Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 4

11-1-2010

On Scientific Research: The Role of Statistical Modeling and Hypothesis Testing

Lisa L. Harlow
University of Rhode Island, lharlow@uri.edu

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 348-358

1538 – 9472/10/$95.00

On Scientific Research: The Role of Statistical Modeling and Hypothesis Testing

Lisa L. Harlow
University of Rhode Island

Comments on Rodgers (2010a, 2010b) and Robinson and Levin (2010) are presented. Rodgers (2010a) initially reported on a growing trend towards more mathematical and statistical modeling; and a move away from null hypothesis significance testing (NHST). He defended and clarified those views in his sequel. Robinson and Levin argued against the perspective espoused by Rodgers and called for more research using experimentally manipulated interventions and less emphasis on correlational research and ill-founded prescriptive statements. In this response, the goal of science and major scientific approaches are discussed as well as their strengths and shortcomings. Consideration is given to how their recent articles intersect or differ on these points. A summary and suggestions are provided regarding how to move forward with scientific inferences.

Key words: Scientific inference, statistical modeling, null hypothesis significance testing.

Introduction

The Focus of Science

The study and practice of science is complex and encompasses various approaches and methods. Central to all of science is the search for basic principles from which phenomena can be explained and predicted. How are the underlying tenets - the golden nuggets of truth - in a scientific field discovered and illuminated? That is one of the main questions of this commentary.

Herbert Simon (1969), a Nobel Laureate in economics and a noted cognitive psychologist, believed that whereas human behavior is inherently simple, the complexity of the environment in which the behavior occurs can prevent or obscure human understanding of the basic processes. Thus, Simon (1969) viewed the main focus of science as finding the simplicity in the complexity of life.

Lisa L. Harlow is a Professor in the Department of Psychology. Dr. Harlow is Past President for the Society of Multivariate Experimental Psychology, Editor of the Multivariate Application Book Series, Associate Editor for *Psychological Methods* Journal and co-Director of Quantitative Training for Underrepresented Groups. Email her at: lharlow@uri.edu.

Four decades later, Michio Kaku (2009), a theoretical physicist and an advocate of making science understandable, reached a conclusion that was not far afield from Simon. Kaku made a comparison with the basic rules of chess and the actual enactment of a multitude of different possible chess games, elaborating that "the rules of nature may also be finite and simple, but the applications of those rules may be inexhaustible. Our goal is to find the rules." (p. 302). Kaku elucidated that the development and testing of basic principles in science "reveals the ultimate simplicity and harmony of nature at the fundamental level" (pp. 302-302), and that testing in science is most often indirect. As a result, it may be more productive to have multiple and varied ways to approach research and inferences in order to arrive at the most salient, underlying, and often latent, truths.

Consistent with the perspective that scientific understanding is not always directly observable, George Lakoff and Rafael Núñez (2000) emphasized the importance of concepts and analogies in what they call "the metaphorizing capacity" (p. 54) for understanding and applying quantitative methods beyond simple arithmetic and counting. These researchers realized the value of considering how a phenomenon is similar to, and different from, other related quantifiable observations. In a comparable view, Brian Hayes (2011) wrote that by breaking down stimuli into small segments and noticing points of contrast and similarity the most salient aspects are revealed. He summarized this process by stating that "the aim is to explore the kinds of patterns that appear frequently in our environment, in the hope of identifying and understanding some characteristic themes or features" (p. 422).

Another perspective was offered by Paul Rozin (2009) who discussed how published and funded research has tended, perhaps mistakenly, to involve results engendered through hypothesis-testing, controlled experiments and building causal evidence. In contrast, Rozin recommended descriptive or other kinds of studies that may have more external validity in varied, real-world settings. Rozin ventured that, "Elegance and clarity are criteria for publication, but there should be a trade-off with novelty and engagement" (2009, p. 437); and further that "a really interesting study with a flaw may be more valuable than a flawless but uninteresting study" (p. 438).

Stefan Hoffmann (2011) suggested that scientific curiosity is fed by having a great deal of background knowledge about a phenomenon, and then noticing anomalies, developing intuitions and finding connections. It is at the intersection of novelty, uncertainty and understanding that brings about scientific curiosity and discovery. Toby Huff (2011) concurred, speaking of how engaging curiosity and overarching synthesis lead to scientific discovery.

Culling together the perceptions of these and other astute thinkers, what appear to be integral for scientific discovery are the inquiring, understanding, seeking, describing, comparing and testing of credible and innovative ideas and relationships that may initially be difficult to discern; and the potential to assess the import and generalizability of findings with rigorous methodological procedures. I would argue that the methods espoused by Robinson and Levin, and Rodgers incorporate much of these elements of scientific discovery, albeit with differing approaches.

Approaches to Scientific Research

A reasonable question to ask is how scientific research should be approached. To accomplish scientific development and discovery, Simonton (2003) argued that it is important to see connections among diverse situations and processes, as well as to have an experimental, problem solving approach. Cronbach (1957) spoke to this seeming duality when discussing the two disciplines of psychology that involved either a correlational or an experimental focus. Each of these researchers is featuring two valuable, although often divergent, aspects of innovative science: naturalistic flexibility and rigorous control. This apparent dichotomy can also be viewed as striving for broad, generalizable external validity, versus strict and controlled internal validity; objectives endorsed in varying degrees by the Rodgers, and Robinson and Levin articles, respectively. Although there are probably as many approaches to scientific investigation as there are researchers, two major methods - null hypothesis significance testing and (correlational) statistical modeling - are the main focus of this commentary.

Null Hypothesis Significance Testing

The traditional approach to research, null hypothesis significance testing (NHST), is supported by Robinson and Levin, and minimized but recognized by Rogers. Briefly, NHST centers on an attempt to reject a null hypothesis of no notable import (e.g., two means are equal, a correlation is zero) and thereby attempting to build evidence for an alternate hypothesis that claims a significant difference or relationship. A noted benefit of NHST is that researchers can clearly specify null and alternate hypotheses and can calculate the probability of obtaining sample results as extreme or more so than are achieved in a relevant and randomly collected sample. Thus, if the probability, or *p-value*, is less than a designated level (e.g., 0.05), researchers can conclude that there is very little chance of obtaining the sample results found if the null hypothesis is true in the larger population from which the sample was drawn. This is particularly helpful if a decision is needed as to whether a specific treatment or intervention should be pursued as a viable option, after conducting a rigorous experiment that had adequate power to detect a significant finding and involved satisfactory design (e.g., random selection and assignment) to rule out possible rival hypotheses or confounds.

Devlin (1998) agreed, pointing out how probability theory is useful when it is necessary to make crucial decisions about whether to endorse a particular treatment or intervention. NHST would be helpful in this regard when there is a need to come to a decision about rejecting a null hypothesis with a specified probability. Others also attested to the benefits of NHST. Mulaik, Raju and Harshman (1997) stated that "as long as we have a conception of how variation in results may be due to chance and regard it as applicable to our experience; we will have a need for significance tests in some form or another" (p. 81). Chow (1996) and Cortina and Dunlap (1997), among others, also applauded the advantage of using NHST to rule out a chance finding in research.

Nonetheless, NHST has been extensively discussed and debated by Robinson and Levin, as well as Rodgers, and in numerous other forums (e.g., Balluerka, Gómez & Hidalgo, 2005; Denis, 2003; Harlow, Mulaik & Steiger, 1997; Kline, 2004; Nickerson, 2000). The better part of criticism regarding NHST appears to center on the exclusive focus of the *p-value* from a statistical test, and the accompanying dichotomizing decision to reject or retain the null hypothesis. Cumming (2012) has spoken at length on the volubility of *p-values* and the practice of NHST. Rice and Trafimow (2010) would likely agree with Cumming in arguing for less concern over Type I errors (i.e., rejecting a null hypothesis when the null hypothesis should not be rejected), and more attention to Type II errors, which refer to the failure to reject a null hypothesis when the alternate, scientific hypothesis may actually have more merit.

Noteworthy is that most, if not all, of the proponents and critics of NHST would also promote the use of additional substantiation over and above, or instead of, evidence of a significant *p-value*. Robinson and Levin advocated for correct applications of statistical hypothesis testing that involve randomized experiments, attention to Types I and II errors, effect sizes and sample size considerations, as well as the use of confidence intervals. Rodgers in turn played down hypothesis testing in favor of what he claimed is a broader, more subsuming and organic modeling approach that has emerged in an almost imperceptible methodological revolution. Before discussing the statistical modeling endorsed by Rogers and eschewed by Robinson and Levin, it is worthwhile to mention the merits of complementary procedures to help corroborate research findings.

Supplementing NHST

Any acknowledged advantages of NHST notwithstanding, current guidelines and research call for additional evidence when making scientific inferences. The recent 6th edition of the American Psychological Association (APA: 2010) publication manual "stresses that NHST is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed...” (p. 33); this viewpoint is consistent with that from Robinson and Levin as well as Rodgers and others.

Seven years before the APA guidelines, Denis (2003) presented a balanced overview of NHST and several possible alternatives. Denis suggested that the use of model testing among two or more reasonable alternatives, using good-enough hypotheses, calculating effect sizes and confidence intervals, and providing graphical displays of the findings are all effective and viable alternatives or supplements to NHST. Neither Rogers nor Robinson and Levin would be likely to take issue with much of this suggestion.

Others call for establishing or replicating a finding before it is accepted. Sawilowsky (2003) cautioned that effect sizes should not be widely published if they are not statistically significant. Filkin (1997) stated that “science seeks to separate fact from fiction by finding evidence” (p. 16); and that “for an idea or theory to be accepted as scientifically proven, it has to be tested in such a way that it can be tested over and over again and the result must always confirm the theory” (p. 20). Carl Sagan (1997) would have agreed with the need for replication; he wrote that the only way to find answers to “deep and difficult questions … [is] by real, repeatable, verifiable observations” (p. 63). Robinson and Levin aptly encouraged conducting “independent replications” to verify whether a significant finding is reliable, a practice also backed by Rodgers.

Consistent with replication, Wilson (1998) affirmed that “scientific evidence is accretionary, built from blocks of evidence joined artfully by the blueprints and mortar of theory … as evidence piles upon evidence and theories interlock more firmly, certain bodies of knowledge do gain universal acceptance” (p. 64). Wilson further highlighted the need for “improving the piecemeal approach science has taken to its material properties” (p. 66). Here, Wilson argued for a multivariate approach, as well as more attention to strong theory to ground scientific research. In this issue of the *Journal of Modern Applied Statistical Methods*, the value of theory was touted by Rodgers as well as Robinson and Levin; however, the usefulness of multivariate methodology was championed by the former but discouraged by the latter researchers.

It is also of interest that discussion about the need to augment NHST is not limited to the topic of abstract methodology, but rather intersects with the content and substance of practice and research. In a recent issue of the journal *Psychotherapy*, Thompson-Brenner (2011) introduced a special section on the role of significance testing in clinical trials. The set of articles illuminated considerations for providing the most accurate information on how best to create effective interventions in clinical practice. In the leading article, Krause (2011a), discussed the limitations of significance testing with randomized clinical trials (RCTs) and called for the inclusion of whole outcome distributions from participants in an RCT. Similar to what Cumming (2012) and others promote, Krause (2011a, 2011b) maintained that the significance test and *p-value*, alone, are not very informative about how to proceed with clinical treatments. Gottdiener (2011) responded by advocating the use of effect sizes and confidence intervals when presenting RCT results and asked researchers to supplement these data with information from case studies that can more specifically delineate treatment effectiveness and failure.

It is noteworthy that Gottdiener - as Wilson (1998) did earlier - also encouraged the study of multiple outcomes, arguing that multivariate data are more apt to provide bases for reliable and valid conclusions regarding treatment success or failure. Wise (2011) provided a compelling discussion on the need for evidence of clinically significant change and the use of a reliable change index, which is similar to a pre-post-intervention *z*-score for participants in an RCT. Here, the convergence and divergence of these proposals with respect to views put forth by Rogers, and Robinson and Levin, are not as clear-cut, except, again, that the former would favor multivariate approaches more readily than the latter researchers.

To round out this discourse on significance testing and its supplements, it is of note that Hagen (1997, 1998), a strong proponent of NHST, also recognized that effect sizes and confidence intervals are meaningful to report. Further, Hagen - who was reportedly "struck by the beauty, elegance, and usefulness of NHST" - went on to acknowledge that "other methods of inference may be equally elegant and even more useful depending on the question being asked" (1998, p. 803). Similarly, whereas Burnham and Anderson (2002) admitted that "for classic experiments (control-treatment, with randomization and replication) we generally support the traditional approaches (e.g., analysis of variance)" (p. viii), largely based on NHST; they more strongly endorsed a modeling perspective. Rozin (2009) would probably agree, stating that hypothesis testing may be more appropriate in fields where there is more knowledge and background. Otherwise, Rozin recommended assessing the nature of the phenomenon and its "generality outside of the laboratory and across cultures" (2009, p. 436), a practice that may be more easily accomplished with modeling. In this regard, it is useful to consider an alternative to NHST, namely, statistical and mathematical modeling.

Statistical and Mathematical Modeling

Rodgers (2010a) argued persuasively for adopting statistical and mathematical modeling, which he claims subsumes the predominant standard of NHST. Rodgers convincingly expressed the benefits and extent of statistical modeling, including such procedures as "structural equation modeling, multi-level modeling, missing data methods, hierarchical linear modeling, categorical data analysis, as well as the development of many dedicated and specific behavioral models." Rodgers further decried the emphasis in NHST on the rejection of a null hypothesis, a practice that, in opposition to Rodgers, was embraced by Robinson and Levin. However, these latter researchers clarify that they view NHST mainly as a screening device (Robinson & Levin, 2010) to illuminate findings worthy of further study, and thus would not be expected to place undue attention on the null hypothesis. Still, as Rodgers pointed out, statistical modeling places the focus on a well-constructed model, as opposed to a null hypothesis, and entails a "powerful epistemological system" of "building and evaluating statistical and scientific models." Rodgers (2010a) further advocated that methodological curriculum should be revised to incorporate a modeling approach, with NHST playing an "an important though not expansive role" (p. 1).

Others would agree with the call for wider use of model testing. Burnham and Anderson (2002) discussed a multi-model approach to understanding and approximating a complex process. Their information-theoretic approach includes comparing a scientific model that has a strong theoretical basis to several reasonable alternative models, while also taking into account parameter estimation, uncertainty and parsimony. In this way, a model or reduced set of models can be retained as the "best approximating model" (p. 2). Their approach represents a balance between over-fitting that would be neither replicable nor externally valid, and under-fitting which would be limiting and lack internal validity. It may seem paradoxical that Robinson and Levin would most likely also go along with the practice of testing multiple models, whereas it could easily be expected that Rodgers would approve of Burnham and Anderson's recommended multi-model testing methodology.

In a similar endorsement, Filkin (1997) described how Stephen Hawking, a renowned physicist, used a method called "sum over histories" to select the most likely approaches or models to understand a specific phenomenon and then to eliminate them one by one until arriving at the most probable solution (p. 272). Likewise, Maxwell and Delaney (2004) presented a convincing and integrative approach to science by proposing the examination of multiple models within a given study, ideally with research based on an experimental design. To varying degrees, Robinson and Levin, as well as Rodgers, would support this emphasis on assessing several viable and relevant models, particularly within the context of rigorous, controlled research.

Congruent with Rodgers' (2010a) focus on statistical modeling that recognizes the role of significance testing, Granaas (1998) claimed that "model fitting combines the NHST ability to falsify hypotheses with the parameter estimation characteristic of confidence intervals" and could still recognize that "effect size estimation is central" (p. 800). In an in-depth and convincing collection of model-based methods, Little, Bovaird and Card (2007) offered a well-articulated treatise on the benefits of statistical modeling, particularly when taking into account various conditions (e.g., mediation, missing data, moderation, multilevel data, multiple time points). I back each of these efforts, which would - at least in part - be supported by Robinson and Levin as to the value of NHST, considering relevant provisos. I would go further to state that statistical modeling may be more effective than NHST in allowing and even encouraging researchers to be more motivated to study, analyze and integrate their findings into encompassing and coherent streams of research. This position would most assuredly be endorsed by Rodgers.

The capabilities aside, it cannot go unnoticed that Robinson and Levin, as well as numerous other researchers (e.g., Baumrind, 1983; Cliff, 1983; Freedman, 1987a, 1987b; Ragosa, 1987) spelled out the possible hazards of statistical modeling, particularly when making unjustifiable causal claims from information that does not stem from longitudinal data or experimental design with adequate controls. Moreover, Kratochwill and Levin (2010), as well as Robinson and Levin, emphasized the importance of randomization, as well as replication and manipulation of the independent variable in order to achieve experimental control and build causal evidence. These authors argued that even single-case intervention designs can be made more rigorous and allow stronger conclusions, particularly by randomizing the assignment, timing and/or replication of interventions.

Shared Variance

Despite the various approaches to conducting scientific research, and the apparently contended methods of NHST and model testing, the articles in this issue by Rodgers, and Robinson and Levin could be said to agree on a number of practices and perspectives, including the merits of randomization and replication, and the cautions against over-interpreting correlations or using causal language when it is not justified. A careful reading of the viewpoints put forth by these authors, who admittedly come from differing epistemological vantages; concur on the importance of each of the following:

- Conducting exploratory / preliminary research that reveals worthwhile avenues to pursue;
- A strong theoretical framework;
- The use of randomization;
- Addressing threats to the validity of research;
- Emphasizing effect sizes and reasonable sample-size considerations;
- Being cautious to not over-interpret correlations;
- Avoiding causal language when not justified;
- Only making meaningful and justified conclusions;
- Encouraging replication;
- Noting the historical importance and development of NHST;
- Recognizing the value of NHST as part of a larger research process;
- Acknowledging the value of both NHST and statistical modeling;
- Realizing that both NHST and statistical modeling can be misused;
- Not disavowing a statistical procedure just because it is sometimes misused; and
- Accruing ongoing knowledge about scientific findings that address relevant problems.

By any yardstick, it would be difficult to deny significant overlap and agreement in the scientific values of Robinson, Levin, and Rodgers.

Just as it would not be accurate to posit hypothesis testing as the exclusive focus on a dichotomous decision between a null hypothesis and a generic alternative hypothesis, there may not be the need for a sharp contrast between the approaches presented by Rogers, and Robinson and Levin. Unlike Schmidt and Hunter (1997) who claimed that "statistical significance testing…never makes a positive contribution" (p. 37), or even McGrath (1998) who ventured that "it is very appropriate to praise the brilliance of NHST, but having done so, perhaps it is time to bury it" (p. 797), a more inclusive approach to science would allow for much of what was advocated by Robinson and Levin as well as Rodgers.

Rodgers (2010a) and Robinson and Levin, among others (e.g., APA, 2010; Wilkinson, et al, 1999), supported a broad and accurate approach that incorporates rigorous considerations (e.g., effect sizes, confidence intervals), alongside either NHST or statistical modeling. Hagen (1998), consistent with Robinson and Levin, and Rodgers, raised another issue by contending that "absence of evidence does not equal evidence of absence" (p. 803). By this Hagen clarified that research that fails to reject a null hypothesis cannot claim that the null hypothesis is true, a point that is sometimes mistakenly made with proponents of both NHST and modeling. In this regard, researchers conducting NHST cannot assert finding proof for the null hypothesis when it fails to be rejected. Similarly, those carrying out statistical modeling cannot overstate the benefit of a model in which the proposed model was not found to be significantly different from the pattern of variation and covariation in the data. Rogers, Robinson and Levin would undoubtedly agree that reasonable alternatives, confounds and considerations need ample deliberation, regardless of scientific approach.

Significant Differences or Type I Errors?

Given the recognized points of convergence, it is informative to at least mention that in this issue, Robinson and Levin, and Rodgers set forth differing or detracting points of view, as evinced in the following:

Robinson and Levin believed that Rodgers presents "a one-sided view of the controversy," and argue that they "have seen frequent misapplication of Rodgers' favored causal modeling techniques." Robinson and Levin further argued against a statistical modeling approach, based largely on the possible misuses associated with such an approach, for example, making unwarranted causal conclusions and overly prescriptive statements when using cross-sectional and correlational data. It is likely that most researchers, including Rodgers, would agree with their encouragement to use hypothesis testing wisely and to supplement with effect sizes and confidence intervals. Similarly, Rodgers and other researchers are apt to endorse their concern with ascribing causality when the research design did not include the necessary controls (e.g., randomization, manipulation, temporal ordering, isolation of effect, repetition).

Whereas Rodgers' (2010b) claim that statistical modeling could serve as a larger framework that subsumes NHST could be acknowledged, some of the writing may be too dismissive. For example, Rodgers charged that NHST does not have status and involves immature and simple science, compared with an epistemological system such as mathematical and statistical modeling. It may be more accurate to state that NHST can focus on more specific research questions, particularly in areas in which there is sufficient background knowledge to make informed and relevant hypotheses (see Rozin, 2009 for more discussion on this point).

Robinson and Levin occasionally made statements that may be overstated or inaccurate, such as using the qualifier "causal" numerous times when referring to modeling procedures or advocates, even when the term "causal" was not necessarily appropriate or endorsed by what was being described. This misattribution of causal language is evident in the title of their article, when referring to "Rodgers' favored causal modeling techniques," when speaking about "causal modeling techniques" and "unfortunate 'causal' nomenclature, "as well as "causal-model researchers," among other instances. Robinson and Levin also provided what they claimed as examples of "unjustified 'causal' excerpts" that are said to have overstated the use of causal language, when the research they describe does not explicitly appear to have done so and where, in some cases, the researchers have cautioned against making causal conclusions. For example, in an article that is critiqued, researchers claimed that "the data in the study are cross-sectional in nature and causal relations cannot be drawn" (Chen, et al, 2009, p. 304) although Robinson and Levin dismissed the stated limitation as "predictable."

Rodgers could also offer more elaboration and careful language when describing relevant examples that would favor modeling research, such as when stating how "selection bias has improperly influenced the interpretation of birth order-intelligence links," on illustrating a "type of sibling control," and on how "findings make a strong statement about both modeling and NHST." When describing each of these examples, there did not appear to be enough information provided to come to the conclusions that Rodgers set forth. Additionally, it would be preferred to use the word "parents" instead of "women" when discussing problems that are "almost completely attributable to the type of women who put their children in day care."

Regarding the use of language, Robinson and Levin occasionally used glib or dismissive terms when describing "the perceived magical quality of SEM allowing researchers to coax causality from correlational data," or referring to "grand prescriptives" in published conclusions. Moreover, these authors chided that cross-sectional and correlational data are "tossed into a statistical modeling analysis and what 'popped out' were causal conclusions", and allude to Rodgers' "seductive subtitle" that could purportedly "cause" researchers to see modeling as "methodological randomization compensating panaceas."

Another point worth noting is that Robinson and Levin, as well as Rodgers, expressed concern about the nature of the articles cited and, conversely, omitted from their respective manuscripts, when almost half of the citations in each manuscript involve one or more of the corresponding authors (i.e., 11 of 24 references are self-citations in Rodgers; and 14 of 33 references in Robinson & Levin similarly involve one or both of the authors). Whereas it is not unusual to cite relevant articles with which one is familiar, there may be some degree of selection bias in what is referenced in both manuscripts.

Are these points indicative of significant differences between Rodgers, and Robinson and Levin, or possibly just Type I errors in some cases? The reader may best decide.

Reconciling Different Approaches to Scientific Inference

Is it possible to come to agreement on how to approach scientific research? As Simon (1969) and Kaku (2009) expounded, whereas the world around us appears complex and unknowable, the role of scientists is to use whatever means are available to see through to the essence or set of truths in a field. These efforts will most likely involve thoughtful theoretical frameworks alongside sophisticated quantitative analysis to uncover what is not easily distinguished on the surface, positions that many scientists, including Rodgers, Robinson and Levin would endorse. Without specifying a precise approach, Devlin, a mathematician, writes that "where the real world is concerned, we have to go out and collect data. We enter the world of statistics" (1998, p. 156). Lakoff & Núñez (2000) affirmed that "mathematics is a magnificent example of the beauty, richness, complexity, diversity, and importance of human ideas" (p. 379), and Galton (1889) eloquently spoke of the wonder of statistics when used judiciously, stating:

> Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue … Science. (p. 62-63)

Advocates of both NHST and statistical modeling would most likely agree with Galton on the overriding splendor of quantitative methods when used responsibly, regardless of the particular approach to scientific research.

Hayes (2011) maintained that scientists may fare well when using statistical, probabilistic models. He argued that, in contrast to using a strictly deductive process and seeking deterministic principals, it is preferable to actively engage with the data by "forming and evaluating hypotheses, building conceptual models, and applying iterative procedures to refine the models or replace them when necessary" (p. 421). This description aptly depicts what Rodgers advocated with statistical modeling, and incorporates what Robinson and Levin encourage with testing hypotheses with randomized experiments "followed by a sufficient number of independent replications until the researcher has confidence that the initially observed effect is a statistically reliable one."

When considering the overall value of hypothesis testing and modeling, Rodgers (2010b) acclaimed that "NHST is a worthy, valuable, and useful tool" and "is still a proper paradigm, but it is a special case of a broader and thus more flexible paradigm." Hagen (1998) also acknowledged, along with Granaas (1998), that statistical modeling may well have advantages over NHST, although knowledge and use of modeling may not be as widely available as NHST, a position endorsed by Rodgers, as well. Certainly, the longer history of NHST as adopted in classrooms and research labs, has found its way into books and scholarly articles in larger volume than that of statistical and mathematical modeling procedures. It could only facilitate the progression of scientific knowledge to encourage more attention to well-tempered modeling to complement the pervasive availability and use of significance testing.

Ultimately, creative science depends on the ability to conduct specifically-focused, controlled studies that involve randomization and allow for causal inference. At the same time, there is a need for more broad-based and overarching statistical modeling that allows more flexible hypothesizing, analyzing and synthesizing of relationships among multiple relevant variables. There need not be an artificial dichotomy between these approaches to scientific research. Indeed, Rodgers (2010b) recognized that hypothesis testing and modeling "can be reconciled and accommodated" (p. 340).

As long as researchers keep in mind what can and cannot be claimed on the basis of their particular studies, the adoption of multiple approaches can only enhance and further the realm of science. A new journal is now available, the *Journal of Causal Inference*, edited by Judea Pearl and others, to encourage a rigorous multidisciplinary exchange of ideas regarding causation in scientific research. It is hypothesized that ongoing and open dialogue among foremost scientific researchers will help clarify the value of maintaining controlled and specific NHST, as well as revolutionary and overarching statistical modeling.

References

APA. (2010). *Publication manual of the American Psychological Association* (*6th Ed.*). Washington, DC: American Psychological Association.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(*2*), 55-70.

Baumrind, D. (1983). Specious causal attributions in the social sciences: The reformulated stepping-stone theory of heroin use as exemplar. *Journal of Personality and Social Psychology*, *45*, 1289-1298.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (*2nd Ed.*). New York: Springer.

Chen, L. H., Wu, C.-H., Kee, Y. H., Lin, M.-S., & Shui, S.-H. (2009). Fear of failure, 2 x 2 achievement goal and self-handicapping: An examination of the hierarchical model of achievement motivation in physical education. *Contemporary Educational Psychology*, *34*, 298-305.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility.* Beverly Hills, CA: Sage.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, *18*, 115-126.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671-684.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Denis, D. (2003). Alternatives to null hypothesis significance testing. *Theory & Science*, *4*, 1-21. Retrieved 9/8/11 from: (http://theoryandscience.icaap.org/content/vol4.1/02_denis.html).

Devlin, K. (1998). *Life by the numbers*. New York: Wiley.

Filkin, D. (1997). *Stephen Hawking's universe.* New York: Basic Books.

Freedman, D. A. (1987a). As others see us: A case study in path analysis. *Journal of Educational and Behavioral Statistics*, *12*, 101-128.

Freedman, D. A. (1987b). A rejoinder on models, metaphors, and fables. *Journal of Educational and Behavioral Statistics*, *12*, 206-223.

Galton, F. (1889). *Natural inheritance.* London: MacMillan.

Gottdiener, W. H. (2011). Improving the relationship between the randomized clinical trial and real-world clinical practice. *Psychotherapy*, *48*, 231-233.

Granaas, M. M. (1998). Model fitting: A better approach. *American Psychologist*, *53*, 800-801.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15-24.

Hagen, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist*, *53*, 801-803.

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

Hayes, B. (2011). Making sense of the world. *American Scientist*, *99*, 420-422.

Hoffman, R. (2011). That's interesting. *American Scientist*, *99*, 374-377.

Huff, T. E. (2011). *Intellectual curiosity and the scientific revolution: A global perspective.* Cambridge: Cambridge University Press.

Kaku, M. (2009). *Physics of the impossible.* New York: Anchor Books.

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

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 124-144.

Krause, M. S. (2011). Statistical significance testing and clinical trials. *Psychotherapy*, *48*, 217-222.

Krause, M. S. (2011). What are the fundamental facts of a comparison of two treatments' outcomes? *Psychotherapy*, *48*, 234-236.

Lakoff, G., & Núñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York: Basic Books.

Little, T. D., Bovaird, J. A., & Card, N. A. (Eds.) (2007). *Modeling contextual effects in longitudinal studies.* Mahwah, NJ: Erlbaum.

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

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist*, *53*, 796-797.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significant tests?*, 65-115. Hillsdale, NJ: Erlbaum.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-302.

Ragosa, D. R. (1987). Causal models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational and Behavioral Statistics*, *12*, 185-195.

Rice, S., & Trafimow, D. (2010). How many people have to die over a type II error? *Theoretical Issues in Ergonomics Science*, *11*(*5*), 387-401.

Robinson, D. H., & Levin, J. R. (2010). The not-so-quiet revolution: Cautionary comments on the rejection of hypothesis testing in favor of a 'causal' modeling alternative. *Journal of Modern Applied Statistical Methods*, *9*(*2*), 332-339.

Rodgers, J. L. (2010a). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1-12.

Rodgers, J. L. (2010b). Statistical and mathematical modeling versus NHST? There's no competition! *Journal of Modern Applied Statistical Methods*, *9*(*2*), 340-347.

Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward? A different perspective. *Perspectives on Psychological Science*, *4*, 435-439.

Sagan, C. (1997). *Billions and billions: Thoughts on life and death at the brink of the millennium.* New York: Ballentine Books.

Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, *2*, 218-225.

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 significant tests?*, 37-64. Hillsdale, NJ: Erlbaum.

Simon, H. A. (1969). *The sciences of the artificial.* Cambridge, MA: The M.I.T. Press.

Simonton, D. K. (2003). Scientific creativity as constrained stochastic behavior: The integration of product, person, and process perspectives. *Psychological Bulletin*, *129*, 475-494.

Thompson-Brenner, H. (2011). Introduction to the special section: Contextualizing significance testing in clinical trials. *Psychotherapy*, *48*, 215-216.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Wilson, E. O. (1998). *Consilience: The unity of knowledge.* New York: Vintage Books.

Wise, E. A. (2011). Statistical significance testing and clinical effectiveness studies. Psychotherapy, *48*, 225-228.