1

CHANGING TIMES

> It is simply that the things that appear to be permanent and dominant at any given moment in history can change with stunning rapidity. Eras come and go.
>
> —George Friedman (2009, p. 3)

This chapter explains the basic rationale of the movement for statistics reform in the behavioral sciences. It also identifies critical limitations of traditional significance testing that are elaborated throughout the book and reviews the controversy about significance testing in psychology and other disciplines. I argue that overreliance on significance testing as basically the sole way to evaluate hypotheses has damaged the research literature and impeded the development of psychology and other areas as empirical sciences. Alternatives are introduced that include using interval estimation of effect sizes, taking replication seriously, and focusing on the substantive significance of research results instead of just on whether or not they are statistically significant. Prospects for further reform of data analysis methods are also considered.

DOI: 10.1037/14136-001
Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, Second Edition, by R. B. Kline

PRÉCIS OF STATISTICS REFORM

Depending on your background, some of these points may seem shocking, even radical, but they are becoming part of mainstream thinking in many disciplines. **Statistics reform** is the effort to improve quantitative literacy in psychology and other behavioral sciences among students, researchers, and university faculty not formally trained in statistics (i.e., most of us). The basic aims are to help researchers better understand their own results, communicate more clearly about those findings, and improve the quality of published studies. Reform advocates challenge conventional wisdom and practices that impede these goals and emphasize more scientifically defensible alternatives.

Reformers also point out uncomfortable truths, one of which is that much of our thinking about data analysis is stuck in the 1940s (if not earlier). A sign of arrested development is our harmful overreliance on significance testing. Other symptoms include the failure to report effect sizes or consider whether results have scientific merit, both of which have nothing to do with statistical significance. In studies of intervention outcomes, a statistically significant difference between treated and untreated cases also has nothing to do with whether treatment leads to any tangible benefits in the real world. In the context of diagnostic criteria, **clinical significance** concerns whether treated cases can no longer be distinguished from control cases not meeting the same criteria. For example, does treatment typically prompt a return to normal levels of functioning? A treatment effect can be statistically significant yet trivial in terms of its clinical significance, and clinically meaningful results are not always statistically significant. Accordingly, the proper response to claims of statistical significance in any context should be "so what?"—or, more pointedly, "who cares?"—without more information.

Cognitive Errors

Another embarrassing truth is that so many cognitive errors are associated with significance testing that some authors describe a kind of **trained incapacity** that prevents researchers from understanding their own results; others describe a major educational failure (Hubbard & Armstrong, 2006; Ziliak & McCloskey, 2008). These misinterpretations are widespread among students, researchers, and university professors, some of whom teach statistics courses. So students learn false beliefs from people who should know better, but do not, in an ongoing cycle of misinformation. Ziliak and McCloskey (2008) put it this way:

> The textbooks are wrong. The teaching is wrong. The seminar you just attended is wrong. The most prestigious journal in your scientific field is wrong. (p. 250)

Most cognitive errors involve exaggerating what can be inferred from the outcomes of statistical tests, or *p* values (probabilities), listed in computer output. Common misunderstandings include the belief that *p* measures the likelihood that a result is due to sampling error (chance) or the probability that the null hypothesis is true. These and other false beliefs make researchers overconfident about their findings and excessively lax in some critical practices. One is the lip service paid to replication. Although I would wager that just as many behavioral scientists as their natural science colleagues would endorse replication as important, replication is given scant attention in the behavioral sciences. This woeful practice is supported by false beliefs.

Costs of Significance Testing

Summarized next are additional ways in which relying too much on significance testing has damaged our research literature. Nearly all published studies feature statistical significance, but studies without significant results are far less likely to be published or even submitted to journals (Kupfersmid & Fiala, 1991). This **publication bias for significance** suggests that the actual rate among published studies of Type I error, or incorrect rejection of the null hypothesis, is higher than indicated by conventional levels of statistical significance, such as .05. Ellis (2010) noted that because researchers find it difficult to get negative results published, Type I errors, once made, are hard to correct. Longford (2005) warned that the uncritical use of significance testing would lead to a "junkyard of unsubstantiated confidence," and Simmons, Nelson, and Simonsohn (2011) used the phrase "false-positive psychology" to describe the same problem.

Publication bias for significance also implies that the likelihood of Type II error, or failure to reject the null hypothesis when it is false in the population, is basically zero. In a less biased literature, though, information about the power, or the probability of finding statistical significance (rejecting the null hypothesis) when there is a real effect, would be more relevant. There are free computer tools for estimating power, but most researchers—probably at least 80% (e.g., Ellis, 2010)—ignore the power of their analyses. This is contrary to advice in the *Publication Manual* of the American Psychological Association (APA) that researchers should "routinely provide evidence that the study has sufficient power to detect effects of substantive interest" (APA, 2010, p. 30).

Ignoring power is regrettable because the median power of published nonexperimental studies is only about .50 (e.g., Maxwell, 2004). This implies a 50% chance of correctly rejecting the null hypothesis based on the data. In this case the researcher may as well not collect any data but instead just toss a coin to decide whether or not to reject the null hypothesis. This simpler, cheaper method has the same chance of making correct decisions in the long run (F. L. Schmidt & Hunter, 1997).

A consequence of low power is that the research literature is often difficult to interpret. Specifically, if there is a real effect but power is only .50, about half the studies will yield statistically significant results and the rest will yield no statistically significant findings. If all these studies were somehow published, the number of positive and negative results would be roughly equal. In an old-fashioned, narrative review, the research literature would appear to be ambiguous, given this balance. It may be concluded that "more research is needed," but any new results will just reinforce the original ambiguity, if power remains low.

Confusing statistical significance with scientific relevance unwittingly legitimizes fad topics that clutter the literature but have low substantive value. With little thought about a broader rationale, one can collect data and then apply statistical tests. Even if the numbers are random, some of the results are expected to be statistically significant, especially in large samples. The objective appearance of significance testing can lend an air of credibility to studies with otherwise weak conceptual foundations. This is especially true in "soft" research areas where theories are neither convincingly supported nor discredited but simply fade away as researchers lose interest (Meehl, 1990). This lack of cumulativeness led Lykken (1991) to declare that psychology researchers mainly build castles in the sand.

Statistical tests of a treatment effect that is actually clinically significant may fail to reject the null hypothesis of no difference when power is low. If the researcher in this case ignored whether the observed effect size is clinically significant, a potentially beneficial treatment may be overlooked. This is exactly what was found by Freiman, Chalmers, Smith, and Kuebler (1978), who reviewed 71 randomized clinical trials of mainly heart- and cancer-related treatments with "negative" results (i.e., not statistically significant). They found that if the authors of 50 of the 71 trials had considered the power of their tests along with the observed effect sizes, those authors should have concluded just the opposite, or that the treatments resulted in clinically meaningful improvements.

If researchers become too preoccupied with statistical significance, they may lose sight of other, more important aspects of their data, such as whether the variables are properly defined and measured and whether the data respect test assumptions. There are clear problems in both of these areas. One is the **measurement crisis**, which refers to a substantial decline in the quality of instruction about measurement in psychology over the last 30 years or so. Psychometrics courses have disappeared from many psychology undergraduate programs, and about one third of psychology doctoral programs in North America offer no formal training in this area at all (Aiken et al., 1990;

Friederich, Buday, & Kerr, 2000). There is also evidence of widespread poor practices. For example, Vacha-Haase and Thompson (2011) found that about 55% of authors did not even mention score reliability in over 13,000 primary studies from a total of 47 meta-analyses of reliability generalization in the behavioral sciences. Authors mentioned reliability in about 16% of the studies, but they merely inducted values reported in other sources, such as test manuals, as if these applied to their data. Such **reliability induction** requires explicit justification, but researchers rarely compared characteristics of their samples with those from cited studies of score reliability.

A related problem is the **reporting crisis**, which refers to the fact that researchers infrequently present evidence that their data respect distributional or other assumptions of statistical tests (e.g., Keselman et al., 1998). The false belief that statistical tests are robust against violations of their assumptions in data sets of the type analyzed in actual studies may explain this flawed practice. Other aspects of the reporting crisis include the common failure to describe the nature and extent of missing data, steps taken to deal with the problem, and whether selection among alternatives could appreciably affect the results (e.g., Sterner, 2011). Readers of many journal articles are given little if any reassurance that the results are trustworthy.

Even if researchers avoided the kinds of mistakes just described, there are grounds to suspect that *p* values from statistical tests are simply incorrect in most studies:

1. They (*p* values) are estimated in theoretical sampling distributions that assume random sampling from known populations. Very few samples in behavioral research are random samples. Instead, most are convenience samples collected under conditions that have little resemblance to true random sampling. Lunneborg (2001) described this problem as a mismatch between design and analysis.
2. Results of more quantitative reviews suggest that, due to assumptions violations, there are few actual data sets in which significance testing gives accurate results (e.g., Lix, Keselman, & Keseleman, 1996). These observations suggest that *p* values listed in computer output are usually suspect. For example, this result for an independent samples *t* test calculated in SPSS looks impressively precise,

$$t(27) = 2.373, p = .025000184017821007$$

but its accuracy is dubious, given the issues just raised. If *p* values are generally wrong, so too are decisions based on them.

3. Probabilities from statistical tests (*p* values) generally assume that all other sources of error besides sampling error are nil. This includes measurement error; that is, it is assumed that $r_{XX} = 1.00$, where r_{XX} is a score reliability coefficient. Other sources of error arise from failure to control for extraneous sources of variance or from flawed operational definitions of hypothetical constructs. It is absurd to assume in most studies that there is no error variance besides sampling error. Instead it is more practical to expect that sampling error makes up the small part of all possible kinds of error when the number of cases is reasonably large (Ziliak & McCloskey, 2008).

The *p* values from statistical tests do not tell researchers what they want to know, which often concerns whether the data support a particular hypothesis. This is because *p* values merely estimate the conditional probability of the data under a *statistical* hypothesis—the null hypothesis—that in most studies is an implausible, straw man argument. In fact, *p* values do not directly "test" any hypothesis at all, but they are often misinterpreted as though they describe hypotheses instead of data.

Although *p* values ultimately provide a yes-or-no answer (i.e., reject or fail to reject the null hypothesis), the question—$p < \alpha$?, where α is the criterion level of statistical significance, usually .05 or .01—is typically uninteresting. The yes-or-no answer to this question says nothing about scientific relevance, clinical significance, or effect size. This is why Armstrong (2007) remarked that significance tests do not aid scientific progress even when they are properly done and interpreted.

New Statistics, New Thinking

Cumming (2012) recommended that researchers pay less attention to *p* values. Instead, researchers should be more concerned with sample results Cumming (2012) referred to as the **new statistics**. He acknowledged that the "new" statistics are not really new at all. What should be new instead is a greater role afforded them in describing the results. The new statistics consist mainly of effect sizes and confidence intervals. The *Publication Manual* is clear about effect size: "For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size" (APA, 2010, p. 34). The qualifier "almost always" refers to the possibility that, depending on the study, it may be difficult to compute effect sizes, such as when the scores are ranks or are presented in complex hierarchically structured designs. But it is possible to calculate effect sizes in most studies, and the effect size void for some kinds of designs is being filled by ongoing research.

Significance tests do not directly indicate effect size, and a common mistake is to answer the question $p < \alpha$? but fail to report and interpret effect sizes. Because effect sizes are sample statistics, or **point estimates**, that approximate population effect sizes, they are subject to sampling error. A confidence interval, or **interval estimate**, on a point estimate explicitly indicates the degree of sampling error associated with that statistic. Although sampling error is estimated in significance testing, that estimate winds up "hidden" in the calculation of p. But the amount of sampling error is made explicit by the lower and upper bounds of a confidence interval. Reporting confidence intervals reflects **estimation thinking** (Cumming, 2012), which deals with the questions "how much?" (point estimate) and "how precise?" (margin of error). The *Publication Manual* offers this advice: "Whenever possible, base discussion and interpretation of results on point and interval estimates" (APA, 2010, p. 34).

Estimation thinking is subsumed under **meta-analytic thinking**, which is fundamentally concerned with the accumulation of evidence over studies. Its basic aspects are listed next:

1. An accurate appreciation of the results of previous studies is seen as essential.
2. A researcher should view his or her own study as making a modest contribution to the literature. Hunter, Schmidt, and Jackson (1982) put it this way: "Scientists have known for centuries that a single study will not resolve a major issue. Indeed, a small sample study will not even resolve a minor issue" (p. 10).
3. A researcher should report results so that they can be easily incorporated into a future meta-analysis.
4. Retrospective interpretation of new results, once collected, is called for via direct comparison with previous effect sizes.

Thinking meta-analytically is incompatible with using statistical tests as the sole inference tool. This is because the typical meta-analysis estimates the central tendency and variability of effect sizes across sets of related primary studies. The focus on effect size and not statistical significance in individual studies also encourages readers of meta-analytic articles to think outside the limitations of the latter. There are statistical tests in meta-analysis, but the main focus is on whether a particular set of effect sizes is estimating the same population effect size and also on the magnitude and precision of mean effect sizes.

The new statistics cannot solve all that ails significance testing; no such alternative exists (see Cohen, 1994). For example, the probabilities associated with confidence intervals also assume that all other sources of imprecision besides sampling error are zero. There are ways to correct some effect sizes for measurement error, though, so this assumption is not always so strict. Abelson

(1997a) referred to the **law of the diffusion of idiocy**, which says that every foolish practice of significance testing will beget a corresponding misstep with confidence intervals. This law applies to effect sizes, too. But misinterpretation of the new statistics is less likely to occur if researchers can refrain from applying the same old, dichotomous thinking from significance testing. Thinking meta-analytically can also help to prevent misunderstanding.

You should know that measuring effect size in treatment outcome studies is insufficient to determine clinical significance, especially when outcomes have **arbitrary (uncalibrated) metrics** with no obvious connection to real-world status. An example is a 7-point Likert scale for an item on a self-report measure. This scale is arbitrary because its points could be represented with different sets of numbers, such as 1 through 7 versus −3 through 3 in whole-number increments, among other possibilities. The total score over a set of such items is arbitrary, too. It is generally unknown for arbitrary metrics (a) how a 1-point difference reflects the magnitude of change on the underlying construct and (b) exactly at what absolute points along the latent dimension observed scores fall. As Andersen (2007) noted, "Reporting effect sizes on arbitrary metrics alone with no reference to real-world behaviors, however, is no more meaningful or interpretable than reporting *p* values" (p. 669). So, determining clinical significance is not just a matter of statistics; it also requires strong knowledge about the subject matter.

These points highlight the idea that the evaluation of the clinical, practical, theoretical, or, more generally, **substantive significance** of observed effect sizes is a qualitative judgment. This judgment should be informed and open to scrutiny, but it will also reflect personal values and societal concerns. This is not unscientific because the assessment of all results in science involves judgment (Kirk, 1996). It is better to be open about this fact than to base decisions solely on "objective," mechanically applied statistical rituals that do not address substantive significance. Ritual is no substitute for critical thinking.

RETROSPECTIVE

Behavioral scientists did not always use statistical tests, so it helps to understand a little history behind the significance testing controversy; see Oakes (1986), Nickerson (2000), and Ziliak and McCloskey (2008) for more information.

Hybrid Logic of Statistical Tests (1920–1960)

Logical elements of significance testing were present in scientific papers as early as the 1700s (Stigler, 1986), but those basics were not organized into a systematic method until the early 1900s. Today's significance testing is actually a hybrid of two schools of thought, one from the 1920s associated with Ronald Fisher (e.g., 1925) and another from the 1930s called the Neyman–Pearson approach, after Jerzy Neyman and Egon S. Pearson (e.g., 1933). Other individuals, such as William Gosset and Karl Pearson, contributed to these schools, but the work of the three principals listed first forms the genesis of significance testing (Ziliak & McCloskey, 2008, elaborate on Gosset's role).

Briefly, the Neyman–Pearson model is an extension of the Fisher model, which featured only a null hypothesis and estimation with statistical tests of the conditional probability of the data, or *p* values. There was no alternative hypothesis in Fisher's model. The conventional levels of statistical significance used today, .05 and .01, are correctly attributed to Fisher, but he did *not* advocate that they be blindly applied across all studies. Doing so, wrote Fisher (1956, p. 42), would be "absurdly academic" because no fixed level of significance could apply across all studies. This view is very different from today's practice, where $p < .05$ and $p < .01$ are treated as golden rules. For its focus on *p* values under the null hypothesis, Fisher's model has been called the ***p* value approach** (Huberty, 1993). The addition of the alternative hypothesis to the basic Fisher model, the attendant specification of one- or two-tailed regions of rejection, and the a priori specification of fixed levels of α across all studies characterize the Neyman–Pearson model, also called the **fixed α approach** (Huberty, 1993). This model also brought with it the conceptual framework of power and related decision errors, Type I and Type II.

To say that advocates of the Fisher model and the Neyman–Pearson model exchanged few kind words about each other's ideas is an understatement. Their long-running debate was acrimonious and included attempts by Fisher to block faculty appointments for Neyman. Nevertheless, the integration of the two models by other statisticians into what makes up contemporary significance testing took place roughly between 1935 and 1950. Gigerenzer (1993) referred to this integrated model as the **hybrid logic of scientific inference**, and Dixon and O'Reilly (1999) called it the **Intro Stats method**. Many authors have noted that (a) this hybrid model would have been rejected by Fisher, Neyman, and Pearson, although for different reasons, and (b) its composite nature is a source of confusion among students and researchers.

Rise of the Intro Stats Method, Testimation, and Sizeless Science (1940–1960)

Before 1940, statistical tests were rarely used in psychology research. Authors of works from the time instead applied in nonstandard ways a variety of descriptive statistics or rudimentary test statistics, such as the **critical ratio** of a sample statistic over its standard error (now called *z* or *t* when assuming normality). An older term for the standard error—actually two times the square root of the standard error—is the **modulus**, described in 1885 by the economist Francis Ysidro Edgeworth (Stigler, 1978) to whom the term *statistical significance* is attributed. From about 1940–1960, during what Gigerenzer and Murray (1987) called the **inference revolution**, the Intro Stats method was widely adopted in psychology textbooks and journal editorial practice as *the* method to test hypotheses. The move away from the study of single cases (e.g., operant conditioning studies) to the study of groups over roughly 1920–1950 contributed to this shift. Another factor is what Gigerenzer (1993) called the **probabilistic revolution**, which introduced indeterminism as a major theoretical concept in areas such as quantum mechanics in order to better understand the subject matter. In psychology, though, it was used to mechanize the inference process, a critical difference, as it turns out.

After the widespread adoption of the Intro Stats method, there was an increase in the reporting of statistical tests in journal articles in psychology. This trend is obvious in Figure 1.1, reproduced from Hubbard and Ryan (2000). They sampled about 8,000 articles published during 1911–1998 in randomly selected issues of 12 different APA journals. Summarized in the figure are percentages of articles in which results of statistical tests were reported.

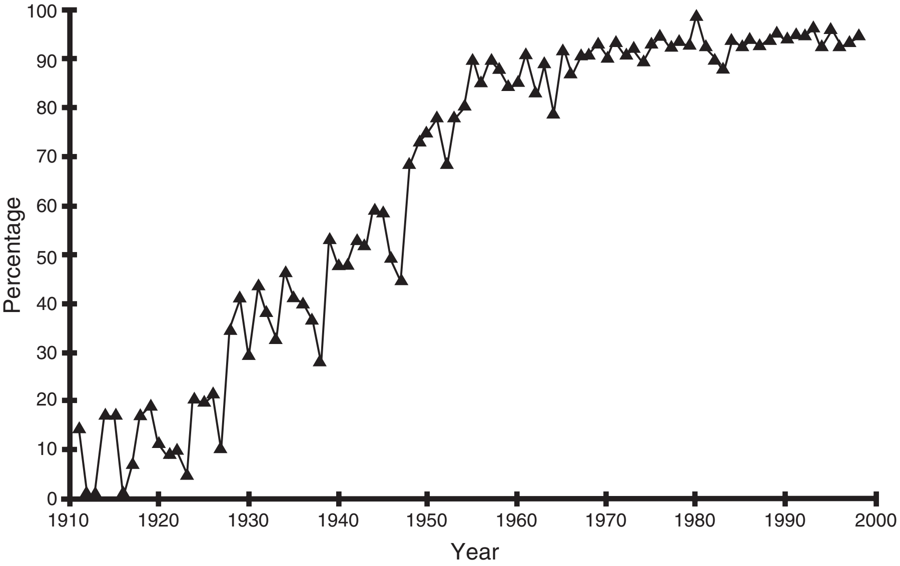

Figure 1.1. Percentage of articles reporting results of statistical tests in 12 journals of the American Psychological Association from 1911 to 1988. From “The Historical Growth of Statistical Significance Testing in Psychology—And Its Future Prospects,” by R. Hubbard and P. A. Ryan, 2000, *Educational and Psychological Measurement, 60,* p. 665. Copyright 2001 by Sage Publications. Reprinted with permission.

This percentage is about 17% from 1911 to 1929. It increases to around 50% in 1940, continues to rise to about 85% by 1960, and has exceeded 90% since the 1970s. The time period 1940–1960 corresponds to the inference revolution.

Although the 1990s is the most recent decade represented in Figure 1.1, there is no doubt about the continuing, near-universal reporting of statistical tests in journals. Hoekstra, Finch, Kiers, and Johnson (2006) examined a total of 266 articles published in *Psychonomic Bulletin & Review* during 2002–2004. Results of significance tests were reported in about 97% of the articles, but confidence intervals were reported in only about 6%. Sadly, *p* values were misinterpreted in about 60% of surveyed articles. Fidler, Burgman, Cumming, Buttrose, and Thomason (2006) sampled 200 articles published in two different biology journals. Results of significance testing were reported in 92% of articles published during 2001–2002, but this rate dropped to 78% in 2005. There were also corresponding increases in the reporting of confidence intervals, but power was estimated in only 8% and *p* values were misinterpreted in 63%.

Some advantages to the institutionalization of the Intro Stats method were noted by Gigerenzer (1993). Journal editors could use significance test outcomes to decide which studies to publish or reject, respectively, those with or without statistically significant results, among other considerations. The method of significance testing is mechanically applied and thus seems to eliminate subjective judgment. That this objectivity is illusory is another matter. Significance testing gave researchers a common language and perhaps identity as members of the same grand research enterprise. It also distinguished them from their natural science colleagues, who may use statistical tests to detect outliers but not typically to test hypotheses (Gigerenzer, 1993).

The combination of significance testing and a related cognitive error is **testimation** (Ziliak & McCloskey, 2008). It involves exclusive focus on the question $p < \alpha$? If the answer is “yes,” the results are automatically taken to be scientifically relevant, but issues of effect size and precision are ignored. Testimators also commit the **inverse probability error** (Cohen, 1994) by falsely believing that *p* values indicate the probability that the null hypothesis is true. Under this fallacy, the result $p = .025$, for example, is taken to mean that there is only a 2.5% chance that the null hypothesis is true. A researcher who mistakenly believes that low *p* values make the null hypothesis unlikely may become overly confident in the results.

Presented next is hypothetical text that illustrates the language of testimation:

> A 2 × 2 × 2 (Instructions × Incentive × Goals) factorial ANOVA was conducted with the number of correct items as the dependent variable. The 3-way interaction was significant, $F(1, 72) = 5.20$, $p < .05$, as were all 2-way

> interactions, Instructions × Incentive, $F(1, 72) = 11.95$, $p < .001$; Instructions × Goals, $F(1, 72) = 25.40$, $p < .01$; Incentive × Goals, $F(1, 72) = 9.25$, $p < .01$, and two of three of the main effects, Instructions, $F(1, 72) = 11.60$, $p < .01$; Goals, $F(1, 72) = 6.25$, $p < .05$.

This text chockablock with numbers—which is poor writing style—says nothing about the magnitudes of all those "significant" effects. If later in the hypothetical article the reader is still given no information about effect sizes, that is **sizeless science**. Getting excited about "significant" results while knowing nothing about their magnitudes is like ordering from a restaurant menu with no prices: You may get a surprise (good or bad) when the bill (statement of effect size) comes.

Increasing Criticism of Statistical Tests (1940–Present)

There has been controversy about statistical tests for more than 80 years, or as long as they have been around. Boring (1919), Berkson (1942), and Rozeboom (1960) are among earlier works critical of significance testing. Numbers of published articles critical of significance testing have increased exponentially since the 1940s. For example, Anderson, Burnham, and Thompson (2000) found less than 100 such works published during the 1940s–1970s in ecology, medicine, business, economics, or the behavioral sciences, but about 200 critical articles were published in the 1990s. W. L. Thompson (2001) listed a total of 401 references for works critical of significance testing, and Ziliak and McCloskey (2008, pp. 57–58) cited 125 such works in psychology, education, business, epidemiology, and medicine, among other areas.

Proposals to Ban Significance Testing (1990s–Present)

The significance testing controversy escalated to the point where, by the 1990s, some authors called for a ban in research journals. A ban was discussed in special sections or issues of *Journal of Experimental Education* (B. Thompson, 1993), *Psychological Science* (Shrout, 1997), and *Research in the Schools* (McLean & Kaufman, 1998) and in an edited book by Harlow, Mulaik, and Steiger (1997), the title of which asks "What if there were no significance tests?" Armstrong (2007) offered this more recent advice:

> When writing for books and research reports, researchers should omit mention of tests of statistical significance. When writing for journals, researchers should seek ways to reduce the potential harm of reporting significance tests. They should also omit the word significance because findings that reject the null hypothesis are not significant in the everyday use of the term, and those that [fail to] reject it are not insignificant. (p. 326)

In 1996, the Board of Scientific Affairs of the APA convened the Task Force on Statistical Inference (TFSI) to respond to the ongoing significance testing controversy and elucidate alternatives. The report of the TFSI (Wilkinson & the TFSI, 1999) dealt with many issues and offered suggestions for the then-upcoming fifth edition of the *Publication Manual:*

1. Use minimally sufficient analyses (simpler is better).
2. Do not report results from computer output without knowing what they mean. This includes p values from statistical tests.
3. Document assumptions about population effect sizes, sample sizes, or measurement behind a priori estimates of statistical power. Use confidence intervals about observed results instead of estimating observed (post hoc) power.
4. Report effect sizes and confidence intervals for primary outcomes or whenever p values are reported.
5. Give assurances to a reasonable degree that the data meet statistical assumptions.

The TFSI decided in the end not to recommend a ban on statistical tests. In its view, such a ban would be a too extreme way to curb abuses.

Fifth and Sixth Editions of the APA's *Publication Manual* (2001–2010)

The fifth edition of the *Publication Manual* (APA, 2001) took a stand similar to that of the TFSI regarding significance testing. That is, it acknowledged the controversy about statistical tests but stated that resolving this issue was not a proper role of the *Publication Manual.* The fifth edition went on to recommend the following:

1. Report adequate descriptive statistics, such as means, variances, and sizes of each group and a pooled within-groups variance–covariance matrix in a comparative study. This information is necessary for later meta-analyses or secondary analyses by others.
2. Effect sizes should "almost always" be reported, and the absence of effect sizes was cited as an example of a study defect.
3. The use of confidence intervals was "strongly recommended" but not required.

The sixth edition of the *Publication Manual* (APA, 2010) used similar language when recommending the reporting of effect sizes and confidence intervals. Predictably, not everyone is happy with the report of the TFSI or the wording of the *Publication Manual.* B. Thompson (1999) noted that only encouraging the reporting of effect sizes or confidence intervals presents a self-canceling mixed message. Ziliak and McCloskey (2008, p. 125) chastised the *Publication Manual* for "retaining the magical incantations of $p < .05$ and $p < .01$." S. Finch, Cumming, and Thomason (2001) contrasted the recommendations about statistical analyses in the *Publication Manual* with the more straightforward guidelines in the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals*, recently revised (International Committee of Medical Journal Editors, 2010). Kirk (2001) urged that the then-future sixth edition of the *Publication Manual* should give more detail than the fifth edition about the TFSI's recommendations. Alas, the sixth edition does not contain such information, but I aim to provide you with specific skills of this type as you read this book.

Reform-Oriented Editorial Policies and Mixed Evidence of Progress (1980s–Present)

Journal editorials and reviewers are the gatekeepers of the research literature, so editorial policies can affect the quality of what is published. Described next are three examples of efforts to change policies in reform-oriented directions with evaluations of their impact; see Fidler, Thomason, Cumming, Finch, and Leeman (2004) and Fidler et al. (2005) for more examples.

Kenneth J. Rothman was the assistant editor of the *American Journal of Public Health* (AJPH) from 1984 to 1987. In his revise-and-submit letters, Rothman urged authors to remove from their manuscripts all references to *p* values (e.g., Fidler et al., 2004, p. 120). He founded the journal *Epidemiology* in 1990 and served as its first editor until 2000. Rothman's (1998) editorial letter to potential authors was frank:

> When writing for *Epidemiology*, you can . . . enhance your prospects if you omit tests of statistical significance. . . . In *Epidemiology*, we do not publish them at all. . . . We discourage the use of this type of thinking in the data analysis. . . . We also would like to see the interpretation of a study based not on statistical significance, or lack of it . . . but rather on careful quantitative consideration of the data in light of competing explanations for the findings. (p. 334)

Fidler et al. (2004) examined 594 AJPH articles published from 1982 to 2000 and 100 articles published in *Epidemiology* between 1990 and 2000. Reporting based solely on statistical significance dropped from about 63% of the AJPH articles in 1982 to about 5% of articles in 1986–1989. But in many AJPH articles there was evidence that interpretation was based mainly on undisclosed significance test results. The percentages of articles in which confidence intervals were reported increased from about 10% to 54% over the same period. But these changes in reporting practices in AJPH articles did not generally persist past Rothman's tenure.

From 1993 to 1997, Geoffrey R. Loftus was the editor of *Memory & Cognition*. Loftus (1993) gave these guidelines to potential contributors:

> I intend to try to decrease the overwhelming reliance on hypothesis testing as the major means of transiting from data to conclusions. . . . In lieu of hypothesis testing, I will emphasize the increased use of figures depicting sample means along with standard error bars. . . . More often than not, inspection of such a figure will immediately obviate the necessity of any hypothesis testing procedures. In such situations, presentation of the usual hypothesis-testing information (*F* values, *p* values, etc.) will be discouraged. I believe . . . that . . . an overreliance on the impoverished binary conclusions yielded by the hypothesis-testing procedure has subtly seduced our discipline into insidious conceptual cul-de-sacs that have impeded our vision and stymied our potential. (p. 3)

Loftus apparently encountered considerable resistance, if not outright obstinacy, on the part of some authors. For example, Loftus calculated confidence intervals for about 100 authors who failed or even refused to do so on their own. In contrast, Rothman reported little resistance from authors who submitted works to *Epidemiology* (see Fidler et al., 2004, p. 124). S. Finch et al. (2004) examined a total of 696 articles published in *Memory & Cognition* before, during, and after Loftus's editorship. The rate of reporting of confidence intervals increased from 7% from before Loftus's tenure to 41%, but the rate dropped to 24% just after Loftus departed. But these confidence intervals were seldom interpreted; instead, authors relied mainly on statistical test outcomes to describe the results.

Another expression of statistics reform in editorial policy are the requirements of about 24 journals in psychology, education, counseling, and other areas for authors to report effect sizes.[1] Some of these are flagship journals of associations (e.g., American Counseling Association, Council for Exceptional Children), each with about 40,000–45,000 members. Included among journals that require effect sizes are three APA journals, *Health Psychology*, *Journal of Educational Psychology*, and *Journal of Experimental Psychology: Applied*. The requirement to report effect sizes sends a strong message to potential contributors that use of significance testing alone is not acceptable.

Early suggestions to report effect sizes fell mainly on deaf ears. S. Finch et al. (2001) found little evidence for effect size estimation or interval estimation in articles published in *Journal of Applied Psychology* over the 40-year period from 1940 to 1999. Vacha-Haase and Ness (1999) found the rate of effect size reporting was about 25% in *Professional Psychology: Research and Practice*, but authors did not always interpret the effect sizes they reported. Results from more recent surveys are better. Dunleavy, Barr, Glenn, and

[1] http://people.cehd.tamu.edu/~bthompson/index.htm, scroll down to hyperlinks.

Miller (2006) reviewed 736 articles published over 2002–2005 in five different applied, experimental, or personnel psychology journals. The overall rate of effect size reporting was about 62.5%. Among studies where no effect sizes were reported, use of the techniques of analysis of variance (ANOVA) and the *t* test were prevalent. Later I will show you that effect sizes are actually easy to calculate in such analyses, so there is no excuse for not reporting them. Andersen (2007) found that in a total of 54 articles published in 2005 in three different sport psychology journals, effect sizes were reported in 44 articles, or 81%. But the authors of only seven of these articles interpreted effect sizes in terms of substantive significance. Sun, Pan, and Wang (2010) reviewed a total of 1,243 works published in 14 different psychology and education journals during 2005–2007. The percentage of articles reporting effect sizes was 49%, and 57% of these authors interpreted their effect sizes.

Evidence for progress in statistics reform is thus mixed. Researchers seem to report effect sizes more often, but improvement in reporting confidence intervals may lag behind. Too many authors do not interpret the effect sizes they report, which avoids dealing with the question of why does an effect of this size matter. It is poor practice to compute effect sizes only for statistically significant results. Doing so amounts to business as usual where the significance test is still at center stage (Sohn, 2000). Real reform means that effect sizes are interpreted for their substantive significance, not just reported.

OBSTACLES TO REFORM

There are two great obstacles to continued reform. The first is inertia: It is human nature to resist change, and it is hard to give up familiar routines. Belasco and Stayer (1993) put it like this: "Most of us overestimate the value of what we currently have, and have to give up, and underestimate the value of what we may gain" (p. 312). But science demands that researchers train the lens of skepticism on their own assumptions and methods. Such self-criticism and intellectual honesty do not come easy, and not all researchers are up for the task. Defense attorney Gerry Spence (1995) wrote, "I would rather have a mind opened by wonder than one closed by belief" (p. 98). This conviction identifies a scientist's special burden.

The other big obstacle is vested interest, which is in part economic. I am speaking mainly about applying for research grants. Most of us know that grant monies are allocated in part on the assurance of statistical significance. Many of us also know how to play the **significance game**, which goes like this: Write application. Promise significance. Get money. Collect data until significance is found, which is virtually guaranteed because any effect that is not zero needs only a large enough sample in order to be significant.

Report results but mistakenly confuse statistical significance with scientific relevance. Sound trumpets about our awesomeness, move on to a different kind of study (do not replicate). Ziliak and McCloskey (2008) were even more candid:

> Significance unfortunately is a useful means toward personal ends in the advance of science—status and widely distributed publications, a big laboratory, a staff of research assistants, a reduction in teaching load, a better salary, the finer wines of Bordeaux. Precision, knowledge, and control. In a narrow and cynical sense statistical significance is the way to achieve these. Design experiment. Then calculate statistical significance. Publish articles showing "significant" results. Enjoy promotion. But it is not science, and it will not last. (p. 32)

Maybe I am a naive optimist, but I believe there is enough talent and commitment to improving research practices among too many behavioral scientists to worry about unheeded calls for reform. But such changes do not happen overnight. Recall that it took about 20 years for researchers to widely use statistical tests (see Figure 1.1), and sometimes shifts in scientific mentality await generational change. Younger researchers may be less set in their ways than the older generation and thus more open to change. But some journal editors—who are typically accomplished and experienced researchers—are taking the lead in reform. So are the authors of many of the works cited throughout this book.

Students are promising prospects for reform because they are, in my experience and that of others (Hyde, 2001), eager to learn about the significance testing controversy. They can also understand ideas such as effect size and interval estimation even in introductory courses. In fact, I find it is easier to teach undergraduates these concepts than the convoluted logic of significance testing. Other reform basics are even easier to convey (e.g., replicate—do not just talk about it.)

PROSPECTIVE

I have no crystal ball, but I believe that I can reasonably speculate about three anticipated developments in light of the events just described:

1. The role of significance testing will continue to get smaller and smaller to the point where researchers must defend its use. This justification should involve explanation of why the narrow assumptions about sampling and score characteristics in significance testing are not unreasonable in a particular study. Estimation of a priori power will also be required whenever statistical

tests are used. I and others (e.g., Kirk, 1996) envision that the behavioral sciences will become more like the natural sciences. That is, we will report the directions, magnitudes, and precisions of our effects; determine whether they replicate; and evaluate them for their substantive significance, not simply their statistical significance.

2. I expect that the best behavioral science journals will require evidence for replication. This requirement would send the strong message that replication is standard procedure. It would also reduce the number of published studies, which may actually improve quality by reducing noise (one-shot studies, unsubstantiated claims) while boosting signal (replicated results).
3. I concur with Rodgers (2010) that a "quiet methodological revolution" is happening that is also part of statistics reform. This revolution concerns the shift from testing individual hypotheses for statistical significance to the evaluation of entire mathematical and statistical models. There is a limited role for significance tests in statistical modeling techniques such as structural equation modeling (e.g., Kline, 2010, Chapter 8), but it requires that researchers avoid making the kinds of decision errors often associated with such tests.

CONCLUSION

Basic tenets of statistics reform emphasize the need to (a) decrease the role of significance testing and thus also reduce the damaging impact of related cognitive distortions; (b) shift attention to other kinds of statistics, such as effect sizes and confidence intervals; (c) reestablish the role of informed judgment and downplay mere statistical rituals; and (d) elevate replication. The context for reform goes back many decades, and the significance testing controversy has now spread across many disciplines. Progress toward reform has been slow, but the events just summarized indicate that continued use of significance testing as the only way to evaluate hypotheses is unlikely. The points raised set the stage for review in the next chapter of fundamental concepts about sampling and estimation from a reform perspective.

LEARN MORE

Listed next are three works about the significance testing controversy from fields other than psychology, including Armstrong (2007) in forecasting;

Guthery, Lusk, and Peterson (2001) in wildlife management; and McCloskey and Ziliak (2009) in medicine.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23,* 321–327. doi:10.1016/j.ijforecast.2007.03.004

Guthery, F. S., Lusk, J. J., & Peterson, M. J. (2001). The fall of the null hypothesis: Liabilities and opportunities. *Journal of Wildlife Management, 65,* 379–384. doi:10.2307/3803089

McCloskey, D. N., & Ziliak, S. T. (2009). The unreasonable ineffectiveness of Fisherian "tests" in biology, and especially in medicine. *Biological Theory, 4,* 44–53. doi:10.1162/biot.2009.4.1.44