Steven Kerr, James Tolliver, and Doretta Petree (1977), for example, surveyed 429 editors and advisory board members of 19 leading management and social science journals to elicit common reasons for manuscript acceptance or rejection. They reported that even when a manuscript was judged to be otherwise competent and of current interest to the field, $p > .05$ results markedly lowered the likelihood of acceptance. Parallel results were found in a survey of 268 manuscript reviewers for Canadian psychology journals (Rowney & Zenisek, 1980). In addition, Atkinson, Furlong, and Wampold (1982) asked 101 consulting editors of two psychology journals to evaluate three versions of a manuscript that differed only with respect to the level of statistical significance attained. The statistically insignificant and almost significant versions were more than three times as likely to be rejected for publication than was the statistically significant one. Epstein (2004) corroborated Atkinson et al.'s findings in a similar study in the field of social work.

Many authors share the same beliefs as those held by editors and reviewers. For instance, 61% of authors who had published empirical

articles in various education and psychology journals in 1988 were of the opinion that only research yielding statistically significant findings would be published (Kupfersmid & Fiala, 1991).

Six previous empirical studies from the social sciences, five employing databases from psychology and one from sociology, suggest the existence of a publication bias against null outcomes. For example, Sterling (1959) examined 362 empirical papers published in the 1955 issues of the *Journal of Clinical Psychology, Journal of Experimental Psychology, Journal of Social Psychology,* and the 1956 issue of the *Journal of Comparative and Physiological Psychology.* He found that only 2.7% of those using significance tests failed to reject the null hypothesis. Smart's (1964) analysis of these same four journals in 1962 showed that 8.7% reported null results. Bozarth and Roberts (1972) determined that only 6% of 1,046 articles using statistical tests in the *Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology,* and *Personnel and Guidance Journal* between January 1967 and August 1970 were unable to reject the null hypothesis. Greenwald's (1975) estimate, based on a content analysis of a single annual (1972) issue of the *Journal of Personality and Social Psychology*, was 12.1%. Sterling et al. (1995), using 1986–1987 data, repeated the earlier study by Sterling (1959) and discovered that only 4.4% of articles failed to reject H_0. The proportion of negative results was higher in sociology. Following a review of the *American Journal of Sociology, American Sociological Review,* and *Social Forces* from July 1969 to June 1970, for instance, Wilson, Smoke, and Martin (1973) found that 19.5% (15/77) of empirical articles had findings which were not statistically significant at the $p \leq .05$ level.

In the business fields, Lindsay (1994) surveyed all empirical budgeting and control articles published in three major accounting journals—*Accounting, Organizations and Society; Journal of Accounting Research*; and *The Accounting Review*—during the period 1970–1987. This procedure identified a total of 38 usable empirical studies, of which 6 (16%) were classified as yielding negative results. And based on a random sample of articles from three leading marketing journals—*Journal of Consumer Research, Journal of Marketing,* and *Journal of Marketing Research*—it was estimated that some 7.8% of the manuscripts published between 1974 and 1989 were unable to reject the null hypothesis (Hubbard & Armstrong, 1992). Moreover, this number has been declining over time; for 1974–1979 it was 11.4%, while for 1980–1989 it was 5.7%.

A recent article incorporating all major disciplines supports the existence of a publication bias in favor of positive and against negative results. Fanelli's (2010, p. 2) study, consisting of a random sample of 1,316

articles from 49 U.S. states (Delaware excepted) and the District of Columbia, showed that only 17.6% of papers told of negative findings. For 5 states this value ranged between 2% and 10%, while investigators from another 8 states recorded zero negative outcomes. Of interest, Fanelli also provides information that more successful researchers—those with a greater publication output—report fewer null results than their less prolific colleagues.

Evidence from the medical literature also suggests a publication bias against null findings. As an example, Simes (1986) reported that whereas the pooled results for published trials for a particular treatment of ovarian cancer showed statistically significant benefits, the pooled results of registered trials (which included both published and unpublished studies) evaluating the same treatment did not. In like manner, Dickersin, Chan, Chalmers, Sacks, and Smith (1987) contacted 318 authors of published clinical trials to see if they had been involved with any unpublished ones. Responses from 156 of them yielded 271 unpublished and 1,041 published trials; while only 14% of the unpublished reports favored the test therapy, this figure was 55% for the published results.

The truth of the matter, as Fanelli (2012) shows, is that negative results are disappearing from the physical and biological sciences, and especially from the management and social sciences, in most countries. This is a great concern for her, owing to the fact that "negative data . . . are crucial to scientific progress, because this latter is only made possible by a collective self-correcting process" (p. 892).

Negative Results Obligate Researchers to Search for $p \leq .05$ Outcomes

On finding null results the investigator decides to persevere with the topic at hand. Greenwald (1975) asked his sample if they were likely to conduct an exact or modified replication of their work following an original full-scale test of their main hypothesis. If the original result was statistically significant, the probability was .36; if an insignificant result was obtained, the probability was .62.

This perseverance, remember, is driven almost solely by the need to offer statistically significant findings to the editor if the manuscript is to have a chance of appearing on the printed page. Under circumstances like these, it is reasonable to speculate that some individuals will be drawn to questionable research behaviors aimed at providing editors-reviewers with what they want to see.

One such questionable behavior in this context is data mining. An activity which consists of tirelessly reworking the data in a search for

statistically significant outcomes, data mining is said to encourage the proliferation of Type I errors (erroneous rejections of the null hypothesis or false-positive results) well in excess of their nominal .05/.01 levels (see Chatfield, 1995; Denton, 1985, 1988; Feige, 1975; Hubbard & Vetter, 1996; Lindsay, 1994, 1997; Lovell, 1983). Or as Leamer (1983, pp. 36–37) put it in an article drolly titled "Let's Take the Con Out of Econometrics," after fitting endless models to data the researcher carefully selects from the bramble of computer output the one she or he enshrines as a rose.

Greenwald (1975, p. 15) believes the underestimation of Type I errors to be "frightening, even calling into question the scientific basis for much published literature." Wilson et al. (1973) had earlier come to this same conclusion with respect to findings in sociology. Fears about the damaging impact of such practices on the integrity of results continue undiminished (cf. C. J. Ferguson & Heene, 2012, p. 558; Pashler & Wagenmakers, 2012, p. 528), culminating in literatures "infested with error" (Holcombe & Pashler, 2012, p. 355).

A variation on this data mining theme is what Norbert Kerr (1998, p. 196) refers to as HARKing (Hypothesizing After the Results are Known), or presenting post hoc (based on one's statistically significant results) hypotheses as if they were a priori in nature. Such a practice, of course, cuts the legs out from under the logic sustaining the H-D model of explanation. Never mind; according to Kerr, HARKing, which raises fortuitous results to center stage, is prevalent in academe. In backing this claim, he tells that a survey of 156 behavioral scientists in social psychology, clinical/community psychology, and sociology revealed that about 45% of them had personally observed, and 55% suspected, instances of HARKing among colleagues. Kerr's results have found emphatic validation. Based on a 20% response rate from a survey of 1,940 management faculty at 104 PhD-granting business schools in the United States, Bedeian, Taylor, and Miller (2010, p. 716) found that 92% reported knowledge of researchers within the previous year developing hypotheses after the results are known.

Moreover, being driven mostly by the need to offer $p < .05$ results supporting one's ideas, HARKing contributes directly to the problem of confirmatory bias, while simultaneously eliminating an opportunity to falsify a theory (Leung, 2011, p. 475). This indicates again that science following a significant difference philosophy tilts toward the production of suspect empirical literatures.

In concert with the above strategies, persevering with a research topic might include the duplicitous reporting by investigators of only those results confirming ($p < .05$) their expectations and/or the suppression of

"inconvenient" data. Fanelli (2009, p. 6) estimates that some 9.5% of researchers are complicit in this regard. Her valuation seems low in comparison with others. Thus, for example, Bedeian et al. (2010, p. 716) discovered that 77.6% of their survey respondents indicated awareness of management researchers selecting only those data supporting hypotheses, while 59.6% claimed to know of scholars discarding observations felt to be inaccurate. John, Loewenstein, and Prelec's (2012, p. 527) online survey of 5,963 psychologists at American universities—response rate 36.1% (2,155)—put these same estimates at 67% and 62%, respectively.

Empirical support for other common deceptive research habits include withholding methodological specifics or results, using another's ideas without permission or giving credit, not revealing data that contradict the author's own earlier work, and publishing the same data or results in two or more outlets (see Bedeian et al., 2010, p. 716, for details). Further sins of omission and commission involve failing to report all dependent measures and all study conditions, ceasing data collection after obtaining the desired result(s), rounding down *p*-values, and claiming to have predicted an unexpected outcome (see John et al., 2012, p. 527, for details).[18]

The ultra-competitive "publish or perish" environment typical of the significant difference school may induce some researchers to go so far as to commit fraud in their efforts to get into the presses, say, by fabricating or falsifying data. Evidence, and not just high-profile cases such as that of the Dutch social psychologist Diederik Stapel (Carpenter, 2012, p. 558), supports this grave concern. For instance, responses from 250 of 663 (38%) accounting faculty who published extensively in that field's 30 top journals found 4% of them admitting to falsifying results, while further believing that 21% of their literature is so tainted (Bailey, Hasselback, & Karcher, 2001, p. 35). Comparable findings emerged from answers to a survey of 1,000 economists (234 returns) administered at the January 1998 meetings of the American Economic Association in Chicago: 4.4% said they have falsified research data, and they thought that about 6% of articles published in the best economics journals are based on such input (List, Bailey, Euzent, & Martin, 2001, p. 166). As another example, in a meta-analysis of 18 surveys, when scholars were asked, anonymously, whether they had ever fabricated or falsified data, some 2% said yes (Fanelli, 2009, p. 6). When survey questions were framed in relation to the behavior of fellow professionals, 14% of respondents said they had personal knowledge of a colleague who had fabricated or falsified data (Fanelli, 2009, pp. 6–7). Replies to John et al.'s (2012, p. 527) poll of psychologists indicates that about 9% of them confessed to falsifying

data. Meanwhile, Bedeian et al. (2010, p. 719) write that 26.8% of U.S. management faculty answering their questionnaire reported knowledge of researchers fabricating outcomes.

John et al. (2012, p. 524) label questionable research practices the "steroids" of scientific competition, artificially bettering the careers of those who adopt them while penalizing scholars who play by the rules. An initial reaction might be that this is an odd and inappropriate analogy. It is, however, a peculiarly apt one conveying as it does the ruthlessness attending a "win at all costs" outlook pervasive in both the sporting and academic arenas.

In view of the sensitivity of the subject matter, few would be surprised if the estimates on the incidence of fabricating/falsifying data and other dubious research practices documented above turn out to be low. On this topic, John et al. (2012, p. 524) suggest that participating in questionable research behaviors "may constitute the de facto scientific norm." Honig, Lampel, Siegel, and Drnevich (2013, p. 2) agree when commenting that rising trends of ethical misconduct are not attributable to individual lapses but tell instead of "systemic problems that are deeply embedded in the institutional fabric of the modern research process." This is not hyperbole; compelling evidence bolsters John et al.'s and Honig et al.'s conjectures. For example, Masicampo and Lalande's (2012, pp. 2272–2273) analysis of the distribution of 3,627 *p*-values found in the August 2008 issues of the *Journal of Experimental Psychology: General, Journal of Personality and Social Psychology,* and *Psychological Science*, along with the preceding 11 issues of each, shows them to be much more common immediately below the arbitrary .05 level than would be expected on the basis of their occurrence in other ranges. This anomaly, which they attribute to undue emphasis on the attainment of $p \leq .05$ outcomes, was present in all three journals. Similar results indicative of the warping influence of the .05 level on the distribution of *p*-values are conspicuous in Gerber and Malhotra's (2008) examination of 46 data-based articles published in the 2003–2005 issues of the *American Sociological Review* and *Sociological Quarterly,* and the 2003–2006 issues of the *American Journal of Sociology*.[19] They concluded that "the hypothesis of no publication bias can be rejected at approximately the 1 in 10 million level" (p. 3).

Findings like these accentuate the insidious effects that editorial-reviewer biases against $p > .05$ results can have on the integrity of a discipline's contributions to knowledge. Perversely, then, the current academic incentive structure rewards the publication of nonreplicable findings (Hartshorne & Schachner, 2012, p. 1), a conclusion so disheartening that

some (e.g., Ioannidis, 2012; Sovacool, 2008; Stroebe, Postmes, & Spears, 2012, p. 681) contest as mythical the idea that science is self-correcting. This same verdict was reached earlier in Section 2.2.5.

2.4.2 Publication Bias and the Credibility of Meta-analyses

Of immediate concern, owing to the publication bias against negative results, meta-analyses (i.e., attempts to quantitatively summarize the empirical literature on a particular issue) conducted under the aegis of the significant difference paradigm may not be as helpful as first imagined. They can even lead us astray. I agree with Sohn (1996, p. 229) that caution should be exercised in crediting meta-analyses as vehicles for knowledge discovery. This is because, as shown quite dramatically above, their databases (the available literatures constituting their input) may be untrustworthy. In fact, sometimes the findings of a meta-analysis may not be replicable (see, e.g., Allison & Faith, 1996; Bullock & Svyantek, 1985; Felson, 1992; Kilpatrick, 1992), thereby eviscerating its purpose.

Proliferation of False Positives

Behaviors motivated by a "hunt" for statistically significant differences (Salsburg, 1985, p. 220), or what Imrey (1994, p. 68) calls *p*-varication, contribute to Rosenthal's (1979, p. 638) well-known *file drawer problem* where "journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (i.e., $p > .05$) results." Needless to say, a field whose literature is corrupted with false-positives is one whose credence is in danger (Simmons et al., 2011, p. 1359).

Regrettably, information based on examinations of various meta-analyses carried out in the social science and medical areas reveals them to be victims of said corruption. Illustrative of this condition, Bakker et al.'s (2012, p. 543) review of 13 meta-analyses encompassing 281 primary studies in a spectrum of fields in psychology uncovered biases and/or an excess of $p \leq .05$ outcomes in 7 of them. Christopher Ferguson and Michael Brannick's (2012, pp. 122–124) conclusions, following an inspection of 48 meta-analyses selected from leading American Psychological Association (*American Psychologist, Developmental Psychology, Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology, Journal of Personality and Social Psychology, Psychological Bulletin*) and Association of Psychological Science (*Perspectives on Psychological Science, Psychological Science*) journals over the period 2004–2009, also

are bothersome. They determined that publication bias was apparent in 25% of these meta-analyses. In a similar vein, Ioannidis and Trikalinos (2007, p. 245) saw exaggerated frequencies of statistically significant results in 6 of 8 biomedical meta-analyses (with 55–155 studies in each one). What confirms the suspiciousness of these overly optimistic findings is the implausibly high occurrence of $p \leq .05$ results given the low levels of statistical power characteristic of meta-analyses undertaken in these disciplines (see, e.g., Bakker et al., 2012, p. 543; Cafri, Kromrey, & Brannick, 2010). Outcomes of this sort bear out Ioannidis's (2011, p. 16) grievance that an epidemic of false claims "is rampant in the social sciences and . . . particularly egregious in biomedicine."

Inflation of Effect Sizes

A related drawback with respect to the publication bias against negative findings and the usefulness of meta-analyses must be raised. Because unpublished (file-drawer) works—thought to be at least 50% of the psychology literature (Bakker et al., 2012, p. 544)—often are underrepresented in meta-analyses, published effect size estimates are inflated. For example, 10 of the 12 education and psychology meta-analyses examined by Mary Smith (1980) showed average effect sizes in published journal accounts to be 33% higher than those reported in theses and dissertations. McLeod and Weisz's (2004) comparison of youth psychotherapy outcomes reported in journals versus unpublished dissertations likewise breeds skepticism. They revealed effect sizes in the former to be twice the size of those in the latter. Continuing, Shadish, Doherty, and Montgomery (1989) surveyed a random sample of 519 members of organizations involved with family and marital psychotherapy outcomes to see if they possessed file drawer studies on the issue. After analyzing 375 responses, they intimated that there may be almost as many of these unpublished works as there are published studies and dissertations. They concluded that population effect sizes of published works are about 10% to 40% larger than those based on unpublished research. Information on the efficacy of psychological, educational, and behavioral treatment casts additional doubt on the veracity of meta-analytic effect size estimates. Lipsey and Wilson's (1993, pp. 1194–1195) dissection of a subset of 92 (from 302) meta-analyses permitting a comparison of effect sizes in published versus unpublished venues in these areas shows published accounts to be some 13% higher than file-drawer appraisals. A simulation study by David Lane and William Dunlap (1978) also shows disquieting results.

They found that editorial biases in favor of .05 and .01 outcomes led to published average effect sizes being one-half to one standard deviation greater than their true values.

Summing up, publication bias against negative outcomes arising from tests of statistical significance directly contributes to the creation of an empirical literature whose believability is suspect. Howard et al. (2009, p. 148) go beyond this in their withering assessment that "because of the widespread use of NHST [null hypothesis significance testing] in psychological research, it is therefore possible that all extant research literatures are systematically misleading and some demonstrably wrong." Similarly, Doucouliagos and Stanley's (2013, p. 332) meta-analysis of findings from a broad range of areas in economics led them to comment that "all summaries of empirical economics . . . must be regarded with some skepticism" and that "reports of economic facts are greatly exaggerated." Presumably, views like these underlie Christopher Ferguson and Moritz Heene's (2012, p. 558) position that seldom have they run across a meta-analysis which has resolved a controversial debate in an area.[20]

2.5 Conclusions

Affiliates of the significant difference model have an unquestioning impression of how knowledge is gained. It is enough to be able to reject the null hypothesis at the $p \leq .05$ level in single studies and then generalize these unique findings to other circumstances and time periods. Further, negative ($p > .05$) results are viewed with disdain. Given these beliefs, it is easy to comprehend how the use of statistical significance testing has come to occupy a vise-like grip on empirical research performed in the management and social sciences. Devoid of such testing, data-based research in these areas is considered to be insufficiently rigorous. The insistence on $p \leq .05$ outcomes leads researchers eager to generate them into improprieties which impugn the validity and reliability of published findings. That just about all empirical articles in the behavioral and management sciences are able to come up with statistically significant results in the face of typically underpowered designs (see the appendix to this chapter) is highly improbable. In some instances, such findings are just too good to be true (Francis, 2012a, p. 585, 2012b, p. 151; Schimmack, 2012, p. 561). This makes it almost certain that fallacious results are entering the literature at troubling rates (Asendorpf et al., 2013, p. 113; Pashler & Harris, 2012, p. 535). And in

areas like biomedicine this manipulation of *p*-values in the quest for publishable results can harm patients (Stang, Poole, & Kuss, 2010, p. 225) and cost lives (Ziliak & McCloskey, 2008).

Followers of the significant difference paradigm endorse wholeheartedly a model of explanation—hypothetico-deductivism (HARKing excepted)—that for the most part has been abandoned by philosophers of science. It is a model that rarely describes the work of practicing scientists as is sometimes falsely portrayed in textbooks.

In short, the significant difference paradigm legislates bad science (Hubbard & Lindsay, 2013b). The philosophical foundations of the significant sameness model, when juxtaposed with those used to rate the merits of the significant difference paradigm—conception of knowledge, model of science, and the role of "negative" results—open a far sturdier route to knowledge development in the management and social sciences. These foundations are laid out in Chapter 3.

Notes

1. The idea that rare occurrences constitute evidence against a hypothesis has a pedigree dating back to the first "significance test" by John Arbuthnot in 1710 concerning the birth rates of males and females in London, and is continued in the work of Mitchell, LaPlace, and Edgeworth, among others. See Baird (1988), Cowles (2001), Gigerenzer et al. (1989, pp. 79–84), Hacking (1965, pp. 74–80), Huberty and Pike (1999, pp. 1–4), and Sauley and Bedeian (1989, pp. 336–338) for synopses of this early history of statistical testing.

2. The reader will appreciate that the data presented in this chapter have been amassed over a period of several years.

3. The arbitrariness of later classifying economics as a management rather than a social science is recognized.

4. Note that all journals except *Professional Geographer* (1949) and *American Journal of Political Science* (1957), with inaugural issues in parentheses, cover the entire 1945–2007 time period.

5. For every year, two randomly chosen issues of management and marketing journals were included in the database.

6. A growing imbalance between empirical and nonempirical/conceptual articles is unfortunate because the latter often have a greater impact on the scholarly community (MacInnis, 2011, p. 151), at least as measured by citations (for details, see Hubbard, Norman, & Parsa, 2010; Yadav, 2010). In addition, some practitioners may find conceptual pieces readable, interesting, and helpful.

7. See also Shirley Martin's (2011) informative telling of the spread of Fisherian methods in experimental dissertations in education under the guidance of Lindquist at the University of Iowa and Johnson at Minnesota during World War II.

8. This last figure for sociology is almost identical with Leahey's (2005, p. 3) estimate that 91% of empirical articles published in the *American Journal of Sociology* and *American Sociological Review* for the period 1995–2000 used tests of statistical significance.

9. In Guttman's (1985, p. 3) words, many researchers don't comprehend the distinction between statistical inference (an aspect of what he calls the scaffolding of science) and science (the building or structure) itself. Thus, they believe that they are doing science when they are dealing merely with part of the scaffolding. Which is to say they are not building anything. In this connection see also Bolles's (1962) telling of the differences between statistical versus scientific hypotheses, as well as Meehl (1967, p. 107, 1990, p. 202).

10. A book edited by Harlow, Mulaik, and Steiger (1997) titled *What If There Were No Significance Tests?* confronts just such an issue.

11. Upsettingly, Hubbard, Brodie, & Armstrong (1992, p. 5) add, Simon's reappraisal of Zielske's data took years to get into print. He was so disenchanted with the "saga," as he named it, of trying to publish his results that he thought it necessary to tabulate in an appendix to his eventual publication the events propelling such an unusual course of action. The contents of the appendix do not paint a picture of scholarly cooperation on Zielske's part. See Chapter 5 for more evidence of this lack of collaboration among researchers.

12. All Google Scholar citations reported in this section were accessed in February 2013.

13. It is unique in another way not addressed by Bottomley and Holden (2001). That they were actually able to get these eight data sets is quite remarkable given that, as detailed in Section 5.5.2, researchers in the social and management sciences are not especially cooperative in such matters.

14. Keuzenkamp (2000, p. 160) proposes that the idea of the "crucial experiment" probably is rare in physics—it may even be a scientific fiction (Hubbard & Lindsay, 2002, p. 393)—never mind in the social sciences.

15. According to Giere (1999, p. 89), the notion that laws of nature must be true statements of universal form had its origins in theology:

> Here there can be no serious doubt that, for Descartes and Newton, the connection between laws of nature and God, the creator and lawgiver was explicit. Nor can there be any doubt that it was Newton's conception of science that dominated reflection on the nature of science throughout the eighteenth century, and most of the nineteenth as well.

Needless to say, God's laws were universal. In addition, Giere continues, it was the secularized version of Newton's characterization of science that prevailed in philosophy of science matters through a substantial part of the 20th century, as exemplified by the logical empiricists.

16. To this end, Keuzenkamp and Magnus (1995, p. 18) joke that the JRSS (*Journal of the Royal Statistical Society*) actually is the JSSR (*Journal of Statistically Significant Results*)!

17. Fisher (1973, p. 45) later backed away from this position: "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." Such advice notwithstanding, the inviolability of the .05 significance level remains to this day.

18. At this juncture I offer my own mea culpas. I have searched for statistically significant results, for example, by experimenting with alternative model specifications, variable transformations, or by recasting one- and two-sided null hypotheses. And I have engaged in HARKing, on one occasion at the behest of a referee who asked "Where are the hypotheses?" in response to my original manuscript submission, which was completely free of them. Few are exempt from the inordinate burden of having to publish the "required" results.

19. That only 46 empirical articles were retained for analysis mirrors the rather stringent criteria necessary for inclusion in Gerber and Malhotra's (2008) "caliper" test for publication bias.

20. Compounding these publication bias deficiencies, in the significant difference model studies included in meta-analyses may suffer from marked construct validity ("comparing apples and oranges") and sample size (insufficient replications on the "same" topic) drawbacks. Or as Bangert-Drowns (1986, p. 388) has it, because studies involved in typical meta-analyses deploy only "roughly similar procedures," problems occur when averaging effects across independent and dependent variables which have been measured quite differently. Therefore, it frequently is unclear whether those studies selected for incorporation in meta-analyses are measuring the same constructs or relationships. This makes the results of conventional meta-analyses even more difficult to interpret and generalize (cf. Farley & Lehmann, 1986, pp. 15–16; Shadish et al., 2002, pp. 446–455). Note that because of the priority assigned to replication research, construct validity, sample size, and publication bias issues are minimal for meta-analyses undertaken in the significant sameness paradigm.

APPENDIX TO CHAPTER 2

An Empirical Regularity Not to Be Proud Of: Inadequate Statistical Power in the Social and Management Sciences

In the Neyman–Pearson variant of the significant difference paradigm, the researcher proposes two hypotheses. One is the null hypothesis, H_0, which the investigator is hoping to reject (nullify) in favor of the other, the alternative (research) hypothesis, H_A. Unfortunately, two kinds of mistakes can happen in deciding between H_0 and H_A. There is the erroneous rejection of H_0 called a Type I error, α. And there is the erroneous acceptance of H_0 called a Type II error, β. Statistical power, then, is defined as 1-β, or the probability of rejecting a false null hypothesis.

When statistical power in a research literature is low or marginal significance tests are prone to yielding null results, and therefore may be viewed as less worthy of being published. At the same time, and acting in the opposite direction, when power is low

> the veracity of even statistically significant results may be questioned, because the probability of rejecting a true null hypothesis may then be only slightly smaller than the probability of rejecting the null hypothesis when the alternative is true. . . . Thus, a substantial proportion of published significant results may be Type 1 errors. (Rossi, 1990, p. 647)

As can be inferred from the above, low power produces an inconsistent and questionable empirical literature wherein the results of some studies are statistically significant while those of others are not. This last point is vital and therefore receives explicit consideration in Section 3.5 of the book, a more suitable context for its discussion.

The power of a statistical test, Jacob Cohen (1969, 1988, p. 4) notes, is predicated on the chosen level of significance, the effect magnitude in the population, and the size of the sample. Accordingly, Cohen (1988) facilitates the calculation of power exhibited in published work by supplying power tables with different significance levels, standardized effect magnitudes, and sample sizes for several commonly used statistical techniques. The latter include the *t* test, the statistical significance of Pearson's *r*, the statistical significance of the differences between correlation coefficients, the sign test, the test for differences between proportions (%'s), χ^2 (chi-square) tests,

the *F* test in analysis of variance (ANOVA), analysis of covariance (ANCOVA), and multiple regression analysis, as well as power tests for multivariate analysis of variance (MANOVA) and multivariate analysis of covariance (MANCOVA). The reader is invited to refer to Cohen (1988) about the details involved in computing power levels for what he calls small, medium, and large effect magnitudes.

Cohen (1969, 1988) offers five guidelines when performing a power analysis: (1) the article as the primary unit of analysis, (2) a significance level of .05, (3) nondirectional tests, (4) only major statistical tests included, and (5) conventional definitions of small, medium, and large effect sizes. In addition, Cohen (1988, p. 56) recommends that the value .80 be used when there is no other rationale for selecting a desired power level.

Table 2A-1 displays the results of 16 retrospective power analyses of articles in leading journals (putatively reflecting best statistical practice) from a variety of behavioral and business disciplines. All of these studies followed Cohen's five recommendations.

Only 2 of the 16 studies displayed in Table 2A-1, both in marketing (Hubbard & Armstrong, 1991; Sawyer & Ball, 1981), are able to meet (in fact, surpass) Cohen's (1988) .80 benchmark for detecting medium effects in the population. This is probably because of the larger sample sizes enjoyed in survey, as opposed to experimental, research in marketing journals.

Across all 16 studies in Table 2A-1, the average probabilities of unearthing small, medium, and large effects are .24, .64, and .86. Minus the two marketing studies the corresponding figures are .22, .61, and .84. Consequently, if medium effect sizes are thought to be the norm in the social and management sciences, an investigator has only about a 60% chance of rejecting a false null hypothesis. This represents a substantial empirical regularity in the significant difference paradigm. Regrettably, it is not one to be proud of.

Table 2A-1 Statistical Power in the Social and Management Sciences

Discipline	*Investigators*	*Sample Sizes*		*Effect Sizes*		
		Articles	*Tests*[a]	*Small*	*Medium*	*Large*
Accounting	Lindsay (1993b)	43	1,871	.16	.59	.83
	Borkowski, Welsh, & Zhang (2001)	96	1,782	.23	.71	.93
Communication	Chase & Tucker (1975)	46	1,298	.18	.52	.79
	Chase & Baran (1976)	48	701	.34	.76	.91
Education	Brewer (1972)	47	373	.14	.58	.78
Management	Mazen, Hemmasi, & Lewis (1987)	84	7,215	.31	.77	.91
	Mazen, Kellog, & Hemmasi (1987)	44	3,665	.23	.59	.83
MIS[b]	Baroudi and Orlikowski (1989)	57	149	.19	.60	.83
Management and psychology	Mone, Mueller, & Mauland (1996)	210	26,471	.27	.74	.92
Marketing	Sawyer & Ball (1981)	23	475	.41	.89	.98

(Continued)

Table 2A-1 (Continued)

Discipline	*Investigators*	*Samples Sizes*		*Effect Sizes*		
		Articles	*Tests*[a]	*Small*	*Medium*	*Large*
	Hubbard & Armstrong (1991)	14	92	.39	.90	.96
Psychology	J. Cohen (1962)	70	2,088	.18	.48	.83
	Chase & Chase (1976)	121	3,373	.25	.67	.86
	Sedlmeier & Gigerenzer (1989)	54	—	.21	.50	.84
	Rossi (1990)	221	6,155	.17	.57	.83
Speech pathology	Kroll & Chase (1975)	62	1,037	.16	.44	.73
Average				.24	.64	.86

[a]The very high number of tests in some of the studies often is because they include all individual correlation coefficients in a correlation matrix.

[b]MIS is Management Information Systems.