Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 3

11-1-2010

Statistical and Mathematical Modeling versus NHST? There's No Competition!

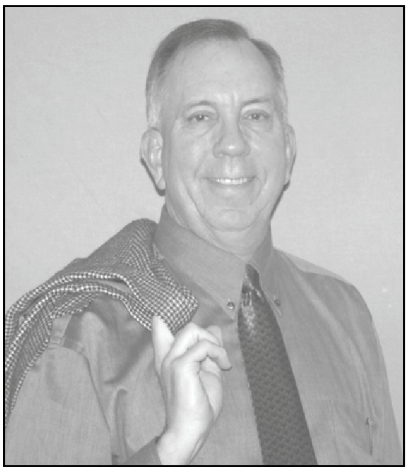
Joseph Lee Rodgers
University of Oklahoma, jrodgers@ou.edu

DOI: 10.22237/jmasm/1288584120

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 340-347

1538 – 9472/10/$95.00

Statistical and Mathematical Modeling versus NHST? There's No Competition!



Joseph Lee Rodgers
University of Oklahoma

Some of Robinson & Levin's critique of Rodgers (2010) is cogent, helpful, and insightful – although limiting. Recent methodology has advanced through the development of 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. These methodological approaches are based on a revised epistemological system, and have emerged naturally, without the need for task forces, or even much self-conscious discussion. The original goal was neither to develop nor promote a modeling revolution. That has occurred; I documented its development and its status. Two organizing principles are presented that show how both perspectives can be reconciled and accommodated. A program of research that could not have occurred within the standard NHST epistemology, without a modeling perspective, is discussed. An historical and cross-disciplinary analogy suggests their view is similar to Galileo's world view, whereas some branches of social and behavioral science may be ready for something closer to a Newtonian perspective.

Key words: NHST (or Null Hypothesis Significance Testing), Modeling, Mathematical ModelsD.

Joe Rodgers is a Quantitative Psychologist in the University of Oklahoma Department of Psychology, where he is a George Lynn Cross Research Professor and the Robert Glenn Rapp Foundation Presidential Professor. He has held visiting teaching/research positions at Ohio State, University of Hawaii, UNC, Duke, the University of Southern Denmark and the University of Pennsylvania. He and his research team have been continuously funded by NIH since 1987 to develop mathematical models of adolescent development, young adult fertility, and family/friendship interactions. His methodological interests include mathematical modeling, resampling theory, linear statistical models, quasi-experimental design methods, EDA and multidimensional scaling. He has been editor of the applied methods journal *Multivariate Behavioral Research* from 2006 to 2011, and is a past-president of the Society of Multivariate Experimental Psychology, the Society for the Study of Social Biology and the APA's Population and Environmental Psychology Division. Email: jrodgers@ou.edu.

Introduction

Null Hypothesis Significance Testing (NHST) has, for many years, been the primary organizational and epistemological system by which we understand statistical practice in behavioral sciences. NHST has been frequently criticized, and in the late 1990s the criticism was sufficient to create substantial contention, explicit calls for NHST to be outlawed or abandoned, the appointment of an American Psychological Association task force to judge its status and to evaluate proper statistical practice, and a great deal of discussion and argumentation, both informally and in published articles. Before and during this same period, a different epistemological system, what I referred to in Rodgers (2010) as a modeling revolution was in development. With little discussion (and most of what would have naturally occurred has been largely drowned out by the clamor over NHST), mathematical and statistical modeling have become the set of organizing principles that has the potential to completely replace NHST as the primary epistemological system. And modeling should replace NHST, for several reasons.

The first is because it is a more natural way for researchers to frame, think about, and conduct research, whereas NHST was a creation of and for statisticians. Second, modeling has more flexibility to support the maturation of both statistical and methodological practice within psychology and other behavioral sciences. Third, modeling includes NHST as a special case, and so NHST has not been replaced or even very much revised as a set of procedure.

Robinson and Levin presented position statements that, in my career, I have taught to my students, and have applied in my research. These principles emerged from a strong and coherent philosophical background, including caution against over-interpretation of correlations, which emerged from John Stuart Mills' (1843) inductive canons of scientific inquiry. Another principle is to use randomization if possible, which emerged from Fisher's (1935) answer to the problem that Mill left open -- how can researchers equate groups, on average, before a manipulation? Yet another is to emphasize the importance of replication; this underappreciated practice serves the purpose of correcting the bad luck that can befall a researcher in "gaming with the devil" (see Box, 1978, p. 144), and is another of Fisher's edicts that helped create the philosophical basis of social/behavioral science methodology.

I could (almost) leave this reply hanging, and emphasize how correct and well-founded are many of the positions stated in their critique. If so, though, I would necessarily conclude with some comments about how none of these principles has any import in evaluating either the status of NHST, or the development of statistical/mathematical modeling, or as criticism of my article, because these principles stand firm in relation to either NHST or statistical/ mathematical modeling. However, if I left my reply here, that would obfuscate my initial intent, which I believe has been mischaracterized.

Two basic principles (and some potential quibbles with the language, to follow) are paramount, and within those principles their criticisms and my position statements will be simultaneously accommodated. The first principle is that NHST is the type of statistical paradigm that naturally applies to a rather immature science, whereas statistical modeling naturally fits a more mature, or at least maturing, science. The second principle is that NHST is subsumed within the modeling perspective. The two paradigms need not compete, as Robinson and Levin implied. Accept the modeling perspective, and it can be sharpened to the special case of the NHST perspective at any time; insist that NHST is the one, only, and proper epistemological position, and the full range and power of structural equation modeling (SEM), multi-level modeling (MLM), and dozens of specialty models are relegated to virtual impotence.

Statistical Modeling Reflects and Supports the Maturation of Social/Behavioral Sciences

The development of statistical and mathematical modeling as an epistemological system didn't occur through high-level mandate or management; it has been a natural and emergent methodological feature of the maturing of psychology (and has parallels within education, economics, sociology, and other social/behavioral sciences). In this sense, it is a

mischaracterization to claim that I "condemn" NHST or that I "perceive vices of statistical hypothesis testing." Most of my article was not prescriptive, despite their suggestions to the contrary; the part that is prescriptive has little to do with liking or condemning NHST. Rather, I described a developmental process that is well advanced, though relatively unexamined in historical perspective. As science has advanced, stronger statements are possible, ones that even in some cases move toward legitimate causal attribution. Nowhere in that previous sentence is there encouragement to assert unjustified causality. Further, to suggest that such unjustified claims occur – even to illustrate with specific examples – does no damage to the position that our science is maturing in that direction. Nor is science necessarily advanced by successful causal claims; sometimes, rather, it advances by identifying past mis-attributions, a process which Robinson and Levin support and appreciate. Ironically, though, certain versions of that process would not likely emerge from an NHST perspective. I described an example from my own research program.

For many years, nearly an entire community of research psychologists has ignored a certain type of selection bias, resulting in the kind of mis-attributed causal process that Robinson & Levin (and I) decry. Scarr and McCartney (1983) made a stark statement concerning this design flaw, which is inherent in literally hundreds (perhaps thousands) of previous published papers: "passive genotype-environment effects arise in biologically related families and render all of the research literature on parent-child socialization uninterpretable" (p. 427). Using a quasi-experimental design that takes advantage of siblings to partially control for selection bias, along with a powerful sibling dataset, my colleagues and I have published a series of articles during the past decade that have separated and quantified the difference between certain types of inherent selection bias and the remaining correlational links, within which the causal attributions are logically expected to exist.

I review several of these studies based on the sibling design and on the children-of-siblings design. (Besides these, other quasi-experimental design innovations exist that also can also be used to separate family-based selection bias from parental and family influence; see D'Onofrio, 2003, for description of the children-of-twins design and Rodgers, et al, 2008, for description of the mother-daughter-aunt-niece design). Rodgers, et al. (2000) showed how selection bias has improperly influenced the interpretation of birth order-intelligence links; at least most (perhaps all) of what has appeared to be birth order effects on intelligence in past research has actually been between-family differences in parental education and IQ, among others (see Rodgers, 2001 for further explanation of this logic, Wichman, et al., 2006, for a modeling demonstration of this phenomenon, and Wichman, et al., 2007, for further elaboration). D'Onofrio, et al. (2008) showed how the link between smoking during pregnancy and child conduct problems is at least partially caused by the kind of women who smoke during pregnancy, thus challenging much of the direct causal attribution.

D'Onofrio, et al. (2009) used a similar design to investigate the relationship between family income and child conduct problems, with similar conclusions. Mendle, et al. (2009) applied this type of sibling control to study the link between father absence and age at first intercourse, and found that much of the apparent direct link between father absence and age at first intercourse has likely been caused by shared genetic factors in the background. Harden, et al. (2009) studied whether population density has a direct influence on antisocial behavior during adolescence, or whether the apparent link is due to selection bias; the latter was more strongly supported. Finally, Jaffee, et al. (2011) showed how placement of infants and young children in day care as an influence on both achievement and behavioral problem scores in childhood is almost completely attributable to the type of women who put their children in day care, leaving very little remaining variance to attribute to the direct influence of the day care experience in and of itself on these child outcomes.

For the purposes of this reply, these findings make a strong statement about both modeling and NHST. Each result above depended on strong design logic combined with a statistical modeling exercise. Further, each study contained within it a number of NHST

results, but the *organizational principles* emerged from a research perspective that required longitudinal and within-family data, strong research designs, powerful measurement tools, and sophisticated statistical models. They would not have likely emerged from an NHST epistemology. Nor are the conclusions that emerge from this type of work necessarily causal; indeed, most of the conclusions above challenge previous causal attributions.

In the tradition of Cook and Campbell (1979) and Shadish, et al. (2002), the researchers' goal, whether in quasi-experimental or experiment research, is to address as many threats as possible to internal validity, the validity of causal attribution, and to admit freely and to self-evaluate in the face of those that remain. Robinson and Levin admitted to this maturational challenge: "Our field of educational psychology is filled with such examples of comparing new innovations with ridiculous strawperson control conditions that no sane researcher would ever consider using." So are psychology, sociology, etc., of course. And so the proper and defensible approach is exactly where they stated it should be, using an appropriate set of methodological tools to draw cautious but legitimate conclusions, and to avoid wasting time asking superficial and uninteresting questions. Hopefully, those methodological tools expand to accommodate improvements, maturation, in the science that they support. Statistical modeling is an example of such expansion.

Statistical Modeling Subsumes NHST

There exists a way to view both NHST and statistical modeling that accommodates both Rodgers (2010) and Robinson and Levin's critique. That accommodation was stated in my original article, but here I shall present this argument in different words. Robinson and Levin presumed I was prescriptively criticizing NHST; that I favor modeling and oppose NHST: "Rodgers (2010) has written a cogent essay on what he perceives as the vices of statistical hypothesis testing and the virtues of statistical modeling." First, my article was intended to be more of an historical account than a desideratum about what should be. Second, I was a strong opponent of outlawing, abandoning, or otherwise providing any type of institutional control over NHST (or any other methodology). I have used the NHST paradigm often, in most of my published research. I have also used modeling approaches, when they appeared to be useful and appropriate.

Many of my publications incorporated both, which leads to my third comment: I do become prescriptive when I describe in detail how the statistical modeling strategy subsumes NHST, because I'm convinced of the value of *both* approaches. Hence, the crux of my reply: NHST is a proper paradigm, but it is a special case of a broader and thus more flexible paradigm. I do not agree there are two competing approaches. One is broader and one is a special case. The modeling approach uses NHST as a fundamental part of the modeling framework. As Rodgers (2010) explained:

> As the two models ... 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 ... are equal to one another). (p. 7)

NHST is a tool, as a way to answer a certain question. I've never understood why researchers would be satisfied with the conclusion to reject H_o or fail to reject H_o, unless the research question was simple enough to warrant such a conclusion. It seems to me that when the research questions become more complex, modeling has the potential to provide more complex answers, and to move scientific epistemology forward substantially further than what can be obtained via NHST.

Minor Issues

There are some mischaracterizations in their critique that require a response (though the majority are accounted for by the two principles in the previous sections). They suggested that "he then goes on to discuss NHST as a hybrid and condemns it;" I did not, though I cited Gigerenzer (1993), who did. They implied that I supported a ban on NHST, when I actually opposed such a ban. They claimed that "Rodgers also condemns the NHST 'jurisprudence model," whereas in fact I teach and promote this way of thinking of NHST. They suggested that "Rodgers mischaracterizes Tukey's 'exploratory data analysis' strategy insofar as the detective nature of that hypothesis-generating approach clearly is not jurisprudence," but I did not link the detective and jurisprudence components – after describing the role of jurisprudence within research, I stated "The researcher is *also* a detective" (p. 3, italics added for emphasis). They failed to make the connection that my section titled "Criticism and Adjustment of NHST" was historical; their first sentence in their section "The Null Hypothesis Hullabaloo" recognized historical goals, but the remainder of that section was not about the "hullabaloo," but rather about their perception that I promulgated it, although I did not. Finally, they suggested that "he gives short shrift to approaches that have defended reasonable and proper applications of statistical hypothesis testing," and cited four articles that would have provided more balance. In fact, I discussed three of those articles.

NHST is a worthy, valuable, and useful tool. It helps researchers to answer a certain question, framed in a certain way. However, its weaknesses are well-known, and often discussed (see Wainer, 1999, for a balanced and interesting account, among dozens of others). Further, as the field of behavioral science matures, it should not stand as the epistemological basis of research methodology within the field of psychological science, because modeling is more useful, flexible, and better supports the future of behavioral science research.

This methodological practice should not be banned or outlawed for two reasons. First, such practice should not be managed at the institutional level (any more than the workers' union should decide to ban hammers or electric saws). Second, NHST has served its value in thousands of scientific settings. It has also been misused, and Robinson and Levin provided support for its proper and legitimate use, in this and other published articles.

Regarding "the 'revolution' about which Rodgers writes is neither quiet nor methodological," they were correct, as I originally asserted. The NHST hullabaloo was anything but quiet. But the modeling revolution was so quiet that apparently many didn't notice, and now aren't sure that it occurred. Robinson and Levin contend that the revolution was not methodological, that the issues are entirely statistical. SEM contains both a structural and a measurement model. Multi-level modeling accounts for clustering, which is often caused by sampling processes. Multilevel modeling also cannot be separated from the design issues that generated the different levels. Analytic procedures that handle missing data require specification of the generating processes – sampling, measurement, etc. – that produced the missing values. In other words, modern statistical models account for design, sampling, and measurement, as well as the formal statistical properties of statistical models. As one example, MacCallum and Tucker (1991) could not have developed their conceptualization separating sampling and model error if they had used an NHST epistemology.

It is perhaps not surprising that those whose way of thinking about the advancement of behavioral science is embedded in the NHST tradition would not recognize the modeling revolution as bringing about the expansion of statistical practice to include many other features of the methodological arena. But such broadening is one among many features of statistical/mathematical modeling that make the use of SEM, MLM, missing data approaches, and other modeling methods exciting and useful. To expand their analogy, there are new dangers created in using models, and their misuse cannot be supported (Cliff, 1983). The danger is analogous to learning how to use electric saws, when hand saws used to be the state-of-the-art. We can either decry electric saws, or teach their proper and safe use. One of premier psychology quantitative journals is called *Psychological*

Methods, and publishes articles on design, sampling, measurement, and statistics, as well as how these different areas overlap and inform one another.

Conclusion

Consider an analogy from the history of science to illustrate the points made in my response. The analogy draws on two popular science books, Sobel (2000) and Gleick (2003). The late 16th and early 17th century occupied a remarkable period of scientific ascendancy in the field of astronomy. In 1543, Copernicus offered the insightful (yet heretical) view that the earth revolved around the sun, rather than vice versa. Galileo was born shortly after, and as Sobel noted, "All his [Galilio's] observations lent credence to the unpopular sun-centered universe of Nicolas Copernicus, which had been introduced over half a century previously, but foundered on lack of evidence" (p. 7). The observations to which Sobel referred were of course obtained with Galilio's new invention, the telescope, through which he observed the moons of Jupiter, the face of earth's moon, and the sunspots moving across the face of the sun. Such observations were, in modern language, exploratory evidence in support of a previously proposed theory.

Although probabilistic reasoning was still in its infancy (and was being developed by Fermat and Pascal in France during the same historical epoch), the epistemological basis of scientific inquiry in astronomy during that period was similar to that in psychology during the 20th century. The NHST paradigm that Robinson & Levin vigorously defended was similar to the one used by Galileo and others during the period of time in which they were collecting information (using telescopes and otherwise). Ultimately, such information of course inductively coheres into theoretical propositions. Galileo offered multiple sources of astronomical evidence for a heliocentric view of the solar system, including the movement of sunspots, the eclipses of the moons of Jupiter, and the tides on earth. Each might be viewed as a separate astronomical significance test of the null hypothesis that the earth was at the center of the universe, a hypothesis that we have ultimately rejected. But astronomy quickly moved on beyond the question of whether the Copernican system could be rejected or not.

Kepler, in 1609 and 1619, published his three laws of planetary motion, and Newton (who was born in 1642, the year that Galilio died), published in 1687 his *Principia*, stating formal mathematical models of motion and the universal law of gravity. These "laws" stepped up to a new epistemological level, using previous observations as the basis for mathematical models that were designed to subsume many previous disparate and separate astronomical observations. (The development of the double-helix model of DNA is another example in a different discipline in which disparate observations were brought together inductively using mathematical modeling.)

To bring these historical references to the current discussion, Robinson and Levin wrote: "we agree that - in the field of education - we have enough theory development studies and need more studies that address practical 'what works' questions." Fair enough. They argued that in many domains of our immature science more knowledge is needed, that more educational and psychological telescopes need to be brought to bear on current problems. Nothing in my own teaching, thinking, or research practice holds anything but praise and agreement for such a position. Indeed, two of my primary courses over the past 30+ years of teaching have been Exploratory Data Analysis and Quantitative Methods in Evaluation Research, where students learn to engage exactly this kind of goal, to address practical "what works" questions.

Then, they stated, "It is our fear that a research approach where the question 'Does the data fit my model?' is far more dangerous than the question 'Is there anything here worth pursuing?'" Again, fair enough. Without knowledge, both scientists and those who consume the science (policymakers, the public, etc.) can be led to the modern equivalent of the geocentric universe, and there is indeed danger in promulgating positions both pro and con in the absence of adequate knowledge, or even with substantial knowledge when that knowledge is at odds with societal expectations (just ask Galilio!). But does such lurking danger excuse statisticians and methodologists from

developing proper tools, perspectives, and whole epistemological systems to support the development and evaluation of such models?

My answer, strongly implied throughout the original article, is indeed not. Both the NHST epistemology they promoted for relatively immature science, and the one that they view as dangerous, the modeling approach, should exist side-by-side within the arena of quantitative methods in both education and psychology. I promoted the development of the latter, not erasing the former. The former can only be criticized when it purports to serve the function of the latter. What is dangerous is asking NHST to provide methodological support beyond that for which it was designed. NHST can answer the question, "Is the null hypothesis plausible, or not?" It was not designed to answer the question, "Which of these two competing mathematical models is preferable in the way that it handles the trade-off between fit and parsimony?" In areas of behavioral science that are ready for more strongly confirmatory research – including the development of mathematical and statistical models that contain both causal and explanatory components (which are, of course, not entirely the same thing) – NHST is naturally expanded into the broader modeling epistemology. That expansion was the subject of my article. The earlier view of NHST as providing epistemological support for important but often separate and disparate individual findings is the topic of Robinson and Levin's criticism. Both stand effectively before criticism.

References

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

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

D'Onofrio, B. M., Goodnight, J. A., Van Hulle, C. A., Rodgers, J. L., Rathouz, P. J., Waldman, I. D., & Lahey, B. B. (2009). A quasi-experimental analysis of the association between family income and offspring conduct problems. *Journal of Abnormal Child Psychology*, 37, 415-439.

D'Onofrio, B. M., Turkheimer, E. N., Eaves, L. J., Corey, L. A., Berg, K., Solaas, M.H., & Emery, R. E. (2003). The role of the children of twins design in elucidating causal relations between parent characteristics and child outcomes. *Journal of Child Psychology and Psychiatry*, 44, 1130-1144.

D'Onofrio, B. M., Van Hulle, C. A., Waldman, I. D., Rodgers, J. L., Harden, K. P., Rathouz, P. J. & Lahey, B. B. (2008). Smoking during pregnancy and offspring externalizing problems: An exploration of genetic and environmental confounds. *Development and Psychopathology*, *20*, 139-164.

Fisher, R. A. (1935). *The design of experiments*.

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*, 311-339. Hillsdale, NJ: Erlbaum.

Gleick, J. (2003). *Isaac Newton*. New York: Pantheon Books.

Harden, K. P., D'Onofrio, B. M., Van Hulle, C., Turkheimer, E., Rodgers, J. L., Waldman, I. D., & Lahey, B. B. (2009). Population density and youth antisocial behavior. *Child Psychology and Psychiatry*, *50*, 999-1008.

Jaffee, S. R., Van Hulle, S., & Rodgers, J L. (2011). Effects of non-maternal care in the first three years on children's academic skills and behavioral functioning in childhood and early adolescence: A sibling comparison study. . *Child Development*, 84, 1076-1081.

MacCallum, R. C., & Tucker, L. R (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin*, *109*, 502-511.

Mendle, J., Harden, K. P., Turkheimer, E., Van Hulle, C. A., D 'Onofrio, B. M., Brooks-Bunn, J., Rodgers, J. L., Emery, R. E., & Lahey. B. B. (2009). Associations between father absence and age of first sexual intercourse. *Child Development*, *80*, 1463-1480.

Mill, J. S. (1843). *A system of logic*. London: John W. Parker.

Rodgers, J. L. (2001). What causes birth order-intelligence patterns? The admixture hypothesis, revived. *American Psychologist*, *56*, 505-510.

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

Rodgers, J. L., Bard, D., Johnson, A., D'Onofrio, B., & Miller, W. B. (2008). The Cross-Generational Mother-Daughter-Aunt-Niece Design: Establishing Validity of the MDAN Design with NLSY Fertility Variables. *Behavior Genetics*, *38*, 567-578.

Rodgers, J. L., Cleveland, H. H., van den Oord, E., & Rowe, D. C. (2000). Resolving the debate over birth order, family size, and intelligence. *American Psychologist*, *65*, 1-12.

Scarr, S. & McCartney, K. (1983). How People Make Their Own Environments: A Theory of Genotype → Environment Effects. *Child Development*, *54*, 424-435

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Sobel, D. (2000). *Galileo's daughter*. New York: Penguin Books.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*(*2*), 212-213.

Wichman, A., Rodgers, J. L., & MacCallum, R. C. (2006) A multilevel approach to the relationship between birth order and intelligence. *Personality and Social Psychology Bulletin*, *32*, 117-127.

Wichman, A., Rodgers, J. L., & MacCallum, R. C. (2007) Birth order has no effects on intelligence: A reply and extension of previous findings. *Personality and Social Psychology Bulletin*, *33*, 1195-2000.