Classification Error in Schooling Practice: The Impact of the "False Positive"

by Edward G. Rozycki

> There are three types of lies—lies, damn lies, and statistics.
>
> —*Mark Twain*

A Decision Simulation

Over several years of university classes I have asked my students, mostly principals and superintendents in a doctoral program, whether they would consider using random drug testing in their schools under the following conditions:

1. The drug test (for brevity's sake, let us call it DT) would correctly identify nineteen out of twenty "true drug abusers" as positives.
2. DT would correctly identify nineteen of twenty "true non-abusers" as negatives.
3. Suppose that, at most, the "true drug abusers" constitute 5 percent of a total of ten thousand K–12 students;
4. The costs of the first use of DT for each student would be borne by the manufacturer; only reapplications of DT, if any, would have to be paid for at the rate of fifty dollars each.
5. To intrude minimally on the educational program, but maintain a hoped-for deterrent presence, only one or at most a few students would be selected each day by lottery to be tested by DT.

Over the years the considerations these experienced educators would entertain, and the conclusions they would reach, were substantially the same: the great majority would say yes to implementing a program of random testing, despite much about it they might find objectionable.

After this first stage of reaching a decision, I would caution them that DT would generate "false positives"—students mistakenly identified by DT as abusers, for whom such a determination might have severe social consequences. With reluctance the educators would insist they

would implement the use of DT, but not push for criminal charges against these young people; rather, they would provide counseling and therapeutic support to the test-identified users.

These politically sensitive principals and superintendents thought that drug abuse by the postulated five hundred students would alarm important members of their communities and justify the risk of falsely identifying some students as positive. As administrators accustomed to tight budgets, they found the no-cost offer particularly enticing: a fifty-dollar-per-student savings.

I pressed them on what they would do to handle possible false positives. After some discussion the consensus was this: have students identified as users retake the test. The test is 95 percent accurate, they argued, so if a student tested twice as a user, the probability was (0.95 x 0.95 = .9025) that the student really was a user.[1] Few retests would be needed: 5 percent, the incidence rate, was a number that kept coming up during the deliberations. Pressed about considerations of nurture and humanity, they forcefully and repeatedly emphasized that their investigation with DT was not intended to be punitive.

Disregard for False Positives

Though not recognizing its importance at the time, some of the more philosophically inclined class members would press me about what the terms "true users" or "true non-users" meant if DT was needed to identify users. What was supposed to be the difference between a "true" user and a DT-identified user, or between a "true" non-user and a DT-identified non-user?

I would reply that the prevalence of drug abuse, the comparison of "ideal" numbers of "true users" to "true non-users," was often obtained by postulation (as in this exercise) or estimated from smaller samples of students obtained by procedures that were too costly or time-consuming to be applied to the whole population. On occasion—more frequently than generally acknowledged—the incidence rate would even be conjured up by guess, intuition, or tradition.

The critical question for the practitioner is this: given that DT has identified a student as a drug user, what is the likelihood that that student is abusing controlled substances? Thinking the only consideration was the accuracy of DT in identifying users, my educator-students were wildly wrong in their estimations—and utterly surprised by the correct answer, which completely reversed the conclusions they had previously reached.

Educators are not unique in their disregard of the effects of classification error. Twenty-five years ago, David M. Eddy found that in trying to evaluate a patient's symptoms, physicians assumed that the relative commonness of a disease in a population should not be used to estimate the

probability that a particular patient has the disease.[2] Gerd Gigerenzer, more recently, reviewed a sequence of examples to show how, even in law, critical mistakes in reasoning about classification have been far from uncommon.[3]

I explained to my educators that they were overlooking a very important thing, prevalence: the relative size of the user group to the entire population. Test accuracy alone was not sufficient to decide the likelihood that test-identified users were real users. Prevalence of abuse has a major influence.

The Proof

Let's begin by reconsidering how DT sorts the students.

a. We have 10,000 K–12 students, of whom 5 percent, or five hundred, are drug abusers. There are 9,500 non-abusers.
b. DT will correctly identify ninety-five of one hundred "true" abusers—that is, four hundred seventy-five of five hundred—as abusers. This group is traditionally called the "true positives" because they are truly abusers and will test "positive" on DT.
c. The other twenty-five true abusers are misclassified as non-users. They are "false negatives" because they will test "negative" on the DT, which is a false characterization of their true abuser status.
d. DT will also correctly identify 95 of 100 "true" non-abusers—that is, 9,025 of the 9,500 "true" non-abusers—as non-abusers. This group is traditionally called the "true negatives" because they are truly non-abusers and will test negative, non-user, on DT.
e. The other 475 true non-abusers are misclassified as users. They are "false positives" because they will test positive on the DT, which is a false characterization of their true non-abuser status.

But the number of non-users is vastly greater than the number of users. This is why prevalence has an effect on the probability of correctly identifying a "true" user. As far as DT is concerned, true users are indistinguishable from and will be confounded with false positives.

Let's do the arithmetic.

The Likelihood of Error

DT will split each group, abuser and non-user, in two in the proportion of five to ninety-five. See the chart on page 196.

Our problem now is to decide whether a student identified as an abuser by DT is a "true" abuser, or a non-user misidentified by DT as an abuser. Members of the two positive groups are indistinguishable. The probability that we have a true positive, given that DT has identified a

	Of "True" Abusers: 500	Of True Non-users: 9,500
Test Positive	475 true positives	475 false positives
Test Negative	25 false negatives	9,025 true negatives

student as positive, is the number of true positives divided by the number of all test positives: that is, 475/(475 + 475), or 1/2.

In other words, we have a fifty-fifty chance of misidentification.[4] My students found this objectionably high. Besides, it would require us to re-test each student at the higher cost. That clinched it: they invariably reversed their decision to implement random testing for the specified situation.[5]

Value Added Issues

This "false positive" effect negatively impacts many other areas of educational decision-making. Lack of space permits a quick review of only one more: value-added or "growth-based" assessment.

Many educational authorities realize that it is unfair to compare schools merely by achievement levels, since those already achieving at high levels can pass any reasonable criterion of success with no effort while those of very low achievement can raise test scores tremendously but still miss the criterion. These authorities have suggested that "value added" or "growth" criteria be the basis of school assessment.[6]

However, grade-level placement in public schools, especially, is a very haphazard process. Parental demands can exert an influence substantially independent of any placement tests, which tend to be given seldom, anyway. If we suppose that each grade level in a particular school requires certain preparatory skills, dispositions, and knowledge for success in the coming year, we can be quite certain that our haphazard admission processes will allow in quite a few "false positives," that is, students who at the point of admission appear no less capable—they walk, they talk, they can fog a mirror—than the students who possess the skills, dispositions, and knowledge to succeed in the class.

Assessing growth thus threatens us with not only unfairness, but statistical damn lies as well, particularly if "true" readiness is based on little more than guess, intuition, or tradition.

Notes

1. Clever. But merely multiplying to get the probability of a repeated trial assumes no error.

2. David M. Eddy, "Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities," in *Judgment under Uncertainty: Heuristics and Biases,* ed. Daniel

Kahneman, Paul Slovic, and Amos Tversky, 249–267 (Cambridge, N.Y.: Cambridge University Press, 1982). See especially pp. 258–259.

3. See Gerd Gigerenzer, "Ecological Intelligence," in *Adaptive Thinking: Rationality in the Real World,* Gerd Gigerenzer, 59–76 (New York: Oxford University Press, 2000). For an interesting exposition on statistical folderol in the drug industry, see R. Brian Attig and Alison Clabaugh, "Clinical Trials and Statistical Tribulations," *Applied Clinical Trials* (February 2008): 42–46, available at <http://actmagazine.findpharma.com/appliedclinicaltrials>.

4. For a well-constructed lesson on this issue, see Jerry Johnson, *Medical Testing,* available at <http://www.math.dartmouth.edu/~mqed/UNR/MedicalTesting/MedicalTesting.phtml>.

See also "How to Improve Bayesian Reasoning without Instruction," in *Adaptive Thinking,* Gigerenzer, 92–123.

5. A member of one of the classes that participated in the simulation wrote a paper expressing his perspective on the issue: see William J. McIlmoyle, "Random Drug Tests for High School Athletes?" available at <http://muse.widener.edu/~egrozyck/EDControversy/McIlmoyleF01.html>.

6. See Ted Hershberg, "Follow Growth, Not Achievement," *Philadelphia Inquirer,* March 3, 2008.

Edward G. Rozycki, Ed.D., is a twenty-five-year veteran of the school district of Philadelphia. He is an associate professor of education at Widener University, Widener, Pennsylvania.