3

PHILOSOPHICAL ORIENTATION—SIGNIFICANT SAMENESS

> Management research places much less emphasis on empirical regularities than we should expect, and that is required, for scholarship that ultimately concerns itself with the real world. (Helfat, 2007, p. 185)
>
> We must not settle for pretend knowledge. (Wells, 2001, p. 497)

3.1 Introduction

Those espousing significant sameness understand that knowledge does not emanate from the rote application of statistical rituals. Allied with this recognition, and portrayed therefore in the first part of Section 3.2, is the crucially important refutation of the myth that the *p*-value is an "objective" measure of evidence in the generation of knowledge. The second part of Section 3.2 offers a more realistic, and accordingly much messier, conception of socially produced knowledge from the significant sameness vantage. It shows that knowledge arises from the conduct of many studies (replications), by many people, over an extended period of time, which may (or may not) win the backing of the scientific community. This section also outlines the role of confidence intervals (CIs) in the acquisition of facts which, in turn, provide the impetus for the creation of theory. The balance of Section 3.2 illustrates the superiority of overlapping CIs versus reliance on *p*-values as a measure of replication success.

Following in Section 3.3 is a discussion of the model of science informing the significant sameness approach, a postpositivist theory called *critical realism*. This model emphasizes *abductive,* as opposed to hypothetico-deductive, reasoning. It is a model accurately reflecting how science progresses.

Additionally, as Section 3.4 reveals, negative results are valued in this paradigm. This is because they mark the boundary conditions of an empirical regularity's expanse. In doing so they can spur theory building by explaining why a limit to a generalization exists. In this same spirit, Section 3.5 makes the case that null results with adequate statistical power are as deserving of publication as their non-null counterparts. Comments summarizing the chapter are made in Section 3.6.

3.2 Conception of Knowledge

The significant sameness paradigm sees the development of knowledge as cumbersome because data rarely speak for themselves (Bamber, Christensen, & Gaver, 2000; Fay, 1996, p. 204). It is ingenuous to view scientific facts as being created by rejecting the null hypothesis in what Gigerenzer (2004, p. 587) calls "mindless statistics." Statistical significance testing is mostly window dressing. To see this, one would think that the ubiquity of such testing presupposes its indispensability in empirical work. Yet it is remarkable how unpersuasive the results of statistical significance tests are; in everyday practice they are not taken seriously (Guttman, 1985, p. 5). For example, Summers (1991, p. 130) challenges his readers to come up with a hypothesis in economics that has fallen into disrepute over the outcome of a statistical test. Likewise, Ziliak and McCloskey (2008, p. 120) are unaware of any advance in economics since World War II that has turned on a test of statistical significance. And Guttman (1977, p. 92) asserts that "no one has yet published a scientific law in the social sciences which was developed, sharpened, or effectively substantiated on the basis of tests of significance." Concerns about the ineffectiveness of statistical significance tests at changing the minds of scholars are found also in Keuzenkamp (2000, p. 164), Lindsay and Ehrenberg (1993, p. 218), and Spanos (1986, p. 660). But if the results of significance tests fail to convince scientists about the veracity of a finding, why use them?

Moreover, because of its revered status among social and management scientists, it is of the utmost importance to contest Fisher's (1973, p. 46) allegation that the *p*-value is an objective measure of evidence against H_0. Drawing on some of my previous work with Murray Lindsay (Hubbard & Lindsay, 2008), Section 3.2.1 shows that several arguments can be marshaled against Fisher's claim.

3.2.1 The P-Value Is Not an Objective Measure of Evidence

P-Values Exaggerate the Evidence Against the Null Hypothesis, H_0

I begin with what is a most telling indictment of the *p*-value as a plausible inferential index, namely, its exaggeration of the evidence against H_0. This, in turn, makes "statistically significant results" relatively easy to attain.

Two-Sided Null Hypotheses. *P*-values exaggerate the evidence against two-sided (point null) hypotheses (Berger & Sellke, 1987), the kind tested all the time in the management and social sciences. A point null hypothesis is expressed as follows, $H_0 : \theta = \theta_0$ versus H_A: $\theta \neq \theta_{0,}$ where θ_0 is a particular value of θ, usually zero. With this as background, using a Bayesian significance test for a normal mean, Berger and Sellke (1987, pp. 112–113) demonstrated that for *p*-values, that is, Pr(x | H_0), of .05, .01, and .001, respectively, the *posterior probabilities* of the null, that is, Pr(H_0 | x), for $n = 50$ are .52, .22, and .034. For $n = 100$ the numbers are .60, .27, and .045. It is clear that these discrepancies between *p* and Pr(H_0 | x) are marked and raise strong doubts over the reasonableness of *p*-values as measures of evidence.

Berger and Delampady (1987) found similarly discrepant results between *p*-values versus posterior probabilities in both normal and binomial situations. This led them to suggest that the use of *p*-values be abandoned when testing precise (point null) hypotheses. Given this discussion, one must agree with Berger and Berry (1988) that the validity of empirical research based on moderately small, including .05, *p*-values is open to challenge.

And besides, except for rare instances (cf. Wainer, 1999), it is impossible to defend in any epistemological sense the practice of point null—or *nil* as Jacob Cohen (1994, p. 1000) would have it—hypothesis testing. "Discovering" in the population that a difference between two means is not *precisely* zero, or that a correlation between two variables is not *precisely* zero, are trivial findings. It is hard to digest the idea that such findings are the lingua franca of empirical social and management science.

Taken literally, point null hypotheses of exactly zero differences between means or exactly zero correlations between variables do not exist in nature. In the real world point null hypotheses always are false, even if only to some small degree, such that large enough samples will lead to their rejection. Or as the celebrated statistician John Tukey (1991, p. 100) explained: "All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'Are the effects different?' is foolish." This view is retold by Lyle Jones and John Tukey (2000, p. 413). But if the point null hypothesis always is false, what's the point of testing a point null hypothesis?

Frequency Distribution of P-Values. A number of studies (e.g., Berger, 2003; Hubbard & Bayarri, 2003; and especially Sellke, Bayarri, & Berger, 2001) commenting on a simulation of the frequency distribution characteristics of *p*-values are illuminating. The simulation is available as an applet at www.stat.duke.edu/~berger.

To illustrate its use, suppose we wish to carry out some tests on the efficacy of an advertising campaign (A-C) designed to increase the awareness among voters of some political candidate. The statistical significance test would be H_0 : A-C = 0 versus H_A : A-C ≠ 0. The simulation revolves around a long series of such tests on normal data (variance known) and records how often H_0 is true for *p*-values in specified ranges, say, approximately equal to .05 or .01. Devastatingly, this simulation of the behavior of *p*-values shows that even when "statistically significant" outcomes near the .05 and .01 levels are obtained, these findings often come from true null hypotheses of no effect or association. In particular, if it is assumed that 50% of the null hypotheses in the (A-C) tests are true, Sellke et al. (2001, p. 63) cautioned:

> 1. Of the subset of [A-C] tests for which the *p*-value is close to the .05 level, *at least* 22% (and generally about 50%) arise from true nulls.
> 2. Of the subset of [A-C] tests for which the *p*-value is close to the .01 level, *at least* 7% (and generally about 15%) arise from true nulls.

So a *p*-value of .05 may provide no evidence against H_0.

P-Values and Sample Size

Small Versus Large Samples. The sample size crucially determines statistical significance levels. As an example, Royall (1986) tells of well-known statisticians whose interpretations of *p*-values in small as opposed to large sample studies are completely at odds. Some statisticians maintain that a given *p*-value in a small sample study is stronger evidence against the null than the same *p*-value in a large-scale study, and vice versa. Seen in this light, a given *p*-value does not possess a fixed, objective meaning, being connected to sample size. Of profound influence, the larger the sample size, the easier the chances of being able to record statistically significant results.

Lindley's "Paradox". Lindley (1957) demonstrated that for any level of statistical significance, *p*, and for any nonzero prior probability of the null hypothesis, $Pr(H_0)$, a sample size can be estimated so that the posterior probability of the null hypothesis, $Pr(H_0 \mid x)$, is 1-*p*. In other words, a null hypothesis that is firmly *rejected* at the conventional .05 level in a Fisherian statistical significance test can nonetheless have 95% *support* from a Bayesian perspective. These diametrically opposed inferences constitute the paradox. As Johnstone (1986, p. 494) writes, the explanation for this conundrum is that regardless of how small the *p*-value, the likelihood ratio $Pr(x \mid H_0)/Pr(x \mid H_A)$ approaches infinity as the sample size gets larger. Therefore, for large *n*, a small *p*-value provides evidence in favor of H_0 instead of against it. The issue of the objectivity and usefulness of the *p*-value as a measure of evidence is hereby dealt a crippling blow.

P-Values and Effect Sizes

The suitability of the *p*-value as a reliable measure of evidence certainly must be called into question when it has little to say about the *effect size* reported in a study (Gelman & Stern, 2006). As it stands, a small sample study with a large effect can produce the same *p*-value as a large sample study with a small effect size. This is seen in Table 3-1, which shows Peter Freeman's (1993) hypothetical data on medical trials wherein all patients receive both treatments A and B and are asked to state their preferences.

The results of trial 1, with its 75% preference rate for A over B, would be taken to reveal a possibly huge endorsement of A's supremacy. Conversely, the results of trial 4, with a 50.1% preference for A, would be considered as enormous evidence that preferences for A versus B are essentially identical. Few investigators would regard the findings of these four trials as being equivalent, yet they all yield a *p*-value of .041. It is for such reasons that Freeman (1993, p. 1443) did an about-face on his opinions of the usefulness of *p*-values:

> This paper started life as an attempt to defend *p*-values. . . . I have, however, been led inexorably to the opposite conclusion, that the current use of *p*-values as the "main means" of assessing and reporting the results of clinical trials is indefensible.

In view of the above, Gibbons's (1986, p. 367) declaration, in an account titled "*P*-values," that "an investigator who can report only a *P*-value conveys the maximum amount of information contained in the sample" simply is wrong. Indeed, Berger, Boukai, and Wang (1997) note that the interpretation of *p*-values can be expected to change drastically from problem to problem.

Table 3-1 Hypothetical Data on Treatment Preferences

Trial	*No. Preferring A*	*No. Preferring B*	*% Preferring A*
1	15	5	75.0
2	114	86	57.0
3	1,046	954	52.3
4	1,001,455	998,555	50.1

Source: Adapted from P. R. Freeman (1993, p. 1446).

P-Values and Subjectivity

Another case of the deficiency of the *p*-value as an unprejudiced measure of evidence is apparent in the decision of whether to use a one-sided or a two-sided test of statistical significance (Goodman & Royall, 1988; Royall, 1997). While two-sided tests are the norm, sometimes researchers are told that they can halve the *p*-value if they anticipate a departure from the null hypothesis in a specific direction. Or as Goodman and Royall (1988) state it, despite the fact that the data are the same, the *p*-value is modified on the basis of the researcher's subjective impressions about the expected outcome of the study. They further mention that similar alterations of *p*-values take place where multiple comparisons are concerned.

P-Values Are Logically Flawed

The logical flaw is that the *p*-value does not compute the probability of the observed data under H_0, but this *as well as the probability of more extreme data.* Because of this, statistical significance tests are influenced by how the probability distribution is spread over unobserved outcomes in the sample space. Otherwise expressed, the *p*-value embodies not only the probability of what was observed, but also the probabilities of all the more extreme events that did not occur.

Numerous statisticians (e.g., Berger & Berry, 1988; Berger & Delampady, 1987; P. R. Freeman, 1993; Goodman, 1999; Royall, 1997; Schervish, 1996) allege that a valid measure of strength of evidence cannot include the probabilities of unobserved outcomes. Jeffreys (1939, p. 316) sums up this illogic about *p*-values as follows: "*What the use of P implies . . . is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure." In this manner, McGrayne (2011, p. 56) records, Jeffreys believed that *p*-values "fundamentally distorted science."

Specification of an Alternative Hypothesis, H_A

Evidence Is Relative. In the case where an alternative hypothesis can be specified, the researcher is able to identify those findings as extreme or greater than the observed event. Therefore, Royall (1997) informs us, it is not low probability under A that makes an observation evidence against A. More properly, it is low probability under A when compared with the probability under a rival hypothesis B, so this makes it evidence against A versus B. This relativistic approach requires a weighing of the evidence between two competing hypotheses, a situation disallowed in Fisherian statistical significance tests. Fisher never saw the need for an alternative hypothesis.

In this context, consider Johnstone's (1986, p. 493) view that the law of likelihood is a better measure of evidence than *p*-values for assessing the believability of two (or more) vying hypotheses. In particular, if the likelihood ratio $\Pr(x \mid H_0)/\Pr(x \mid H_A)$ is larger than 1, then the evidence favors H_0 over H_A, and vice versa. Regrettably, Fisher's disjunction applies only to $\Pr(x \mid H_0)$; it has nothing to say about $\Pr(x \mid H_A)$. The *p*-value is a tail-area probability and not a likelihood ratio.

We're Interested in the Alternative (Research), Not the Null, Hypothesis. Making explicit an alternative hypothesis is not only a way of covering values more extreme than those observed on a null hypothesis. The alternative (research) hypothesis is the one investigators are concerned with. Berkson (1942, p. 326) saw this well ahead of others when critiquing Fisher's paradigm of null hypothesis testing:

> In the null hypothesis schema we are trying only to nullify something. . . . But ordinarily evidence does not take this form. With the corpus delicti in front of you, you do not say, "Here is evidence against the hypothesis that no one is dead." You say, "Evidently, someone has been murdered."

Statistical tests are more likely to be useful when they focus on the research hypothesis, rather than being preoccupied with rejection of the null hypothesis. Unfortunately, Fisher's statistical framework denies the existence of an alternative (research) hypothesis. In attempting to rectify this state of affairs, it is sometimes argued that Fisherian statistical significance testing has an implicit alternative (research) hypothesis that is simply the complement of the null. Yet as Hubbard and Bayarri (2003, p. 172) remind us, this argument is difficult to formalize. Questions arise concerning exactly what is the complement of an $N(0, 1)$ model. Is it the average differing from 0, the variance differing from 1, the model not being Normal? Fisher considered only the null model and wanted to see whether the data were congruent with it.

The above account demonstrates from a variety of positions that, contrary to Fisher's notification, adopted all too willingly by management and social scientists eager to establish their scholarly status, the *p*-value is anything but an objective and credible measure of evidence.[1] To repeat, objective, value-free measures of evidence are fallacies; subjective judgment *always* will be required in the statistical analysis and interpretation of data, be it classical (Fisher/Neyman–Pearson) or Bayesian in nature (see Berger & Berry, 1988; Birnbaum, 1962; Chatfield, 1985, 2002; Freedman, 1999; Johansson, 2011; Leamer, 1983; Lindsay, 1995; Perlman & Wu, 1999). Despite this, there are those in the social (e.g., Frick, 1996) and management (e.g., R. Kent, 2007, p. 38) sciences who persist in interpreting the *p*-value as an objective measure of evidence.

3.2.2 Knowledge Development—Significant Sameness

Scientific facts do not somehow arise from the outcomes of statistical significance tests. They are, rather, intellectually constructed assertions whose relation to the external world is neither immediate nor certain (Ravetz, 1971). Every research conclusion is the outcome of an imperfect process which is shaped by the formulation of the problem (following from its theoretical underpinnings); the data that are collected or made available (contexts); the method of analysis; and the skills, biases, and epistemology that influence the many decisions an investigator makes while carrying out the research and interpreting its output (Hubbard & Lindsay, 2013a, p. 1379). This is why methodological authorities continually warn that the results of a single study, no matter how well designed and statistically significant the outcomes, are virtually meaningless in the cultivation of high-level understanding and causal explanations of phenomena (Nelder, 1986, p. 112; Popper, 1959, p. 45; Yates, 1951, p. 33). Fisher (1966, p. 13), of course, was aware of this: "We thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon."

As Ravetz (1971) explains it in his rich work on the production of knowledge, scientific facts must surmount three hurdles: (1) a *social test of significance* (the scientific community views the result as deserving of additional research), (2) *empirical stability* (the result can be empirically replicated), and (3) *empirical invariance* (the result is capable of being generalized to applications other than those peculiar to its own construction). Therefore, although scientific representations are constructed socially, over time it is possible to differentiate between those representations which are or are not reasonably congruent with reality because of science's empirical nature (Giere, 1999). The significant sameness paradigm is altogether consistent with this *social* view of science. These last remarks are expanded upon below.

In contrast to the widely held view of those endorsing the significant difference paradigm, the attainment of scientific knowledge is not accomplished by solitary people (Bauer, 1994, p. 52; Eichenbaum, 1995,

p. 1620; Tukey, 1980, p. 23). Or as management theorists Pfeffer and Sutton (2006b, p. 46) declare: "Knowledge isn't generated by lone geniuses who magically produce brilliant new ideas in their gigantic brains. This is a dangerous fiction." Rather, the production of scientific knowledge in any field is a communal endeavor in search of the widest possible consensus of rational opinion (Hackett, 2005, p. 788; Ziman, 1978, p. 3). As such, knowledge creation *takes a great deal of time and the conduct of many, many studies*.

Locke's (2007) work is important here. He wishes to make the case for inductive theory building—not H-D—in social research, a case I obviously applaud.[2] In doing so Locke reinforces the view that this process takes a long period of time:

> The social sciences accepted the hypothetico-deductive model [and] . . . this meant that researchers often had to pretend that they had theories before they had a firm basis for any. This method makes for quick and often short-lived theories; in contrast, true inductive theorizing takes many years, even decades, and, I believe is far more likely to withstand the test of time. (p. 872)

As shown below, Locke offers three examples of the labyrinthine manner in which theories become inductively grounded to support his argument.

The first is Beck's (Beck, 1993; D. A. Clark & Beck, 1999) cognitive theory of depression, begun in 1956, which did not arrive full blown, but which meandered through "torturous paths" to reach its final form. The second is Bandura's (1986) social cognitive theory. Here, Locke (2007, p. 877) quotes Bandura (2005, p. 29) as saying:

> Theory building is for the long haul, not for the short winded. The formal version of the theory, that appears in print, is the distilled product of a lengthy interplay of empirically based inductive activity and conceptually based deductive activity.

The third is Locke's own research, with Latham, on goal setting theory, a theory ranked first in importance among 73 competitors by organizational behavior academicians (Miner, 2003). Locke (2007, p. 879) notes that it was only after 25 years of work, embracing some 400 studies by himself, Latham, and others, that they felt comfortable enough to develop a theory on goal setting (see Locke & Latham, 1990, 2002). Locke's experience meshes with Haig's (2013a, p. 137) estimate that it might take anywhere from 3 to 30 years for social scientists to construct explanatory theories to account for regularities/phenomena of interest.

Even in disciplines like physics, Ziman (1978, p. 40) relates, the time element in acquiring knowledge is unavoidable:

> [The researcher] learns how easy it is to persuade oneself of the validity of a model which later turns out to be false, and comes to realize that even in very strongly mathematical and well-defined scientific issues it may take a long time, much criticism and the death of many promising conjectures . . . before a reliable theory is well-based and thoroughly acceptable.

This explains why the content of undergraduate physics textbooks is approximately 90% true, while that of the primary physics journals is 90% false (Ziman, 1978, p. 40).

Assembling scientific knowledge calls for great *patience* (Gribbin, 2004, p. 462). From this aspect, theory development is better understood as a lengthy, arduous *process* (Faust, 1984, p. 131; Hull, 1988) rather than as a *product* (Kaplan, 1964, p. 409; Weick, 1995). Contemporary philosophy of science sees investigators beginning with very low-level theory, known to be defective and perhaps false, which undergoes active elaboration within a research network or program (Hubbard & Lindsay, 2013b, pp. 1395–1396; T. C. Jones & Dugdale, 2002). This is why talk of "confirmation" or "refutation" is meaningless since theories are, at best, only abstractions of reality (Suppe, 1977). It also validates Hacking's (1983, p. 15) position that accepting and rejecting theories plays a minor role in science.

In line with the above, as Kuhn (1970, p. 24) tells it, most researchers in the established sciences are not involved with inventing and testing new theories. Rather, they spend their careers engaged in "mopping up" or "normal" science activities. Which is to say they are occupied with the relatively prosaic jobs of extending the facts and predictions, along with the further articulation, of the dominant paradigm they labor within. The truth is that the reconstruction of important scientific findings seen in textbooks betrays the sense of how progress really is achieved because usually they fail to mention the many studies and false detours that led in the end to the acceptance of a theory by the scientific hierarchy (Blachowicz, 2009; Gower, 1997; T. C. Jones & Dugdale, 2002; Ladyman, 2002; Levy, 2010; Nash, 1963; Ravetz, 1971; Roberts, 1989). The untidiness inherent in scientific advance is captured perfectly by Livio (2013) in his recent book *Brilliant Blunders From Darwin to Einstein,* describing the monumental knowledge breakthroughs made by Charles Darwin, Albert Einstein, Fred Hoyle, Lord Kelvin, and Linus Pauling. Or as Rosenbaum (2002, p. 11) astutely summarizes this process:

> Scientific questions are not settled on a particular date by a single event, nor are they settled irrevocably. We speak of the weight of evidence. Eventually, this weight is such that critics can no longer lift it, or are too weary to try. Overwhelming evidence is evidence that overwhelms responsible critics.

Reiterating, only those raised in the significant difference school of thought see knowledge procurement as instantaneous, brought about by the unquestioning application of formal statistical protocols looking for $p \leq .05$ results in single-shot studies.

In practical terms, then, the question to be answered is: How, exactly, are scientific facts arrived at in the significant sameness paradigm? This time-consuming task requiring multiple studies (replications) is described below.

3.2.3 Point Estimates and Confidence Intervals (CIs)

The fixation with *p*-values deflects attention from gauging the size of the phenomenon under scrutiny, the latter being a crucial need in science (A. W. F. Edwards, 1992; Hubbard & Lindsay, 2008, 2013b; Lindsay, 1995; Tukey, 1969; Ziliak & McCloskey, 2008). How should these sizes be assessed? In the significant sameness paradigm, this is done in research programs (not by isolated authors) reporting sample statistics, effect sizes, and the confidence intervals (CIs) around them. Over time, through replication research, the CIs of additional (new) results about the phenomena in question can be compared with the increasingly robust baselines created by their myriad predecessors to see whether they are consistent with them (Hubbard & Lindsay, 2013b, p. 1394). In turn, these assembled magnitudes become the grist for the meta-analyst's mill (Eden, 2002, p. 842). The results of these more directed, fact-focused meta-analyses would permit the emergence of a rational consensus of opinion among researchers about the outcomes in any given area. This is totally at variance with the ethos of the governing significant difference paradigm, where editorial-reviewer insistence on novelty vitiates the attainment of cumulative knowledge, and hence the possibility of consensus, inviting and rewarding instead an anything-goes outlook with respect to empirical results (Andreski, 1972, p. 16; Pfeffer, 1993, pp. 612, 616; Simmons, Nelson, & Simonsohn, 2011, p. 1359). If a chemist claimed that she or he had performed a study in a "regular" environment showing that water freezes at 58°F, she or he would be laughed out of the academy. Likewise with an astronomer who writes of demonstrating that Newton's laws of motion and gravitation are false when applied to celestial bodies. But in an anything-goes world, with little in the way of credible yardsticks for judging the validity of results, all findings are equally admissible, provided that they come with the $p \leq .05$ seal of approval.

Why the use of CIs? To begin with, CIs supply all the information contained in a significance test and more (Natrella, 1960). For instance, CIs underline the desirability of estimation over testing. They indicate, also, the precision or reliability of the estimate via the width of the interval. Moreover, because they are couched in the same metric as the point estimate, CIs are easy to interpret and provide evidence on the substantive, as opposed to statistical, significance of a result. And while I do not condone this usage, a CI can be employed as a statistical significance test; a 95% CI not including the null value (usually zero) is equivalent to rejecting the hypothesis at the .05 level. In addition, the CI is a frequentist measure, that is, part of statistical orthodoxy.[3] From a pedagogical standpoint, therefore, the transition from emphasizing CIs rather than *p*-values ought to be a relatively straightforward one. Consequently, one is left wondering why CIs—a procedure that Tukey (1960) viewed as probably the most important among all types of statistical methods we know—are not routinely used, reported, and interpreted.[4]

Of singular interest from a significant sameness viewpoint, use of CIs promotes the acquisition of cumulative knowledge. It does so, in what Geoff Cumming and Sue Finch (2001), Bruce Thompson (2002), and Roger Kirk (2003) maintain is a largely unexplored but critical topic in the social sciences, by obligating the researcher to think meta-analytically about estimation, replication, and comparing intervals across studies. This is in keeping with the advocation that overlapping CIs be adopted as the criterion for a successful replication.[5] Overlapping CIs suggest credible estimates of the same population parameter(s). Fortunately, there have been some useful recent contributions in this area (see, e.g., Cumming, 2012; Cumming & Finch, 2001, 2005; Cumming & Maillardet, 2006; Fidler, Thomason, Cumming, Finch, & Leeman, 2005; Goldstein & Healy, 1995; Huberty & Lowman, 2000; Schenker & Gentleman, 2001; F. L. Schmidt, 1996; Smithson, 2003; B. Thompson, 2002; Tryon, 2001). So central, in fact, is the idea of overlapping CIs as a measure of replication success that its championing calls for greater explanation.

3.2.4 Overlapping CIs as a Definition of Replication "Success"

Significant Difference

A custom prevails in the social and management sciences of relying on the outcomes of significance tests to determine the success or failure of a replication. A replication success is defined as a result that was statistically significant ($p \leq .05$) in the initial study and continues to be so (in the same direction) in the follow-up (as criticized by, e.g., Bayarri & Mayoral, 2002; Humphreys, 1980; J. Miller, 2009; Ottenbacher, 1996; Rosenthal, 1990). This tradition is inimical to the development of empirical regularities because it ignores the pernicious influence of low statistical power, a condition endemic in the business and social sciences as referenced in Chapter 2. That is, two studies may each have similar quantitative relationships (slopes) or effect sizes, but a statistically significant coefficient is not found in the replication because statistical power is too low. Unfortunately, many researchers do not understand the link between power and the probability of obtaining a successful replication (Busche & Kennedy, 1984; Hubbard & Armstrong, 1994; Lindsay, 1993b; Ottenbacher, 1996; Tversky & Kahneman, 1971; Utts, 1991) because sampling variation and/or small *n* rarely is seen as a possible reason for disparate results (Gelman & Stern, 2006; Lindsay, 1993b; Ottenbacher, 1996).

Consider the following example from Frank Schmidt (1996) who found that a large study in the area of personnel selection using 1,428 subjects obtained a correlation of 0.22 between a single clerical test and job performance. Based on the median sample size of $n = 68$ found in the personnel psychology literature, Schmidt made 21 random draws (without replacement) from this larger study. Table 3-2 shows the correlation coefficients for the smaller studies. Only 8 of these (38%) reached statistical significance at the conventional .05 level. If observers had access to all 21 studies, they would probably say that these results are mixed, thus stimulating additional research to uncover likely explanations for the discrepancies found in the literature. But the bias against publishing negative results makes it highly unlikely that the discipline will have access to many of the other 13 studies reporting nonsignificance. As stated earlier, this leads to an inflated estimate of the population effect size. The average effect size for studies attaining statistical significance is 0.33, which is 50% larger than the real population effect size.

This example shows that the statistical significance test procedure is an unreliable criterion for certifying replication success.[6] Conflicting results are inevitable—even in situations where the only difference among studies is sampling error and/or differences in sample size. The consequences, however, are by no means trivial. The usual response is to search for additional moderator variables (interactions) to explain the "contradictory" findings. Yet these more elaborate models fare no better over a series of studies. A vicious cycle of proposing ever more complex models takes place until the research community abandons the area for lack of

Table 3-2 Random Draws (*N* = 68) Without Replacement From a Larger Study on Clerical Testing and Job Performance (*N* = 1,428) Possessing a Correlation of r = 0.22

		95% Confidence Interval	
Draw Number	*Correlation*	*Lower*	*Upper*
19	0.39*	0.19	0.59
5	0.38*	0.18	0.58
14	0.37*	0.16	0.58
8	0.36*	0.15	0.57
3	0.31*	0.09	0.53
16	0.29*	0.07	0.51
6	0.27*	0.05	0.49
17	0.26*	0.04	0.48
11	0.23	0	0.46
20	0.22	–0.01	0.45
21	0.21	–0.02	0.44
13	0.21	–0.02	0.44
9	0.20	–0.03	0.43
18	0.17	–0.06	0.40
7	0.15	–0.08	0.38
2	0.14	–0.09	0.37
15	0.14	–0.09	0.37
4	0.12	–0.12	0.36
12	0.11	–0.13	0.35
1	0.04	–0.20	0.28
10	0.02	–0.22	0.26

Source: Adapted from Schmidt, Frank L. (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers," *Psychological Methods,* 1 (1), 121. Copyright © 1996 by the American Psychological Association. Adapted with permission.

**p* < .05 (two-tailed).

progress. This is why Schmidt (1996, p. 120) correctly emphasizes that "significance testing in psychology and other social sciences has led to frequent serious errors in interpreting the meaning of data . . . that have systematically retarded the growth of cumulative knowledge."

A second way in which null hypothesis significance testing impedes the establishment of empirical generalizations and theory building is now addressed. Two studies, with both reaching statistical significance for the coefficient of interest (with similar signs), nevertheless may have quantitative relationships that are clearly at odds and/or display widely varying effect sizes. Such results would indicate that the relationship is different and/or that some other variable in need of identification is operating. Yet the focus on *p*-values will not necessarily provide this information, intimating once more that it is a poor criterion of replication success.

Gendall, Hoek, and Brennan's (1998) article is enlightening in this respect. They wanted to see if a one-dollar cash incentive would yield a higher response rate to a mail questionnaire compared with a no-incentive control group. Table 3-3 shows that after two reminders there was a 68.2% (386/566) response for the control group and a 76.6% (431/563) cooperation rate for the one-dollar group, the difference being statistically significant ($z = 3.11$; $p < .002$). Now review the hypothetical attempt to confirm the original finding. Table 3-3 conveys that in the replication the responses of the control and one-dollar groups were 32.7% (185/566) and 54.4% (306/563), respectively. Again, this difference is statistically significant ($z = 6.78$; $p < .0001$). Normally, a result like this is taken to be evidence of a successful replication. From a significant sameness perspective, however, the discussion in the next section will show that a profitable opportunity to expand our learning will be missed if one stops here.

Significant Sameness

A significant sameness outlook utilizing CIs produces very different conclusions for the two studies above. With respect to the Schmidt (1996) data, Table 3-2 shows that the 95% CI for each correlation coefficient overlaps with the other studies—even for the largest and smallest correlations. Overlapping CIs suggest that the studies are in agreement with one another, contrary to the false impression left by the traditional "nose counting" approach that only 8 studies support the hypothesis while 13 do not. In this manner, use of CIs offers the prospect of unifying an otherwise seemingly fragmented literature caused by adopting the *p*-value criterion of replication success. Significant sameness addresses

Table 3-3 Hypothetical Replication of Gendall et al.'s (1998) Study on Incentives and Mail Survey Response Rates

	Gendall et al. Study			*Replication*		
	Sample	*Responses*	*Proportion*	*Sample*	*Responses*	*Proportion*
No incentive	566	386	.682	566	185	.327
$1 incentive	563	431	.766	563	306	.544
Differences in proportions			.084			.217
$z = 3.11$; $p < .002$				$z = 6.78$; $p < .0001$		
Confidence intervals around differences in the no-incentive and $1 incentive groups						
Lower 95% CI:			.033			.162
Upper 95% CI:			.135			.272

Source: Adapted from Hubbard, Raymond and R. Murray Lindsay (2013a), "From Significant Difference to Significant Sameness: Proposing a Paradigm Shift in Business Research," *Journal of Business Research*, 66 (September), p.1383 with permission from Elsevier.

commonalities in data sets, the road to generalization. Seen in this light, it would be fascinating to examine how differently published articles would read had the attention been on CIs and not *p*-values (cf. C. Poole, 2001).

CIs, of course, cannot specify the value of the population effect size. This necessitates conducting a meta-analysis. Such an analysis of the data in Table 3-2 demonstrates that there is only one population value—0.22—and that the differences in sample correlations are due only to sampling error.

I return now to the Gendall et al. (1998) mail survey incentive study. While it is apparent that the use of a one-dollar incentive boosted the frequency of responses over the control group (as indicated by the significance test), the significant sameness approach reveals a *failure* to replicate. This happens in two ways. First, the proportion of responses is different in the control and one-dollar groups for the two investigations. This is observed in Table 3-3 by the non-overlapping 95% CIs for the Gendall et al. study (3.3%–13.5%) and the replication (16.2%–27.2%). Second, the percentage increase in cooperation rates for the control and one-dollar groups is markedly greater in the replication. In Gendall et al.'s study there was a 12% increase in replies (.084/.682), while this figure was 66% for the follow-up study. This much better relative performance for the latter occurred even though, in absolute terms, responses in the original study were far higher. These points signify that there are important distinctions between the two studies that need to be explored.

For instance, the impressively high average survey return rate of 72% found in Gendall et al. (1998) was for a New Zealand sample, while the corresponding figure for the American sample in the replication was only 44%. Perhaps this evidently greater willingness by New Zealanders to answer questionnaires is because they are less inundated by them than are Americans. Or possibly the respondent characteristics (beyond those of nationality) and/or content of the survey in the American replication were different from those in the original research. The crux of the matter is that the New Zealand and American investigations vary in ways that call for explicit attention. But this call will likely go unheeded because a "successful" replication, as defined by *p*-values, ignores these differences by erroneously declaring the two works to be equivalent.

In sum, the *p*-value is an unreliable criterion for deciding replication success and encouraging true learning. Significant sameness—does the same relationship (i.e., effect size, quantitative model) hold across many sets of data—must become the new yardstick of replication success. And the use of overlapping CIs around sample statistics and effect sizes is the vehicle for deciding this, a strategy not without its detractors.

Criticism of the Overlapping CI Criterion—and Rejoinder

Still another reason for promoting the use of overlapping CIs across different studies is to spotlight and/or sidestep the baneful effects of low statistical power common in traditional significance testing. Ball and Sawyer (2013, pp. 1389–1390), however, are critical of the use of overlapping CIs as the means for establishing significant sameness. Yet as the following discussion explains, in well-designed studies the proper interpretation of CIs does not result in misleading inferences (Hubbard & Lindsay, 2013b, pp. 1393–1394).

Observe, for instance, quadrant 1 in Figure 3-1 depicting overlapping CIs in a well-designed, high-powered test. Here, the results are informative: The parameters are from the same population. The situation in quadrant 2, showing overlapping CIs in a low-powered test, affords a less tidy interpretation. Under these conditions the investigator cannot infer that the results of the two studies are the same; all that can be concluded is that there is no evidence to show that they differ. But a low-powered study is *not* a well-designed one. This low power revealed in the (pronounced) width of the intervals signals the need for collecting additional data to supply more precise estimates of the parameter(s) of concern. Beyond this, a case can be made that low-powered research should not be published. In contradistinction, non-overlapping CIs in a low-powered test strongly suggests the existence of a real difference between the parameters of the two studies, despite neither being estimated reliably (quadrant 4).

Ball and Sawyer (2013, p. 1390) note further that when exceptionally high-powered tests are involved, usually via the use of enormous sample sizes (think data mining), the CIs of two investigations may not overlap even though their population parameters appear for all intents and purposes to be alike (quadrant 3). The validity of their criticism is acknowledged; when sample sizes become huge, CIs reduce to points.

Given the above, it must be said that in applying the concept of overlapping CIs as the criterion of a replication success, we are at pains to underline that this should not be done reflexively (Hubbard & Lindsay, 2013b, p. 1394). Although focusing on the precision of magnitudes' CIs is infinitely better than using *p*-values when it comes to weighing the plausibility of knowledge claims, no thoughtless application of a rule of thumb should ever substitute for the exercise of subjective judgment when making inferences from data. These inferences must be tempered by concerns such as whether a result looks "reasonable," that is, whether it is aligned with previous results and background knowledge (unless a

Figure 3-1 Interpreting Replication Outcomes: Confidence Intervals (CIs) and Statistical Power

	Statistical Power	
	High	**Low**
Overlapping CIs	[1] Informative: similar results across studies. Parameters are from the same population.	[2] Not informative: need to collect additional data.
Non-overlapping CIs	[3] Mixed: parameters may or may not be from different populations. Here, as always in science, background knowledge and subjective judgment are required to assess whether practical or theoretical differences exist.	[4] Informative: parameters are from different populations.

Source: Adapted from Hubbard, Raymond and R. Murray Lindsay (2013b), "*The Significant Difference Paradigm Promotes Bad Science,*" *Journal of Business Research*, *66* (September), p. 1394 with permission from Elsevier.

boundary condition has been reached), seems to have been arrived at competently, and so on. Science resists formalism, although the value of formal methods is recognized (Hubbard & Lindsay, 2013b, p. 1394).

3.3 Model of Science—Critical Realism

Whereas the significant difference model revolves around theory *testing,* the significant sameness paradigm concentrates on theory *development.* Of note, the idea of significant sameness is totally in accord with an increasingly influential postpositivist metatheory, attributed to Bhaskar (1978, 1979), called *critical realism.* This philosophy unfolded in part through the study of researcher practices and so is based on what scientists actually do (Haig, 2013b, p. 7; M. L. Smith, 2006). It is beginning to attract the attention of scholars in economics (Lawson, 1997, 2003), education (S. Clegg, 2005), geography (Yeung, 1997), management (Rousseau, Manning, & Denyer, 2008; Tsang & Kwan, 1999; Van de Ven, 2007), marketing (Easton, 2002; Hubbard & Lindsay, 2013a; S. D. Hunt, 2003), and sociology (Danermark, Ekström, Jakobsen, & Karlsson, 2002), among others.

While including aspects of both, critical realism offers an alternative philosophy to those found wanting—positivism/empiricism on the one hand and relativism/interpretivism on the other (Sayer, 2000, p. 2). That is, it provides a bridge across the philosophical divide between the quantitative and qualitative research camps (Lindsay & Hubbard, 2011). This is done by pointing to the importance of context while at the same time eschewing "naïve" realism. As such, the critical realist approach capitalizes on the virtues of both quantitative and qualitative research capabilities, seeing no dichotomy between them.

Briefly, this philosophy asserts that, first, the world exists independently of our knowledge of it. Second, science aims at developing genuine, but always imperfect and subject to change, knowledge about the world. That knowledge is produced socially, however, and as a consequence is theory laden, does not make it theory determined (Sayer, 2000, p. 47). Third, all theories concerning knowledge claims must be critically evaluated; knowledge is not immune to empirical check (Sayer, 1992, p. 5). It is via an exacting appraisal of competing theories that the scientific college, over time, is able to retain and improve on those which do a better job of approximating reality, while discarding those which do not. In short, not all knowledge is equally fallible (M. L. Smith, 2006).

In common with critical realism, the significant sameness model is philosophically grounded in abductive (explanatory) reasoning. Abduction—subsequently named inference to the best explanation (Harman, 1965) or retroduction (Lawson, 2003, pp. 145–146)—is a vital concept introduced by the American pragmatist philosopher Charles S. Peirce.[7] As pointed out in Charles Hartshorne and Paul Weiss's (1934) anthology of Peirce's work, "Abduction consists of studying facts and devising a theory to explain them" (1903, Vol. 5, p. 90). Like detective work, abduction is a method of inference for generating plausible (best) explanations for the facts we possess. It is to be understood that the facts referred to here are of the stubborn variety.[8]

Empirical regularities or facts can be predicted by theory (T→E) or can precede theory (E→T). The former avenue glorifies hypothetico-deductivism and the significant difference philosophy. The significant sameness model, however, emphasizes that by means of inductive enumeration the empirical regularities must come *first*. In this bottom-up (E→T) interpretation of research, the discovery of empirical generalizations *fuels* (high-level) theory development rather than vice versa (Ehrenberg, 1993b; Haig, 2013a, 2013b, 2014; Hubbard & Lindsay, 2013a, 2013b; Lynch, Alba, Krishna, Morwitz, & Gürhan-Canli, 2012). The rationale behind this is that the purpose of theory is to explain and systematize lower-level findings and generalizations (Rosenberg, 1986); consequently, before any theory explaining a process can be developed or tested, that process needs to be understood through the identification of repeatable facts, phenomena, or regularities. Fiske (1986, p. 75) considers this to be the first step to achieving progress (see also Ehrenberg & Bound, 1993; Hacking, 1983, ch. 9; Haig, 2014, ch. 2; Ladyman, 2002, p. 48). This is because, the rank and file of empirical social and management research notwithstanding, we are unlikely to be successful in trying to explain particular data sets, or other isolated events in individual studies, since they tend to be ephemeral and affected by idiosyncratic boundary conditions that are extremely difficult to ascertain. Put another way, much theory testing in these areas concerns entities and their interrelationships whose viability is far from established, a precarious exercise to say the least. This also accounts for the claim made repeatedly here that data seldom speak for themselves, as well as the attendant proviso that the results of single studies must be treated with circumspection.

In light of these handicaps it makes sense to uphold the overlooked view that fruitful theoretical interpretation typically occurs *after* a pattern (fact) has been empirically determined from the parsing of many data sets (studies). Thus, a preferable strategy is to anchor theory development around the discovery of repeatable facts or regularities in the behavior of phenomena; their relative stubbornness *demands* an explanation. Keuzenkamp (2000, p. 221) seconds this rationale, noting that facts can inspire economic theory. As does Ormerod (1997, p. 210), who recommends that theory building by economists should originate from the facts, and not from abstractions about how a rational world ought to operate. So, too, does Summers (1991, p. 140) when advising that economic theories should seek to explain regularities. Critical realists share this position. Lawson (2003, p. 146), for instance, acknowledges that finding causal mechanisms "presupposes" the existence of partial or demi-regularities. Danermark et al. (2002, p. 116) concur: "We claim that social scientific research is about identifying demi-regularities and from them trying to find explanations." Management (Rousseau et al., 2008, pp. 481, 487), marketing (Sharp & Wind, 2009, p. 122), and psychology (Haig, 2013a, p. 136, 2013b, p. 9) theorists harbor analogous thoughts. Finally, no less an authority than Sherlock Holmes reckoned, "It is a capital mistake to theorize in advance of the facts" (cited in N. L. Kerr, 1998, p. 201).

To bring to the fore this little-appreciated point, more Nobel prizes have been awarded to those discovering stubborn facts or empirical generalizations than have been granted for the construction of theories (Haig, 2005, p. 384). It also underlines Blaug's (1992, p. 134) conviction that scientific progress comes only from the maximization of facts, Hunter's (2001, p. 157) insistence that in the advancement of science, facts are at least as important as ideas, Ziman's (1978, p. 6) and Imrey's (1994, p. 65) desire that scientific knowledge should consist of firmly established facts and principles, Hambrick's (2007, p. 1349) proposal that we should be willing to start with the compilation of facts, Keuzenkamp's (2000, p. 22) urging that the goal of econometric inference should be discovering regularities that are simple and descriptively accurate, and Alba's (1999, p. 2) and Lehmann, McAlister, and Staelin's (2011, p. 157) calls to place greater stress on the obtainment of facts.

Following Ehrenberg (1993b), accounting for stubborn facts necessitates a sequential process of data-theory (or concrete-abstract in critical realist language) interactions over an expanding range of different conditions. At this stage much will be understood, for example, that some factors do not matter, whereas others do, providing for considerable familiarity with the phenomenon in question and for the possibility of linking the result with findings and/or theories in other areas. This process encourages further conjecture (T), requiring the testing of new implications (E). The cycle keeps repeating itself (i.e., E→T→E→T . . .), thereby leading to an increase in our depth of understanding of phenomena (cp. Eisenhardt, 1989; Eisenhardt & Graebner, 2007). As noted, it is the latter, rather than some one-to-one correspondence between theory and observations, that convinces researchers about the feasibility of theoretical explanations.

Furthermore, the detection of empirical regularities as a paramount driving force behind theory building not only applies to the management and social sciences, but has been of major importance in the physical sciences as well (R. Brown, 1963, p. 136). For example, Bernard Cohen (1985, p. 125) acknowledges that the detailed recordings of the Danish astronomer Tycho Brahe greatly influenced Newton's thinking and led eventually to the Newtonian revolution. Indeed, Wigner (1964, p. 995) maintained in his Nobel laureate address that recognizing the need to specify regularities within a certain range of conditions may be the greatest discovery in physics:

> It is often said that the objective of physics is the explanation of nature, or at least of inanimate nature. . . . It is clear that physics does not endeavor to explain nature. In fact, the greatest success of physics is due to a restriction of its objectives: *it only endeavors to explain the regularities in the behavior*

> *of objects*. This renunciation of the broader aim, and the *specification of the domain* for which an explanation can be sought, now appears to us as an obvious necessity. In fact, the specification of the explainable may have been the greatest discovery of physics so far. (emphasis added)

Given that the procurement of empirical regularities and their later generalization is a major influence on theory development, the widespread preoccupation in the business and behavioral literatures on theory testing, following the H–D model of science, makes little sense (Barwise, 1995, p. G32; Cook & Campbell, 1979, p. 24; Hubbard & Lindsay, 2002, p. 393; Toulmin, 2001, p. 213). Much more attention needs to be devoted to identifying empirical regularities, thus making the case that external validity considerations are as important as their internal cousins. Shadish, Cook, and Campbell (2002) share this position, writing that earlier statements by Donald Campbell and Julian Stanley (1966) that internal validity is the sine qua non of experimentation have been misinterpreted by readers for many years. In an effort to set the record straight, Shadish et al. announce, "Let us be clear: *Internal validity is not the sine qua non of all research*" (p. 98), and go on to say that "in this book, methods for studying external validity now receive the extensive attention that our past work gave to internal validity" (p. xvii).

Finally, while the importance of acquiring empirical generalizations has been emphasized to counteract its chronic neglect to date, this is not considered to be the foremost goal of science. Rather, the development of empirically grounded theory (explanation) and causal understanding of phenomena, along with practical applications of theory, are what is desired. However, the attainment of predictable regularities is seen as a critical means of achieving these goals within a framework of abductive inference that many (e.g., Haig, 2005, 2014; Hubbard & Lindsay, 2013a, 2013b; Ketokivi & Mantere, 2010; Lipton, 2004; Psillos, 1999; Rozeboom, 1997) testify is, in fact, an accurate description of how researchers actually behave in progressive science.[9] And it is why, in the final telling, for those researchers favoring a critical realist position, replications are essential to disciplinary advance (Koole & Lakens, 2012, p. 609).

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

In stark contrast to the publication bias against negative outcomes emblematic of the significant difference paradigm, such results are *welcomed*, provided they can be reproduced (Ehrenberg, 1975), when the emphasis is placed on significant sameness. This is because contradictory results can play an important role in identifying the *boundary conditions* of an empirical generalization's applicability (C. M. Christensen & Carlile, 2009; Cortina & Folger, 1998; Hubbard & Vetter, 1996; Lindsay & Ehrenberg, 1993).

Going against the grain of many researcher beliefs, theories do not produce universal covering laws of the kind suggested by Hempel (1965). In Hempel's world an event is explained by showing how it can be deduced from some universal or general law(s) conjoined with a set of initial conditions. But, as Ladyman (2002, p. 202) spells out, a critical requirement for the deduction to be valid is the empirical truth of the universal law(s). However, owing to Hume's "problem of induction"—just because all *observed* A have the property B does not entail that *all* A have the property B—we cannot guarantee the universality of laws.[10] As discussed earlier, there are no immutable laws in the sense that they exhibit absolute invariance. The positivists' goal of generating infallible knowledge is impossible.

For universal laws to arise would require a totally closed system in which an invariable sequence of "if A, then B" empirical events could be observed readily (Lawson, 2003, p. 5; Sayer, 1992, pp. 122–124). Because closure of this kind cannot be met in natural systems, other than in simplified experimental situations, even physical laws must be conditional. The idea that the natural sciences, physics included, deal with phenomena that are not context dependent is a myth (Holtzman, 1986, p. 348; Shadish et al., 2002, p. xv). Instead, "laws" are best viewed as *restricted* generalizations possessing extensive empirical support (Chalmers, 1999, p. 216; Giere, 1999, p. 93; S. Gordon, 1991, p. 34; Toulmin, 2001, p. 111; Turner, 1967, pp. 251–252).

When comparisons are made between the physical and social sciences, what frequently goes unsaid is that researchers in the former are able to apply theory in the real world because they work very hard at specifying the relevant conditions under which the effect will and will not come about. This is a laborious task that demands numerous carefully conducted studies to delineate the existence of boundary conditions affecting a result. Boyle's Law, for example, says that at a given temperature the volume and absolute pressure of a gas vary inversely. This law does not hold, however, when the density of a gas increases or when the temperature changes. In a similar vein, Snell's Law of Refraction does not hold for certain temperature ranges, light rays, or for transparent crystals (R. Brown, 1963). The point is, all major laws in the physical sciences—including those of Ampère, Charles, Coulomb, Faraday, Kepler, Newton, and Ohm—are similarly constrained (Giere, 1999, p. 90; Losee, 2001, pp. 191–192).[11]

While these empirical generalizations may be of a restricted nature, they have nevertheless been found to hold in sufficient instances as to render them of great practical value. Scott Gordon (1991, p. 603) elucidates on this:

> When necessary, a scientist will, without a qualm, use "Avogadro's number," which, though it has been computed from a limited set of specific cases, asserts that *all* gases, at equal temperature and pressure, contain 6.023 x 10^{23} molecules per gram molecular weight. In the *Handbook of Chemistry and Physics* there are literally hundreds of thousands of such universal numerical statements for particular elements and compounds: boiling points, melting points, solubilities, densities, X-ray diffraction angles, etc., most of which are not even given with ± qualifiers.

And should an anomalous result occur when using these generalizations, a clear message is sent that some boundary has been reached concerning its relevance.

The notion of limited empirical regularities heeds, and is explainable by, the tenets of critical realism, where Bhaskar (1978, p. 50) advises that the tangible impact of causal laws must be viewed as "tendencies" rather than inevitabilities. The reason for this is that the portrayal of laws as universal empirical regularities is both ontologically and epistemologically naïve. Social (and physical) reality is not transparent and cannot be drawn down to the realm of observable phenomena in closed systems. It is, rather, an ontologically "deep" stratified and differentiated open system composed of three separate domains: real, actual, and empirical (Bhaskar, 1978, p. 56; Sayer, 2000, pp. 11–12).

The *real* domain is the deepest, and here reside objects of inquiry (e.g., consumers) and structures of internally related objects (e.g., buyers and sellers) with causal powers or mechanisms capable of producing events in the world (e.g., exchange relationships). To invest causal powers to objects and structures is to comment on what they will or can do under appropriate circumstances because of their intrinsic natures or "ways of acting" (Bhaskar, 1978, p. 14). It is these causal mechanisms, working alone or in concert with others, if and when triggered, that generate events in the *actual* domain. And some of these events may be observed in the *empirical* domain. The fundamental aim of science is to develop causal explanations by finding or imagining necessary generative mechanisms to account for the nonrandom patterns of events that are observed empirically (Danermark et al., 2002; M. L. Smith, 2006; Sobh & Perry, 2006). Put succinctly by Little (1993, p. 185), "the central explanatory task for social scientists is to uncover causal mechanisms."

But since the causal mechanisms which objects and structures possess are only tendencies which may be reinforced, modified, or inhibited in complex interactions with other objects' and structures' mechanisms in an open social world, they may or may not reveal themselves empirically (Danermark et al., 2002, p. 163). Therefore, whatever empirical materializations happen will surely not be in the form of invariable configurations. The social world is not some highly aggregated deterministic model, but rather one in which causal relationships are contingent. This, parenthetically, forms the basis for Lawson's (2009, p. 765) criticism that insistence on mathematical deductivist modeling "is why modern economics has continually failed on its own terms."

Yet to deny the existence of universal empirical generalizations in no way prevents the occurrence of restricted ones. It is true, as the above examples show, that the chances of detecting regularities are higher in the more closed systems characteristic of the natural sciences (Sayer, 1992, pp. 121–123, 2000, pp. 14–15) than in the open social world. However, it is possible for social systems (e.g., social and work organizations, families, buyer-seller relationships, health care systems) to be quasi-closed, thereby permitting the discovery of approximate (Sayer, 1992, p. 124), demi- (Lawson, 1997, pp. 204–221, 2003, pp. 105–107), phenomenal (Little, 1993, p. 187), sufficient (Ormerod, 1997, p. 210), or stylized (Helfat, 2007, pp. 186–189; Keuzenkamp, 2000, pp. 220–221) regularities. Moreover, some of these are "often very stable" (Danermark et al., 2002, p. 165), "repeatable events" (Haig, 2013a, p. 143).

Further, the identification of boundary conditions provides a powerful heuristic leading to higher-order theoretical synthesis (Brinberg & McGrath, 1985; Lynch, 1999; Magnani, 2001; McGuire, 1983; Poincaré, 1908/2004; M. S. Poole & Van de Ven, 1989; Van de Ven, 2007). The history of science is replete with examples of how the recognition of anomalies has motivated scholars to refine theories or to develop new and better ones in the course of understanding why the boundaries apply (Ehrenberg, 1975; N. L. Kerr, 1998, p. 210; Trusted, 1979). For this reason, Clayton Christensen and Michael Raynor (2003b) state that negative results represent a triumph, not a failure.[12]

Historically, management and social scientists have been unconcerned with this ponderous task of ascertaining the scope and limits of a finding (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986; Hunter, 2001; Tsang & Kwan, 1999; Wells, 2001)—the "hard slog" as Ehrenberg (2004, p. 41) put it. This is unfortunate because some (e.g., R. Brown, 1963, pp. 149, 154; Ehrenberg, 1993a, p. 385; Ehrenberg & Bound, 1993, p. 191; Gage, 1996, p. 14; Harvey, 1969, p. 111; Kincaid, 1996,

p. 3; H. A. Simon, 1990, p. 2) see no reason, in essence at least, why empirical generalizations of the same limited sort found in the physical sciences cannot be uncovered also in the social sciences.[13]

Of direct relevance on this matter is Hedges's (1987) intriguing analysis of the empirical cumulativeness—the degree of agreement among replicated studies—of research published in the physical and social sciences. Using the same statistical methods (weighted least squares, Birge ratios) yielding numerical indexes for purposes of comparison, Hedges was able to determine the consonance of experimental results in physics and psychology. The physics data consisted of 13 quantitative reviews of the mass and lifetime of stable (against strong decay) particles from the Particle Data group. Hedges's corresponding data for psychology came from 13 meta-analyses encompassing six sub-areas of the discipline: sex differences in spatial ability; sex differences in verbal ability and field articulation; the effects of open education on attitude toward school, mathematics achievement, reading achievement, and self-concept; the effects of desegregation on educational achievement; the validity of student ratings of college faculty; and the effects of teacher expectancy on IQ. No doubt surprising many people, Hedges announced that "the evidence presented here suggests that social science research may not be overwhelmingly less cumulative than research in the physical sciences" (p. 453). His conclusion is hugely encouraging.

In fact, there is additional vindication for the hopes aired by those believing in the possibility of discovering empirical regularities in the social and managerial sciences. As noted above, and discussed further below, regularities of varying strengths have been found in these fields.

As a case in point, the psychophysics area of psychology has Weber–Fechner's Law (Gescheider, 1976; Teigen, 2002; Turner, 1967), which states that $\Delta I/I = k$. Here, I is the intensity or size of an initial stimulus, ΔI is the amount of change in a stimulus required to produce a *just noticeable difference,* and k is an empirical constant. An example might be increasing the volume of music until it is perceived to be noticeably louder (ΔI) than the original volume (I) and recording k. Just like laws in the physical sciences, this one is of restricted applicability; it works quite well when applied to human sense modalities over the intermediate range of intensities, but breaks down at the extremes. The Flynn effect is another good example of an empirical generalization in psychology.[14] As Haig (2013a, p. 138) and Wikipedia (2015) describe, this effect reveals the striking regularity that IQ scores have increased by some three points each decade from about 1930 to the present in 20 countries, from areas as diverse as North America, Europe, Asia, and Australasia. Geography boasts the rank-size rule for the population size of cities in various countries. It is formulated as $p_i = p_1/i$, where p_i is the size of the population of the ith city when all cities are ordered 1, 2, 3, 4, . . . n in terms of descending population sizes, and p_1 is the population of the largest city. The rank-size rule follows a negatively sloping straight line when plotted on a log-log scale and holds decently for the United States (Uncles & Wright, 2004, p. 6). Political science offers the "iron law of oligarchies," positing that social organizations tend to form hierarchical orders (S. Gordon, 1991, pp. 35–36). Or see Berelson and Steiner's (1964), albeit lax, inventory of generalizations about human behavior found in the social sciences as a whole.

When it comes to the management disciplines, we have the Boston Consulting Group's Law of Experience, which predicts that as cumulative output doubles, unit costs are reduced by a fixed percentage. For example, if a company has, say, an 85% experience curve—where experience is defined as the combined effects of learning, volume, investment, and specialization—this means that each time cumulative production doubles, per unit costs drop to 85% of their preceding level. Grant (1998) observes that this "law" holds in a number of studies from the manufacture of bottle caps and refrigerators to insurance policies and long-distance phone calls. There is, too, the predictable manner in which innovations occur within companies (Christensen & Raynor, 2003a).

In the marketing area, regularity is seen in the profiles of innovators. Robertson (1971), for example, provides information from 21 independent studies spanning a wide range of product categories and populations describing these profiles. Innovators usually have higher income and education levels, possess a more venturesome personality, and are much more likely to be opinion leaders than non-innovators. Other candidates for empirical regularities in marketing include the retail gravitation effect, the market share–ROI (return on investment) relationship, and the 80–20 rule describing how the majority of sales in some industries (e.g., airlines, beer, bank revenues, college donations) come from a minority of the market (Hubbard & Lindsay, 2002; Kerin & Sethuraman, 1999; Sheth & Sisodia, 1999).

Andrew Ehrenberg's work is exemplary when it comes to establishing empirical generalizations within a marketing context, or elsewhere in the management and social sciences for that matter. For instance, Ehrenberg (1988) demonstrated the progress achieved in modeling brand purchase frequencies and repeat buying patterns using the negative binomial distribution up through the increasingly more general findings on individual purchase incidences and brand choice made possible by the comprehensive Dirichlet model. The appeal of this model is its ability to account, parsimoniously (only the market share of each brand is required as input), for many empirical patterns of buyer behavior, including those described by the duplication of purchase law and the double jeopardy (DJ) phenomenon (Hubbard & Vetter, 1996, p. 154). The DJ effect, which says that brands with smaller market shares not only are bought by fewer customers in a given time period (penetration level) but also are bought less often (average frequency of purchase), has undergone extensive replication confirming its applicability. For example, Ehrenberg and Bound (1993) note that DJ holds for over 50 different products (including convenience and shopping goods, differentiated and undifferentiated items, as well as tangible products and services). DJ also applies to different distribution channels, different countries, different time periods, and so on. They further tell of a few exceptions (boundary conditions) where DJ doesn't hold, or holds only partially. Dan Vetter and I submit that Ehrenberg's (e.g., Ehrenberg & Bound, 1993; Ehrenberg, Goodhardt, & Barwise, 1990) work "provides perhaps the quintessential example of the value of systematic replication and extension research in producing robust and generalizable results" (Hubbard & Vetter, 1996, p. 155).

Furthermore, economics has the Law of Demand, the inverse relationship between the price of a product and the quantity consumed at that price, which Blaug (1992, p. 139) hails as "one of the best corroborated statistical 'laws' of economics." Then there is Pareto's Law of Income Distribution stating that the latter was essentially the same in all countries (S. Gordon, 1991, p. 35; Strathern, 2001, pp. 212–213).

Still another example of the resilience of empirical relationships in economics is found in Thomas Piketty's (2014) improbable best-seller *Capital in the Twenty-First Century.* In this book Piketty analyzes over time the capital/income ratios (βs) exhibited by various countries. Written as $\beta = s/g$, this index is the ratio of a country's savings to growth rates. For instance, if s = 10% and g = 2%, then β = 5. What this means is that if a country saves 10% of its national income each year and the annual growth rate of this same income is 2%, then in the long run the country will have amassed capital worth 5 years of national income. Clearly, as β gets larger we see a growing inequality in incomes favoring the "haves" over the "have-nots" and the potential havoc that could ensue as a result. Piketty (p. 26) shows that for Europe, aggregate private wealth was worth 6–7 years of national income from 1870 to 1910, fell to 2–3 years by 1950 (because of two world wars and the Great Depression), and rose uniformly to some 4–6 years in 2010. The pattern for the United States was about 3 years in 1770, 5 years in 1910, less than 4 years in 1950, and 4.5 years in

2010. As Piketty's gathering of the facts shows beyond doubt, the ratio of private wealth to income has increased steadily since 1950 and now rivals 19th century values. It may well eclipse those historical benchmarks.

Interestingly, the above discussion puts the lie to the objections of those of a more humanistic or qualitative persuasion that empirical generalizations cannot be detected in the social sciences. The real problem is that the dominant paradigm researchers operate within places no weight on their establishment. In addition, it would be remiss to leave the impression that qualitative researchers somehow are absolved of the responsibility for seeking valid generalizations. They are, in fact, just as culpable in this regard as their quantitative colleagues (Freese, 2007; Golden, 1995; Hubbard & Lindsay, 2002; J. A. Maxwell, 1992; Wells, 2001; Wilk, 2001). In any case it must be reminded that critical realists see no absolute split between quantitative and qualitative research methods. They accept fully the need for involving *both* perspectives in the discovery and understanding of regularities (Rousseau et al., 2008, pp. 486–487; Sobh & Perry, 2006, p. 1195; Van de Ven, 2007, p. 70).

Once empirical generalizations are identified, the goal of science is devising explanations for them. And this is where abductive reasoning comes into prominence because causal explanations, as opposed to causal descriptions, must be *invented* (Bunge, 1997). Most causal mechanisms, be they physical (e.g., gravity) or social (e.g., the market's "invisible hand") are hidden, and therefore must be conjectured or visualized. This is why Bunge (1997, p. 423) says: "Imagine what would have happened if Newton had abstained from positing unobservables, such as mass and gravitation, and from postulating laws, focusing instead on observable properties and events and their statistical correlations." The apparent discovery of the Higgs boson, or so-called God particle, is a very recent example of support for the existence of a heretofore hidden causal mechanism for understanding why matter has mass (Heilprin, 2012, p. 5A).

Moreover, causal explanations for some observed regularities exist in the business and social sciences. These explanations range from the embryonic to the fully fledged. An example of the former is the DJ phenomenon. To date, the causal mechanism(s) underlying this widespread occurrence remains little more than a simple description put forward originally by the sociologist McPhee (1963) that DJ will arise whenever competitive items (e.g., market shares, movies, politicians), through varying degrees of exposure to an audience, differ in their popularity (see Ehrenberg et al., 1990, p. 85, for a fuller telling; Habel & Lockshin, 2013, for recent developments). The Flynn effect portraying an approximately linear increase in IQs over a number of decades has several, some mutually supportive, explanations. Among potential candidates receiving ongoing attention are the impacts of improved nutrition, greater environmental complexity, better education, smaller family sizes, heterosis (the birth of genetically superior children from mixing the genes of parents), and increases in test-specific skills (Wikipedia, 2015). Via his fundamental inequality $r > g$ (where r denotes the average annual rate of return on capital, and g is the growth rate of the economy), Piketty (2014, pp. 25–26) supplies the rationale underlying the divergence in capital/income ratios mentioned earlier. As is traditionally the case, when $r > g$ then inherited wealth must necessarily outstrip national income, possibly to toxic levels. As a final example, also in economics, sophisticated (mathematical) competing mechanisms to account for downward-sloping demand curves have been available for some time. These include the income and substitution effects (Slutzky equation), the law of diminishing marginal utility and utility maximization, and revealed preference theory (see, e.g., Green, 1976, pp. 62–69, 81–83, 121–127; Phlips, 1974, pp. 16–26, 40–45).

It has been shown that negative results can play a valuable role in the significant sameness paradigm. This is especially so when they are backed up with adequate statistical power. Section 3.5 speaks further to this issue.

3.5 The Statistical Power of "Negative" ($p > .05$) Results

An investigation by Scott Armstrong and me is informative here (Hubbard & Armstrong, 1992). Based on a content analysis of 32 randomly selected annual issues each of the *Journal of Consumer Research* (*JCR*), *Journal of Marketing* (*JM*), and *Journal of Marketing Research* (*JMR*) over the period 1974–1989, we uncovered some 54 articles reporting negative (i.e., $p > .05$) results.[15] We inspected the statistical power found in these 54 papers.

Remember that the power of a statistical test (i.e., the probability of rejecting a false null hypothesis) depends on the selected level of significance (usually .05), the effect size in the population, and the sample size. Tests with insufficient statistical power are likely to yield null findings and thus be seen as being less entitled to publication. Conversely, statistically nonsignificant results buttressed by high power are potential contributions to knowledge (Fagley, 1985), suggesting, perhaps, that a boundary condition on some phenomenon has been met. This is because with high-powered negative results the probability of a Type II error (erroneous acceptance of the null hypothesis) must be low (Rossi, 1990, p. 646). Therefore, it is necessary to determine the statistical power of "insignificant" results.

We adopted Jacob Cohen's (1988) methods and recommendations, as outlined in the appendix to Chapter 2, in an attempt to calculate the statistical power of the 54 articles in our sample reporting nonsignificant outcomes. It was found that 11 of these could not be power-analyzed, 6 because they used methods for which power tests were unavailable and 5 because not enough information to do so was provided. The remaining 43 articles allowed power analyses of 410 statistical significance tests, or an average of 9.5 tests per article. Of these 43 papers, 19 were from *JCR,* 8 from *JM,* and 16 from *JMR.*

Table 3-4 portrays the frequency and cumulative percentage distributions of the average power of the 43 articles to detect small, medium, and large effect sizes in the population. Recall that Cohen (1988) advises that the value .80 be employed when there is no other basis for establishing a satisfactory power level. Following this advice, the mean power levels of these articles are seen to be high; the probabilities of detecting small, medium, and large effect magnitudes per article are .35, .89, and .99, respectively. Corresponding figures based on the 410 individual statistical significance tests are .36, .87, and .98.

Over the period 1974–1989, all three journals showed consistent power levels for discerning small, medium, and large effects. For *JCR* these figures were .40, .87, and .98; for *JM* they were .29, .86, and .99; for *JMR* they were .33, .92, and .99.

On average across all three journals, the 43 statistically nonsignificant articles showed reasonable probabilities of detecting even small effects. Fourteen of these (32.6%) revealed a 50-50 chance or better of doing so, and 2 were able to exceed the recommended level of .80. If medium effects typify marketing's literature, these articles, on average, had almost a 90% chance of distinguishing them. Ten of the papers (23.2%) did not reach the nominated .80 power yardstick (although 9 of these revealed power in the .60 to .79 range), 33% met or exceeded the .99 power level, and none were below .50. All 43 articles reached or outgained the 80% power benchmark to uncover large effect sizes. In fact, 32 (74.4%) of them had a .99 probability or more of rejecting a false null hypothesis (Table 3-4).

The average power levels of publications recording statistically nonsignificant findings were consistent over the 1974–1989 time period enclosing our work. Splitting these time periods by decades, for 1974–1979 the ability to detect small, medium, and large effects across all three journals was .29, .88, and .99. For 1980–1989, these values were .42, .90, and .99.

Table 3-4 Power of 43 Published Marketing Studies With Statistically Nonsignificant ($p > .05$) Results: 1974–1989

	Small Effects		*Medium Effects*		*Large Effects*	
Power	*Frequency*	*Cumulative %*	*Frequency*	*Cumulative %*	*Frequency*	*Cumulative %*
.99–			14	100.0	32	100.0
.95–.98			5	67.4	8	25.6
.90–.94			8	55.8	2	7.0
.80–.89	2	100.0	6	37.2	1	2.3
.70–.79	3	95.3	5	23.2		
.60–.69	5	88.3	4	11.6		
.50–.59	4	76.7	1	2.3		
.40–.49	0	67.5				
.30–.39	2	67.5				
.20–.29	14	62.8				
.10–.19	13	30.2				
Number of articles (43)	.35		.89		.99	
Number of tests (410)	.36		.87		.98	

Source: Adapted from Hubbard & Armstrong (1992, p. 133).

An argument might be made, of course, that one of the reasons these 43 articles with null results were published in the first place is precisely because of their overall high levels of statistical power, the imputation being that *unpublished* manuscripts with $p > .05$ findings were noticeably underpowered. It was not possible for us to evaluate directly the soundness of this argument because, to the best of our knowledge, the power levels of published and unpublished marketing (or other management or social science) papers with null results have never been examined.

It must be pointed out, however, that indirect evidence indicates that concerns about statistical power played a minor part in the publication decision. Authors of published works obtaining, or failing to obtain, statistically significant results do not seem to formally include power considerations in their research designs (J. Cohen, 1990). We found that none of the 54 articles with insignificant findings displayed power calculations (Hubbard & Armstrong, 1992, p. 134). Indeed, as mentioned earlier, 5 of these studies were published without supplying information needed to compute levels of statistical power. Sawyer and Ball (1981) questioned authors of empirical works published in five issues of *JMR* (November 1978 to November 1979) about how they decided on sample sizes. They reported that tangible calculation of statistical power is uncommon among researchers; only 4 of 28 respondents (14%) computed power before data collection, while 2 others did so afterward. Our content analysis of these same five *JMR* issues showed that none of the 59 articles using statistical significance tests told about power calculations, and only 4 alluded to the topic of inadequate sample sizes (Hubbard & Armstrong, 1992, p. 134).

Research papers with null results, but which show that they meet or exceed Cohen's (1988) advocated .80 power level, are capable of yielding helpful information. Sufficiently powered studies failing to reject H_0 can provide strong evidence of a negligible effect size in the population. On the other hand, statistically significant outcomes with high power, often realized with large samples, may mask trivial effect magnitudes. Seen in this light, null results can be useful. The publication of (adequately powered) null outcomes also can be expected to deter researchers from reentering blind alleys. Acknowledging the potential contribution to knowledge of "negative" findings, editorials in the *Journal of Clinical Neuropsychology* (Rourke & Costa, 1979), the *New England Journal of Medicine* (Angell, 1989), and the *Journal of Cerebral Blood Flow and Metabolism* (Dirnagl & Lauritzen, 2010) have confirmed an openness to publish well-designed manuscripts with null results. Social and managerial science journals can benefit by adopting a similar posture.

3.6 Conclusions

It has been shown from numerous perspectives that the *p*-value is not an objective measure of evidence against the null hypothesis, H_0. But even if it was, so what? That the differences between two means or correlations between two variables are not exactly zero hardly qualifies as valuable scientific knowledge. Yet this is what adherents of the significant difference school are all too willing to settle for.

More broadly, it must be said that the notion of the scientific method, in the sense that dutifully abiding by a prescribed set of methodological steps yields knowledge, is apocryphal. Bauer (1994, p. 128) clarifies this when writing that if all the sciences were united by application of the scientific method, then sociology and chemistry would differ only because of their subject matters.[16] Yet it is well known that sociology and chemistry are poles apart in terms of their scientific status. To repeat, it takes a great deal of time and many studies (replications focusing on the estimation of effect magnitudes, including their CIs and degree of overlap) by many people for the research authorities to filter the wheat from the chaff (Bauer, 1994, pp. 48–52). So from the above it cannot possibly be that following the rules of the so-called scientific method is what distinguishes science from non-science. Despite such observations, members of the significant difference paradigm cling resolutely to this illusion. Patrons of the significant sameness camp do not.

Among other things, this chapter has drawn attention to the importance of replication research in the quest for scientific knowledge. Chapter 4 underscores this claim.

Notes

[1] Additional information on the fallibility of the *p*-value as a dependable measure of evidence—for example, its logical flaws and susceptibility to differences in interpretation due to alternative experimental designs—is available in Hubbard and Lindsay (2008, pp. 76, 78–80).

[2] From the questions raised, for example, in Locke (2007, p. 879), it is clear that he and Latham were engaged in abductive (see Section 3.3) as well as inductive reasoning. Locke's not acknowledging this explicitly is understandable for there is some confusion among philosophers of science about the distinction between abductive and inductive inference. In particular, some philosophers view abduction as a form of induction (see, e.g., Ladyman, 2002, p. 219; Okasha, 2002, p. 30).

[3] Introduced by Neyman (see, e.g., 1934, 1937), the construction of frequentist CIs is predicated on the use of random samples. There will be many (most?) occasions in the significant sameness paradigm where employing probability samples is not feasible (it is

important to emphasize that this stricture applies equally in the significant difference paradigm). Does this, then, prohibit the use of CIs? The answer is yes when adopting the dominant *statistical* model of generalization, where one is inferring from sample *to* population. When the less ambitious, but more practical, goal of *empirical* generalization—generalizing *across* many well-defined (sub)populations—favored by the significant sameness approach and its replication strategy is taken up, this is not a pressing concern. These issues are treated at length in Chapters 4 and 6.

4. Disturbingly, Estes (1997, pp. 330–331) allots a major part of this failure, based on several decades of editing journals, to the fact that a great many (psychology) researchers do not understand the theoretical and computational bases of such a measure. Evidence supports Estes's hunch. Responses from 473 authors of journal articles in psychology, behavioral neuroscience, and medicine caused Belia, Fidler, Williams, and Cumming (2005, p. 395) to conclude that many of them have "fundamental and severe misconceptions" about how CIs can be used to make inferences from data.

5. See also Wilkinson and the American Psychological Association Task Force on Statistical Inference (1999, p. 599).

6. See Hubbard and Lindsay (2013a, p. 1383, 2013b, p. 1394) and Kline (2004, p. 74) for further examples of the problems of using *p*-values to decide replication success.

7. Stigler (1999, p. 192), a well-known statistician, lists Peirce as one of the two greatest American scientific minds of the 19th century (physicist J. Willard Gibbs being the other).

8. In this connection, see Haig's (2014) excellent treatment of ATOM, or Abductive Theory Of Method.

9. For example, abduction has played a vital role in scientific advances in the past, as witnessed in Darwin's theory of evolution, Einstein's work on Brownian motion (Okasha, 2002, pp. 31–33), and Keynes's ideas in economics (Skidelsky, 2009, p. 58). It continues to do so, with Magnani (2001, p. 94) waxing that its contributions "have guaranteed the philosophical centrality of abduction in present-day cultural, scientific and technological developments."

10. This reprimand is captured by the well-known example that no number of empirical confirmations of white swans precludes the subsequent appearance of a black one (as in Australia).

11. This is why Starbuck's (2006, p. 167) recommendations in his *The Production of Knowledge* in the social sciences are off the mark:

> 1. Journals should refuse to publish studies that purport to contradict the baseline propositions. Since the propositions are known laws of nature, valid evidence cannot contradict them.
> 2. Journals should refuse to publish studies that do no more than reaffirm the baseline propositions. Known laws of nature need no more documentation.

In the first place, as just told, even laws in the hard sciences are circumstantial. Second, the management and social science literatures have little in the way of baselines against which to judge the authenticity of a new result. Hence the appeal to endorse a policy of significant sameness. Third, even if reasonable baselines existed, a result at odds with them may signal the discovery of a boundary condition, which itself can be extremely helpful.

12. It is not being suggested that every study with null or negative results be published or considered as meaningful. A methodology for judging when to accept a null hypothesis

needs to be formulated. See, in this respect, Cook, Gruder, Hennigan, and Flay (1979), Fagley (1985), Frick (1996), Hubbard and Armstrong (1992), Lindsay (1993b, 1994), Schimmack (2012), and especially Cortina and Folger (1998) and Hauck and Anderson (1986). This topic is revisited in Section 3.5.

13. There are those who disagree with such assertions. Among them are proponents of the Austrian/neo-Austrian school of economics who disdain quantitative estimates. If the latter have such fleeting shelf lives as the neo-Austrians seem to imply, then, of course, it is pointless to make them. Obviously, I do not share this belief, particularly as I call for *multiple* estimates (and their CIs) of the phenomenon at hand, rather than relying on the typical one-off result.

14. I am indebted to Brian Haig for suggesting this example.

15. Deciding whether an article did or did not reject the null hypothesis involved the use of Sterling's (1959, footnote 2, pp. 31–32) detailed criteria. For example, when a dominant hypothesis was apparent, this decision was relatively straightforward. When an article involved two or more variables and it wasn't clear which were the more important, if H_0 was not rejected for at least 50% of these variables, the study was classified as H_0 not rejected and vice versa. A similar logic was employed for works with multiple studies within a given article.

16. Shapin (1995, pp. 294, 306) drops comparable hints about the gullibility of such thinking among some members of the social sciences.