The Epistemology of Mathematical and Statistical Modeling

A Quiet Methodological Revolution

Joseph Lee Rodgers
University of Oklahoma

A quiet methodological revolution, a modeling revolution, has occurred over the past several decades, almost without discussion. In contrast, the 20th century ended with contentious argument over the utility of null hypothesis significance testing (NHST). The NHST controversy may have been at least partially irrelevant, because in certain ways the modeling revolution obviated the NHST argument. I begin with a history of NHST and modeling and their relation to one another. Next, I define and illustrate principles involved in developing and evaluating mathematical models. Following, I discuss the difference between using statistical procedures within a rule-based framework and building mathematical models from a scientific epistemology. Only the former is treated carefully in most psychology graduate training. The pedagogical implications of this imbalance and the revised pedagogy required to account for the modeling revolution are described. To conclude, I discuss how attention to modeling implies shifting statistical practice in certain progressive ways. The epistemological basis of statistics has moved away from being a set of procedures, applied mechanistically, and moved toward building and evaluating statistical and scientific models.

Keywords: mathematical models, statistical models, null hypothesis significance testing (NHST), Sir Ronald Fisher, teaching methodology

A relatively silent methodological revolution has occurred over the past several decades, almost without discussion. Instead, null hypothesis significance testing (NHST) has received virtually all of the attention. Statistical and mathematical modeling—which subsume NHST—have developed into a powerful epistemological system, within which NHST plays an important though not expansive role. To most practicing researchers in psychology, as well as in most statistics and methodology textbooks, NHST is the primary epistemological system used to organize quantitative methods. Instead, researchers and textbook authors should be discussing how to create and compare behavioral models that are represented mathematically and evaluated statistically.

Models are all around us, and the term *model* has many meanings. For example, I built *model* airplanes as a child. We see male and female fashion *models* in magazines. Psychologists study role *models*. These disparate concepts share in common that they are all simplifications of a complex reality. The airplane model does not fly, though its physical appearance matches one that does. The runway model epitomizes beauty and fashion. Role models engage in idealized (though perhaps unrealistic) behavior. In the same sense, our language is also a model. Speaking or writing creates a verbal instantiation of a complex reality. Our language can be nuanced and evocative, as when Elizabeth Barrett Browning wrote about love. Yet language has inherent limitations that necessarily simplify the complexity that it describes. For example, the word *love* must not have much precision, if I can simultaneously *love* barbecue, John Steinbeck, and my wife (although precision is added by context, of course). Clearly, language structures have both flexibility and limitations as models of the complicated reality they describe, and so do mathematical models, the topic of this article.

What are mathematical models? To Neimark and Estes (1967, p. v), "a mathematical model is a set of assumptions together with implications drawn from them by mathematical reasoning." Why not just use verbal or conceptual models? Bjork (1973) listed the advantages of mathemati-

This article is a revision of work originally presented as a presidential address to the Society of Multivariate Experimental Psychology at their annual meeting in October 2005. The ideas presented in this article have benefited from discussions over years with many different individuals (none of whom share credit for any weaknesses but who collectively can share in whatever strengths come forward from this work): Mark Appelbaum, David Bard, David Budescu, Will Beasley, Jack Carroll, Linda Collins, Elliot Cramer, Bob Cudeck, Lise DeShea, Linda Wanstrom, Scott Gronlund, Chris Hertzog, Lyle Jones, Hans-Peter Kohler, Sung-Hyuck Lee, Robert MacCallum, Scott Maxwell, Jack McArdle, Gary McClelland, Matt McGue, Jorge Mendoza, Keith Muller, Jay Myung, Alan Nicewander, Charles Reichardt, David Rindskopf, Jacci Rodgers, David Rowe, Frank Schmidt, Robert Terry, Rick Thomas, Paul Thompson, Vaida Thompson, Larry Toothaker, Tom Wallsten, Keith Widaman, and Forrest Young. With each person named above I have had discussions and conversations—some as long as 30 years ago—that I still remember and that influenced my thinking about the issues treated in this article. There are undoubtedly others who influenced my thinking as well but whose influence I do not remember directly; I offer my apologies (for forgetting) and appreciation to them as well.

Correspondence concerning this article should be addressed to Joseph Lee Rodgers, Department of Psychology, University of Oklahoma, Norman, OK 73019. E-mail: jrodgers@ou.edu

© 2010 American Psychological Association 0003-066X/10/$12.00
Vol. 65, No. 1, 1–12 DOI: 10.1037/a0018326

Joseph Lee Rodgers

cal models compared with verbal models: The mathematical models are more readily falsifiable, they force theoretical precision, their assumptions can be more easily studied, they promote data analysis, and they have more practical applications. In psychology and other behavioral sciences, methodologists and applied researchers describe the world in part through mathematical models. This article treats modeling, and also NHST, which modeling has quietly subsumed.

I have several related goals. First, I describe the history of NHST and reinterpret recent NHST controversy. The modeling revolution that was already in process, naturally and quietly, obviated the need for much of the controversy. Second, I describe conceptually what a mathematical model is, and I use several illustrations to show how one can be constructed. Third, I distinguish between using existing statistical modeling procedures and building mathematical models to evaluate statistically. Fourth, I discuss several implications of the modeling revolution for methodological pedagogy in psychology. In conclusion, I discuss ultimate implications of this quiet revolution.

A Brief History of NHST, the Recent Controversy, and the Trial

The Development of NHST

Sir Ronald Fisher originally conceptualized and developed NHST. At the computational basis of the analysis of variance (ANOVA) tradition, from which NHST emerged, is the concept of variability. Fisher (1921) "analyzed the variance" underlying individual differences in crop yields in relation to different fertilization regimes, and the ANOVA was born. Though Fisher's development of ANOVA is credited with starting NHST, Fisher also helped define the modeling revolution. However, both methodologists and applied researchers were so enamored with ANOVA and the associated NHST paradigm that Fisher's modeling contributions were relatively neglected for several decades.

Ultimately, the development of NHST was shared between Fisher on the one hand and Jerzy Neyman and E. S. Pearson on the other, with famous contention. Their arguments reflected different epistemological goals. Fisher developed his approach to NHST to answer scientific questions and to evaluate theory. Neyman and Pearson had more pragmatic and applied goals, to draw conclusions in difficult decision-making settings (e.g., in quality-control problems). Fisher viewed the hypothesis testing process as incremental, driven by replication, improving with each NHST decision, and potentially self-correcting. The Neyman–Pearson perspective, on the other hand, emphasized the importance of each individual decision.

The current version of NHST practiced in the applied research arena and presented in statistics courses is neither purely Fisherian nor purely Neyman–Pearsonian. For example, Fisher never referred to an alternative hypothesis or to the alpha level. As Christensen (2005) noted, Fisher used "the distribution of the (null) model and we examine whether the data look weird or not" (p. 121). On the other hand, "NP [Neyman–Pearson] testing is designed to optimally detect some alternative hypothesis" (Christensen, 2005, p. 123) but is not related to "proof-by-contradiction." Christensen (2005, p. 123) concluded, "NP testing and Fisherian testing are not comparable procedures" (also see Hubbard & Bayarri, 2003; Huberty, 1987; and Maxwell & Delaney, 2004).

Over time, the field of statistics merged the two different NHST methods into one. Modern statistics textbooks define NHST as a combination of the two approaches, emphasizing the main points of each—the null hypothesis and *p* value coming from the Fisherian tradition, and the alternative hypothesis, alpha level, and power emerging from the Neyman–Pearson tradition. Over a decade ago, Gigerenzer (1993, p. 314) called current practice "an incoherent mismash of some of Fisher's ideas on the one hand, and some of the ideas of Neyman and E. S. Pearson on the other hand"[1] (also see J. F. Box, 1978; Stigler, 1986; and Cowles, 1989).

In modern rendering of the NHST paradigm, the philosophical/epistemological basis begins by assuming the null hypothesis that a chance process generated the data. If the data patterns are inconsistent enough with the distributional predictions of the null, it is rejected in favor of a specified alternative. The alternative, the researcher's hypothesis, is always compared with what

[1] It is important to note, however, that sophisticated interpretations exist that conceptually unite Fisherian and Neyman–Pearsonian thinking. For example, Berger (2003) provided a Bayesian perspective that achieves this unity, and he opened his article with the comment that "Ronald Fisher . . . and Jerzy Neyman disagreed as to the correct foundations for statistics, but often agreed on the actual statistical procedure to use" (p. 1).

would be observed by chance under the null. Such focus on the null hypothesis originally engendered disagreement between Fisher and Neyman–Pearson and is the NHST feature that has been most frequently criticized in the years since.

Criticism and Adjustment of NHST

General criticism of NHST began in earnest in the 1950s. Early critics included Jones (1952), Rozeboom (1960), Bakan (1966), and Lykken (1968; also see Jones & Tukey, 2000, and Rozeboom, 1997, for interesting historical comparisons). Over the next 25 years, many scholars commented on and criticized NHST (see reviews in Nickerson, 2000, and various chapters in Harlow, Mulaik, & Steiger, 1997). The anti-NHST movement was especially stimulated by Cohen's (1994) cogent and engaging criticisms, and it reached a crisis in the mid-1990s.

What were the basic criticisms? Broadly stated, critics of NHST have viewed the conclusions generated by the NHST paradigm as both intellectually trivial and scientifically sterile, an opinion captured in Cohen's (1994) title: "The Earth is Round ($p < .05$)." Other specific realizations of this concern include the following:

1. All (point) null hypotheses can be rejected if the sample size is large enough.
2. Rejecting the null does not provide logical or strong support for the alternative.
3. Failing to reject the null does not provide logical or strong support for the null.
4. NHST is backwards, because it evaluates the probability of the data given the hypothesis, rather than the probability of the hypothesis given the data.
5. Statistical significance does not imply practical significance.

Dozens of articles evaluating and criticizing NHST have discussed these and similar issues (for relatively recent treatment and review, see MacCallum, 2003, and Nickerson, 2000).

In response to these criticisms were a number of innovative and creative approaches that were designed to fix one or more of the weaknesses of NHST. Past fixes have included the good-enough principle (Serlin & Lapsley, 1985), Harris's (1997) expansion of Kaiser's (1960) approach using three alternatives as outcomes (reject the null, fail to reject the null, or decide that there is not enough information), and Tryon's (2001) proposal of an inferential approach based on confidence intervals. Steiger (2004) described three adjustments to NHST in ANOVA settings, including tests of contrasts, confidence intervals of standardized effect size, and testing null hypotheses other than those of zero effect. But Cohen (1994, p. 1001) dismissed all such efforts: "Don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist."

Putting NHST on Trial

In the field of quantitative methodology, the 20th century ended with a bang, not a whimper. The criticism of NHST had built gradually and then exploded in its intensity during the middle of the 1990s. Extreme frustration with the weaknesses of NHST could be detected in the writing during this period, much of which was stark: NHST "never makes a positive contribution" (Schmidt & Hunter, 1997, p. 37); NHST "has not only failed to support and advance psychology as a science but also has seriously impeded it" (Cohen, 1994, p. 997); NHST is "surely the most boneheadedly misguided procedure ever institutionalized in the rote training of science students" (Rozeboom, 1997, p. 335).

Many students in introductory statistics courses have been taught the NHST jurisprudence model as a pedagogical tool. An analogy is developed between the courtroom and a research project, in which the null hypothesis is put on trial. Just as the defendant is presumed innocent until proven guilty, the null is assumed correct (innocent) until it is rejected (proven guilty) in favor of some more plausible alternative. The burden of proof rests on the researcher, as prosecutor, to demonstrate the "guilt" of the null, against a skeptical scientific community, the jury. The researcher is also a detective. As Tukey (1977, p. 3) noted, "Exploratory data analysis is detective in character. Confirmatory data analysis is judicial or quasi-judicial in character."

The field of quantitative methodology put this very set of trial-like procedures themselves on trial in the late 1990s, following Schmidt's (1996, p. 116) proposal that "we must abandon the statistical significance test." An American Psychological Association (APA) Task Force on Statistical Inference was convened to evaluate the status of NHST, and many methodologists and research psychologists were involved in public and private evaluation of the merits of doing away with NHST. The task force concluded that NHST was broken in certain respects but that it should not be outlawed: "Some had hoped that this task force would vote to recommend an outright ban on the use of significance tests in psychology journals. Although this might eliminate some abuses, the committee thought that there were enough counterexamples . . . to justify forbearance" (Wilkinson & the Task Force on Statistical Inference, 1999, pp. 602–603).

The Quiet Methodological Revolution

A basic thesis of this article is that the heated (and interesting) NHST controversy during the 1990s was at least partially unnecessary. In certain important ways, a different methodological revolution precluded the need to discuss whether NHST should be abandoned or continued. This quiet revolution, a modeling revolution, is now virtually complete within methodology. But within the perspective of the diffusion of innovations, this revolutionary thinking is only beginning to spread to the applied research arena and graduate training in quantitative methods. Identifying the revolution is one mechanism that will promote its diffusion. The methodological revolution to which I refer has involved the transition from the NHST paradigms developed by Fisher and Neyman–Pearson to a paradigm based on building, comparing, and evaluating statistical/

mathematical models. To illustrate, I describe how two basic summary statistics are viewed within each perspective.

The mean and variance have done yeoman service to psychology and other behavioral sciences for many years, yet their weaknesses are well-known (including, especially, their extreme sensitivity to outliers). Tukey (1977) developed a widely celebrated paradigm shift called *exploratory data analysis* (EDA), which is defined, at least in part, as a correction in statistical practice to overreliance on the mean and variance. (Note that EDA is *not* the revolution to which I refer, although it does overlap.) For the purposes of the current article, it is important to recognize how the mean and variance have shifted in concept during the modeling revolution. Within the NHST tradition they were often referred to as "descriptive statistics" (or even "mere descriptive statistics"; Tukey, 1977, p. vii). Postrevolution, the mean and the variance are simple mathematical models of an observed distribution of data.[2] The EDA perspective suggests that they are not "the" models, but rather that each is one of many models of the center and spread of a distribution of data. It is important to note that the *computation* of the mean and variance (or median and mean absolute deviation) are identical under each perspective. It is the conceptualization that has shifted, and this shift has important implications.

The null hypothesis is the leverage point to distinguish NHST from modeling. The philosophical basis of NHST begins with the null hypothesis. The researcher's hypothesis (embedded somewhere within the alternative) is compared with a chance process (the null). Emerging from the modeling revolution, the basic NHST question is asked in a slightly different way, and a completely different answer is provided. In fact, the answer emerges from the opposite direction. Within the modeling tradition, the relevant question is, "Does the researcher's model work to achieve its goals?" and the answer to the question "Compared with what?" is, "Compared with other models that are reasonable competitors." Estes (2002) summarized the procedure that emerges from this revised perspective:

> To determine whether the assumptions of the given model are necessary for successful prediction, some kind of comparative testing is needed. In practice, investigators have turned to evaluations of relative success by comparing a model with alternative models on each of a succession of tests. . . . if on any test the reference model generates the less accurate account of the test data, it is disconfirmed; otherwise, it is retained as the provisionally correct model. (p. 7)

In the modeling perspective, the researcher's hypothesis is disconfirmed, rather than the whole focus being on the null hypothesis that characterizes NHST. In modeling, the null hypothesis invoking pure chance as an explanatory process is deemphasized, although the null model from the NHST paradigm is sometimes (but not nearly always) one reasonable competitor. McDonald (1997) treated this reverse orientation:

> After the introduction of . . . structural models . . . , it soon became apparent that the structural modeler has, in some sense, the opposite intention to the experimentalist. The latter hopes to "reject" a restrictive hypothesis of the absence of certain causal effects in favor of their presence—rejection permits publication. . . . The former wishes to "accept" a restrictive model of the absence of certain causal effects—acceptance permits publication. (p. 202)

Similarly, Abelson (1997, p. 133) also referenced the epistemological shift that occurred as a result of the modeling revolution, though notably his discussion of what he refers to as " 'Goodness-of-Fit' Testing" is buried in an article otherwise devoted to contention over NHST: "Most tests of null hypotheses are rather feckless and potentially misleading. However, an additional brand of sensible significance tests arises in assessing the goodness-of-fit of substantive models to data."

To fully appreciate the import of this shift, we must recognize that the epistemological focal point has completely shifted. Within the epistemological tradition emerging from NHST, the focus is the *null hypothesis*, H_0, which we assume until it can be rejected in favor of a reasonable (but broad and relatively unspecified) alternative. The null hypothesis provides the leverage against which we understand the world. But even then, when we reject H_0 as implausible, we have not addressed the question, What is the new H_0 for future research?[3]

The postrevolution focal point is no longer the null hypothesis; it is the *current model*. This is exactly where the researcher—the scientist—should be focusing his or her concern. Statisticians discuss the merits of classical versus Bayesian inference, of exploratory versus confirmatory approaches, or of meta-analytic approaches (see, e.g., Berger, 2003; Howard, Maxwell, & Fleming, 2000). But researchers, who are scientists, should be focusing on building a

[2] As noted by a reviewer of this article, some methodologists and statisticians might be uncomfortable with the mean and variance being defined as "models." In a more sophisticated and mathematical view, models are defined as families of probability distributions, with different members of the family characterized by the particular combination of parameter values (see Myung, 2003). In this perspective, the population mean and variance are examples of such parameters. The sample mean and variance are used to estimate their population counterparts and, in combination with distributional assumptions, to draw conclusions about population processes. Without criticizing this statistical perspective, however, postrevolutionary thinkers have expanded the definition of a model. For example, Tukey's (1977) EDA vision represented an effort to broaden this parametric perspective, and he certainly did view both the mean and variance as simple mathematical models. In this sense, it matters to distinguish exploratory and confirmatory goals, and the more complex definition of a model described in this footnote clearly refers to confirmatory efforts. At an exploratory level, simpler definitions like those within the text of this article suffice.

[3] The Bayesian perspective provides an alternative conceptual approach to this decision-making problem (and also frames its own cogent philosophical criticism of NHST; see, among others, Pruzek, 1997; Rindskopf, 1997; Trafimow, 2003, and the response by Lee & Wagenmakers, 2005). The Bayesian approach—in which a set of prior probabilities are updated by data into posterior probabilities, which then can become priors for future research—has many supporters. Its philosophical grounding also overlaps with the modeling revolution (but as with EDA, the modeling revolution is conceptually separate from the Bayesian perspective).

model, embedded within well-developed theory: "The major task in any science is the development of theory" (Schmidt, 1992, p. 1177). The null hypothesis has always been a creation of the statistician. But for the scientist–researcher, his or her own research hypothesis is the natural focus. This simple and straightforward observation leads to a radical shift in both philosophy and practice: Rather than testing null hypotheses that our data arise from a chance process, we should be developing mathematical models and evaluating them statistically.[4]

If we are working in confirmatory mode (contrast with exploratory mode; Tukey, 1977), evaluating a model or theory, we specify a model (or class of models) and define appropriate competitors. For illustration, assume two nested models that are being compared, using empirical data.[5] The best model is the one that fits the data best in relation to its complexity, and it wins the contest. If it is not our original model, we admit it and attempt to interpret. The postrevolution question is, "Does the fit of the more complex model increase enough compared with that of the less complex model that the cost of additional complexity is worth it?" Or, "Is the increase in fit in moving from the simpler to the more complex model more than would be expected by chance?" Of course, the art and science of this enterprise involve how we operationalize the concepts "enough" and "better," how we measure "increase in fit," how we define "chance," and how we measure "complexity." In a sense, the null and alternative hypotheses still exist, but the null now has flexibility. However, we should stop calling the simpler model the "null" (which has substantial historical baggage, typically referring to the absence of any but chance processes) and refer to it as the "reduced" or "simpler" model.

I conclude this section with analytic examples representing the transition from NHST to mathematical and statistical modeling. Prerevolution NHST methods include general linear model (GLM) analyses such as ANOVA (and the t test), regression, and analysis of covariance (ANCOVA), each applied to test standard null hypotheses. It is notable, however, that these GLM methods have a postrevolution interpretation as well, one within which they are used to compare, evaluate, and choose between competing models; this perspective is explicitly developed in the next section. A prototypical modeling approach, the development of which has helped define the modeling revolution, is structural equation modeling (SEM). Log-linear modeling, a method that emerged from categorical data analyses based on chi-square tests, also has a strong modeling orientation. Several other analytic procedures can be implemented from a standard NHST perspective but are naturally viewed from the modeling perspective as well, including hazards modeling and time-series (autoregressive integrated moving average, or ARIMA) modeling. Another exemplary modeling method is multilevel modeling (or hierarchical linear modeling), which expands and recasts Fisher's ANOVA by explicitly separating sources of variance into the levels within which they occur and defining models of the correlated error structure caused by nested levels (see Curran, 2003, who treats the relationship between multilevel models and structural equation models).

So what is a model? How do we build and evaluate a mathematical model? The next section is directed to answering those questions.

Models as Simplified Reality Simplified as Models

The title of this section has at its heart the word *reality*, with two different subtitles emanating from that focal point. A prerequisite to understanding the modeling revolution is an appreciation of what a (mathematical) model is and how it relates to reality. These concepts are not complex. The definition of a model has two important, characterizing, features:

1. A model matches the reality that it describes in some important ways.
2. A model is simpler than that reality.

Pearl (2000, p. 202) defined a *model* as "an idealized representation of reality that highlights some aspects and ignores others." A *mathematical* model is one that captures these two features within one or more mathematical equations. Luce (1995, p. 2) suggested that "mathematics becomes relevant to science whenever we uncover structure in what we are studying." There is an important tension embedded within this definition. As the model matches reality better, it necessarily becomes less simple. Or, as it becomes simpler, it necessarily loses some of its match to reality. How does a researcher resolve this conceptual tension? The answer is the same both before and after the modeling revolution: statistically, of course.

It is worth noting that this definition is simple and understandable, yet broad. For example, the X-15 model airplane that youth of my generation built was a model in

[4] I felt the early bubbling of the modeling revolution in graduate school when LISREL was first released (e.g., Jöreskog & Sörbom, 1979). In my 1979 PhD comprehensive exams in the L. L. Thurstone Psychometric Laboratory at the University of North Carolina, Professor Jack Carroll required us to read and critique a LISREL analysis published in a major journal that interpreted results exactly backwards. A large test statistic and small p value were interpreted as good (i.e., supporting the researcher's hypothesis), when in fact, in the postrevolution context on which LISREL analyses were based, they actually showed just how poorly the researcher's model fit the data. In other words, the researchers were using postrevolution machinery but prerevolution thinking. Ironically, this mistake would not likely occur in the pages of a modern journal, but most of our teaching methods in psychology statistics classes today would still lead a student to make exactly this mistake.

[5] The text refers to nested models for simplicity of exposition. The problem of how to compare nonnested models with one another is a burgeoning topic within the methodology and statistics literature. A starting point for this assessment is the measurement of a model's complexity using metrics other than those that count degrees of freedom. Preacher's (2006) fitting propensity is one example, and information-theoretic approaches such as Akaike's information criterion (AIC; Akaike, 1973) and the minimum description length (MDL; e.g., Myung, 2000) are others. Careful consideration of the mathematical and statistical complexity of comparing nonnested models is beyond the scope of the current article but is being treated seriously in the methodology literatures.

exactly this sense. It matched the real X-15 airplane in many physical aspects yet was obviously simpler. The mechanical engineer's circuitry blueprint of the X-15 was also a model of the same plane, designed to fit a different part of the reality. Similarly, those who build models of adolescent sexual behavior can simplify in different directions. To some researchers, an outcome variable measuring sexual motivation is a function of genes, proteins, and hormones. To others, it is a function of religiosity and parental influence.

It has been popular to suggest that all models are wrong. G. E. P. Box (1979) said, "Models of course are never true, but fortunately it is only necessary that they be useful" (p. 2). The definitions above solve this concern from a slightly different perspective: The ultimate correctness of models is not typically at issue, because they "of course are never true." Rather, a particular model is supposed to be simple and to match only certain specified parts of the complex reality. Further, as Box's comment makes explicit, models must be useful, and their utility is defined (at least partially) in relation to whether they accurately match the part of the reality they attempt to describe.

As illustration, I return to the arithmetic mean, viewed by postrevolution thinkers as a simple mathematical model of a distribution of data. To begin, consider the arithmetic mean in relation to the two features in the definition of a model. The mean obviously simplifies the data, because it substitutes a single summary measure for *N* data points. It also fits the data in certain senses, because it is the center of mass of the data, and it also has optimality properties in relation to the data that it summarizes. Suppose that our goal is to build a model that explains the development of cigarette smoking among adolescents. On the basis of a nonrandom sample of 306 college students collected in 2003, Rodgers and Johnson (2007) reported the mean number of cigarettes smoked during one day by the 109 who called themselves "current smokers" as 3.3. The standard deviation of this distribution was 6.2. These summary statistics are not competing models, either conceptually or statistically, because they summarize different and complementary features of a distribution. Some models are not designed to compete with one another. But in other settings, the competition between simple summary statistics as basic models is both interesting and critical. For example, given the potentially skewed nature of the cigarette distribution, should one use a mean or a median to characterize the center of this distribution? The median is 0.2 cigarettes per day, much lower than the mean because a number of these "current smokers" were not currently smoking; for 41 of them, their average cigarette consumption per day was 0. Thus, the median in a sense competes with the mean as a mathematical model explaining cigarette use, on both conceptual and empirical grounds.

But in more sophisticated research settings, especially when the equation defining a smaller model is embedded (nested) within that specifying a larger model, then the natural goal for a researcher is to ask whether the added complexity is justified by how well the larger model fits empirical data. Then the two models compete with one another on fit and parsimony. We can build more complex and interesting models of cigarette smoking. For example, consider the simple conceptual model: "Smokers in a given year include both past smokers and new smokers, and the rate of transition into the smoking category is an important component of the process." Mathematically, we can operationalize that model using the following equation:

$$P(\mathrm{smoker})_t = P(\mathrm{smoker})_{t-1} + T{*}P(\mathrm{nonsmoker})_{t-1}. \qquad (1)$$

P(smoker) and *P*(nonsmoker) measure the proportion of adolescents who are smokers and nonsmokers, respectively, and *T* is a parameter of the model called the "transition rate," estimating the likelihood that a nonsmoker will become a smoker between time $t - 1$ and time *t*. At this point there are no statistical issues involved. Rather, issues of logical coherence, mathematical tractability, and face validity have helped us frame a mathematical model. A critic might note that there is no obvious role for social influence within the model, or advertising influence, or nicotine addiction. The model simply says that some fraction of nonsmokers become smokers during an interval, and *T* estimates the transition rate.

The next logical step would be to posit a more complex model that has a closer match to reality and to statistically evaluate whether the increased complexity is worth the loss of simplicity. Suppose we believe adolescents who have best friends who smoke to be differentially likely to become smokers compared with those whose best friends do not smoke:

$$\begin{aligned} &\mathrm{P}(\mathrm{smoker})_t = P(\mathrm{smoker})_{t-1} \\ &\quad + T1{*}P(\mathrm{nonsmoker\ with\ smoking\ best\ friend})_{t-1} \\ &\quad + T2{*}P(\mathrm{nonsmoker\ with\ nonsmoking\ best\ friend})_{t-1}. \qquad (2) \end{aligned}$$

Equation 2 is our "current model." Its competitor is the simpler Equation 1. Conceptually, we ask whether distinguishing adolescents with smoking versus nonsmoking best friends is an important feature of the reality we are modeling. Statistically, we ask whether the additional match to reality of Equation 2 is worth the cost of losing simplicity compared with Equation 1. Obviously we will need a measure of a model's "match to reality" (usually called a "fit statistic") and a measure of a model's "simplicity" (usually defined using degrees of freedom).

This example can be contrasted to a similar evaluation within the NHST perspective. For example, we could use a two-group *t* test to statistically evaluate whether adolescents whose best friends are smokers try their first cigarettes at a younger age than those whose best friends are nonsmokers. This design, like the one underlying the example in the previous paragraph, also emerges from a useful and interesting (and more common) research question, one with policy relevance. But the model comparison in Equations 1 and 2 has certain advantages. Those were dynamic models that accounted explicitly for age or time. They referenced more features of the reality that is being described by the models (e.g., an adolescent network and interaction between adolescents) that are only implicit in the ANOVA. Finally, the parameters in Equations 1 and 2 are structurally meaningful.

Within the modeling perspective, there are several basic and well-known ways to evaluate both fit and complexity within each model and to compare the two models statistically. Fit statistics evaluate how closely the predicted values generated by each model approximate the empirical reality contained in a set of measurement values or data (or, equivalently, how small the residuals are after the model is fit to the data). Equation 1 is a special case of (i.e., it is nested within) Equation 2, because *T*1 and *T*2 can be constrained to equality, and then Equation 2 is equivalent to Equation 1. As a result, Equation 2 must approximate reality at least as well as Equation 1. Thus, a fit statistic for the second model will necessarily be better (theoretically, no worse) than a fit statistic for the first. But is the increase in fit enough to decide that Equation 2 is the better model? Statistical theory provides an answer to this question. Fit statistics that can help us evaluate and compare models include the chi-square, the root mean square error of approximation, and the Akaike information criterion (AIC), among others. As the two models defined above are evaluated, no chance-level null hypothesis is posited, nor is an alternative constructed, at least not in the sense that those concepts are usually treated. However, traditional statistical concepts are used in this comparison, such as a test statistic (e.g., chi-square values), a sampling distribution (the theoretical chi-square), and an alpha level (to tune the trade-off between fit and parsimony). Further, the NHST perspective is embedded within this statistical evaluation in the sense that there is a null hypothesis built into the model comparison (i.e., whether the population parameters that *T*1 and *T*2 estimate are equal to one another). This is the sense in which the modeling perspective subsumes NHST.

If the models are special cases of the GLM (e.g., regression, ANOVA, or ANCOVA models), then there is a traditional link to standard statistical analysis. In this case, the uniformly most powerful test statistic to compare the models is Fisher's *F*. Further, the standard representation, $F = MS$ (between) / MS (within), can be algebraically rewritten as a model comparison between two linear models, one of which is nested within the other:

$$F(df_{\mathrm{r}} - df_{\mathrm{f}}, df_{\mathrm{f}}) = \frac{(E_{\mathrm{r}} - E_{\mathrm{f}})/(df_{\mathrm{r}} - df_{\mathrm{f}})}{E_{\mathrm{f}}/df_{\mathrm{f}}}, \qquad (3)$$

where E = error sum of squares, df = degrees of freedom, and the subscripts r and f refer to the reduced model and the full model, respectively. Cramer (1972), Appelbaum and Cramer (1974), and Maxwell and Delaney (2004, p. 77) developed and elaborated this formulation, and Yuan (2005) presented a modern discussion of the relation between test statistics and fit indices.

There is considerable conceptual value in studying the structure of Equation 3. In the numerator of the numerator, $E_{\mathrm{r}} - E_{\mathrm{f}}$, the fits of the full and reduced models are compared (note that this part of Equation 3 is often represented using R^2 instead of error-sum-of-squares terminology: $R^2_{\mathrm{f}} - R^2_{\mathrm{r}}$). The difference in fit is algebraically traded off against the difference in parsimony, measured as $df_{\mathrm{r}} - df_{\mathrm{f}}$, by dividing the former by the latter. The simpler model estimates fewer parameters and therefore leaves more data points free to test the fit of the model to the data. Conceptually, therefore, the ratio in the numerator of Equation 3 compares the models by measuring the improvement in fit per degree of freedom. Finally, the ratio in the denominator defines the total fit versus parsimony trade-off available in the full model, so that the equation measures the amount of available fit versus parsimony trade-off that is actually achieved.[6] Further, this conceptual interpretation for trading off fit and parsimony also has a well-known and carefully studied sampling distribution—the theoretical *F* distribution developed by Fisher—under a set of also well-known and carefully studied assumptions.

Historically, it is important that ANOVA and other linear models, even when cast into the NHST framework, have always had a model comparison interpretation underlying them. Prerevolution, the *F* measures whether the difference between group means is greater than would be expected by chance. Postrevolution, the *F* measures whether the larger model fits the data better per degree of freedom than the smaller model, in broad linear model settings. It is notable that the mathematics is algebraically (computationally) identical in the two perspectives, but the epistemological import has shifted substantially. Fisher, whose development of ANOVA motivated the NHST paradigm, also provided the basis for modeling. In fact, he developed a framework for the modeling perspective even before NHST became the basic 20th-century epistemology, and he stated (Fisher, 1925, pp. 8–9) that modeling comprises three steps, specification, estimation, and goodness of fit.

In the next section, I address one of the causes of the past unhealthy preoccupation with NHST, which is the focusing on statistical analyses as a set of procedures. Postrevolution, statistical analysis supports the development and evaluation of models.

Behavior Fit to Designed Models Designed to Fit Behavior

The word *models* is at the center of this section heading, with two different perspectives emanating from that central word. At least part of the problem in the diffusion

[6] In some textbooks and many software systems such as SAS, an even larger model is used in the denominator, the "full–full" model. This model's fit and parsimony are defined in relation to the largest referenced model. Thus, for example, if there are six independent variables (IVs) in a regression equation, and one of the four-IV models is being compared with a nested three-IV model, the numerator of Equation 3 will explicitly compare the four-IV model with the three-IV model in terms of the trade-off between fit and parsimony of the two models of interest. But the denominator will use the error-sum-of-squares (or R^2) of the six-IV model in the denominator to account for the trade-off between fit and parsimony available in the largest model.

of postrevolution thinking is that within the field of psychology, we practice research methods and teach statistics as though statistical modeling is just a set of procedures. After a semester of either undergraduate or graduate statistics, students are typically conversant with the prerevolution perspective, and many applied researchers also think of their statistical problems from the procedural perspective implied by the NHST epistemology. Typical prerevolution questions include the following: "How do I run a chi-square?" "What is the best procedure, a Kruskal–Wallis test or a standard ANOVA?" and "Let me tell you about my data, and you can tell me what procedure to run."[7] Statistics is the technical methodological language that scientists use to communicate with one another. It has embedded within it the philosophical basis of how we conduct our science, including statements about ethical scientific practice and causality. If statistical procedures provide foundational support for psychological research, how can otherwise sophisticated scientists treat philosophy, ethics, causality, and our basic language as "mere procedures?" In this section, I discuss elevating this conceptualization, in the two different ways implied by the section title.

Two different roles exist for postrevolution statistical models. The first is a model-comparison framework based on the application of existing statistical methods like ANOVA and SEM, used and applied by researchers who study and develop models of behavior. The second involves the development of mathematical models to match topics of explicit interest to researchers. Within this second framework, substantive scientists study behavior and from that process develop mathematical models specific to their research domain. Here, statistical methods are used to compare and evaluate these mathematical models. Both approaches are important for a successful methodology within psychology.

SEM has been built into a powerful analytic method and is a prototype of the first approach to postrevolutionary modeling. SEM is a quantitative method that exists independently of any particular application. The success of SEM depends on the extent to which it is applied in many research settings. Such methods can be (and often are) treated as "procedures," though this perspective is limiting. Similarly, multilevel modeling, log-linear modeling, and ARIMA modeling can be, and often are, applied procedurally.

Alternatively, some scholars build their own models to fit specific behavioral domains. Atkinson and Shiffrin (1968) developed a mathematical model of short-term and long-term memory. Shepard (1982) presented a mathematical model of musical perception based on the same type of double-helix structure that underlies the DNA molecule. Becker (1991) presented mathematical models of intrahousehold allocations, which treat the family as an economic unit. Rodgers and Doughty (2001) built probabilistic stopping-rule models of human reproduction in which childbearing decision making was conditioned on the previous sex composition of the family. Other examples abound. These models explicitly corresponded to the psychological, economic, or demographic behavior they were describing (see Rodgers, Rowe, & Buster, 1998, for further discussion). This second postrevolution approach builds the model to fit specific behavioral characteristics and requirements. Such models are not likely to be useful to anyone outside the particular behavioral domain. Their success is measured by how well they are observed by experts in a particular subject area to predict and explain tightly constrained behaviors. Roberts and Pashler (2000) defined a number of ways to evaluate the success of theoretical models in addition to good fits, including a model's flexibility, its ability to rule out plausible but theoretically indefensible alternatives, and its ability to predict unusual or surprising occurrences (also see Rodgers & Rowe, 2002, for commentary).

The cigarette smoking models in the previous section illustrate this second approach. The smoking models discussed there are examples of a class of models called "epidemic models of the onset of social activities" (EMOSA models). They have been developed to describe the explicit social contagion process through which adolescents influence one another to engage in certain adolescent behavior, including smoking and drinking (Rodgers, 2003; Rowe & Rodgers, 1991) and sexual behavior (Rodgers & Rowe, 1993; Rowe, Rodgers, & Meseck-Bushey, 1989). These models are operationalized as equations, each term of which explicitly matches a piece of the reality that the equations are designed to fit. Rodgers et al. (1998) contrasted an EMOSA model of adolescent sexual development with traditional linear models:

> [These EMOSA models] are generated from processes that are more realistic than those for which linear models account. For example, no one believes that an adolescent deciding whether to engage in sexual intercourse for the first time explicitly evaluates their biological drive, weights it by a constant, adds that to a weighted measure of peer pressure, subtracts a weighted measure of religiosity, and makes a decision about sexual behavior (or anything else) on the basis of the computation. Yet we understand that adolescents may approximately act as though they are making this computation, and therefore the model is predictive (even if not descriptive). On the other hand, our . . . nonlinear dynamic model of adolescent sexuality . . . used mathematical equations derived by assuming that adolescents interact with other adolescents, and pair up with opposite-sex partners. . . . That . . . model reflected a simplification of a social process that we presumed to actually occur. (pp. 1096–1097)

[7] This problem is illustrated through the following anecdote, an approximately verbatim exchange I had a few years ago with a dissertation-level student from another department whom I had never met: I answered a knock on my office door and found a young man who was a graduate student. "Dr. Rodgers, one of my committee members told me I should run a multidimensional scaling analysis and that you could show me how." I said, "Well, that's a fairly extensive conversation. In fact, I teach a whole graduate-level course in the theory and methods of scaling." The student replied, "Well, I only have about 10 minutes, but if you could just show me how to do it, that's all I really need. I don't need all that other theory stuff."

The parameters in these models are what Wright (1921, 1960) and Pearl (2000) referred to as "causal" or "structural" parameters, meaning that they should be interpreted as having substantive import.

Roald Hoffman (2003), a Nobel Prize-winning chemist, suggested that "uncomplicated models . . . play a special role in the acceptance and popularity of theories among other theorists. . . . The models become modules in a theoretical erector set, shuttled into any problem as a first (not last) recourse, . . . taking one piece of experience over to another" (p. 225). As theoretical understanding matures, Hoffman suggested, good theories become portable and stimulate experimentation and replication:

> What excitement there is in person A advancing a view of how things work, which is tested by B, used by C [who] . . . tests the limits of the theory, which leads to D (not C) finding that [relevant] . . . agent, whereupon a horde of graduate students of E or F are put to making slight modifications! (p. 225)

Clearly, Hoffman's imaginary model, focused on "how things work," is designed to match a physical reality.

Putting the Heart Before the Course

The "quiet revolution" has been relatively silent for several (overlapping) reasons. First, there is the inertia associated with NHST; it is difficult to change a zeitgeist, especially one with deep epistemological status. Second, the basic concepts associated with modeling are still spreading through the applied research community. Third, the modeling perspective is not treated in most statistics or methods classrooms. In this section, I discuss moving the modeling revolution into the classroom.

There are three epistemological approaches described in this article and potentially used in teaching statistics in the classroom. All three are critical and should be incorporated into the basic graduate curriculum in psychology departments. The first perspective is the prerevolution approach used in the majority of psychology classrooms, which is based on NHST. The second is the model development and model comparison approach, applied in relation to existing statistical methods such as SEM, multilevel modeling, and categorical data analysis. This pedagogy is realized when students take classes in which models compete against one another using statistics that account for the fit–parsimony trade-off. The third perspective occurs when students are taught how to develop their own mathematical models and to evaluate them statistically. The first approach— basic NHST—must continue to be taught at introductory levels (though accounting for its relationship to the second and third perspectives, which has seldom occurred in the past at the introductory level). Including both the second and third perspectives as core curriculum within graduate training programs will be necessary to further spread the modeling revolution to applied researchers.

Some textbook writers have already contributed to the implementation of this transition. Judd and McClelland (1989) and Maxwell and Delaney (1990, 2004 [2nd ed.]) anticipated the modeling revolution.[8] Many of the basic concepts are relatively stable between pre- and postrevolutionary thinking, including computational procedures associated with the mean and variance, the median and mean absolute deviation, as well as ANOVA, regression, and the GLM. At the level of the details, not much has changed (which is a partial explanation for why the modeling revolution has remained quiet). But the philosophical basis has changed, and thus we should add new elements to our pedagogy. The two textbooks mentioned above have contributed to that philosophical transition.

How has the philosophy changed? First, the null no longer has a very important role in this process. Even (especially) Fisher would support this change. In fact, he was never especially enamored with ANOVA or its broader NHST potential; rather, he called it only a "convenient method of arranging the arithmetic" (Fisher, 1934, p. 52). The null hypothesis of no effect, of purely chance processes generating the patterns in the data (what some call the "nil" hypothesis) is just one of many competitors with the larger model. Is this null model of any interest? Only if the researcher suggests that it is (see Wainer, 1999, for examples of a few settings that motivate the value of the nil/null). Second, shifting from "testing the null" to "building a model" puts us right where most of our language is anyway, when we distinguish between exploratory and confirmatory approaches, between developing or evaluating a model. Third, the treatment of *p* values should be shifted, so that students do not finish their statistics training believing that all small *p* values (and large values of the test statistic) are necessarily good and that large *p* values are by definition bad. Next, the concept of model residuals—the lack of fit of a particular model—should be developed fully. (For example, residuals can be subject to additional analysis, as emphasized within the EDA perspective.) Especially, the role of degrees of freedom, and other measures of the complexity of a model, should be brought forward within the modern teaching agenda. Finally, our methodological pedagogy should be developed so that—at least in more advanced methods courses, after mastery of basic principles—students are taught how to apply statistical concepts to evaluate their own models, from both of the

[8] I exchanged e-mails with Gary McClelland and Scott Maxwell during the development of this article. They recommended several other texts that they felt had also anticipated in certain ways the modeling revolution, including those by Estes (1991), Lockhart (1998), and Kenny (1987). Others that I have identified that fall into the category as well are those by Lunneborg (1994) and Abelson (1995). Obviously, the early 1990s was a period of progress that reflected the beginnings of the modeling revolution. I note that not all of these texts strongly express or reflect the modeling revolution, but they all are similar in spirit and perspective. Further, each contributes to breaking out of the standard "statistical modeling as NHST" paradigm and thus helps move the process forward. Little progress occurred after the mid-1990s (perhaps because so much attention was given to the NHST controversy). An exception is Kline's (2004) recent book, which provides many cogent comments about the history of NHST and about revising NHST. I have not yet had a chance to read the book by Jaccard and Jacoby (released December 30, 2009), which promises to overlap substantially with ideas presented in this article. Note that the text by Judd and McClelland was revised and re-released in 2008.

second and third epistemological perspectives developed earlier in this section.

Future textbook writers must necessarily help with this transition. Embedding statistical analysis models within successful and well-understood modeling methods (e.g., SEM, multilevel modeling) is advantageous (the second epistemological approach above). But there is no reason to rely only on existing analytic models, in either applied research or teaching. A different class of models—nonlinear models, chaotic models, differences models, and so forth—can be built so that the mathematical features of the models match the empirical reality being described. Explicit pedagogy should be given to measuring the simplicity, flexibility, and fit of a mathematical model (see, e.g., Myung, Forster, & Browne, 2000; Preacher, 2006). Treatment of power should be expanded, because the probabilistic specification of power shifts between the NHST perspective and the modeling perspective (e.g., MacCallum, Browne, & Sugawara, 1996). A number of suggestions contained within the response by the APA Task Force on Statistical Inference (Wilkinson & the Task Force on Statistical Inference, 1999) are relevant to the broadening of statistical pedagogy, including a focus on effect sizes and confidence intervals (e.g., Steiger, 2004).

A logical teaching curriculum—implemented across several classes—would begin with treatment of exploratory and descriptive procedures such as the mean and variance, not just as descriptive statistics but as models of the data (including attention to the resulting residuals and graphical methods for portraying data). At this level, the idea of simple models competing with one another (e.g., the mean vs. the median) can be introduced. The next natural transition would involve treatment of traditional statistical models that must be mastered by research scientists, including ANOVA, regression, and other linear models. For at least the short term, these models can first be treated from the NHST sums-of-squares approach, though infusing that treatment with an introduction to modeling is critical.[9] Then, treating structural equation models, multilevel models, log-linear models, and nonlinear dynamic models should be the next step, and at this point the modeling perspective is necessary pedagogy (and will be more natural if the modeling perspective has been a part of the pedagogy from the beginning). Finally, explicit pedagogy should be developed to teach students how to build their own models of the particular reality in which they have interest. Currently, this type of teaching—when it occurs at all—usually occurs through mentorship within graduate (and postgraduate) training. Introduction to this third epistemological approach should be included within graduate-level textbooks, and dedicated texts should be developed as well.

How should psychology quantitative methods textbooks look to support this process? The introduction needs to change, fundamentally. The philosophy of focusing on the model of interest should be a thread running through the whole text. The treatment of the null and alternative hypotheses, of Type I and Type II errors, and of power needs to change to accommodate the focus on the researcher's model, rather than the null (nil) hypothesis. The concept of degrees of freedom should expand. Fisher viewed degrees of freedom as statistical currency, available to be used as payment in the estimation of parameters (J. F. Box, 1978). This view can reinforce postrevolution pedagogy. Finally, treatment of EDA and the mathematics of transforming variables, Tukey's (1977) ladder of re-expression, residual analysis, and linking exploratory and confirmatory approaches as based on the common goal of building models should more completely infuse both introductory and advanced statistics textbooks. The textbooks mentioned above are actively contributing to these types of transitions.

Modeling as Art as Modeling

A final advantage of shifting our epistemology from a mechanistic NHST perspective to one that involves developing and evaluating scientific models has been (barely) below the surface of this whole discussion. We might call it the softening of our statistical and methodological practice, in the sense of both increased flexibility and increased reliance on creativity. The language of NHST is filled with rigidity and conflict: "rejection;" "null;" "power;" "fail to reject." In prerevolutionary thinking about statistics, we cannot even "go fishing." Meehl (1990) discussed the concept of "verisimilitude," which includes within it concepts related to the beauty, as well as the utility, of a model. Thissen (2001), in his presidential address to the Psychometric Society, quoted Pablo Picasso: "Art is a lie that enables us to realize the truth" (p. 475). Abelson (1995, p. 14) suggested that "if a single explanatory principle explains several results, the story is . . . not only coherent, it is *elegant*" (italics in original). Hoffman (2003) suggested that "many theories are popular because they tell a rollicking good story, one that is sage in capturing how the world works" (p. 224).

Engaging in science as a creative process requires thinking scientifically in creative ways. The application of NHST as a mechanical set of procedures precludes creativity. Building and evaluating statistical and mathematical models encourages creativity. To support the broad episte-

[9] It is an important point in support of the perspective promoted within this article that NHST should disappear as an epistemological system but be retained as a piece of the modeling perspective. For example, standard NHST is used to evaluate individual parameters within the development of a model; in these tests, the null hypothesis of the parameter value equaling zero is tested, for example, against a nonzero alternative. Testing null hypotheses as a mechanism within the model comparison framework to help evaluate the researcher's model is not what the modeling revolution is replacing. What methodologists criticize—and what the modeling revolution has changed through natural evolution within the methodological arena—is the conceptualization that NHST exists as a single overall statistical paradigm, the dominant epistemological system. Unlike NHST, the modeling paradigm (which subsumes NHST and includes it as a piece of the overall process of building and evaluating a model) has broad epistemological import in helping us understand the nature and functioning of psychological processes that we wish to understand, which solves many, perhaps most, of the concerns that have been stated for many years by those critical of NHST.

mological shift within the field of psychology from the first to the second perspective is the primary goal of this article, a process well underway, but one that requires further specification, organization, and attention.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Erlbaum.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akadémiai Kiadó.

Appelbaum, M. I., & Cramer, E. M. (1974). Some problems in the nonorthogonal analysis of variance. *Psychological Bulletin, 81,* 335–343. doi:10.1037/h0036315

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York, NY: Academic Press.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437. doi:10.1037/h0020412

Becker, G. S. (1991). *A treatise on the family* (2nd ed.). Cambridge, MA: Harvard University Press.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science, 18,* 1–32. doi:10.1214/ss/1056397485

Bjork, R. A. (1973). Why mathematical models? *American Psychologist, 28,* 426–433. doi:10.1037/h0034623

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association, 74,* 1–4.

Box, J. F. (1978). *R. A. Fisher: The life of a scientist.* New York, NY: Wiley.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59,* 121–126. doi:10.1198/000313005X20871

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003–066X.49.12.997

Cowles, M. (1989). *Statistics in psychology: An historical perspective.* Hillsdale, NJ: Erlbaum.

Cramer, E. M. (1972). Significance tests and tests of models in multiple regression. *The American Statistician, 26,* 26–30.

Curran, P. E. (2003). Have multi-level models been structural equation models all along? *Multivariate Behavioral Research, 38,* 529–569. doi:10.1207/s15327906mbr3804_5

Estes, W. K. (1991). *Statistical models in behavioral research*. Hillsdale, NJ: Erlbaum.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin and Review, 9,* 3–25.

Fisher, R. A. (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *Journal of Agricultural Science, 11,* 107–135.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1934). Discussion of "Statistics in agricultural research" by J. Wishart. *Journal of the Royal Statistical Society, Supplement, 1,* 26–61.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol 1. Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145–174). Mahwah, NJ: Erlbaum.

Hoffman, R. (2003). Why buy that theory? In O. Sacks (Ed.), *The best American science writing: 2003* (pp. 222–227). New York: HarperCollins.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5,* 315–332. doi:10.1037/1082–989X.5.3.315

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician, 57,* 171–177.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher, 16,* 4–9.

Jaccard, J., & Jacoby, J. (2009). *Theory construction and model building skills: A practical guide for social scientists.* New York, NY: Guilford Press.

Jones, L. V. (1952). Tests of hypotheses: One-sided vs. two-sided alternatives. *Psychological Bulletin, 49,* 43–46. doi:10/1037/h0056832

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5,* 411–414. doi:10.1037/1082–989X.5.4.411

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models* (J. Magidson, Ed.). Cambridge, MA: Abt Books.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach.* San Diego, CA: Harcourt Brace Jovanovich.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review, 67,* 160–167. doi:10.1037/h0047595

Kenny, D. A. (1987). *Statistics for the social and behavioral sciences.* Boston, MA: Little, Brown.

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

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112,* 662–668. doi:10.1037/0033–295X.112.3.668

Lockhart, R. (1998). *Introduction to statistics and data analysis for the behavioral sciences.* New York, NY: Freeman.

Luce, D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology, 46,* 1–26. doi:10.1146/annurev.ps.46.020195.000245

Lunneborg, C. E. (1994). *Modeling experimental and observational data.* Belmont, CA: Duxbury Press.

Lykken, D. E. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159. doi:10.1037/h0026141

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38,* 113–139. doi:10.1207/S15327906MBR3801_5

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149. doi:10.1037/1082–989X.1.2.130

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Belmont, CA: Wadsworth.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McDonald, R. P. (1997). Goodness of approximation in the linear model. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 199–219). Mahwah, NJ: Erlbaum.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141. doi:10.1006/obhd.1997.2687

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. doi:10.1006/jmps.1999.1283

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47,* 90–100. doi:10.1016/S0022–2496(02)00028–7

Myung, I. J., Forster, M., & Browne, M. W. (2000). Guest editors' introduction: Special issue on model selection. *Journal of Mathematical Psychology, 44,* 1–2. doi:10.1016/jmps.1999.1273

Neimark, E. D., & Estes, W. K. (1967). *Stimulus sampling theory.* San Francisco, CA: Holden-Day.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301. doi:10.1037/1082–989X.5.2.241

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, England: Cambridge University Press.

Preacher, K. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41,* 227–260. doi:10.1093/jpepsy/jsm107

Pruzek, R. M. (1997). An introduction to Bayesian inference and its applications. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 287–318). Mahwah, NJ: Erlbaum.

Rindskopf, D. M. (1997). Testing "small," not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319–332). Mahwah, NJ: Erlbaum.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367. doi:10.1037/0033–295X.107.2.358

Rodgers, J. L. (2003.) EMOSA sexuality models, memes, and the tipping point: Policy and program implications. In D. Romer (Ed.), *Reducing adolescent risk: Toward an integrated approach* (pp. 185–192). Newbury Park, CA: Sage.

Rodgers, J. L., & Doughty, D. (2001). Does having boys or girls run in the family? *Chance, 14,* 8–13.

Rodgers, J. L., & Johnson, A. (2007). Nonlinear dynamic models of nonlinear dynamic behaviors: Social contagion of adolescent smoking and drinking at aggregate and individual levels. In S. M. Boker & M. J. Wenger (Eds.), *Data analysis techniques for dynamical systems* (pp. 213–242). Mahwah, NJ: Erlbaum.

Rodgers, J. L., & Rowe, D. C. (1993). Social contagion and adolescent sexual behavior: A developmental EMOSA model. *Psychological Review, 100,* 479–510. doi:10.1037/0033–295X.100.3.479

Rodgers, J. L., & Rowe, D. C. (2002). Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler (2000). *Psychological Review, 109,* 599–604. doi:10.1037/0033–295X.109.3.599

Rodgers, J. L., Rowe, D. C., & Buster, M. (1998). Social contagion, adolescent sexual behavior, and pregnancy: A nonlinear dynamic EMOSA model. *Developmental Psychology, 34,* 1096–1113. doi:10.1037/0012–1649.34.5.1096

Rowe, D. C., & Rodgers, J. L. (1991). Adolescent smoking and drinking—are they epidemics? *Journal of Studies on Alcohol, 52,* 110–117.

Rowe, D. C., Rodgers, J. L., & Meseck-Bushey, S. (1989). An "epidemic" model of sexual intercourse prevalences for Black and White adolescents. *Social Biology, 36,* 127–145.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428. doi:10.1037/h0042040

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–392). Mahwah, NJ: Erlbaum.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181. doi:10.1037/0003–066X.47.10.1173

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129. doi:10.1037/1082–989X.1.2.115

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83. doi:10.1037/0003–066X.40.1.73

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review, 89,* 305–333. doi:10.1037/0033–295X.89.4.305

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182. doi:10.1037/1082–989X.9.2.164

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Thissen, D. (2001). Psychometric engineering as art. *Psychometrika, 66,* 473–486.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110,* 526–535. doi:10.1037/0033–295X.110.3.526

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371–386. doi:10.1037/1082–989X.6.4.371

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4,* 212–213. doi:10.1037/1082–989X.4.2.212

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. doi:10.1037/0003–066X.54.8.594

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557–580.

Wright, S. (1960). Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics, 16,* 189–202.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40,* 115–148.