From **Brunelli, Lina & Cicchitelli, Giuseppe (editors). Proceedings of the First Scientific Meeting (of the IASE).** Università di Perugia (Italy), 1994. Pages 199-211. This pdf file is from the IASE website at http://www.stat.auckland.nz/~iase/publications/proc1993. Copies of the complete Proceedings are available for 10 Euros from the ISI (International Statistical Institute). See http://isi.cbs.nl/sale-iase.htm for details.

UNDERSTANDING PROBABILITY AND STATISTICAL INFERENCE THROUGH RESAMPLING

Clifford Konold
Scientific Reasoning Research Institute
Hasbrouck Laboratory, University of Massachusetts
Amherst, MA 01003, USA

1. Introduction

During the past four years, I have been developing *curricula* and computer software for teaching probability and data analysis at the introductory high-school and college level. The approach I've taken emphasizes the use of real data, where "telling a story" takes priority over testing hypotheses, and in which mathematical formalism is kept to a minimum (see Cobb, 1992; Scheaffer, 1990; Watkins *et al.*, 1992).

A major question I have considered is how probability ought to be integrated into this new data-rich curriculum. There are two major reasons for keeping probability in the data-analysis (or statistics) curriculum. First, at some point in the process of constructing theories that account for patterns in data, it is important that students consider alternative explanations. Among these is the possibility that some outcome of interest resulted from chance. Second, probability is an important concept in its own right (Falk and Konold, 1992). It comprises a world view and should not be viewed a necessary evil that must be faced if students are to understand statistical inference.

In deciding how to relate probability and data analysis, I have adopted an approach Julian Simon began advocating in the late 1960s. Simon describes his approach as having grown out of his frustration watching graduate students do silly things when trying to test a statistical hypothesis (Simon and Bruce, 1991). He began designing physical experiments from which he hoped they could build up sound probabilistic understandings. These eventually developed into a resampling approach that promised a more intuitive take on probability and data analysis, and which made the connections between the two fields more apparent. Of course, Simon didn't invent Monte Carlo methods, nor the randomization tests he would come to employ, but he was among the first to see their educational potential, and long before the computer was widely available.

Rather than elaborate Simon's argument here, I briefly describe two software tools we've developed, highlighting aspects that emphasize the relation between probability and data analysis. I also report some results from our primary test site, a high school in Holyoke, Massachusetts.

2. Modeling a problem with Prob Sim

Most educational probability-simulation software comprise several ready-made models (e.g., coin model, die model). Students load the appropriate model, draw samples, and then see results displayed. The software thus offers empirical demonstrations of various facts and principles, such as the law of large numbers and the binomial distribution. The software we have designed, "Prob Sim", includes no ready-made models. Rather, the student must build the model, specifying the appropriate sampling procedure and analyses in order to estimate the probability of some event. The process of building a simulation model is at least as important as, if not more important than, drawing the appropriate conclusions from the results. To illustrate, I'll describe one of our activities entitled "LAPD" (see Konold, 1993, for another example).

Students begin by reading excerpts from *The New York Times* account (March 18, 1991) of the beating of Rodney King by officers of the Los Angeles Police Department. After reporting that "at least 15 officers in patrol cars converged on King", the article broaches one of the issues that made this incident explosive: "In what other police officers called a chance deployment, all the pursuing officers were white. The force, which numbers about 8,300, is 14% black".

Students are asked to build a model of this situation to estimate the probability of finding no blacks in a random sample of 16 officers. This problem has generated lively discussions in our test sites. When students care deeply about a problem, they are more willing to persist through difficulties. Moreover, they learn that application of probability theory is not limited to rolling dice and blindly drawing socks out of dresser drawers.

2.1 *Building a model*

I demonstrate below the various stages through which students progress in modeling this situation, illustrating the steps with screen shots from Prob Sim. In Fig. 1, a mixer has been filled according to the information provided in the article. There are a total of 8,300 elements, 1,162 of them labeled *B* (Black) and 7,138 labeled *N* (Not black). The non-replacement option has been selected to preclude the possibility of having the same

element (officer) in a sample more than once.

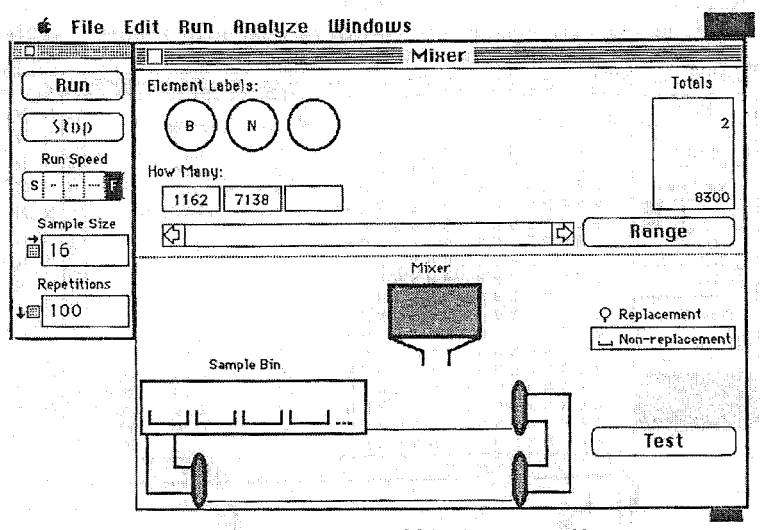

Figure 1. *A sampling model for the LAPD problem*

The Run controls on the far left shows the sample size set at 16, and number of repetitions set (somewhat arbitrarily) at 100. After the Run button is pressed, the computer draws 16 elements from the mixer without replacement, repeating this a total of 100 times. This is analogous to looking at 100 occasions when 16 randomly-selected officers appeared on a crime scene. The sampling process is animated in the lower part of the Mixer window. The results of each repetition are displayed in a Data Record window (not shown).

After the block of 100 repetitions has been drawn, results can be analyzed as shown in Fig. 2. The Analysis window in this case shows the number (and proportion) of occurrences of each of the 17 possible unordered outcomes. Ten percent of the samples had 0 *B*'s. Thus, a first estimate of the probability that a "chance deployment" would include no black police officers is .10.

2.2 *Repeating the experiment*

In many textbook examples of simulation, the experimental component ends here. But it is important not only to estimate the probability of some event, but to get some sense of the range (or variability) in that estimate given the number of repetitions. Indeed, the notion of chance is not

apparent in a probability value per se, but in the variability of result over

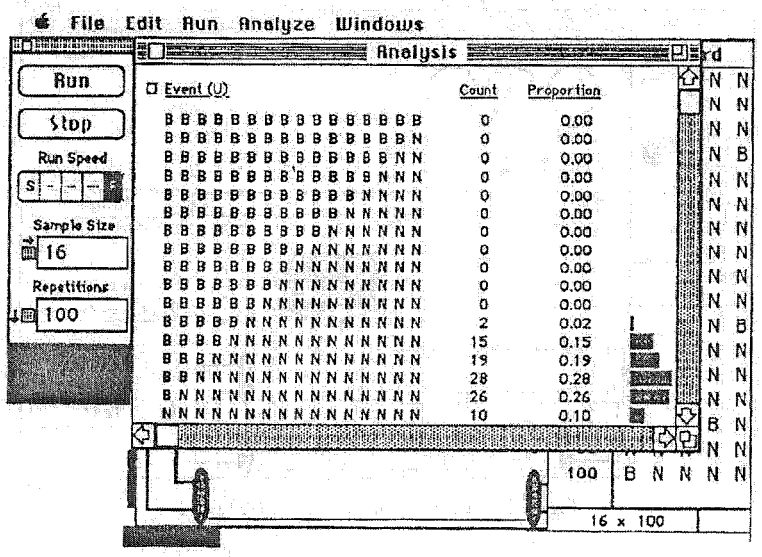

Figure 2. *The analysis window showing the results of 100 repetitions*

replications of an experiment. To replicate an experiment in Prob Sim, the student has only to press the run button again. Another random sample is drawn, and the Analysis window updates to show the new results.

Students replicate the experiment about 10 times, plotting the results on a line plot which helps to emphasize the variability in the process. They then pool their results to come up with a final probability estimate. For the *10 results plotted in Fig. 3, the pooled estimate of P (0 blacks) is .109.*

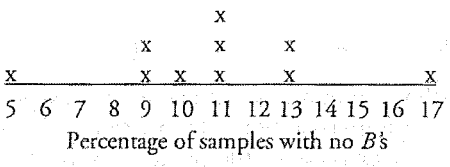

Figure 3. *Line plot of result of 10 replications*

In comparing results obtained in other groups, students discover that there is considerably less variability among the pooled estimates of the different groups than there is in the 10 estimates they got in their own group. Students thus see empirical demonstrations of the law of large numbers long before encountering a formal description. I've used such tasks in short two-week workshops in which I've never formally introduced the law of large numbers. According to performance on post-test items, these students show significant gains in understanding the effects of sample size on the variability of a sampling distribution.

When they can, students also work out theoretical probabilities. Through comparing these to their empirical values, they gain confidence in the simulation process as well as an appreciation for the power of mathematical theory and formalism.

2.3 *Questioning assumptions and making decisions*

In the process of model building, various assumptions are made. Students are asked, for example, to consider how the fact that police officers often arrive in pairs might affect the probability in question if partners tend to be of the same race. If they can, they build and run models consistent with these new assumptions and discuss the implications of their results. It is through addressing the assumptions incorporated into any model that students begin to understand what the process of modeling involves. In order to determine the predictive value of the simulation results, students must decide if and how various considerations affect the real situation. In some cases they can alter the model to take into account an additional factor (e.g., have police arrive on the scene in same-race pairs). If they can't do this, they can often predict the direction of biases introduced by their simplified model. Through this process of comparing a model to the target situation, students come to realize that though they can't avoid simplifying assumptions, the more aware they are of the limits of a particular model, the more informative the data from that model becomes.

At the end of the lab, students discuss the problem of deciding, based on their information, whether it was just chance that none of the responding officers were black. And, of course, the probability of the event is only one of the relevant pieces of information. More important are the subsequent testimony of the officers and the sensibility of any conspiracy theories that explain why black officers weren't on the scene. In other words, the probability of some event is never the only or last word in making some decision related to that event.

Such discussions seldom occur until instruction moves beyond the realm of coins and dice. Prob Sim is an important tool in supporting discussion of

this type, because it can be used to model engaging problems too complex for students to tackle formally. Prob Sim's simplicity, and the speed with which sufficient data can be generated, gives students the time to design a sampling model, collect adequate data, draw conclusions, and discuss implications of the findings. In the next section, I show how we build on ideas introduced in modeling probabilistic situations when we move on to data analysis.

3. Data analysis using DataScope

We designed a data-analysis program called "DataScope" for use in introductory courses stressing exploratory data analysis in which students work with real data. Our primary objective is for students to learn basic data-analysis techniques for exploring relationships among various variables. By using multiple-variable data sets, we give students the opportunity to explore questions of particular interest to them. DataScope encourages students to make initial judgments of relationship by visually comparing plots. This is demonstrated below using data obtained from a questionnaire administered to 84 students in two high schools in western Massachusetts: Amherst Regional and Holyoke High. Amherst is a small college town, while Holyoke is a larger, industrial city. The information collected on each student included gender, age, birth order, family size, marital status of parents, religious activity, rating of school performance, educational level of parents, curfew times, working hours and wages, and time spent on homework. Students spend about two weeks exploring various questions in this data set, among them the question of whether holding an after-school job adversely affects school performance.

DataScope encourages exploration by allowing students to form subgroups of some variable based on the values of some other, presumably related, variable. For example, Fig. 4 shows the box plots for hours of homework (HWHRS) and hours spent working (JOBHRS) for Amherst and Holyoke students (in the case of JOBHRS at Amherst, the median is the same as the 1*st* quartile, as indicated by the double-thick line). The number of cases in each box plot is displayed to the right of the plot. This is important to include as students will frequently draw conclusions without considering the number of cases represented by each box plot. The plots below suggest that Amherst students spend more time on homework than working a part-time job, and that the opposite is true of Holyoke students. *This trend is consistent with commonly-held stereotypes of the two towns.*

It is tempting to conclude from Fig. 4 that those with jobs spend less time on homework than those without jobs. However, when HWHRS are

"grouped" by the categorical variable JOB, students discover the reverse appears to be true – students with a job studied an average of three hours-per-week more than those without a job.

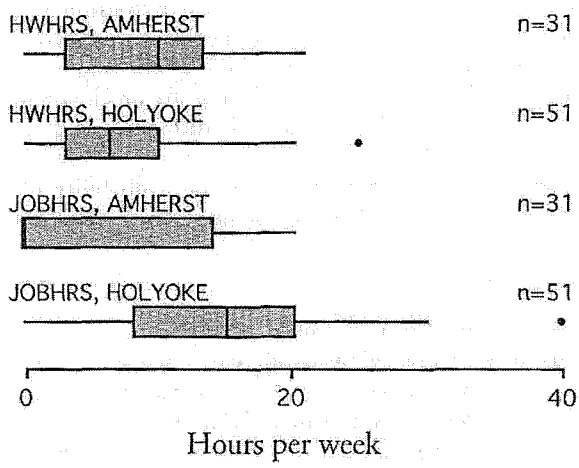

Figure 4. *Box plots of homework hours and job hours for Amherst and Holyoke students*

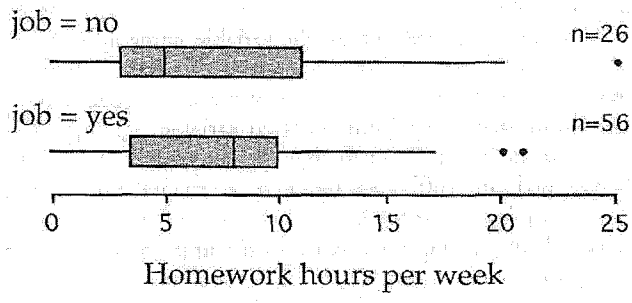

Figure 5. *Box plots of homework hours for student with and without jobs*

At this point, students reconsider their expectations and develop theories that might explain these data. One possibility that those who work also study more because they have learned to effectively manage their time. Or maybe the students with jobs are older and have more homework assigned. Some of these explanations can be investigated by looking at other variables

in the data set. However, among the explanations to consider is the possibility that the difference was due to the "luck of the draw" – that with a better or larger sample, no difference would be found. In my experience, students do not spontaneously raise this possibility, even when they have been using resampling techniques, as described above to estimate probabilities. They will judge differences between two medians as important in one case, and unimportant in another, apparently making the judgment based on the distance between the two medians, as it appears on the computer screen. They do not spontaneously evaluate this difference with respect to the variability (i.e., to the IQRs as shown in the box plots). Therefore, before showing them how we might determine the probability of a difference occurring by chance, I typically must remind them that this is a possibility.

3.1 *Randomization tests in DataScope*

Below is a demonstration of how DataScope is used to estimate the one-tailed probability of observing a difference at least as large as the one observed in the sample above. The method is based on the randomization procedure as originally developed by Fisher (see Barbella *et al.*, 1990). This involves randomly reordering one of the variables (without replacement) to estimate the probability of the observed difference under the null hypothesis. With DataScope, the student can first do this "manually", to develop a sense of what the computer is doing. A "reorder" command randomly reorders the values of one of the selected variables (in this case, the values for HWHRS). That is, the values of HWHRS are randomly reassigned to cases. Once reassigned, the variable name appears in the data table with the symbol ® on both sides to remind the student that the column has been randomized (another command will restore the original order). A box plot of this randomly ordered variable, again grouped by the job variable, can be viewed. Given that the values of HWHRS have been randomly assigned, any difference between the medians of the two groups (with or without jobs) is due to chance. In the example shown in Fig. 6, this difference is -1 (sub-tracting the median of the upper plot from the median of the lower plot).

Once students understand the procedure, the computer can be instructed to repeatedly reorder the variable and compute the difference between medians of the job and no-job groups, recording these in a new data table as illustrated in Fig. 7. In this case the computer has been instructed to draw 100 random samples. The first few differences that were obtained are shown in Fig. 7 in the "Resampling" data table. The first value is the observed difference, -3. In the background you can see the variables

selected (V1 and G1) in the primary data table. The values in HWHRS are being randomly reordered.

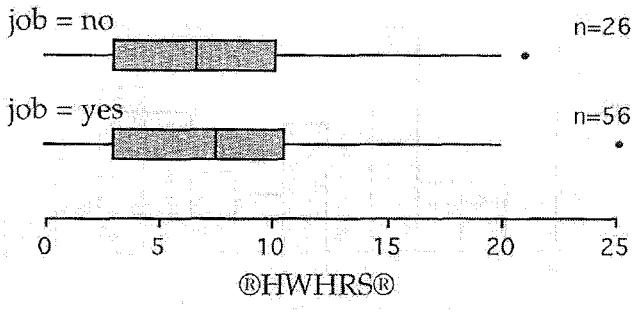

Figure 6. *Homework hours randomly assigned to job and no job groups*

HS Survey 90

PARENTS	COLLEGE	$ONYOU	®HWHRS®	JOB
			V1	G1
			10	no
			7	yes
			15	yes
			10	yes
			10	no
			10	yes
			3	yes
			11	yes
			20	yes
			8	yes
			15	no
			15	yes
separated	undecided	10	15	yes

Resampling of HWHRS

	D.M.
1	-3
2	-3
3	-1.5
4	-2.5
5	-1.5
6	-5
7	0
8	0

Figure 7. *Table of differences between median HWHRS for job and no-job groups in successive resampling runs.*

After the 100 random differences have been obtained, the results can be displayed in a histogram as shown in Fig. 8. In this instance, 25 of the samples had a difference at least as large as -3. Thus, an estimate of the one-tailed p value is .25. Additional repetitions could be conducted for a more precise estimate. This same procedure is used in DataScope to test the

statistical significance of a value of *r*, and of frequency counts in a 2 X 2 table.

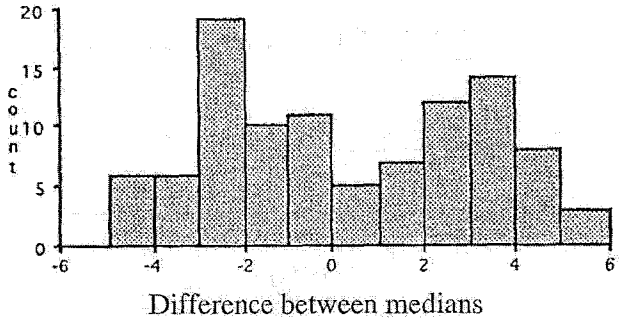

Figure 8. *Sampling histogram displaying results of 100 resamplings*

4. Educational outcomes

I have found, as have Simon and Bruce (1991), that students are enthusiastic about probability and statistical inference when approached through resampling. But, do students using this approach learn more than they do in a traditional course? Simon *et al.* (1976) compared student performance in courses taught using resampling vs. conventional methods. Given problems that could be solved using methods taught in either course, students in courses using the resampling approach consistently outscored students using the traditional approach.

Many or most students who take an introductory course will never need to conduct a statistical test or determine a probability precisely. They do, however, as members of a complex and increasingly technological society, require a basic understanding of uncertainty and the savvy to evaluate "research" claims in the mass media. Accordingly, though students should be able to solve problems using methods they have been taught, it is even more important that they understand basic concepts which underlie these methods. Konold and Garfield (1993) have developed items to assess understanding of these basic ideas. Below is one of our problems, adapted from an item in Falk (1993, p. 111).

The Springfield Meteorological Center wanted to determine the accuracy of their weather forecasts. They searched their records for those days when the forecaster had reported a 70% chance of rain. They compared these forecasts to records of whether or not it actually rained on those particular days.

The forecast of 70% chance of rain can be considered *very* accurate if it rained on:

a. 95% - 100% of those days.
b. 85% - 94% of those days.
c. 75% - 84% of those days.
d. 65% - 74% of those days.
e. 55% - 64% of those days.

This problem was designed to assess whether a student understands that a probability is a *quantitative* measure of uncertainty (or frequency of occurrence). It is difficult to imagine how, without this understanding, a student could correctly interpret a *p* value, e.g., that if we estimate the probability in the LAPD problem as .10 this means that roughly 10% of the time when a random sample of 16 officers is selected, there will be no blacks among them.

Fig. 9 shows the frequency of responses to the various options in the weather problem by 199 students we have administered this item to before instruction. Forty-three students (32%) selected the correct range 65-74, which includes within it the normative value of 70%. The majority of students (36%), selected the highest range, suggesting that they expect rain to occur nearly all of the time when it has been forecast with a 70% probability. This answer may be based on an approach to uncertainty that is prevalent among adults across a range of problems. Because it is based on the belief that the objective in probabilistic situations is to predict the *outcome* of a *single* trial, I have referred to this as the "outcome approach" (Konold, 1989).

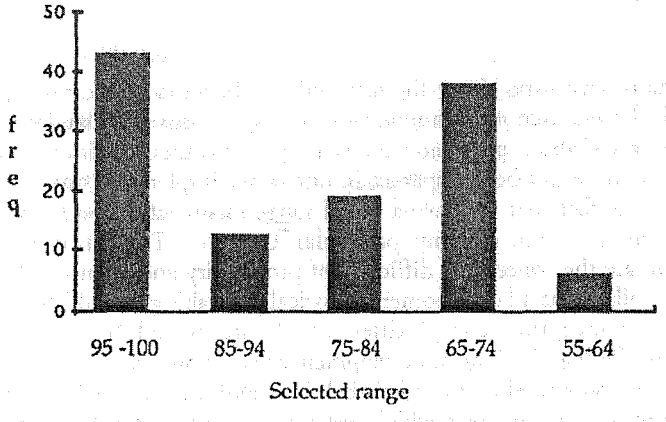

Figure 9. *Frequency of responses before instruction of 199 students to the weather problem*

One of my instructional objectives, therefore, is for students to realize that a probability value typically tells us little or nothing about results in the short run, but a great deal about results in the long run. Fig. 10 compares the results on this same problem before (black) and after (gray) instruction. Correct responses increased only 6% with instruction. The results are similar across the majority of our assessment items. At a deeper level, many students after instruction using resampling appear unaware of the fundamental nature of probability and data analysis.

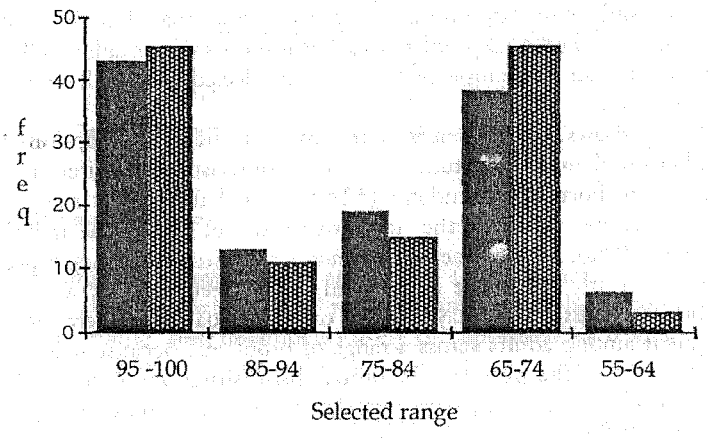

Figure 10. *Frequency of responses before (black) and after (gray) instruction of 199 students to the weather problem.*

I remain optimistic about the instructional benefits of the resampling approach. I have used it in individual tutoring sessions with students at a variety of levels who express momentary insight into the logic it employs. If the benefits have not been apparent in our larger implementations, it may be due to the fact that we have a broad range of objectives and have not devoted enough time to any particular concept. Too, it is easy to underestimate the conceptual difficulty of probability and chance. The fact that probability was a late bloomer, historically speaking, should help alert us to the conceptual complexities which are belied by the simple formalisms. Perhaps looking at development over a single course is too small a unit of analysis in the case of probability, and we should be thinking about series of courses over which we can expect to effect and observe conceptual change.

Bibliography

Barbella P., Denby L. and Landwehr J.M. (1990), Beyond exploratory data analysis: The randomization test, *Mathematics Teacher*, 83, pp. 144-149.

Cobb G.W. (1992), Teaching statistics: More data, less lecturing, in L. Steen (ed.), *Heeding the Call for Change*, (MMA Notes # 22, pp. 3-43), Mathematical Association of America.

Falk R. (1993), *Understanding probability and statistics: Problems involving fundamental concepts*, Wellesley, MA.: AK Peters (to appear).

Falk R. and Konold C. (1992), The psychology of learning probability, in F.S. Gordon and S.P. Gordon (eds.), *Statistics for the twenty-first century* (MMA Notes # 26, pp. 151-164), Mathematical Association of America.

Konold C. (1993), Teaching probability through modeling real problems, *Mathematics Teacher* (to appear).

Konold C. (1989), Informal conceptions of probability, *Cognition and Instruction*, 6, pp. 59-98.

Konold C. and Garfield J. (1993), *Statistical Reasoning Assessment. Part 1: Intuitive Thinking*, Scientific Reasoning Research Institute, University of Massachusetts, Amherst.

Scheaffer R. L. (1990), Toward a more quantitatively literate citizenry, *The American Statistician*, 44(1), pp. 2-3.

Simon J.L. and Bruce P. (1991), Resampling: A tool for everyday statistical work, *Chance: New Directions for Statistics and Computing*, 4(1), pp. 22-32.

Simon J.T., Atkinson D.T. and Shevokas C. (1976), Probability and statistics: Experimental results of a radically different teaching method, *American Mathematical Monthly*, 83, pp. 733-739.

Watkins A., Burrill G., Landwehr J. and Schaeffer R. (1992), Remedial Statistics?: The implications for colleges of the changing secondary school curriculum, in F.S. Gordon and S.P. Gordon (eds.), *Statistics for the twenty-first century* (MMA Notes # 26, pp 45-55), Mathematical Association of America.

Note

The curriculum materials and research described in this article were supported by a grant from the National Science Foundation (Grant #MDR-8954626). The opinions expressed here are the author's and not necessarily those of the Foundation.