

An Exercise Where Students Demonstrate the Meaning of “Not Statistically Significantly Different”

Charlotte Rappe Zales Joseph C. Colosi

Statistical analysis is a fundamental tool of science, but students find it difficult to grasp (Bradstreet 1996). In addition, many students find statistics a source of great anxiety (Schacht & Stewart 1992). In science lab classes, students learn about the scientific process by running experiments, making measurements, and analyzing their results. As part of the learning process, students learn to use statistics, such as *t*-test and Chi square, to analyze the results of experiments. However, they often do not comprehend the statistical concepts that make sense of their results. Students may be able to apply the statistical procedure and make the correct decision without having any idea what their statements mean. For example, when students must decide if observed differences are caused by the treatment effect or are just due to chance based on sampling, they cannot appreciate their experimental results without understanding how sampling effects can result in different measured values. This problem is especially a concern with computerized statistics packages with which students enter the data and magically receive “the answer.”

The essential concept that often eludes students is the near certainty that two samples taken from the same population will have different means. If students do not realize this fact, they cannot understand how we could declare that different means of two or more treatments could be considered

“not statistically significantly different.” Here we present a simple, quick exercise where students see from their own efforts what this statement means. Additionally, this exercise demonstrates the bell curve distribution, the relationship of a sample to the population it is supposed to estimate, the concept of probability, the effect of increasing accuracy and precision with increasing sample sizes, and the *t*-test. All of these concepts are demonstrated in a nontechnical manner based on measurements the students make and values they calculate. This exercise has worked well with freshman nursing majors taking microbiology, freshman biology majors in introductory biology, upper level nonmajors in a nonmajors biology class, and M.Ed. students taking a research course.

The Exercise

The Sample

A large forsythia shrub on campus provides the leaves that are the material for the exercise. Without an explanation of the theory about leaf size on different parts of plants, students are asked to pick five leaves at random from the north side and five leaves from the south side of the shrub, keeping the two sets of leaves separate. A sketch on the chalkboard before leaf collection shows students how to measure the leaf length from the tip of the blade to the end of the petiole (Figure 1).

The Frequency Distribution

Using one set of leaves, e.g. the north leaves, we ask the students for the shortest and longest values, explaining that we want to establish about 10 length intervals which span the range of values. For example, if

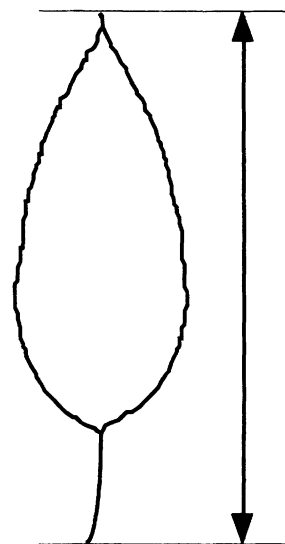


Figure 1. Diagram of a forsythia leaf, which illustrates measuring the leaf length from the tip of the blade to the end of the petiole.

the shortest and longest leaves collected are 4 and 11 cm respectively, length intervals of 1 cm each are established. We survey the students orally for the number of leaves each has in each length interval. We display the results as a frequency distribution on the chalkboard and then construct a frequency histogram on an overhead transparency (Figure 2). In each of the 10 times we have done this exercise, the histogram has had a bell curve distribution (Figure 3a). To guide students, we ask the following set of questions:

Questions: Are the leaves evenly distributed among length intervals? Where do most of the values fall in the distribution?

Is there a pattern to the distribution of values?

Charlotte Rappe Zales is Assistant Professor of Education in the Department of Educational Leadership and Administration at Immaculata College, Immaculata, PA 19345. **Joseph C. Colosi** is Associate Professor of Biology in the Department of Natural Sciences at Allentown College of St. Francis de Sales, Center Valley, PA 18034.

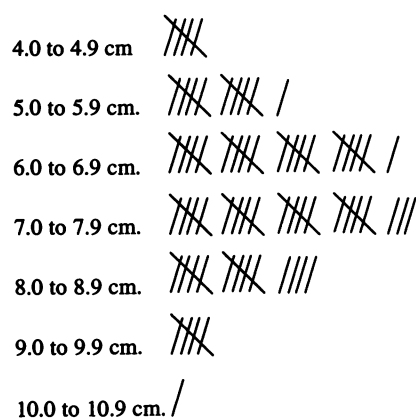


Figure 2. Frequency distribution of number of leaves in each length interval from the north side of the forsythia shrub.

These questions are answered by observing the graph. The leaves are not evenly distributed; most have lengths in the middle of the range forming a bell curve. We tell them that a bell curve distribution is the most common one observed for biological measurements.

The Effect of Sample Size

To illustrate how a sample of leaves estimates the population of leaves on the north side of the shrub, students calculate the mean of their samples of five north leaves, and we calculate the mean of the entire north group. We prepare a frequency table of the means and write the totals corresponding to each length interval above the bars on the histogram, labeled $n=5$ (Figure 3b).

Questions: How is the distribution of the samples of size $n=5$ different from the distribution of the individual leaf lengths?

How well do the means of the $n=5$ samples estimate the mean of the north group?

Students notice that more of the samples are clustered toward the center.

Then, we pair students, and each pair calculates the average length of the 10 north leaves that they have together. Means for $n=10$ are added to the histogram (Figure 3c). Next, we form groups of four students, who calculate the average length of the 20 north leaves that they have together, and $n=20$ means are added to the histogram (Figure 3d). If class size permits, groups of eight students calculate the average length of the 40 north leaves that they have as a group,

and $n=40$ means are recorded (Figure 3e). We discuss changes in the distribution of the samples after each calculation. We calculate the percent of matches of the calculated means with the mean of the north group.

Questions: What is the trend in distribution of sample means as the sample size gets larger?

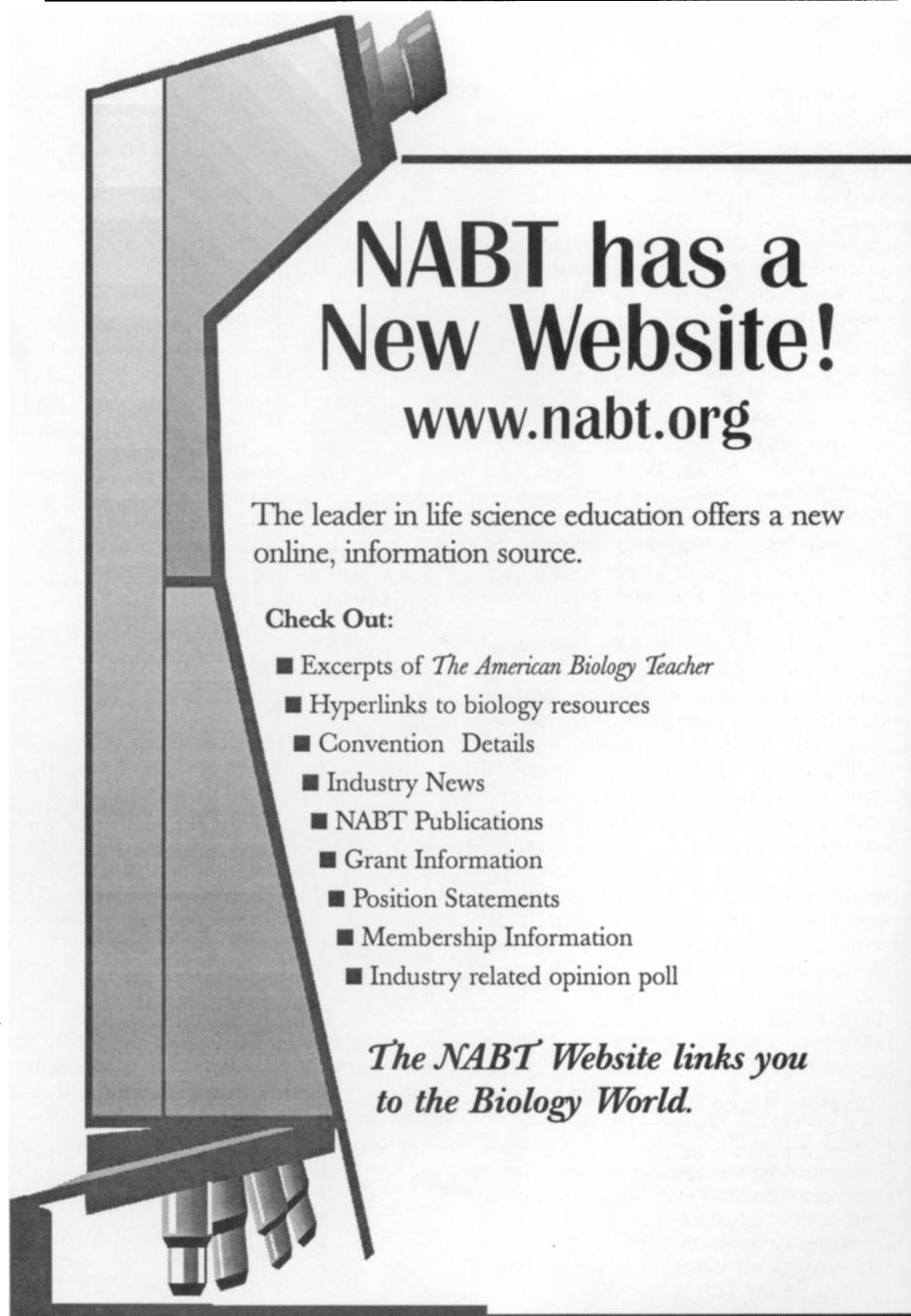
How do larger samples estimate the overall mean compared to smaller ones?

Students notice that as n gets larger, the percent of matches increases, that is, more of the calculated means are

in the same length interval as the mean of the total sample. We explain that when we take a sample, we are trying to estimate the population, and the larger each sample, the more closely we approach the mean of the population. We point out that as n gets larger, there is a smaller range of variation among the sample means.

Chance of Different Samples from the Same Population

We explain that when we take samples, every sample may vary from the overall mean. This variation always



NABT has a New Website!

www.nabt.org

The leader in life science education offers a new online, information source.

Check Out:

- Excerpts of *The American Biology Teacher*
- Hyperlinks to biology resources
- Convention Details
- Industry News
- NABT Publications
- Grant Information
- Position Statements
- Membership Information
- Industry related opinion poll

The NABT Website links you to the Biology World.

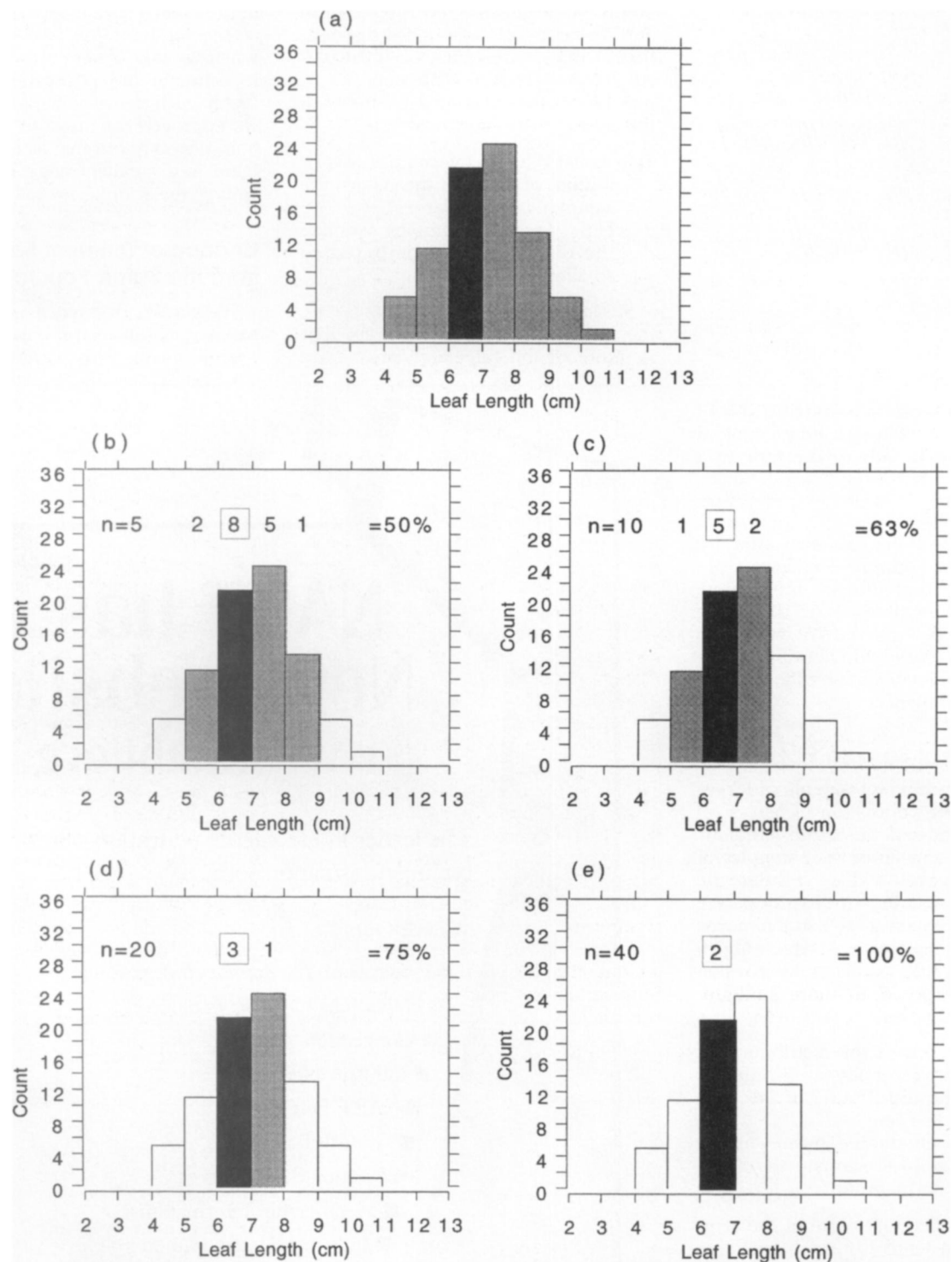


Figure 3. Frequency histogram of number of leaves in each length interval from the north side of the forsythia shrub with increasing sample sizes. Shaded bars indicate length intervals that include means of the indicated sample size. Above the histogram are the number of means that fall in each length interval for the indicated sample size. Black shaded bar indicates length interval that includes mean of the entire north group, which is 6.6 cm. The number in the box indicates the number of sample means that fall in the interval containing the mean of the entire north group. Percent value is the percent of means that fall in the interval containing the mean of the entire north group.

- (a) Individual leaves.
- (b) Samples of $n=5$.
- (c) Samples of $n=10$.
- (d) Samples of $n=20$.
- (e) Samples of $n=40$.

occurs in sampling; the variation is due to chance.

We point to the largest and smallest $n=5$ sample means. The largest and smallest $n=5$ sample means are invariably very different from each other. However, students know that these came from the same set of leaves. They see for themselves that there is a probability by chance alone of drawing two samples from the same population whose means differ from each other.

We ask a series of questions so that the students discover that the probability of a difference is less and less likely for larger and larger differences between means of samples.

Questions: For $n=5$, what proportion of the means differs from the overall mean by 1 cm or more? What proportion of means of $n=5$ differs by 2 cm or more? 3 cm or more? (See Table 1 for answers.)

Comparing Sample Means

We state that statistics are tools that help us make decisions. Statistics tell us how much difference has to exist between two means before we are confident that the difference is not due to chance sampling. Statistics apply probability to the differences we observe and calculate how likely it is for this difference to occur because of chance sampling.

We remind students that this experiment began with collecting leaves from the north and south sides of the shrub and discuss that it is well known that sun leaves have smaller areas and are thicker than shade leaves (Raven, Evert & Eichorn 1992). Objects that face the south are exposed to more sunlight than objects that face north. We

inferred that north-facing leaves should have larger surface area than south-facing leaves and have crudely gauged area by leaf length.

We calculate the mean of the south group and note that the means of the north and south groups are different (Figure 4). In this example with a class of 16 students, the 80 north leaves had a mean length of 6.6 cm, and the 80 south leaves had a mean length of 6.3 cm. We remind students that this difference may be due to sampling or to a biological effect of location. We need an objective way of deciding.

Question: What does the difference between the means of the north-group and south-group leaves tell us about the difference between the populations of north and south leaves?

We state that one of the statistical tools that is available for comparing the means of two samples is the *t*-test. The *t*-test compares the differences between the means of two samples based on how much variation there is within each sample. The more variation there is in each sample, the greater the difference we need between the means of the samples before we can declare that the difference is statistically significant. At some point, we say we know it is possible to get this much difference because of sampling, but the difference we observe is large enough that it is unlikely to have occurred because of sampling, and we label the difference "statistically significant."

Before we make a decision based on statistical analysis, we must decide how certain we need to be about our conclusions. We explain that there is

Table 1. Proportion of means that differ from overall mean.

Difference from Overall Mean			
	1 cm or more	2 cm or more	3 cm or more
$n=5$ samples	8/16	1/16	0/16

a probability associated with how different two samples can be which have been drawn from the same population. For many situations, a reasonable and commonly used probability to allow is 5%, which means that there will be a 1 in 20 chance that the difference is the result of sampling. We ask if the probability of getting our difference because of sampling is less than 5%; if it is, we conclude that our difference most likely did not occur because of sampling.

If, depending on the context of the decision, we need to be more certain that the difference was not the result of sampling, we can set the probability lower than 5%. For example, we could set the probability at 1%, which means that there will be a 1 in 100 chance that the difference is the result of sampling. As we insist on being more certain that the difference is not the result of sampling, we need to observe larger differences between the means before we can state that this difference is unlikely to have occurred because of sampling. However, regardless of the probability level that is selected, we cannot forget that there is still a risk that the difference was indeed the result of sampling.

We select 5% for our probability. Using a computerized statistics pack-

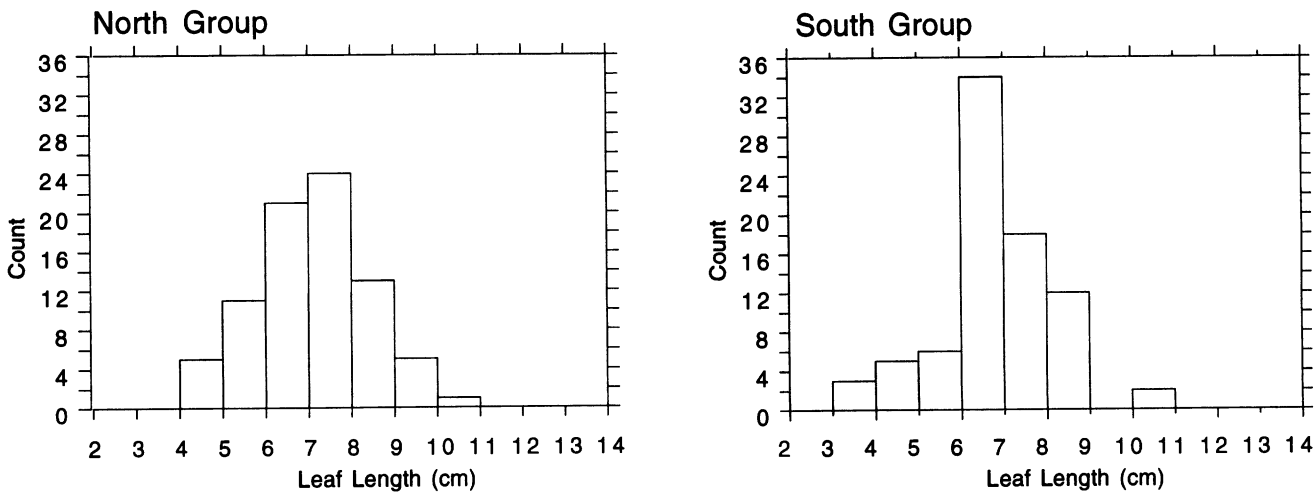


Figure 4. Frequency histograms of the number of leaves in length intervals from the north side and the south side of the forsythia shrub.

Table 2. Unpaired *t*-test for lengths of leaves from the north and south sides of a forsythia shrub.

	Mean Difference	dF	t-Value	P-Value
North, South	.300 cm	158	1.415	.1591

age, students perform a *t*-test on their individual data, and we perform a *t*-test on the entire set of data, comparing the means of the north and south groups of leaves. We examine the *P*-value to see if it is less than 5%. In the example we are presenting here, for the entire set of data, we obtain a *P*-value of .1591 (Table 2), which indicates that the probability of getting the observed difference because of sampling is approximately 16%.

Question: As a result of the *t*-test, what does the difference between the means of the north-group and south-group leaves tell us about the difference between the populations of north and south leaves?

We conclude that the difference is due to sampling rather than due to the biological effect of location. When we state that our sample means are "not statistically significantly different," we mean that the two samples have numerically different means, but they do not vary from each other enough for us to be confident that the difference was not the result of sampling.

Occasionally, a student's individual data result in a *t*-test whose significance differs from the *t*-test of the entire set of data. This occurrence brings home the point of the chance of samples from the same population being different; in a set of small samples, some may appear to be statistically significant whether there is a biological cause for the difference or not. This occurrence also reinforces the value of large samples.

Discussion

The problem of statistics for students, especially with plug-in-the-number-and-get-the-answer computer programs, is that they can do the problems without understanding what the results mean. Perhaps this is a universal problem of mathematically based

knowledge. Dykstra, Boyle and Monarch (1992) point out that students in introductory physics courses often gain the ability to do the problems without gaining the conceptual framework needed to understand physics. One attempt to help students use intuitive reasoning to understand statistics is to replace the formulas for statistical tests by computer resampling to derive the probability due to chance associated with observations (Peterson 1991). Another offers a number of "gimmicks" that help students learn statistics without the anxiety; these are interactive, attention-getting techniques, such as creating a data set based on a cartoon (Schacht & Stewart 1992, p. 329).

Educational research has demonstrated that when students are actively involved in learning, they have higher achievement (Langlois & Zales 1991). Educators in various content areas have verified this. For example, graphing activities in an economics class helped students grasp concepts of supply and demand (Cohn 1995), and group projects in an introductory biology course resulted in creative solutions to problems (Goodwin, Miller & Cheetham 1991). Findings of a national survey of colleges and universities indicate a large increase in the number of courses that include active learning (El-Khawas 1995). When students are grappling with abstract concepts, the teacher can abet the learning process by making the concept concrete (Woolfolk 1995). Students using real data find it more interesting and appreciate that analyzing the data is an essential part of research (Thompson 1994).

Our method of helping students to understand statistical reasoning walks them through a simple problem and focuses on the few central concepts of statistics. The concepts are made concrete, and the students are actively involved. This exercise provides a clear picture of the bell curve distribution from the data that the students gather.

It shows students the relationship between a sample and the population and the fact that a particular sample may not represent the population well. It also shows the effect of sample size on statistical results. Students have actually "seen" the most important concept: that there is a probability of drawing two samples with different means from the same population. The exercise can enable students to distinguish between two samples whose means are different and two samples whose means are statistically significantly different.

References

- Bradstreet, T. E. (1996). Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician*, 50, 69-78.
- Cohn, C. L. (1995). Graphing-to-learn in economics. *College Teaching*, 43(3), 110-111.
- Dykstra, Jr., D. I., Boyle, C. F. & Monarch, I. A. (1992). Studying conceptual change in learning physics. *Science Education*, 76(6), 615-652.
- El-Khawas, E. (1995). *Campus Trends 1995: New Directions for Academic Programs* (Higher Education Panel Report, Number 85). Washington, DC: American Council on Education. (ERIC Document Reproduction Service No. ED 386 089)
- Goodwin, L., Miller, J. E. & Cheetham, R. D. (1991). Teaching freshmen to think—Does active learning work? *BioScience*, 41(10), 719-722.
- Langlois, D. E. & Zales, C. R. (1991). Anatomy of a top teacher. *The American School Board Journal*, 178(8), 44-46.
- Peterson, I. (1991). Pick a sample. *Science News*, 140, 56-58.
- Raven, P. H., Evert, R. F. & Eichorn, S. E. (1992). *Biology of Plants* (5th ed.). New York: Worth Publishing.
- Schacht, S. P. & Stewart, B. J. (1992). Interactive/user-friendly gimmicks for teaching statistics. *Teaching Sociology*, 20, 329-332.
- Thompson, W. B. (1994). Making data analysis realistic: Incorporating research into statistics courses. *Teaching of Psychology*, 21(1), 41-43.
- Woolfolk, A. E. (1995). *Educational Psychology* (6th ed.). Boston: Allyn and Bacon.