# Null Science

## Psychology's statistical status quo draws fire

By BRUCE BOWER

Geoffrey R. Loftus, a psychologist at the University of Washington in Seattle, experiences "a certain angst" about his discipline these days. Over the past 30 years, he has built a successful scientific career and now edits the journal MEMORY AND COGNITION. From this lofty vantage point, Loftus sees with dismay a research landscape dotted with dense stands of conflicting data that strangle theoretical advances at their roots.

Findings reported by one set of investigators often fail to hold up in independent studies and rarely lead to breakthrough models of how minds work, Loftus remarks. This conceptual muddle, in his view, reflects a deeply flawed approach to doing science. Most researchers strap on a statistical straitjacket that offers enough flexibility to fire off publishable rounds of data but prevents anyone from heaving any thunderbolts of psychological insight.

"What we do, I sometimes feel, is akin to trying to build a violin using a stone mallet and chain saw," Loftus says. "The tool-to-task fit is not very good, and we wind up building a lot of poor-quality violins."

Loftus' musical analogy resonates deeply with many psychologists. In fact, a growing number openly criticize what they see as their field's statistical shortsightedness. Discontent has focused particularly on the practice known as null hypothesis testing, or significance testing.

In a significance test, the investigator typically gathers data to test the prediction that key experimental measures bear no relationship to one another. For example, such a null hypothesis might posit that the average amount and intensity of behavior problems in two groups of children occur independently of the presence or absence of marital distress in the youngsters' families.

Psychologists usually hope to reject the null hypothesis, based on a 5 percent or lower level of significance. Many see this level as indicating that they will be wrong no more than 5 percent of the time when they claim that two conditions are linked. Such significance levels signal to them that the measured variables probably do bear a relationship to one another.

At that point, researchers engage in a kind of 5 percent solution and proffer their favored explanations for a finding—say, by concluding that misbehavior mushrooms in children who grow up with battling parents.

Critics view this practice, and the assumptions underlying it, as unjustified. Significance testing simply establishes the probability of obtaining a certain data set, they argue, assuming from the start that the null hypothesis is true.

**"The shared secret of psychological researchers is that we don't take our own data too seriously when reaching theoretical judgments."**
**—John E. Richters**

Thus a 5 percent significance level in the study described above indicates to them that an errant statistical link between children's misbehavior rates and discord in their parents' marriages would occur only 1 in 20 times, if these variables indeed operate independently. From this perspective, significance levels—no matter how low they go—say nothing about the likelihood of any proposed explanation for statistically significant results.

"The shared secret of psychological researchers is that we don't take our own data too seriously when reaching theoretical judgments," contends John E. Richters, head of the disruptive disorders program at the National Institute of Mental Health in Rockville, Md.

"Even the brightest people use empirical research mainly to keep their careers going. When I talk to them in private, they express much more sophisticated views about mental functioning than what you see in their published reports."

Nonetheless, many of the same folks treat significance testing as a handy way to convert behavioral observations into objective scientific conclusions, notes psychologist Patrick E. Shrout of New York University. In complementary fashion, peer reviewers and editors at top journals routinely reject papers that do not boast significance levels of 5 percent or lower.

Growing discontent over the dominance of significance testing contributed to the American Psychological Association's formation of a 12-member task force on statistical inference. An initial 2-day meeting of the task force last December yielded a brief preliminary report.

The task force refuses to rule out significance testing, noting that it is one of many statistical methods available to researchers. Instead, it calls for descriptions of research data that go beyond assurances of having met or surpassed a set significance level.

For instance, the report notes that a study of a new treatment for agoraphobia, fear of open spaces, should include graphic displays of the sample population's range of responses to the treatment. This information would enable researchers to calculate the likely range of responses in the larger population of agoraphobia sufferers. It would also allow them to estimate the likelihood that the treatment would produce comparable results in repeated studies, assuming the treatment truly offers relief from agoraphobia.

The task force also recommends that researchers hone their theories in small pilot studies before unleashing significance tests on larger groups, strive for simplicity in research designs, and evaluate carefully the plausibility of data generated by computer programs.

Loftus sees little reason to hang onto null hypothesis testing, even in the limited way proposed by the APA task force. Researchers who hope to mold better theories from their data need to reach for other tools, such as metanalysis, he argues.

Metanalysis statistically combines a large number of studies of some phenomenon, such as sex differences in performing spatial tasks, and calculates the overall extent to which such differences occur. This approach has become popular in the past decade.

Another statistical strategy, known as planned comparisons, requires the researcher to predict how volunteers will respond to shifting experimental conditions. Such a study might examine whether the time needed to solve mental arithmetic problems rises proportionally after volunteers drink 0, 1, 2, 3, or 4 ounces of alcohol. The suitability, or fit, of this hypothesis in light of average response times at each level would then be calculated.

Substantial measurement errors exist in any research design, Loftus adds, because a multitude of uncontrollable influences

impinges on a person's nimbleness with numbers or whatever else the experimenters decide to study.

"Social scientists have embraced null hypothesis tests because they provide the appearance of objectivity," he contends. "But such objectivity is not, alas, sufficient for insight. [It provides] only the illusion of insight, which is worse than providing no insight at all."

For the past decade, Gerd Gigerenzer, director of the Max Planck Institute for Psychological Research in Munich, has made much the same argument in talks to psychologists at universities in the United States and Europe. Researchers who attend these lectures tend to abandon technical defenses of null hypothesis testing fairly quickly, Gigerenzer remarks. In private conversations with him, they call the practice a necessity for getting papers published and for reaching the academic Promised Land of tenure.

"Null hypotheses are set up and tested in an extremely mechanical way reminiscent of compulsive hand washing," Gigerenzer maintains. "There is widespread anxiety surrounding the exercise of informed personal judgment in matters of hypothesis testing. It's a question of intellectual morality."

This situation stems from what Gigerenzer calls an "inference revolution" that occurred in U.S. psychology between approximately 1940 and 1955. During that period, textbooks, universities, and journal editors jointly embraced a statistical process that mechanized the way in which researchers generated hypotheses from data.

The resultant inference apparatus consisted of an incoherent patchwork of procedures from warring schools of statistical thought, Gigerenzer argues. Significance testing, developed more than 60 years ago by British statistician Ronald A. Fisher as a means of identifying fertilizer compounds that produced the largest crop yields, was adopted as a basic experimental practice. Even Fisher noted in his writings, though, that the technique does not weigh the merits of alternatives to the null hypothesis.

To that end, psychology's hybrid statistics recruited several methods championed by Jerzy Neyman and Egon Pearson, contemporaries of Fisher who considered null hypothesis testing to be a meaningless exercise. Neyman and Pearson called for experimenters to specify at least two alternative hypotheses; to calculate how frequently they would expect to identify correctly a proposed hypothesis that is indeed true; and to conduct tests repeatedly with different random samples.

Finally, Gigerenzer holds, hybrid statistics encouraged researchers to assume that an instance of null hypothesis rejection illuminates the value of their or any other alternative explanation for the findings. Inspiration for this belief came from Thomas Bayes, who more than 200 years ago devised a formula to estimate an individual's expectation that an event will occur or a theory will prove true, based on several relevant observations (such as the base rate of a particular event).

Fisher, Neyman, and Pearson unanimously rejected Bayes' assumption that probability calculations apply to single events and instead focused on the long-term frequency of events.

**"Null hypotheses are set up and tested in an extremely mechanical way reminiscent of compulsive hand washing."**
**— Gerd Gigerenzer**

Denying such conflicts over the nature of probability allowed psychologists to institutionalize a single, "objective" form of statistical inquiry that in practice leaned most heavily on significance testing, according to Gigerenzer. Experimental ingenuity and informed choices of statistical tools withered, in his view, despite periodic pleas from several prominent psychologists over the past 40 years to give null hypothesis testing the boot.

"We should ban the ritualistic, mindless use of statistics, whether it revolves around significance testing or any other technique," Gigerenzer says.

Psychology's problems go beyond statistics, argues Richters. The discipline conducts research on the implicit assumption that the variables of interest in a study, such as child misbehavior and marital distress, maintain the same relationship to one another across all individuals. In other words, whatever's happening, it works the same way for everyone.

Null hypothesis testing quantifies this assumption, even if many psychologists privately hold more sophisticated views, Richters contends in the June DEVELOPMENT AND PSYCHOPATHOLOGY. In particular, significance tests assess links among only a few measures, which may respond to a multitude of other influences, and exalt group averages while mathematically obscuring individual differences.

The twists and turns of individual development, however, ensure that different sets of influences can yield similar outcomes, he holds. For instance, some kids may become unruly as a result of an impulsive temperament, poor self-esteem, a violent home life, and little support from adults; others may turn violent through a combination of high intelligence, resourcefulness, poor parental monitoring, and acceptance into a youth gang.

Conversely, similar traits can contribute to an array of consequences in people's lives. Consider extroversion. Depending on all sorts of personal qualities and external influences, an extroverted kid may end up pursuing a career in acting, sales, politics—perhaps anything other than writing.

The welter of influences on real-world behavior makes it relatively easy to nix the null and cook up a moderate correlation between almost any measures of interest, Richters contends. Teasing out the ingredients in varying recipes for a behavioral style or trait, such as a child's persistent misbehavior or lack of concentration, is quite another matter.

Richters plans to study a small number of children and their families to probe for sets of risk factors that, depending on the youngster, promote antisocial behavior. He will then study larger groups of children, each recruited on the basis of a particular array of risk factors for violent and criminal activity.

Despite having attracted a crescendo of criticism, significance tests have a place in psychology, remarks Yale University's Robert P. Abelson, a cochair of the APA task force. Successful demolition of a null hypothesis can enable researchers to defend surprising results against charges that random factors account for the data, he asserts.

Abelson advises researchers to apply null hypothesis tests to particularly powerful phenomena that seldom occur by chance. He acknowledges that single studies in many areas of psychology currently have little hope of corralling such effects.

Psychologist Jerome Kagan of Harvard University sees no need to ban null hypothesis tests. At the same time, he discerns no great value in them, aside from their usefulness in evaluating such highly controversial findings as those suggesting the existence of extrasensory perception.

"When you're theoretically barren, you rush to statistical methodology," Kagan says. "If you have a powerful theory that predicts something of importance, you don't need significance testing."

Psychology remains a young science, he adds. Conceptual leaps may ensue from the study of how various types of brain activity relate to performance on sophisticated mental and behavioral tests.

In the meantime, theories that make bold, precise predictions will encourage a more informed use of statistical methods by psychologists, Gigerenzer adds.

"Data without theory have a low life expectancy, like a baby without a parent," he says. "We need to develop more theoretical courage." □