8-1-2006

Why We Don't Really Know What "Statistical Significance" Means: A Major Educational Failure

Raymond Hubbard
Drake University

J. Scott Armstrong
University of Pennsylvania, armstrong@wharton.upenn.edu

Postprint version. Published in *Journal of Marketing Education*, Volume 28, Issue 2, August 2006, pages 114-120.
Publisher URL: http://dx.doi.org/10.1177/0273475306288399

Why We Don't Really Know What "Statistical Significance" Means:

A Major Educational Failure*

Raymond Hubbard
College of Business and Public Administration
Drake University
Des Moines, IA 50311
Phone: (515) 271-2344
E-mail: Raymond.Hubbard@drake.edu

J. Scott Armstrong
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104
Phone: (215) 898-5087
E-mail: Armstrong@wharton.upenn.edu

July 13, 2005

* The authors have benefited from discussions on this topic with Stuart Allen, M.J. Bayarri, James Berger, Eric Bradlow, Steven Goodman, and Rahul Parsa. Any remaining errors or shortcomings are our responsibility.

Why We Don't Really Know What "Statistical Significance" Means: A Major Educational Failure

ABSTRACT

The Neyman–Pearson theory of hypothesis testing, with the Type I error rate, α, as the significance level, is widely regarded as statistical testing orthodoxy. Fisher's model of significance testing, where the evidential *p* value denotes the level of significance, nevertheless dominates statistical testing practice. This paradox has occurred because these two incompatible theories of classical statistical testing have been anonymously mixed together, creating the false impression of a single, coherent model of statistical inference. We show that this hybrid approach to testing, with its misleading $p < \alpha$ statistical significance criterion, is common in marketing research textbooks, as well as in a large random sample of papers from twelve marketing journals. That is, researchers attempt the impossible by simultaneously interpreting the *p* value as a Type I error rate and as a measure of evidence against the null hypothesis. The upshot is that many investigators do not know what our most cherished, and ubiquitous, research desideratum—"statistical significance"—really means. This, in turn, signals an educational failure of the first order. We suggest that tests of statistical significance, whether *p*'s or α's, be downplayed in statistics and marketing research courses. Classroom instruction should focus instead on teaching students to emphasize the use of confidence intervals around point estimates in individual studies, and the criterion of overlapping confidence intervals when one has estimates from similar studies.

Keywords: α levels; *p* values; $p < \alpha$ criterion; Fisher; Neyman–Pearson; (overlapping) confidence intervals

For many scholars the significance test is the glue that binds together the entire research process. The test of statistical significance largely dictates how we formulate hypotheses; design questionnaires; organize experiments; and analyze, report, and summarize results. It is viewed not only as our chief vehicle for making *statistical* inferences, but for drawing *scientific* inferences, too. That is, the test of significance is regarded as playing an important epistemological role. As Lindsay (1995) notes with dismay, computing such a test has come to be equated with scientific rigor, and is considered the touchstone for establishing knowledge. Gigerenzer et al. (1989, p. 108) share Lindsay's sentiments: "What is most remarkable is the confidence within each social-science discipline that the standards of scientific demonstration have now been objectively and universally defined." This test, in short, is no mere statistical "technique," but instead is seen to lie at the heart of the way in which we conceptualize and conduct research. Or as Cicchetti (1998, p. 293) tersely put it, the focus on significance testing often is considered "…as an end, in and of itself."

To see the validity of the above account it is only necessary to look to our own experiences as graduate students and educators. We were (almost) all taught that the significance testing paradigm is *the* way to *do* sound research. Indeed, most of us trained in this paradigm have no idea of how research was carried out prior to its rise to dominance, and would be hard-pressed to visualize what future research would look like if the paradigm collapsed.

Others (e.g., Sawyer and Peter 1983) have noted that marketing researchers misinterpret the outcomes of significance tests. For example, such tests are erroneously believed to indicate the probability that (1) the results occurred because of chance, (2) the results will

replicate, (3) the alternative hypothesis is true, (4) the results will generalize, and (5) the results are substantively significant.

Our paper is not concerned with these misinterpretations, serious as they are. Rather, we maintain that misconceptions among researchers regarding statistical significance tests are far deeper than earlier works suggest. Specifically, we argue that *researchers are confused over the very meaning of "statistical significance" itself*. This inability to comprehend the exact nature of the criterion we so earnestly, and routinely, seek above all others to adjudicate knowledge claims underscores that something is seriously wrong in statistics and marketing research education. The present paper explains, and demonstrates the consequences of, a major educational breakdown—the failure to correctly teach generations of students *precisely* what "statistical significance" means. In doing so, we show that significance testing is a mechanistic ritual so thoroughly misunderstood as to be largely bereft of meaning. And worse, this emphasis on significance testing in the classroom and textbooks has diverted attention from superior data analysis strategies designed to promote cumulative knowledge growth. The end result is that our literature is comprised mainly of uncorroborated, one-shot studies whose value is questionable for academics and practitioners alike.

The paper is organized as follows. First, we describe how the wholesale confusion over the meaning of statistical significance has been caused by mixing together in statistics and methodology textbooks two different classical statistical testing models—Fisher's and Neyman–Pearson's. This necessitates a brief outline of some key differences between them, which, in turn, leads to a discussion of the problematical $p < \alpha$ criterion as a measure of statistical significance. Second, we indicate how the authors of marketing research textbooks often mistakenly define and interpret p values and α levels, treat them interchangeably,

invoke the $p < \alpha$ yardstick, and thereby obscure the meaning of statistical significance. Third, we show via a random sample of articles from twelve marketing journals how these mistakes carry over into the empirical literature. Fourth, we offer some advice regarding data analysis. This includes a short section for those intent on using significance tests. Better yet, however, we suggest replacing such tests with estimates of sample statistics, effect sizes, and their confidence intervals in single studies. We also recommend the criterion of overlapping confidence intervals for determining the equivalence (or otherwise) of estimates across similar studies.

WHY THE CONFUSION OVER THE MEANING OF "STATISTICAL SIGNIFICANCE"?

Some authors (e.g., Gigerenzer, Krauss, and Vitouch 2004; Goodman 1993; Hubbard and Bayari 2003; Royall 1997) allege that the principal reason why researchers cannot accurately define what is meant by "statistical significance" is because many statistics and methodology textbooks are similarly confused over the exact meaning of this concept. This is because these texts inadvertently mix together *two incompatible measures of "statistical significance" into an anonymous hybrid*, thereby creating the illusion of a single, harmonious theory of statistical inference. One is Fisher's evidential *p* value and the other is the Type I error rate, α, of the Neyman–Pearson (N–P) school. The distinction between *evidence (p's)* and *errors (α's)* is not a matter of splitting hairs. As Hubbard and Bayarri (2003) noted it reflects the pronounced differences between Fisher's views on significance testing and *inductive inference*, and N–P ideas on hypothesis testing and *inductive behavior*. But because statistics and methodology textbooks tend to combine elements from both schools of thought, something that neither Fisher nor N–P would have agreed to, there is confusion over what "statistically significant at the .05

[or other] level" really means. We briefly discuss some key differences between the Fisherian and N–P camps below.

Fisher's Significance Testing and Neyman–Pearson's Hypothesis Testing Paradigms[1]

The *p* value from Fisher's *significance testing* procedure measures the probability of encountering an outcome (x) of this magnitude (or larger) conditional on a true null hypothesis of no effect or relationship, or Pr ($x \mid H_0$). Thus, a *p* value is a measure of inductive evidence against H_0, and the smaller the value, the greater the evidence. Fisher saw statistics as playing a vital part in inductive inference, drawing conclusions from the particular to the general, from samples to populations. He held that knowledge is created via inductive inference, and for him the evidential *p* value had an important role in this process.

The N–P theory of *hypothesis testing*, which began assuming the mantle of statistical orthodoxy over Fisher's significance testing paradigm after World War II (Royall 1997), is quite different from the latter. It is not a theory of statistical *inference* at all. N–P summarily dismissed the concept of inductive inference, and focused instead on statistical testing as a mechanism for making decisions and guiding *behavior*. Whereas Fisher specified only the null hypothesis (H_0), N–P introduced two hypotheses, the null and the alternative (H_A), and their approach invites a decision between two distinct courses of action, accepting H_0 or rejecting it in favor of H_A. Mistakes occur when choosing between accepting H_0 or H_A. According to N–P, the significance level, or Type I error, α, is the false rejection of H_0, while a Type II error, β, is the false acceptance of H_0. N–P statistical testing is aimed at error minimization, and is not concerned with gathering evidence. Furthermore, this error minimization is of a *long-run* variety; unlike Fisher's approach, N–P theory does not apply to an *individual study*. Consider, finally, that Fisher's evidential *p* value is a data-dependent *random variable*. This is in contrast to N-P's α,

which must be *fixed* in advance of gathering the data so as to constrain the probability of a Type I error to some agreed-upon value.

The Hybrid Testing Paradigm

Fisher (1955, p. 74) complained, justifiably, that his significance test had become "assimilated" into the N–P hypothesis testing framework. Because of this assimilation, most empirical work in marketing and the social sciences, echoing what is presented in the textbooks, is carried out roughly as follows: The investigator specifies the null (H_0) and alternative (H_A) hypotheses, the Type I error rate/significance level, α, and (supposedly) calculates the power of the test (e.g., *z*). These steps are congruent with N–P orthodoxy. Next, the test statistic is computed, and in an effort to have one's cake and eat it too, a *p* value is determined. Statistical significance is then established by using the problematical $p < \alpha$ criterion; if $p < \alpha$, a result is deemed statistically significant, if $p > \alpha$, it is not.

The end result of this assimilation of Fisher's and N–P's methods is that, despite being completely different entities with completely different interpretations, the *p* value is now associated in researchers' minds with the Type I error rate, α. And because both concepts are tail area probabilities, the *p* value is erroneously interpreted as a frequency-based "observed" Type I error rate, and at the same time as an incorrect (i.e., $p < \alpha$) measure of evidence against H_0 (Goodman 1993; Hubbard and Bayarri 2003).

There are problems with the interpretation of the $p < \alpha$ criterion. For example, when formulated as "reject H_0 when $p < \alpha$, accept it otherwise," only the N–P claim of $100\alpha\%$ false rejections of the null with ongoing sampling is valid. That is, the specific value of *p* itself is irrelevant and should not be reported. In the N–P decision model the researcher can only say whether or not a result fell in the rejection region, but not *where* it fell, as might be shown by a

p value. So, if α is fixed at the .05 level before the study is conducted, and the researcher gets, *after the fact*, a *p* value of, say, .0023, this exact value cannot be reported in an N–P hypothesis test. As Goodman (1993) points out, this is because α is the probability of a *set* of potential outcomes that may fall anywhere in the tail area of the distribution under the null hypothesis, and we cannot know ahead of time which of these particular outcomes will occur. This is not the same as the tail area for the *p* value, which is known only after the outcome is observed.

For the same reasons it is not admissible to report what Goodman (1993, p. 489) calls "roving alphas," i.e., *p* values that take on a limited number of categories of Type I error rates, e.g., $p < .05$, $p < .01$; $p < .001$, etc. As discussed, a Type I error rate, α, must be fixed before the data are collected, and any attempt to later reinterpret values like $p < .05$, $p < .01$, etc. as *variable* Type I error rates applicable to different parts of any given study is not allowed. Further complicating matters, these variable Type I error "*p*" values are also interpreted in an evidential fashion when $p < \alpha$, e.g., where $p < .05$ is called "significant," $p < .01$ is "highly significant," $p < .001$ is "extremely significant," and so on. Because of the confusion created among researchers by the $p < \alpha$ rule of thumb, Hubbard and Bayarri (2003) called for its abolition in textbooks and journal articles.

CONFUSION OVER "STATISTICAL SIGNIFICANCE" IN MARKETING RESEARCH TEXTBOOKS

We examined a convenience sample of fourteen marketing research textbooks to determine whether their methodological leanings were N–P, Fisherian, or some combination thereof. In no case did these authors explicitly acknowledge the intellectual heritage underlying their discussions of statistical testing. The anonymous treatment of such testing was the norm.

Therefore, in Table 1 we assigned these texts to one of five categories on an N–P-to-Fisherian continuum of statistical testing.

Insert Table 1 about here

The text by Kinnear and Taylor (1991) presents a strictly N–P approach. They discuss Type I and II errors, the power of a test, and refer to α as the significance level. Moreover, *p* values are absent in their account. Nevertheless, they cross over to Fisher's camp when they speak of "evidence," something that is denied in N–P theory.

Hair, Bush, and Ortinau (2003), Tull and Hawkins (1993), and Zikmund (1997) also employ an N–P approach, with discussions covering Type I and II errors, the power of a test, and α as the significance level. But in all three cases the authors unwittingly mix N–P and Fisherian methods when *p* values, without explaining their appearance and meaning, infiltrate the empirical examples.

Six of the fourteen texts—Aaker, Kumar, and Day (2001), Churchill and Iacobucci (2002), Cooper and Schindler (2006), Malhotra (2004), McDaniel and Gates (2002), and Parasuraman, Grewal, and Krishnan (2004)—present N–P testing methods. But they also blend ideas from both camps when discussing, to varying extents, *p* values. Malhotra (2004), and Cooper and Schindler (2006), hedge their bets by offering the researcher the choice of either of the (unstated) Fisherian or N–P options. And they, too, recommend the $p < \alpha$ rule of thumb.

McDaniel and Gates (2002, p. 537) subscribe to the $p < \alpha$ criterion in statistical testing, and also incorrectly define the *p* value as "The exact probability of getting a computed test statistic

that was largely due to chance." Parasuraman et al. (2004), on the other hand, in common with Cooper and Schindler (2006), misinterpret the *p* value as a Type I error rate.

Textbooks by Burns and Bush (2000), Crask, Fox, and Stout (1995), and Lehmann, Gupta, and Steckel (1998) are basically non-committal in terms of their Fisherian versus N–P allegiances. Both Burns and Bush (2000) and Crask et al. (1995), for example, contain no discussions of α as the significance level, Type I and II errors, or the power of a test. Burns and Bush (2000) nevertheless champion the misleading $p < \alpha$ statistical testing criterion. And Crask et al. (1995) reveal something of a preference for the N–P camp when discussing statistical testing at the 5% and 10% "risk levels." Lehmann et al. (1998) bow in the direction of N–P. For example, they briefly address Type I and II errors, but do not speak to the power of a test or refer to α (or *p* values) as the significance level. They simply talk of results being "statistically significant" at the .05 or .01 levels.

Finally, Sudman and Blair's (1998) text is mostly Fisherian in nature. There is a complete absence of N–P terminology. Like Lehmann et al. (1998), they are neutral in their discussion of the .05 and .01 "significance levels," invoking neither *p*'s nor α's. Sudman and Blair (1998) do, however, use (unexplained) *p* values in their numerical examples.

It is clear from the above that marketing research textbooks typically contain an anonymous mixture of competing Fisherian and N–P ideas about statistical testing, as well as some of the problems that inevitably accompany this. Most of them emphasize formal N–P theory, but this unintentionally erodes when *p* values and α levels are treated interchangeably without offering any explanation as to their very different origins and interpretations. As shown in the following section, this same hybrid of Fisherian and N–P testing is seen in leading marketing journals. Only this time, it is the former's influence that is dominant.

CONFUSION OVER "STATISTICAL SIGNIFICANCE" IN MARKETING JOURNALS

We investigated how the results of statistical tests are reported in marketing journals. More specifically, *two* randomly selected issues of each of twelve marketing journals—the *European Journal of Marketing (EJM, 1971), International Journal of Market Research (IJMR, 1966), Journal of the Academy of Marketing Science (JAMS, 1973), Journal of Advertising Research (JAR, 1960), Journal of Consumer Research (JCR, 1974), Journal of Macromarketing (JMM, 1981), Journal of Marketing (JM, 1936), Journal of Marketing Education (JME, 1979), Journal of Marketing Research (JMR, 1964), Journal of Retailing (JR, 1960), Marketing Letters (ML, 1990),* and *Marketing Science (MS, 1982)*—were analyzed for every year indicated in the parentheses through 2002 in order to determine the number of empirical articles and notes published therein.[2] This procedure yielded a sample of 4,344 empirical papers. The latter were then inspected to see whether statistical tests had been employed in the data analysis. It was discovered that 3,021 of the 4,344 empirical works, or 69.5%, did so. Moreover, the incidence of empirical papers using statistical significance testing has grown steadily over time. Thus, for example, 37.4% of empirical papers used significance tests during 1960–1969, a number increasing monotonically for 1970–1979 (65.5%), 1980–1989 (76.6%), 1990–1999 (80.4%), and 2000–2002 (85.3%).

Although the evidential *p* value from a significance test violates the orthodox N–P model, the last line of Table 2 shows that *p* values are commonplace in marketing's empirical literature. Conversely, α levels are in short supply.

Of the 3,021 papers using statistical tests, fully 1,660, or 54.9%, employed "roving alphas," i.e., a discrete, graduated number of *p* values interpreted variously as Type I error rates and/or measures of evidence against H_0, usually $p < .05$, $p < .01$, $p < .001$, etc. In other words, these

p values are sometimes viewed as an "observed" Type I error rate meaning that they are not pre-assigned, or fixed, error levels as would be dictated by N–P theory. Instead, these "error rates" are determined solely by the data. Further clouding the issue, these same *p* values will be interpreted simultaneously in a quasi-evidential manner as a basis for rejecting H_0 if $p < \alpha$. In short, these "roving alphas" can assume a number of incorrect and contradictory interpretations. We also plead guilty to the charge of having abused roving alphas in this way.

A further 254 (8.4%) chose to report "exact" *p* values, while an additional 367 (12.1%) opted to present various combinations of exact *p*'s with either "roving alphas" or fixed *p* values. Conservatively, therefore, 2,281, or 75.5%, of empirical articles in a sample of marketing journals report the results of statistical tests in a manner that is incompatible with N–P doctrine. Another 79 (2.6%) studies were not sufficiently clear about the disposition of a finding (beyond statements such as "this result was statistically significant at conventional levels") in their accounts.

This leaves 661 (21.9%) studies as eligible for the reporting of "fixed" level α values in the fashion intended by N–P. Unfortunately, 246 of these 661 studies reported "fixed *p*" rather than fixed α levels. After subtracting this group, only 415 (13.7%) studies remain eligible. Of these 415, some 346 simply refer to their published results as being "significant" at the .05, .01 levels, etc. No information about *p* values or α levels is provided. Finally, only 69 of 3,021 empirical papers using statistical tests, or 2.3%, explicitly used α levels.

Insert Table 2 about here

This meshing of *p*'s and α's is not only wrong from a conceptual and methodological perspective, but also has a pronounced impact on the results of statistical tests. While α can indeed be fixed at some prespecified (e.g., .05) level, this same constraint does not apply to *p* values. This can be seen by accessing an applet at ***www.stat.duke.edu/~berger*** which simulates the frequentist performance of *p* values. More specifically, the applet simulates via ongoing normal testing the proportion of times that the null hypothesis is true for a given *p* value. Thus, if the researcher wishes to see the proportion of times H_0 is true for $p = .05$, a small range such as .049 to .050 must be chosen. The simulation then carries out a long series of tests, and calculates how often the null is true and false whenever the *p* value is in the .049 to .050 range. The researcher must also state the proportion of null hypotheses chosen to be true in the sequence of simulated tests. For instance, suppose we conduct a long series of tests examining the responsiveness of sales revenues to varying advertising outlays. Suppose, further, we specify that H_0 is true for one-half of these advertising outlays; then of all the tests yielding a *p* value of around .05, the final percentage of true nulls is *at least* 22% and as high as 50%. The implications for applied research are chilling: 22% to 50% of the times we see a *p* value of .05 reported in the literature, it is in fact coming from the null hypothesis of no effect.

We see only marginal value in significance testing, no matter the variety. However, for those who insist on using statistical testing we offer the following advice. At a minimum, researchers should make a conscious effort to determine whether their concerns are with controlling errors or collecting evidence. If the former, as in quality control experiments, then the N–P approach is best for guiding behavior. But this should be accompanied by a serious attempt to calculate the costs associated with committing Type I and II errors, rather than the habitual adoption of $\alpha = .05$ and the absence of a power analysis to detect effect sizes in the population. Moreover, if

this option is chosen, it is imperative that the α level be fixed before the study begins, and that the reporting of nonsensical "roving alphas" ceases. Finally, under no circumstances invoke the $p < \alpha$ criterion of statistical significance.

If the goal of the research is evidential in nature (which will be most of the time), then the use of Fisher's *p* value is appropriate. Whenever possible, report exact *p* values to once again avoid the "roving alphas" dilemma. Further, do not employ the $p < \alpha$ significance criterion; a *p* value is not an error probability. But a better strategy for data analysis is to focus on estimation, not testing. This is discussed below.

(OVERLAPPING) CONFIDENCE INTERVALS—AN ALTERNATIVE TO "STATISTICAL SIGNIFICANCE"

Rather than relying on significance testing, researchers should instead report the results of sample statistics, effect sizes, and their confidence intervals (CIs). CIs are far more informative than a yes-no significance test. First, they emphasize the importance of estimation over testing. Scientific progress almost always depends on arriving at credible estimates of the magnitude of effect sizes; and a CI yields a range of estimates deemed likely for the population. Second, the width of the CI provides a measure of the reliability or precision of the estimate. Third, CIs make it far easier to determine whether a finding has any substantive, as opposed to statistical, significance. This is because they are couched in the same metric as the estimate itself, and thus the plausibility of the values in the interval are easy to interpret within the context of the problem. Fourth, unlike statistical significance tests which are vulnerable to Type I error proliferation, CIs hold the true error rate (.05, .01, etc.) to the chosen level (Schmidt 1996). Fifth, if need be, a CI can be used as a significance test. For example, a 95% CI that does not include the null value (usually zero) is equivalent to rejecting the hypothesis at the .05 level.

Finally, and of critical importance, the use of CIs promotes cumulative knowledge development by obligating researchers to think meta-analytically about estimation, replication, and comparing intervals across studies (Thompson 2002). It allows for the possibility of unifying a seemingly fragmented literature. Unfortunately, the preoccupation with obtaining statistically significant results frustrates cumulative knowledge development. This is because, Ottenbacher (1996) points out, a "successful" replication is typically defined as a null hypothesis that was rejected in the original investigation is again rejected (in the same direction) in a follow-up study. But this is too stringent a benchmark. Rather than using statistical significance to denote a successful replication, we advocate the criterion of *overlapping CIs* around point estimates across similar studies. Overlapping CIs indicate credible estimates of the same population parameter.

To illustrate the superiority of this strategy for developing cumulative knowledge, we selected real correlational data present in Schmidt (1996) on personnel selection. But we renamed the variables to suit an educational scenario. Suppose there are four articles, each in this case with sample size $n = 68$, dealing with the correlation between the number of hours spent studying and GPAs. The correlation coefficients, r's, and 95% CIs for these four articles are as follows: (1) $r = 0.39$ (CI = 0.19 to 0.59); (2) $r = 0.29$ (CI = 0.07 to 0.51); (3) $r = 0.14$ (CI = –0.09 to 0.37), and (4) $r = 0.11$ (CI = –0.13 to 0.35). The first two studies are significant at $p < .05$, while the last two are not.

When using significance testing and "nose counting" as evaluative criteria, a traditional review of this literature would conclude that it is made up of *contradictory* results; half the investigations support the hypothesis of a relationship between the number of hours studying and GPAs, and half do not. But this conclusion would be incorrect. In fact, all four articles

corroborate one another because they all show a positive relationship between study-hours and GPAs, even though two of them are not significant. This is revealed by the fact that their CIs all overlap, even for the highest and lowest correlations. This literature is consistent, not contradictory. Use of overlapping CIs fosters cumulative knowledge growth, while the emphasis on significance testing thwarts it.

But to be able to perform this kind of analysis requires that the articles are indeed dealing with "similar" studies. And this is why Hubbard and Armstrong (1994) stress the crucial need for *systematic* replication with extension research programs aimed at discovering empirical generalizations, or the missing bedrock of marketing knowledge that Leone and Schultz (1980) called for.

Another worrisome problem, given the publication bias against insignificant results (Hubbard and Armstrong 1992), is that reported estimates of the effect size in the population will be inflated. For example, if the two "negative" results papers above never see print, then the average effect size will be given as $r = 0.34$, when it is only $r = 0.23$.

CONCLUSIONS

The mixing of measures of evidence (*p*'s) with measures of error (α's) is commonplace in classrooms, textbooks, and scholarly journals. The upshot is that many researchers have an unsure grasp of what "statistical significance" really means. Is it captured by *p* values, α levels, the $p < \alpha$ criterion, or any and all of the above? Such confusion makes ritualistic significance testing largely vacuous. Gigerenzer et al. (2004, p. 395) said as much with respect to psychology research: "The collective illusions about the meaning of a significant result are embarrassing to our profession." Yet a similar environment prevails in marketing.

While this situation is regrettable, it is also understandable. It was caused by the anonymous blending of two schools of classical statistical testing, each with incompatible measures of statistical significance, into what textbooks continue to misrepresent as a single, uncontroversial theory of statistical inference.

The solution to this problem necessitates changes in graduate classroom instruction, and the textbooks that sustain it. With this in mind, we offer two recommendations. First, if statistical significance testing is to be featured in the curriculum, the differences between the Fisher and N-P paradigms require explanation. Students need to be better informed about exactly what is meant by "statistical significance." All too often we rely on computer printouts reporting a thicket of significance levels without fully understanding the reasoning behind them. Second, and better yet, we should be taught to provide confidence intervals around sample statistics and effect sizes, and examine whether the relevant CIs overlap across similar studies in systematic replication with extension research programs. This would facilitate meta-analyses aimed at building a cumulative knowledge base in marketing. At present, our empirical literature is made up of mostly unverified, one-shot studies, fueled by an emphasis on significance testing. It is past time for a serious overhaul in statistics and marketing research education.

REFERENCES

Aaker, David A., Vinay Kumar, and George S. Day. 2001. *Marketing research*. New York: Wiley. Seventh Edition.

Burns, Alvin C. and Ronald F. Bush. 2000. *Marketing research*. Upper Saddle River, NJ: Prentice Hall. Third Edition.

Churchill, Gilbert A., and Dawn Iacobucci. 2002. *Marketing research. Methodological Foundations*. New York: Harcourt. Eighth Edition.

Cicchetti, Domenic V. 1998. Role of null hypothesis significance testing (nhst) in the design of neuropsychologic research. *Journal of Clinical and Experimental Neuropsychology* 20: 293–95.

Cooper, Donald R., and Pamela S. Schindler. 2006. *Marketing Research.* New York: McGraw–Hill.

Crask, M., R.J. Fox, and R.G. Stout. 1995. *Marketing research: Principles and application.* Englewood Cliffs, NJ: Prentice Hall.

Fisher, Ronald A. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society, B,* 17: 69–78.

Gigerenzer, Gerd, Stefan Krauss, and Oliver Vitouch. 2004. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In *The Sage handbook of quantitative methodology for the social sciences*, edited by D. Kaplan, 391–408. Thousand Oaks, CA: Sage Publications.

Gigerenzer, G., Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Kruger. 1989. *The empire of chance.* Cambridge: Cambridge University Press.

Goodman, Steven N. 1993. *p* values, hypothesis tests, and likelihood. Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* 137 (March): 485–96.

Hair, J.F., R.P. Bush, and D.J. Ortinau. 2003. *Marketing research within a changing information environment.* New York: McGraw–Hill. Second Edition.

Hubbard, Raymond, and J. Scott Armstrong. 1992. Are null results becoming an endangered species in marketing? *Marketing Letters* 3 (April): 127–136.

Hubbard, Raymond, and J. Scott Armstrong. 1994. Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing* 11 (June): 233–48.

Hubbard, Raymond, and M.J. Bayarri. 2003. Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing (with comments). *The American Statistician* 57 (August): 171–82.

Kinnear, Thomas C., and James R. Taylor. 1991. *Marketing research: An applied approach.* New York: McGraw–Hill. Fourth Edition.

Lehmann, Donald R., Senil Gupta, and Joe H. Steckel. 1998. *Marketing research.* New York: Addison–Wesley.

Leone, Robert P., and Randall L. Schultz. 1980. A study of marketing generalizations. *Journal of Marketing* 44 (Winter): 10–18.

Lindsay, R. Murray. 1995. Reconsidering the status of tests of significance: An alternative criterion of adequacy. *Accounting, Organizations and Society* 20: 35–53.

Malhotra, Naresh K. 2004. *Marketing research: An applied orientation.* Upper Saddle River, NJ: Prentice Hall. Fourth Edition.

McDaniel, C., and R. Gates. 2002. *Marketing research: The impact of the internet.* Cincinnati, OH: South–Western. Fifth Edition.

Ottenbacher, Kenneth J. 1996. The power of replications and replications of power. *The American Statistician* 50: 271–275.

Parasuraman, A., Dhruv Grewal, and R. Krishnan. 2004. *Marketing research.* Boston: Houghton Mifflin.

Royall, Richard M. 1997. *Statistical evidence: A likelihood paradigm.* New York: Chapman and Hall.

Sawyer, Alan G., and J. Paul Peter. 1983. The significance of statistical significance tests in marketing research. *Journal of Marketing Research* 20 (May): 122–33.

Schmidt, Frank L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1: 115–29.

Sudman, Seymour, and Edward Blair. 1998. *Marketing research: A problem-solving approach.* New York: McGraw–Hill.

Thompson, Bruce. 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher* 31 (April): 25–32.

Tull, Donald S., and Del I. Hawkins. 1993. *Marketing research: Measurement & method.* New York: Macmillan. Sixth Edition.

Zikmund, William G. 1997. *Exploring marketing research.* New York: Dryden Press. Sixth Edition.

TABLE 1

STATISTICAL TESTING IN MARKETING RESEARCH TEXTBOOKS: UNSTATED METHODOLOGICAL ORIENTATIONS

Strictly Neyman–Pearson Approach (No Discussion of p Values)	*Neyman–Pearson Approach (No Discussion of p Values—But They Appear in Examples)*	*Neyman–Pearson Approach (But Also Discuss p Values)*	*Nominally Neyman–Pearson Approach*	*Basically Fisherian p Value Approach*
These texts discuss α as the significance level, Type I and II errors, the power of a test, etc. Example: *Kinnear/Taylor (1991)* But they switch to Fisher when talking of the "evidence" in a study. Neyman–Pearson theory denies evidential interpretations; it prescribes only behaviors.	These texts discuss α as the significance level, Type I and II errors, the power of a test, etc. In addition, they introduce *p* values/significance probabilities in numerical examples, but without explaining them. Examples: *Hair/Bush/Ortinau (2003)* *Tull/Hawkins (1993)* *Zikmund (1997)*	These texts discuss α as the significance level, Type I and II errors, the power of a test, etc. In addition, some texts attempt an explanation of *p* values. Examples: *Aaker/Kumar/Day (2001)* Only text that tries to explain differences between *p*'s and α's. Does not acknowledge the incompatibility of *p*'s and α's. Essentially invokes the $p < \alpha$ criterion in statistical testing. *Churchill/Iacobucci (2002)* Does not distinguish between *p*'s and α's. *Cooper/Schindler* (2006) Invokes the $p < \alpha$ criterion in statistical testing. Incorrectly defines *p* value as a Type I error rate. *Malhotra (2004)* Advocates use of both *p*'s and α's. Invokes the $p < \alpha$ criterion in statistical testing. *McDaniel/Gates (2002)* Incorrectly defines *p* value. Invokes the $p < \alpha$ criterion in statistical testing. *Parasuraman/Grewal/Krishnan (2004)* Incorrectly defines *p* value as a Type I error rate.	These texts briefly allude to Neyman–Pearson orthodoxy. Examples: *Burns/Bush (2000)* Does not discuss Type I and II errors, the power of a test, or α levels. Nevertheless, invokes the $p < \alpha$ criterion in statistical testing. *Crask/Fox/Stout (1995)* Does not discuss Type I and II errors, the power of a test, and either α levels or *p* values as the significance level. But does discuss testing at the 5% and 10% "risk levels." *Lehmann/Gupta/Steckel (1998)* Briefly mentions Type I and II errors. Does not discuss the power of a test, α levels, or *p* values. Talks instead of "statistically significant" at the .05, .01, etc. levels.	These texts avoid all reference to Neyman–Pearson theory. They do not discuss Type I and II errors, the power of a test, or α as the significance level. Examples: *Sudman/Blair (1998)* Falsely equates hypothesis tests with significance tests. Basically adopts the Fisherian significance testing approach, $p(x \mid H_0)$, without invoking *p* values. Refers only to .05, .01, etc. significance levels. Do use *p* values in numerical examples, but without explaining them.

TABLE 2

THE REPORTING OF RESULTS OF STATISTICAL TESTS

	"Roving Alphas"(R)		*Exact p Values (E_p)*		*Combination of E_p's With Fixed p Values and "Roving Alphas"*		*"Fixed" Level Values*							
							P's		*"Significant"*		*α 's*		*Unspecified*	
Journal	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
EJM	54	46.2	12	10.3	21	17.9	10	8.5	14	12.0	2	1.7	4	3.4
IJMR	40	35.7	25	22.3	8	7.1	13	11.6	18	16.1	2	1.8	6	5.4
JAMS	186	54.7	28	8.2	53	15.6	29	8.5	32	9.4	9	2.6	3	0.9
JAR	120	43.0	40	14.3	22	7.9	36	12.9	53	19.0	2	0.7	6	2.2
JCR	327	77.3	6	1.4	49	11.6	17	4.0	13	3.1	3	0.7	8	1.9
JM	168	47.7	38	10.8	55	15.6	21	6.0	49	13.9	8	2.3	13	3.7
JME	49	31.8	32	20.8	31	20.1	18	11.7	9	5.9	8	5.2	7	4.5
JMM	21	43.8	9	18.8	12	25.0	4	8.3	1	2.1	1	2.1	—-	—
JMR	399	60.5	36	5.5	45	6.8	48	7.3	90	13.6	18	2.7	24	3.6
JR	164	60.3	12	4.4	34	12.5	21	7.7	34	12.5	4	1.5	3	1.1
ML	75	49.3	10	6.6	32	21.1	18	11.8	11	7.2	6	3.9	—	—
MS	57	50.9	6	5.4	5	4.5	11	9.8	22	19.6	6	5.4	5	4.5
Totals	1,660	54.9	254	8.4	367	12.1	246	8.1	346	11.5	69	2.3	79	2.6

FOOTNOTES

[1] For a fuller account of these distinctions see Gigerenzer, Krauss, and Vitouch (2004), Goodman (1993), Royall (1997), and especially Hubbard and Bayarri (2003).

[2] With three exceptions, the dates in parentheses are the initial year the journal was published. It was not possible to locate the first four years of the *EJM* (then known as the *British Journal of Marketing)*, nor the first seven years of the *IJMR* (until recently the *Journal of the Market Research Society).* Given the nature of the data being collected in the study, it was unnecessary to go back prior to 1960 for the *JR*. Also, data for the *EJM* and the *IJMR* extend only through 2000.