May 10, 2010

Metric Mania

By JOHN ALLEN PAULOS

In the realm of public policy, we live in an age of numbers. To hold teachers accountable, we examine their students' test scores. To improve medical care, we quantify the effectiveness of different treatments. There is much to be said for such efforts, which are often backed by cutting-edge reformers. But do wehold an outsize belief in our ability to gauge complex phenomena, measure outcomes and come up with compelling numerical evidence? A well-known quotation usually attributed to Einstein is "Not everything that can be counted counts, and not everything that counts can be counted." I'd amend it to a less eloquent, more prosaic statement: Unless we know how things are counted, we don't know if it's wise to count on the numbers.

The problem isn't with statistical tests themselves but with what we do before and after we run them. First, we count if we can, but counting depends a great deal on previous assumptions about categorization. Consider, for example, the number of homeless people in Philadelphia, or the number of battered women in Atlanta, or the number of suicides in Denver. Is someone homeless if he's unemployed and living with his brother's family temporarily? Do we require that a women self-identify as battered to count her as such? If a person starts drinking day in and day out after a cancer diagnosis and dies from acute cirrhosis, did he kill himself? The answers to such questions significantly affect the count.

Second, after we've gathered some numbers relating to a phenomenon, we must reasonably aggregate them into some sort of recommendation or ranking. This is not easy. By appropriate choices of criteria, measurement protocols and weights, almost any desired outcome can be reached. Consider those ubiquitous articles with titles like "The 10 Friendliest Colleges" or "The 20 Most Lovable Neighborhoods." Such articles would be more than fluff if they answered critical questions. Are there good reasons the authors picked the criteria they did? Why did they weigh the criteria in the way they did? If changes in the criteria were made, would the rankings of the friendliest colleges or most lovable neighborhoods be vastly different?

Since the answer to the last question is usually yes, the problem of reasonable aggregation is no idle matter. Recently released e-mail from employees at the credit-rating agency Standard & Poor's indicated a wish to "discuss adjusting criteria" for rating securities and "massage the subprime and alt-A numbers to preserve market share." The criteria adjustments were analogous to the adjustments that would put an area of abandoned buildings onto the list of the most lovable neighborhoods.

These two basic procedures — counting and aggregating — have important implications for public policy. Consider the plan to evaluate the progress of New York City public schools inaugurated by the city a few years ago. While several criteria were used, much of a school's grade was determined by whether students' performance on standardized state tests showed annual improvement. This approach risked putting too much weight on essentially random fluctuations and induced schools to focus primarily on the topics on the tests. It also meant that the better schools could receive mediocre grades becausethey were already performing well and had little room for improvement. Conversely, poor schools could receive high grades by improving just a bit.

Medical researchers face similar problems when it comes to measuring effectiveness. Consider the temptation to use the five-year survival rate as the primary measure of a treatment for a particular disease. This seems quite

reasonable, and yet it's possible for the five-year survival rate for a disease in one region to be 100 percent and in a second region to be 0 percent, even if the latter region has an equally effective and cheaper approach.

This is an extreme and hypothetical situation, but it has real-world analogues. Suppose that whenever people contract the disease, they always get it in their mid-60s and live to the age of 75. In the first region, an early screening program detects such people in their 60s. Because these people live to age 75, the five-year survival rate is 100 percent. People in the second region are not screened and thus do not receive their diagnoses until symptoms develop in their early 70s, but they, too, die at 75, so their five-year survival rate is 0 percent. The laissez-faire approach thus yields the same results as the universal screening program, yet if five-year survival were the criterion for effectiveness, universal screening would be deemed the best practice.

Because so many criteria can be used to assess effectiveness — median or mean survival times, side effects, quality of life and the like — there is a case to be made against mandating that doctors follow what seems at any given time to be the best practice. Perhaps, as some have suggested, we should merely nudge them with gentle incentives. A comparable tentativeness may be appropriate when devising criteria for effective schools.

Arrow's Theorem, a famous result in mathematical economics, essentially states that no voting system satisfying certain minimal conditions can be guaranteed to always yield a fair or reasonable aggregation of the voters' rankings of several candidates. A squishier analogue for the field of social measurement would say something like this: No method of measuring a societal phenomenon satisfying certain minimal conditions exists that can't be second-guessed, deconstructed, cheated, rejected or replaced. This doesn't mean we shouldn't be counting — but it does mean we should do so with as much care and wisdom as we can muster.

John Allen Paulos, a professor of mathematics at Temple University, is the author most recently of "Irreligion."