

I

INTRODUCTORY CONCEPTS

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
Prepublication version of Chapter 1.

1

CHANGING TIMES

Somehow the wondrous promise of the earth is that there are things beautiful in it, things wondrous and alluring, and by virtue of your trade you want to understand them.

—Mitchell Feigenbaum, Chaos theorist
(quoted in Gleick, 1988, p. 187)

This book has a home page on the Internet; its address is

<http://www.apa.org/books/resources/kline>

From the home page, readers can find supplemental readings about advanced topics, exercises with answers, resources for instructors and students, and links to related Web sites. Readers can also download data files for some of the research examples discussed later.

In 1996, the Board of Scientific Affairs of the American Psychological Association (APA) convened the Task Force on Statistical Inference (TFSI). The TFSI was asked to respond to the longstanding controversy about statistical significance tests and elucidate alternative methods (Wilkinson & TFSI, 1999). Some hoped that the TFSI would recommend a ban on statistical tests in psychology journals. Such a ban was also discussed in recent special sections or issues of *Psychological Science* (Harris, 1997a), *Research in the Schools* (McLean & Kaufman, 1998), and an edited book by Harlow, Mulaik, and Steiger (1997), the title of which asks the question, “What if there were no significance tests?” Problems with and alternatives to statistical tests were also discussed in a recent special issue of the *Journal of Experimental Education* (B. Thompson, 1993a).

Serious discussion of a ban reflects the culmination of many years of disenchantment with statistical tests. In fact, this controversy has escalated decade by decade and has crossed various disciplines as diverse as psychology and wildlife sciences (e.g., D. Anderson, Burnham, & W. Thompson,

2000). Talk of a ban on statistical tests would probably come as a surprise to a casual observer, to whom it would be plain that results of statistical tests are reported in most empirical articles published in psychology and related disciplines in the last 50 years (Hubbard & Ryan, 2000). The same observer stepping into almost any university classroom for courses in behavioral science statistics at either the undergraduate or graduate level would find that these tests have been the main subject matter for the last 20 years or more (Aiken, West, Sechrest, & Reno, 1990; Frederich, Buday, & Kerr, 2000).

Nevertheless, by the late 1990s the number of voices in the behavioral sciences decrying the limitations of statistical tests started to reach a critical mass. This was apparent in the formation of the TFSI by the APA to look at this controversy. It is also obvious in the fact that about two dozen research journals in psychology, education, counseling, and other areas now require the reporting of effect sizes in submitted manuscripts (Fidler & B. Thompson, 2001).¹ Two of these are flagship journals of associations (American Counseling Association, Council for Exceptional Children) each with more than 50,000 members. One of the most recent APA journals to make this requirement is the *Journal of Educational Psychology*. The requirement to report effect sizes sends a powerful message to potential contributors to these journals that use of statistical tests alone is insufficient, and the number of journals making it is bound to increase. Editorial policies in prominent journals can be an important bellwether for reform (Sedlmeier & Gigerenzer, 1989; Vacha-Haase, 2001). Indeed, Kaufman (1998) noted that the controversy over the use of statistical tests is the major methodological issue of our generation. So perhaps our casual observer might sense that change is coming after all.

GOALS AND PLAN OF THE BOOK

This book aims to help readers stay abreast of ongoing changes in the ways we analyze our data and report our findings in the behavioral sciences. These readers may be educators, applied researchers, reviewers of manuscripts for journals, or undergraduate or graduate students in psychology or related disciplines. It is assumed that many readers (like the author) were trained in traditional methods of data analysis; that is, the use of statistical tests as the primary (if not only) way to evaluate hypotheses. Readers who are currently students are perhaps at some advantage because their views and skills may not yet be so narrow. However, even very experienced researchers who have published many articles may have wondered whether

¹B. Thompson keeps a list of journals that require the reporting of effect sizes at <http://www.coe.tamu.edu/~bthompson/journals.htm>

there are not better ways to evaluate research hypotheses, or whether it is actually necessary to include results of statistical tests in articles (e.g., Kaufman, 1998). Readers already convinced of the limitations of statistical tests should find in this book useful arguments to reinforce their viewpoint. Readers not sharing this view at present will hopefully find some interesting ideas to ponder.

This book does not debate whether we in the psychological community should change the way we use statistical tests. Tryon (2001) and others have noted that more than 50 years of trying to remediate misuses of statistical tests by discussing their technical merits has not been productive. This book assumes instead that developments in the field already point toward a diminishing role for statistical tests. Consequently, the goals of this book are to help readers understand (a) the controversy about and limitations of statistical tests, (b) strengths and weakness of some proposed alternatives to statistical tests, and (c) other methods related to a reduced role for statistical tests, such as meta-analysis. Of primary importance for the second and third points just listed is effect size estimation which, as mentioned, is now required by many journals. The estimation of average effect size across a set of studies in the same general area is also key part of most meta-analyses. Another major focus of this book involves interval estimation, especially the construction of confidence intervals around observed effect sizes.

Part I of the book is concerned with fundamental concepts and summarizes the debate about statistical tests. Chapter 2 reviews principles of sampling and estimation that underlie confidence intervals and statistical tests. Chapter 3 outlines arguments against the continued use of statistical tests as our primary means to evaluate hypotheses. This discussion assumes that although there is nothing inherently wrong with statistical tests, what they actually do makes them unsuitable for perhaps most types of behavioral research. It is also argued that research progress in psychology has been hindered by our preoccupation with statistical tests.

Part II consists of four chapters that emphasize effect size estimation in *comparative studies* that compare at least two different groups or conditions. Chapter 4 reviews the general rationale of effect size estimation and introduces basic parametric effect size indexes for continuous outcome variables, including standardized mean differences and measures of association. Also considered is the comparison of groups at the case level with relatively simple statistics based on proportions of scores above or below certain reference points. The critical problem of evaluating substantive (theoretical, clinical, or practical) significance versus statistical significance is also discussed. Chapter 5 introduces nonparametric effect size indexes for comparing groups on categorical outcomes, such as relapsed versus not relapsed. Chapters 6 and 7 concern effect size estimation in one-way designs with at least

three conditions and factorial designs with two or more independent variables. Many empirical examples are presented in chapters 4 to 7.

Presentations about effect size estimation are often chock full of equations. This is because many effect size indexes can be computed in more than one way, such as from group descriptive statistics or test statistics. To reduce the overall number, only the most essential equations are given in chapters 4 to 7. Some of these equations are for primary researchers who have access to the original data, but others are also handy in secondary analyses based on summary statistics often reported in printed or on-line documents. Information about additional ways to compute effect size indexes is available in technical books about meta-analysis, such as Cooper and Hedges (1994b).

Part III includes two chapters that cover topics related to reform of methods of data analysis in the social sciences. Chapter 8 deals with principles of replication and meta-analysis. The latter has become an increasingly important tool in both the social and health sciences for synthesizing results across a research literature. Its emphasis on effect sizes in primary studies instead of results of statistical tests avoids some of the limitations of the latter. Researchers working in areas with sufficient numbers of studies for meta-analysis thus need to understand its potential strengths and limitations. Chapter 9 surveys two other alternatives to traditional statistical tests that are often overlooked in psychology, statistical resampling—which includes the method of bootstrapping—and Bayesian estimation.

RETROSPECTIVE

Comprehensive historical accounts of the long-standing controversy about statistical significance tests can be found in Gigerenzer (1993), Huberty and Pike (1999), and Oakes (1986):

Hybrid Logic of Statistical Tests (1920-1960)

The basic logical elements of what is today often referred to as null hypothesis significance testing (NHST) were present in scientific papers as early as the 1700s (Stigler, 1986). These elements were not formally organized into a systematic method until the early 1900s, however. The method of NHST in its contemporary form is actually a hybrid of two different schools of thought, one from the 1920s associated with R. Fisher (e.g., 1925), and another from the 1930s called the Neyman-Pearson approach, after J. Neyman and E. S. Pearson (e.g., Neyman & E. S. Pearson, 1933). Other individuals contributed to these schools, such as K. Pearson and A. Wald (Hogben, 1957), but the work of the three principals listed above forms the genesis of NHST.

Briefly, the Neyman-Pearson model is an extension of the Fisher model. Fisher's approach featured only a null hypothesis and subsequent estimation of the conditional probability of the data under it with statistical tests. The probabilities generated by statistical tests are commonly called p values. There was no alternative hypothesis in Fisher's model. The conventional levels of statistical significance used today, .05 and .01, are generally attributed to Fisher, but he apparently did not advocate that these values be applied across all studies (Cowles & Davis, 1982). Anyhow, 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 application of fixed levels of α across all studies characterize the Neyman-Pearson model. The last characteristic just listed is perhaps the main source of the rigid application of the .05 or .01 levels of statistical significance that is today's practice. For the same reason, the Neyman-Pearson model has been described as the *fixed-p approach* (Huberty, 1993). The Neyman-Pearson model also brought with it the conceptual framework of power and associated decision errors, Type I and Type II. A modern power analysis is in spirit and fact based on the Neyman-Pearson model, not the Fisher model.

To say that advocates of the Fisher model and the Neyman-Pearson model exchanged few kind words about each other's approach is an understatement. Their long-running debate was acrimonious. Nevertheless, the integration of the two models by statisticians and authors other than Fisher, Neyman, and E. S. Pearson into what makes up contemporary NHST took place roughly between 1935 and 1950 (Huberty, 1993). Gigerenzer (1993) refers to this integrated model as the *hybrid logic of scientific inference*, and P. Dixon and O'Reilly (1999) call it the "Intro Stats" method because this is the approach outlined in virtually all contemporary textbooks for introductory statistics in the behavioral sciences. Many authors have noted that (a) the hybrid logic that underlies modern NHST would have been rejected by Fisher, Neyman, and E. S. Pearson, although for different reasons; and (b) its composite nature may be a source of confusion about what results from statistical tests really mean.

Institutionalization of the "Intro Stats" Method (1940-1960)

Before 1940, statistical tests were used in relatively few published articles in psychology. Authors of works from this time instead used in non-standard ways a variety of descriptive statistics or rudimentary test statistics. However, from roughly 1940-1960 during what Gigerenzer and D. Murray (1987) called the *inference revolution* in psychology, the "Intro Stats" method

was widely adopted in textbooks, university curricula, and journal editorial practice as basically *the* method to test hypotheses. Gigerenzer (1993) identifies two factors that contributed to this shift. One is the move in psychology away from the study of single cases, such as in operant conditioning studies of individual animals, to the study of groups. This change occurred roughly from 1920-1950. Another is what Gigerenzer (1993) and others call the *probabilistic revolution* in science, which introduced indeterminism as a major theoretical concept in areas such as quantum mechanics and genetics to better understand the subject matter. In psychology, though, it was used to mechanize the inference process through NHST, a critical difference as it turns out.

After the widespread adoption of the “Intro Stats” method, there was a dramatic increase in the reporting of statistical tests in journal articles in psychology and related fields. This trend is obvious in Figure 1.1, reproduced from Hubbard and Ryan. These authors sampled about 8,000 articles published between 1911 and 1998 in randomly selected issues of 12 different APA journals. Summarized in the figure are the percentages of articles in which statistical tests were used in the data analysis. This percentage is about 17% from 1911-1929. It increases to around 50% in 1940, continues to rise to about 85% by 1960, and exceeds 90% since the 1970s. The time period of the most rapid increase in use of NHST, about 1940-1960, corresponds to the inference revolution in psychology.

Some advantages to the institutionalization of NHST were noted by Gigerenzer (1993). The behavioral sciences grew rapidly after 1945, and its administration was made easier by the near-universal use of statistical tests. For example, journal editors could use NHST outcomes to decide which studies to publish or reject, respectively, those with or without statistically significant results, among other considerations. The method of NHST is mechanically applied, and thus seemed to remove subjective judgment from the inference process. That this objectivity is more apparent than real is another matter (more about this point later). The method of NHST also gave behavioral researchers a common language and perhaps identity as members of the same grand research enterprise. It also distinguished them from their counterparts in the natural sciences, who may use statistical tests to detect outliers but not typically to test hypotheses (Gigerenzer, 1993). The elevation of any method to dogma has potential costs, some of which are considered next.

Increasing Criticism of Statistical Tests (1940-present)

There has been controversy about statistical tests for over 70 years, or as long as they been around (Kirk, 1996). Some examples of early critical works include Boring (1919), Berkson (1942), Rozeboom (1960), a book by

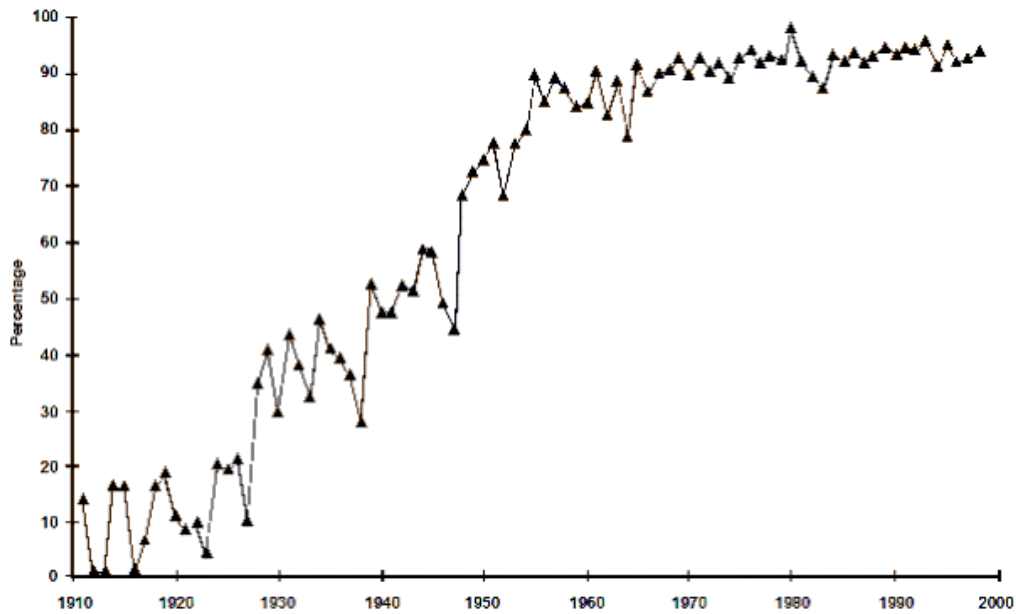


Figure 1.1. Percentage of articles reporting results of statistical tests in 12 journals of the American Psychological Association from 1911 to 1998. From “The Historical Growth of Statistical Significance Testing in Psychology—And Its Future Prospects,” by R. Hubbard and P. A. Ryan, 2000, *Educational and Psychological Measurement*, 60, p. 665. Copyright 2001 by Sage Publications. Reprinted with permission.

Hogben (1957), and edited books by Morrison and Henkel (1970) and Kirk (1972). Overall, the numbers of published works critical of NHST has been increasing exponentially since the 1940s. D. Anderson et al. searched the research literatures in ecology, medicine, business/economics, statistics, and the social sciences for works that questioned the scientific utility of statistical tests. Presented in Figure 1.2 are the total numbers of such works across all surveyed disciplines. Relatively small numbers were published from 1940-1960. However, the numbers of critical articles has increased rapidly since the 1970s, and just about 200 were published in the 1990s in psychology and the other disciplines surveyed by D. Anderson et al.

Summarized next are some of the major arguments against the continued widespread use of statistical tests in the behavioral sciences; they are considered in more detail in later chapters:

1. The p values generated by statistical tests are widely misunderstood. These misunderstandings include the belief that p values measure the likelihood of sampling error, replication, and the truth of the null or alternative hypothesis. These false beliefs may not be solely the fault of users of statistical tests, however. This is because the logical underpinnings of contemporary NHST are not entirely consistent.

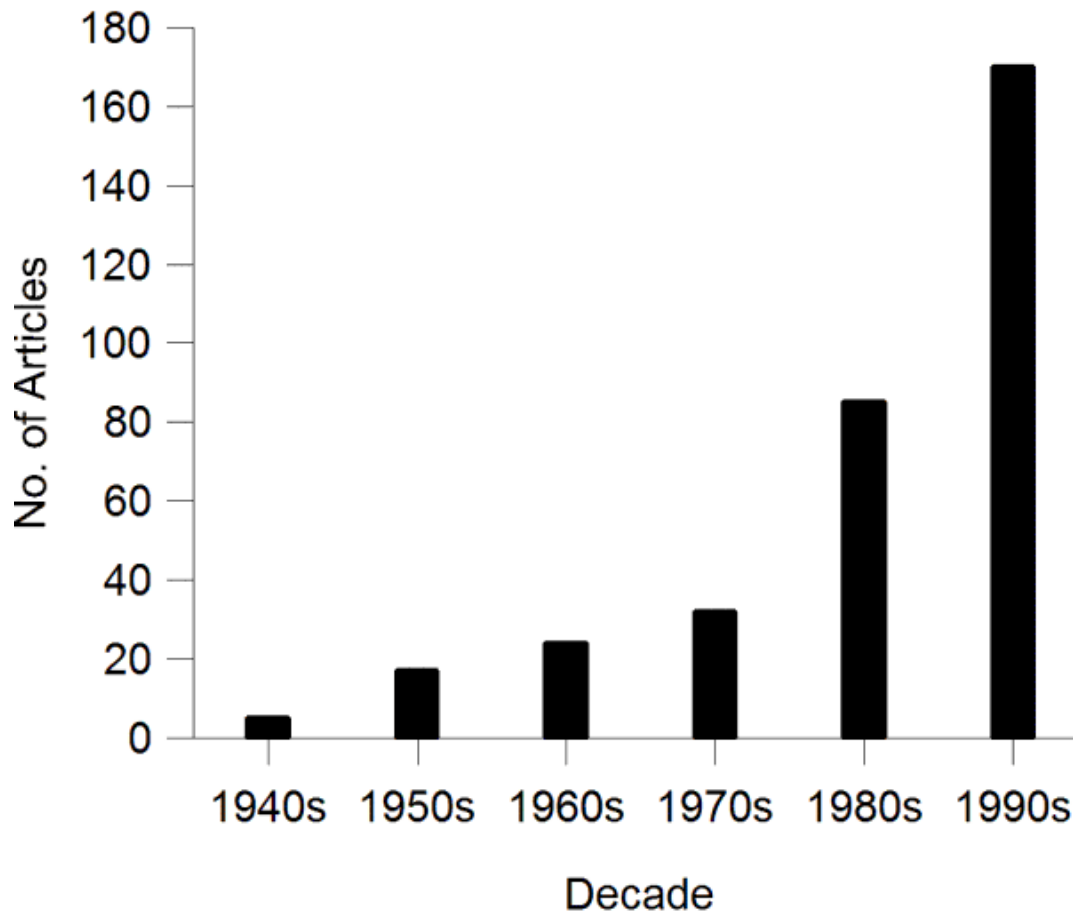


Figure 1.2. Total numbers of articles in ecology, medicine, business–economics, statistics, and the social sciences that question the utility of statistical tests. From “Null Hypothesis Testing: Problems, Prevalence, and an Alternative,” by D. R. Anderson, K. P. Burnham, and W. L. Thompson, 2000, *Journal of Wildlife Management*, 64, p. 913. Copyright 2000 by The Wildlife Society. Adapted with permission.

2. Mistaken beliefs about what statistical tests tell act as a collective form of cognitive distortion that has hindered research progress in the behavioral sciences. This is apparent by the failure to develop a stronger tradition of replication compared to the natural sciences, a lack of relevance of much of our research, and wasted research effort and resources. Critics do not generally suggest that statistical tests are the sole cause, but their excessive use exacerbates these problems.
3. It is likely that p values from statistical tests in many (if not most) behavioral studies are not very meaningful, especially when implausible null hypotheses are specified or distributional assumptions do not hold. If p values are suspect, so are decisions based on them.

4. The information actually provided by a statistical test is very specific, so much so that statistical tests do not typically tell researchers what they really want to know.
5. Statistical significance says nothing directly about the size of an effect or whether it has theoretical or practical import. Effect size magnitude, substantive significance, and whether a result replicates are what we really want (and need) to know.
6. For all the problems of statistical tests, though, there is no magical alternative (J. Cohen, 1994). That is, proposed alternatives, such as effect size estimation and interval estimation in individual studies and the use of meta-analysis to synthesize these results across studies, have their own potential problems. Thus, alternatives to statistical tests should not be uncritically endorsed.

The Failure of Early “Suggestions” to Report Effect Sizes (1994-Present)

One way to compensate for some of the limitations of statistical tests is to report supplemental information, such as a measure of effect size magnitude. Kirk (1996) notes that the idea of effect size estimation is hardly new: It can be found in the work by K. Pearson in the early 1900s, and one of most widely-used effect size statistics in the analysis of variance—estimated eta squared ($\hat{\eta}^2$), also called the correlation ratio (R^2)—is attributed to R. Fisher. The reporting of effect sizes is also generally advocated by contemporary critics of statistical tests.

The fourth edition of the APA’s publication manual (APA, 1994) for the first time encouraged, but did not require authors to report effect sizes along with results of statistical tests. Unfortunately, results of several empirical surveys of post-1994 volumes in over 20 different journals indicate that this encouragement has had relatively little impact (Vacha-Haase, Nilsson, Reetz, Lance, & B. Thompson, 2000). For example, Kirk (1996) examined the 1995 volumes of four different APA journals. The proportions of empirical articles reporting effect sizes across the four journals ranged from 12-77%. The figure of 77% seems impressive, but Kirk (1996) noted that authors in this particular journal were more likely to use regression techniques, which automatically generate correlation effect sizes such as R^2 . Rates of effect sizes in empirical articles in other journals are about 25%, but authors did not always interpret the effect sizes that they reported (e.g., B. Thompson & Snyder, 1998; Vacha-Haase & Ness, 1999). In a broader survey of reporting practices of articles published in the *Journal of Applied Psychology*, Finch, Cumming, and Thomason (2001) found little evidence of reform in the reporting of results of statistical tests over the years 1940-

1999. That there would so little change in reporting practices since the inference revolution in psychology is surprising, especially given that relatively inexpensive personal computers have made available to applied researchers many sophisticated statistical methods. An analogy would be putting the engine from a modern car in the body of a car from the 1940s: Due to limitations of its dated chassis, the car may not actually go any faster.

The Rise of Meta-Analysis and Meta-Analytic Thinking (1976-Present)

Since its introduction in the late 1970s (Glass, 1976; R. Rosenthal, 1976), meta-analysis has become an important tool for research synthesis in several disciplines. Meta-analysis is described in chapter 8, so only its impact on the controversy about statistical tests is outlined here. The typical meta-analysis in the social sciences estimates the central tendency and variability in standardized effect sizes across a set of studies of the same phenomenon, such as the relative effectiveness of treatment over control. This focus on effect size and not statistical significance in individual studies encourages the reader of a meta-analytic article to think outside of the limitations of the latter. There are also now several examples where meta-analytic results show that conclusions based on whether null hypotheses are rejected in individual studies have been wrong (e.g., Rossi, 1997).

The increasing use of meta-analysis has also encouraged *meta-analytic thinking*, to which Cumming and Finch (2001, p. 555) and B. Thompson (2002b) attribute the following characteristics:

1. An accurate appreciation of the results of previous studies is seen as essential.
2. A researcher should view their own study as making a modest contribution to that body of previous research.
3. A researcher should report results so that they can be easily incorporated into a future meta-analysis. This includes the reporting of effect sizes and confidence intervals.
4. Retrospective interpretation of new results, once collected, via direct comparison with prior effect sizes.

Meta-analytic thinking is likely to become ever more predominate. It is also incompatible with using statistical tests as the primary inference tool.

Report of the TFSI and the APA's Fifth Edition of the *Publication Manual* (1999-present)

The report of the TFSI dealt with a wide range of methodological and statistical issues (Wilkinson & TFSI). It also offered suggestions for the then-upcoming fifth edition of the APA's *Publication Manual* (APA, 2001). Some of the TFSI's main recommendations concerning data analyses are summarized next:

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

However, the TFSI decided not to recommend a ban on statistical tests in psychology journals. In its view, such a ban would be a too extreme way to curb abuses of statistical tests (Wilkinson & TFSI, pp. 602-603).

The fifth edition of the APA's *Publication Manual* (APA, 2001) takes a similar stand. That is, it acknowledges the controversy about statistical tests, but it also states that it is not a proper role of the *Publication Manual* to resolve this debate (pp. 21-22). It goes on to recommend the complete reporting of the results of statistical tests, which would include the value of the test statistic, its degrees of freedom, and either the level of alpha (α) applied across all tests, such as $p < .05$, or the exact p value from the output of a computer program, such as $p = .012$. Other recommendations about the statistical analyses (pp. 21-26) include

1. Report adequate descriptive statistics, such as means, variances, and sizes of each group and a pooled within-groups variance-covariance in a comparative study or a correlation matrix in a regression analysis. This information is necessary for later meta-analyses or secondary analyses by other researchers.
2. Effect sizes should "almost always" be reported (p. 25). Several examples of effect size indexes are listed, many of which are dis-

cussed later in this book and by Kirk (1996) and Borenstein (1998), among others. The absence of effect sizes is also cited as an example of a study defect (p. 5). However, authors are still not required to report them.

3. The use of confidence intervals is “strongly recommended,” but not required (p. 22).

Predictably, not everyone is happy with the report of the TFSI or the fifth edition *Publication Manual*. For example, B. Thompson (1999) noted that only encouraging the reporting of effect sizes or confidence intervals presents a self-canceling mixed message. Sohn (2000) lamented the lack of clear guidelines in the report of the TFSI for changing data analysis practices that may improve the relevance of psychology research. Finch et al. welcomed the TFSI report, but contrasted the somewhat ambiguous recommendations about statistical analyses in the APA’s current *Publication Manual* against the relatively simple set of guidelines for manuscripts submitted to biomedical journals by the International Committee of Medical Journal Editors (1997). Kirk (2001) also welcomed the TFSI report, but suggested that the next (6th) edition should contain a much more detailed section on the recommendations of the TFSI. He also noted the relative absence of examples in the current *Publication Manual* of how to appropriately report statistics. See TFSI (2000) for responses to some of these criticisms.

Interviews by Fidler (2002) with some of the principals shed light on why the fifth edition *Publication Manual* does not require reporting of effect sizes. There are some situations where it is difficult or impossible to compute effect sizes. This is especially true for some complex repeated measures designs or multivariate designs. Thus, there was a reluctance to mandate a requirement that in some research contexts could not be met. However, it is possible to calculate effect sizes in perhaps most behavioral studies. It is also true that the effect size estimation void for some kinds of designs is being filled by ongoing research, and is or soon will be filled

PROSPECTIVE

I believe that the events just described indicate a future in which the role of traditional statistical tests in behavioral research will get smaller and smaller. This change will not happen overnight, and statistical tests are not about to disappear in the short term. Indeed, it is expected in the meantime that researchers will still have to report the results of statistical tests in their manuscripts. This is because, to be frank, their manuscripts may be rejected if they contain no statistical tests. However, researchers should give much

less interpretive weight to outcomes of statistical tests than in the past. Specific recommendations follow.

1. Researchers should not view a statistically significant result as particularly informative. For example, they should not conclude that such results are automatically noteworthy or that they are likely to replicate.
2. Researchers should also not discount a statistically nonsignificant result. For example, they should not conclude that failing to reject the null hypothesis means that the population effect size is zero. This false belief may be responsible for the overlooking of possibly beneficial effects in health research (R. Rosenthal, 1994).
3. Effect sizes should always be reported, and confidence intervals should be constructed about them whenever possible. However, real reform does *not* involve computing effect sizes only for statistically significant results. This would amount to “business as usual” where the statistical test is still at center stage (Sohn, 2000). Real reform also means that effect sizes are interpreted and evaluated for their substantive significance, not just reported.

Other recommendations are given later in the book, many of which do not involve the use of statistical tests at all. This is consistent with a vision of the future in behavioral research that I and others advocate (e.g., B. Thompson, 2002b): Most studies in the future will *not* use statistical tests as the primary decision criterion, and those that do will concern only very specific problems for which variations of NHST may be appropriate, such as equivalence testing or inferential confidence intervals (chap. 3, this volume). It is also envisioned that the social sciences will become more like the natural sciences. That is, we will report the directions and magnitudes of our effects, determine whether they replicate, and evaluate them for their theoretical, clinical, or practical significance, not just their statistical significance (Kirk, 1996).

VOICES FROM THE FUTURE

As mentioned, it may be easier for younger researchers who are not as set in their ways to respond to the call for reform in our methods of data analysis and inference. However, 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 in this book. Students are also promising prospects for reform because they are, in my experience and

that of others (Hyde, 2001), eager to learn about limitations of traditional statistical tests. They can also understand, with proper instruction, ideas such as effect size and confidence intervals, even in introductory statistics courses. In fact, it is this author's experience that it is easier to teach undergraduates these concepts than the convoluted logic of NHST. Other basics of reform are even easier to teach, such as the need for replication.

Presented next are some example responses to the question, "What is the most important thing you learned in this class?" on a recent final examination in introductory psychology statistics I gave. These words from future behavioral researchers also reiterate some of the major goals of this book. May we all be so wise.

- Null hypothesis significance testing is not the only or necessarily the best way to test hypotheses.
- Just because a finding is statistically significant doesn't mean that it's important or reliable.
- If you increase the sample size enough, any result will be statistically significant. This is scary.
- To be skeptical of research papers that put a large emphasis on statistical significance.
- Statistical significance does not mean practical significance. Effect size, power, means, and standard deviations should be included in research reports. There needs to be in the social sciences a better understanding of statistical tests so that we can make better, more informed choices.

It is said that if we do not make the future, others will do it for us. It is time to start building our future by moving beyond statistical tests and leaving behind other, old ways of doing things. We need to explore other possibilities for testing our hypotheses, ones that may lead to a more productive future for research in psychology and related disciplines. I hope this book will challenge, encourage, and support readers to think and act along these lines.

CONCLUSION

The controversy about statistical tests in psychology was briefly described as were the events leading up to it. This history gives the context for the 1999 report of the Task Force on Statistical Inference and the 2001 fifth edition of the *Publication Manual* of the American Psychological Association. It also indicates that the continued use of statistical tests as the sole

way to test hypotheses and make inferential decisions in the behavioral sciences is unlikely. The points raised above set the stage for reviewing in the next chapter some fundamental statistical concepts. These concepts are also crucial for understanding the limitations of statistical tests and characteristics of some proposed alternatives, such as interval estimation and effect size estimation.

RECOMMENDED READINGS

- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren and C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.
- Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, 61, 213-218.
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.