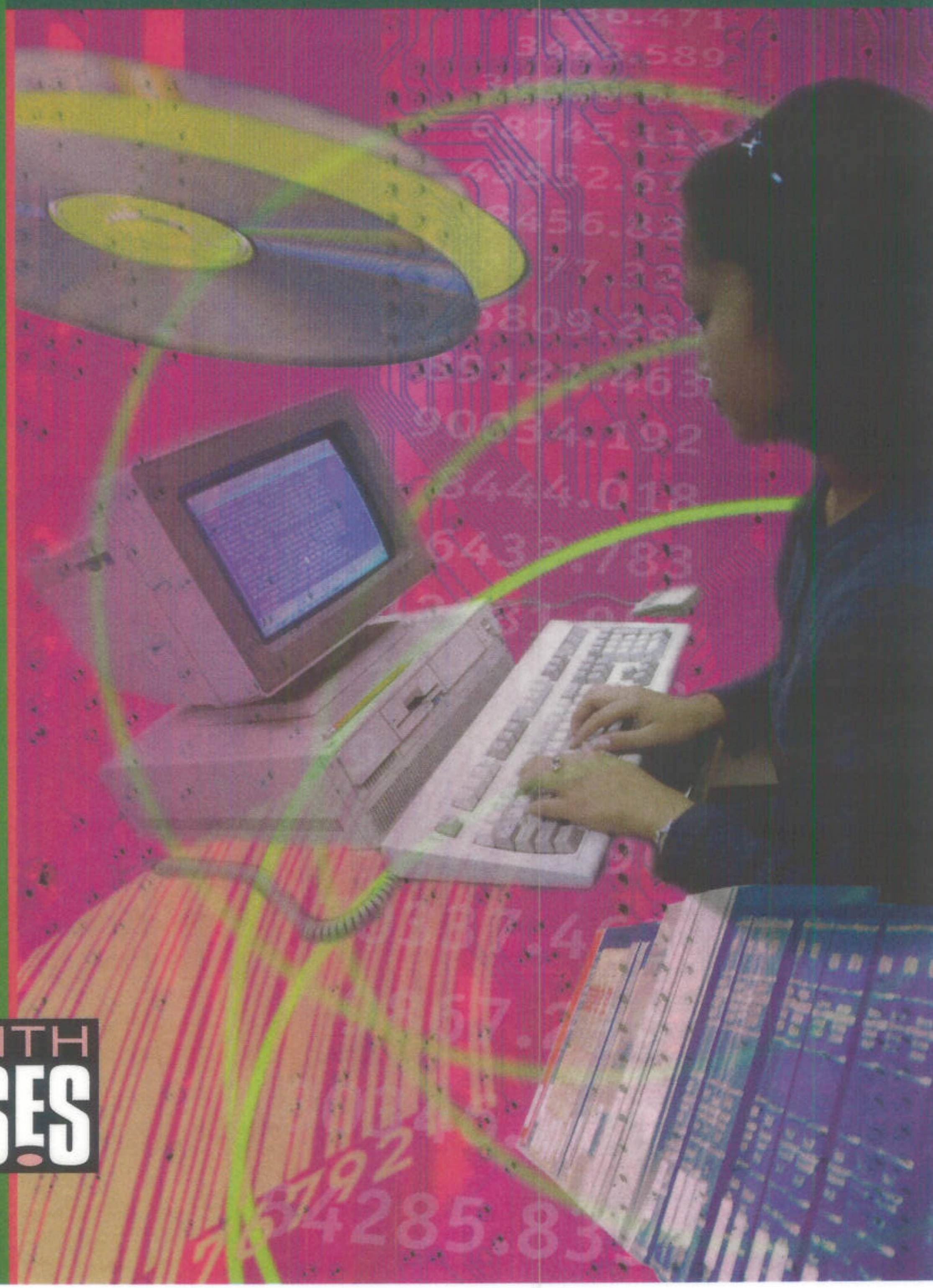
SKILLS BOOK

Database BASICS

Activities to explore the use of data and databases



TEACH WITH DATABASES

TEACH WITH DATABASES

DATABASE BASICS

SKILLS BOOK

BY

BARBARA SPOONER

AND

JAY BARRACATO

PUBLISHED BY

NATIONAL SCIENCE TEACHERS ASSOCIATION

WITH SUPPORT FROM

U.S. ENVIRONMENTAL PROTECTION AGENCY

NSTA Stock Number PB142X
Library of Congress Catalog Card Number 98-84910

ISBN 0-87355-170-2

Printed on recycled paper in the United States of America by Automated Graphic Systems

The National Science Teachers Association is an organization of science education professionals and has as its purpose the stimulation, improvement, and coordination of science teaching and learning.

Table of Contents

Scientists rely on technology to enhance the gathering and manipulation of data.

—*National Science Education Standards*

Student inquiries should culminate in formulating an explanation or model. Models should be physical, conceptual, and mathematical.

—*National Science Education Standards*

Science and technology are essential social enterprises, but alone they can only indicate what can happen, not what should happen.

—*National Science Education Standards*

Introduction to Data and Databases

Scientific literacy enables people to use scientific principles and processes in making personal decisions and to participate in discussions of scientific issues that affect society.

—National Science Education Standards, page ix

Data are all-important in today's society. Data are collected for every conceivable event—from the daily weather report, to the results of sporting events, to daily events in the workplace and school. However, the power of data comes not from simply listing it, but from using using it to make predictions, develop decisions, and better understand our world. Traditionally the language of statistics has been daunting to many students and teachers alike. Now, new methods which rely greatly on graphical rather than algebraic techniques are opening the way for extensive student analysis of data. In addition, many sources of data are increasingly available to the public, as computerized and on-line databases become the norm.

These new sources of data and methods of data analysis are creating many new learning opportunities for all grade and skill levels. The purpose of this book is to present activities which will familiarize students with the presentation, manipulation, and analysis of data. As students perform these activities, they will become familiar with the language of data analysis and, hopefully, transfer these skills to all of their classroom activities.

WHAT ARE DATA?

Data are bits of information that are collected and stored in some way. Data are always collected to serve some purpose. For instance, many stores now ask customers for zip codes or area codes when making a purchase. The information is not enough to identify the customer, but it does provide the store with an idea of where their strongest and weakest markets are located. Many simple questions we are asked during the day are designed to provide data about our actions and preferences.

One piece of information by itself is not particularly useful. Large amounts of data when gathered, however, help establish existing trends and predict future trends. In order to perform these skills, it is important for the data to be organized. Databases organize data so that users can easily find what they need.

WHAT IS A DATABASE?

A database is a collection of similar types of information. While today "database" usually refers to data stored within a computer program, they can take any form. All databases have several common attributes, including an organizational system, searching methods, and the ability to sort information.

Phone books provide a great example of a database. Phone books contain names, addresses, and phone numbers. In addition, many phone books include information on businesses. Each type of information, like a name or address, is referred to as a field. Each field holds only one piece of data. A record contains several such fields. A phone book record, for example, holds the name, phone number, and address fields. A database is the collection of all these records.

In addition to simply listing data, databases allow us to perform two extremely valuable functions: searching and sorting. Searching a database involves looking through the records for a particular value in a particular field. Because residential phone books are organized alphabetically by last name, a person looking for someone's phone number would first search the name field. Once the name field is found, all the information in that record is available, including the field which contains the phone number. Business directories, such as the Yellow Pages, sort by business type or category. Once the listing for a business is found, all of the records—including phone numbers—are available.

Computers can aid the search process, especially when a database contains a considerable amount of information. Most computer databases allow data to be searched or sorted by any field. Imagine an interactive phone book that would let you instantly list all the phone numbers of your softball team, and then just as easily all the pizza delivery places in town.

In addition to searching and sorting data, many computer databases can cross-reference information. A familiar example is a library's check-out system. Part of the database holds the records of the library's holdings (books and other materials). Another part of the database contains the records of people with checkout privileges. The database constantly cross-references records of which patrons have which books. Librarians can use the cross-reference function to identify all the books checked out by an individual or a group of people. Databases with a cross-referencing function allow for this type of elaborate search.

Another reason computerized databases are useful is that they allow for Boolean searches. Boolean searches allow the database user to establish search criteria in more than one field. While Boolean logic and set theory is beyond the scope of this book, database users should be aware of two basic searching methods: the "AND" and the "OR." A search using the AND operator will return records that meet the conditions in two fields specified in the search. A search using the OR operator will return records that meet the conditions set in either of the fields searched. This same result can be accomplished by doing sequential searches, where a search is performed on the results on another search. An example of a sequential search would be checking the card catalog for all the books on a particular subject, then using that list to check which books are presently in the library. Both searches and sorts are highly effective in helping us reach our goal of analyzing existing trends and predicting future trends.

MANIPULATING DATA

All too often, students are distracted from the importance of a set of data by the rigors of trying to interpret the data. For data analysis to be of use, it must enhance understanding rather than create a distraction. For these reasons, data analysis skills should be taught before the technique is applied to data from which students are expected to draw important concepts. For example, graphing techniques are usually taught early in the school year. The objective is to learn solid technique before actually working with data.

Once a set of data has been searched or sorted, a student must choose the most appropriate way to analyze the data. That method depends both upon the format of the data and the desired interpretation. The analysis methods outlined in this book are important tools for working with scientific data, and are not especially mathematically daunting. The goal is to make sense of data without bogging students down with algebra. However, in order to look at the ways data can be manipulated, a mathematical vocabulary must be established. Common mathematical terminology is therefore used and defined throughout the text. It is necessary that the student become familiar with the appropriateness and limitations of data analysis methods, so that they can be confident in how they interpret their data.

CONFIDENCE IN DATA

Many of the techniques described in this book also help determine the confidence that can be placed in data. Error analysis can be confusing to students, but the graphical methods in Section Two, pages 17-33, provide easy techniques for identifying and reporting variations in data. While the graphical methods do not have the rigor of algebraic methods of analyzing error, they do provide the students with a quicker, simpler method of interpreting data. Section Three, pages 34-46, further explores the concept of confidence from the standpoint of decision-making.

ANALYZING DATA

One of the most important steps in analyzing data is to ensure the correct data is collected. The problem most often encountered by students is struggling to draw an interpretation from data which does not include the answer they need. To avoid this pitfall, students first need to clearly state their hypotheses, including the relevant variables and the statistical methods which can answer it. The hypothesis must be one that can be supported by the statistical methods used. When a well-written problem and hypothesis are supported by a valid data set, analyzing the data becomes simple.

The statistical model (choice of analysis method) chosen for an investigation should reflect the type of information needed. Data can be analyzed to show trends within the data (interpolate), predict trends past the ends of the data (extrapolate), to show if a cluster of points is the same or different from another cluster of points, or to show relationships between variables (correlations).

About the *Teach With Databases* Series

The *Teach with Databases* series was developed by the National Science Teachers Association with support from the Environmental Protection Agency. It uses actual databases from the public and private sectors to teach students a broad range of data skills, and guides them through real-life applications of these skills. The scope of these data helps students understand the global importance of data gathered on a local level. The series addresses a wide variety of science disciplines, including chemistry, biology, geology, environmental science, and Earth science. Most of the concepts are applicable for more than one discipline, and can be utilized in several areas of the curriculum. The series was designed to provide teachers with the greatest level of flexibility, so that activities presented in any of these books might be adapted or expanded as needed.

The databases and their corresponding activities might be used as stand-alone units, science fair products, attention-grabbers before an in-depth study of a particular topic, or as a source of data for a class project. Additionally, *Database Basics* can be used to prepare students to work with data from their own investigations, other students' investigations, or any other databases they might need to use.

How to Use This Book

Because even the most advanced high school students may be database novices, it is recommended they first learn good data analysis techniques. Indeed, mathematics is an essential part of science inquiry. The *National Standards* recommends the use of hand tools, measuring instruments, calculators, and computers for the "collection, analysis, and display of data" (page 175). The concepts explored in *Database Basics* give students a sound foundation for future activities requiring more rigorous data analysis techniques.

To make it easier to determine which basic skills students should have before beginning a database activity, the books in the series are cross-referenced to *Database Basics*. Every activity in the Teachers' Guides highlights one or more *Basics* activities that can be conducted as preparation. And every activity in *Basics* lists the skills students will obtain through that activity. In addition, *Database Basics* can be used as a stand-alone curriculum unit, or adapted to prepare students to work with any data set.

While activities in this book can be used as a stand-alone units, each section is designed to augment skills and concepts learned in the previous section. The best preparation for working with a database would use one or more activity from all three sections, tailored to fit students' interests and educational needs.

The book contains three sections: "Creating Databases," "Working with Data Sets," and "Making Decisions Based on Data." These categories of skills correspond with the process of using a database: collection, analysis, and display of data. Specific skills taught in each section are listed on the opposite page. They are also closely linked to the *Standard's* call for students to gain experience in "summarizing data, using language appropriately, developing diagrams and charts, explaining statistical analysis, ... constructing a reasoned argument, and responding appropriately to critical comments" (page 176).

The first section, "Creating Databases," helps students develop their understanding of databases. Activities introduce the concepts of audience, organization, static vs. dynamic, and basic surveying techniques. "Working with Data Sets" explores the formulas used for describing and analyzing data. It provides background material for teachers new to statistical analysis, and gives students a variety of methods for evaluating data. The third section of this book, "Making Decisions Based on Data," explores concepts in oral, written, and rhetorical communication. It gives students tools for communicating the results of their research, choosing appropriate graphs and charts, and forming logical decisions.

The activities use a combination of techniques to supply students with data. In some activities, students perform a brief survey to gather data. To maintain flexibility, data-collecting is typically limited to a single homework assignment which can be expanded if class time, interests, and skill level permit. Other activities provide ready-to-use, reproducible data sets so that attention can be focused on analysis and decision techniques. Most of these data sets are based on actual data. The premise of the *Databases* series is to provide students access to real-life data.

Outline of Data Skills Taught in *Database Basics*

SECTION 1
Creating Databases

Identifying ways databases are used for everyday purposes

Understanding the organization of a common database

Understanding why one database may use several organizational methods

Considering the advantages and disadvantages of computer and Internet databases

Understanding how databases are organized to suit an intended audience

Composing the parameters of a database based on its intended use and audience

Designing a strategy to collect data

Designing a method of storing and maintaining a database

Making database information accessible and easy to use

Creating written instructions for database users

Comparing categories of data

Identifying sampling error

Identifying trends in data

Understanding very large and very small numbers

Extrapolating data

Understanding dilutions, solutes, and solvents

Understanding toxicity

Working with parts per million

Working with concentrations

Working with data indicators

SECTION 2
Working With Data Sets

Drawing comparisons in absolute terms

Comparing different types of data

Understanding and quantifying mass

Describing data using statistical measurements: median, mode, and average

Identifying trends in data

Using the following data analysis techniques:

a. Range
b. Average
c. Standard deviation
d. Median score
e. Quartile
f. Examining distribution with stem-and-leaf plots
g. Examining distribution with box-and-whiskers plots
h. Identifying outliers
i. Establishing relationships between two variables
j. Examining distribution with scatter plots

SECTION 3
Making Decisions Based on Data

Identifying appropriate methods to graphically display data

Communicating scientific research

Explaining data using graphics

Identifying and researching the most relevant data

Preparing statistical summaries from a set of data

Understanding how statistics can be used for multiple purposes

Using statistics to present an argument

Making oral presentations based on database information

Using a scientific organization method to reach a logical decision

Learning how to weigh data when forming a decision

Identifying gaps in information

Making decisions without complete data

Understanding how public perception of risk may affect a decision

ABOUT THE AUTHORS

Barbara Spooner taught three years of math at the Emma Willard School in Troy, New York, and Edison High School in Alexandria, Virginia. Her experience covers all levels of high school math from pre-algebra to advanced placement calculus. She also has spent countless hours outside of class tutoring students individually. She graduated from Williams College with a B.A. in mathematics. She earned her Masters in Education at Harvard University where she concentrated in technology in education. While at Harvard, she developed the math section of an interdisciplinary curriculum implemented in Great Falls, Montana.

Jay Barracato considers himself a science teacher. Over the past six years at Patuxent High School in Calvert County, Maryland, he has taught chemistry, biology, Earth science, physics, and environmental science. Jay does independent work at Chesapeake Biological Laboratory on trace level analysis of chlorinated compounds in marine environments, and has involved his high school classes in raising hyperthermophilic *archaea* (bacteria similar to those found at hydrothermic vents). Jay has participated in the Maryland Governors Academy and the University of Maryland Graduate Fellows Program. His students use statistics to analyze real data they collect, and he uses on-line databases as a source of data in his classroom.

ACKNOWLEDGEMENTS

Many people helped develop *Teach With Databases*. First, Kathy Hogan of EPA, Ron Slotkin of EPA, Greg Crosby of USDA, Keith Wheeler of Global Rivers Environmental Education Network, and Steve Vandas of USGS provided advice on the content and concepts to be included. Kathy Hogan then designed a survey for science educators, and 92 NSTA members graciously responded with their plans and practices for using online data in their classrooms. Kathy Hogan rounded up the valuable EPA publications included with the *Teach With Databases* package.

During manuscript development, authors Barbara Spooner and Jay Barracato received able assistance from reviewers and colleagues, including Kathy Hogan, Greg Crosby, Rob Sugar, and Tom Wysocki. Christina Frasch contributed Activity 10, The Steps of Decision Making.

Project advisors and reviewers at EPA's Office of Prevention, Pesticides, and Toxic Substances were Georgianne McDonald and Odelia Funke. They coordinated the EPA technical review of *Teach With Databases*, and provided advice and encouragement throughout the development process.

Teach With Databases is published by the National Science Teachers Association, Gerry Wheeler (Executive Director), Phyllis Marcuccio (Associate Executive Director for Publications). *Teach With Databases* was developed and produced by NSTA Special Publications, Shirley Watt Ireton (Director of Special Publications), Chris Findlay (Associate Editor), Anna Marie Gillis (Associate Editor), Michelle Triestman (Assistant Editor), and Christina Frasch (Program Assistant). Project Editors for *Teach With Databases* were Christina Frasch for *Database Basics*, and Michelle Treistman for *Toxics Release Inventory Teacher's Guide*. Michelle Eugeni, Eric Knaub, and Doug Messier also provided assistance. *Database Basics* was designed by Sharri Harris Wolfgang of AURAS Design with production assistance from Robin Cather of AURAS Design, and printed by Automated Graphic Systems.

Exploring the Library Database

TIME MANAGEMENT

One classroom period for a library visit. The remainder of the activity can be adjusted to fit your schedule.

MATERIALS

Student Sheet, page 3

BACKGROUND

Organization is the primary trait shared by all databases. The method of organization, however, varies greatly. Ask your students what factors might affect how a database is organized. By precisely organizing materials, the library makes it possible to find specific texts without searching every shelf.

Even databases containing similar types of information may be organized differently. Libraries, for example, use either the Dewey Decimal System or the Library of Congress system—each has different benefits and drawbacks. (Dewey Decimal is simpler, while Library of Congress can better handle large quantities of books. Small libraries tend to use Dewey; larger libraries usually use Library of Congress.) Ask your school librarian to explain to the class which catalog system is used by the school library. Have students research the two systems, and find out the strengths and weaknesses of each. Discuss why different libraries use different organizational methods. How might the potential audience affect how a library is organized? How might it affect how another database is organized?

This activity helps students think about how and why data are organized. Also, because libraries are a database with which students are already familiar, the activity encourages students to identify the databases they may use on a daily basis. The writing/drawing homework exercise reinforces their learning and has students develop a definition of a database for future reference.

PROCEDURE

1. Schedule a class period in your school's library. During the first part of the class, the librarian can discuss how the library "database" is managed. How does its organization help users? Have students consider advantages and disadvantages of this method. Why might a method of organization work well in one situation, but not another?

2. During the second half of the library visit, have your students imagine that a stranger needs help finding a book. She has never used a library, and doesn't understand the word "database." Have each student write down a definition for a database, and write a paragraph explaining why a library is a database. They can then create a simple flow chart or diagram that explains how the library's search system is organized. Fig. 1.1, on page 2, shows a sample flow chart.

3. After students have finished their charts, have them exchange them with their classmates. Can their classmates follow them?

4. Ask students to complete a final draft of their diagram for homework, adding any necessary definitions or clarifications. They could also write one-page letters to the stranger to elaborate on their diagrams and explain the library's organization.

ANSWERS TO STUDENT SHEET

1. A database is a collection of data organized especially for rapid search and retrieval. School libraries organize books and other materials to aid student research.

ACTIVITY SUMMARY

Students use a common database (a library), develop a definition of a database, and explore database organization.

DATABASE SKILLS

Identifying ways databases are used for everyday purposes

Understanding the organization of a common database

Understanding why one database may use several organizational methods

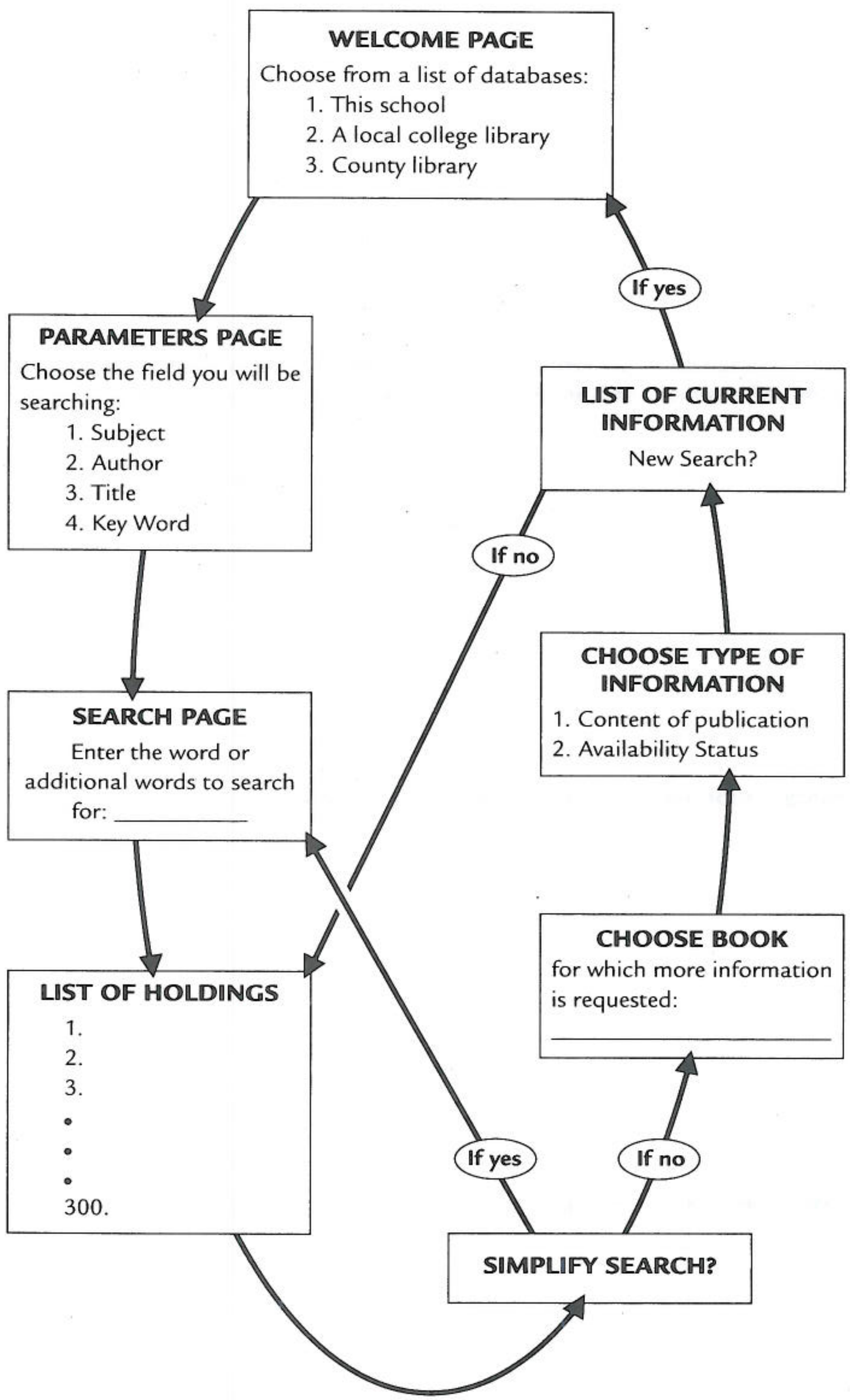
Considering the advantages and disadvantages of computer and Internet databases

Understanding how databases are organized to suit an intended audience.

NOTE: The activity calls for students to create written and graphical descriptions of databases. This constructivist format was developed from the *National Science Education Standards*, "Content Standards: 9–12," which recommends that "student explanations become a baseline for instruction as teachers help students construct explanations aligned with scientific knowledge" (page 172). Further information on using written and graphical techniques to supplement everyday science lessons can be found in NSTA's *How to...Write to Learn Science*. (See the Resources Section in Appendix E, page 52, for more information.)

FIGURE 1.1

Sample Flow Chart of a Library Search System



2. One advantage is that information is accessible from multiple locations, both on-site and remote. Another advantage is that computer systems can be searched by keyword, as well as the more traditional categories of subject, author, and title. Also, computer systems can be instantly updated—one glance at a computer database can tell a user quickly whether the target book is checked out and when it is due back. However, computer systems are costly. Smaller libraries might not have the money to create and maintain a computerized database. Some patrons may be intimidated by computer technology, or be more comfortable with the traditional card catalog. And all computer systems have the potential to crash—an unfortunate scenario for a busy library.

3. Answers will vary, but may include date of publication, intended audience, and length.

4. Answers depend on which libraries students investigate, and the type of organizational method the libraries use. (This question might be used as an extra credit or extension exercise, or a longer-term homework project.)

Exploring the Library Database

1. Write a brief definition of a database. Why is your school library a database?

2. Today, a great number of U.S. libraries have converted their card catalogs into computer databases. Write a few paragraphs about how the computer system improves on the card catalog. What are the drawbacks?

3. List several nontraditional ways books and other library materials might be organized. Feel free to be creative!

4. Compare library sites online, in school, and in local libraries for ease of use, content, organization, and maintenance costs.

Creating a Database

ACTIVITY SUMMARY

Students design and create their own database as a class project.

DATABASE SKILLS

Composing the parameters of a database based on its intended use and audience

Designing a strategy to collect data

Designing a method of storing and maintaining a database

Making database information accessible and easy to use

Creating written instructions for database users

TIME MANAGEMENT

One class period to brainstorm the parameters of the database.

Two weeks outside of the classroom to research colleges.

One class period to compile data into a database.

MATERIALS

Student Sheet, page 6

BACKGROUND

A crucial component of database design is determining what information will be included, and what won't be. These limits, called parameters, are the set of values or constraints that characterize a set of information. Parameters keep databases from becoming unwieldy. Another important aspect of any database is whether it is static or dynamic. Static databases do not allow users to add or change any information—the database will always remain the same. Dynamic databases give users the option of revising data.

As students develop their own database, help them decide what type of parameters are appropriate. They can base their decisions on the time they have to do research, the availability of information, and/or the needs of the people who will be using their database. Encourage your class to give some thought to whether their database should be static or dynamic. Which is more useful? Which is easier to maintain?

In this activity, students will learn how to define and choose the limits of a database, research potential database components, and develop a method for combining and organizing data. The procedure below is based on college selection, but can be adapted depending upon students' interests and skill levels. Another option is to create a database around a topic students have been studying in class. You can also adapt this procedure to introduce future activities on a particular database.

PROCEDURE

1. Have the class brainstorm factors which may affect a student's choice of college. (Factors include cost, location, academic resources, handicapped accessibility, and campus aesthetics.)

2. Determine the parameters of the database. In this case, the parameters will be the types of schools that are included. Will the class include colleges from around the country or stick with local schools? Do they want to limit the database to four-year colleges, or include two-year colleges, community colleges, trade schools, and universities?

3. Have the class develop a list of specific schools they will include in their database.

4. Divide the class into groups of four or five students, assigning each group several schools to research. (Or let each group select their schools, making sure that no two groups research the same schools.)

5. Out-of-class research might take one or two weeks, depending upon how large the database will be. During this time, students should research the schools, and gather information about the criteria they chose in Step 1. Sources of information include college admission departments, the colleges' Internet sites, local book stores, libraries, and your school's academic advisors.

6. Once all the information has been gathered, have the class decide what

medium they will use. Should the database be stored on a computer or written on paper?

7. Before the database is built, the class should also decide how the information will be organized. Once they have answered all these questions, have them build their database.

8. Lead a discussion about the best way for people to use the database. First-time users might need some initial guidance. One way to set up a database is to have database users fill out a questionnaire to help them rank their priorities. Users could then use these priorities to search the database. Have students compile an introduction, questionnaire, or other type of guide to using the database.

9. Place the introduction and database in a location—such as the school library or guidance office—where they are accessible to students selecting colleges. If your school has an Internet site, you might also be able to create a link to the database.

10. (Optional) Have the class interview students who used the college database. Did they find it helpful? What features did they like about the database? What suggestions do they have to make the database easier to use or more helpful? Have the class revise the database to address this feedback.

ANSWERS TO STUDENT SHEET

1. Answers will vary, but should display a basic understanding of static and dynamic databases. An example of a static database is one which records water quality in a particular year; there is no need for revision once all the data are collected. An example of a dynamic database might be one which lists car safety records; it could be updated each year with new car models and the latest safety records.

2. Answers will vary. Examples include:

a. a database maintained by an environmental advocacy group (It would likely have unlimited accessibility so that information can be shared and their positions furthered.)

b. a database with information about books published by a consortium of 100 different companies (Publishers themselves might be able to update their own information whenever necessary.)

c. a database with data collected from only one investigation, survey, experiment, etc. (Because this information is so specific, the database would likely be updated only by the group that maintains it.)

3. Answers will vary.

Creating a Database

1. Imagine the chaos if anybody could, on a whim, add or remove books from your town library! Static databases are created at a specific point in time, and the data are never updated. A dynamic database is updated periodically. Give an example of each, and decide whether your college database should be static or dynamic. Explain your decision.

2. If you chose to make your college database dynamic, who would be able to update it? The information in some databases may be altered by anyone, others may be altered only by the group that maintains it. Database creators must decide who will have access to change information. Consider the following types of access (in terms of updating and/or changing data):

a. unlimited accessibility: general public is able to update any and all information
b. slightly limited access: only some groups or individuals are able to update information
c. limited accessibility: only the group that maintains the database is able to update information

In the space below, list three hypothetical databases that are good examples of each of these different forms of access.

3. As the Internet becomes an increasingly viable place to house databases, the public is re-examining the limits of personal privacy. What constraints, if any, do you think there should be on information available over the Internet? Why?

Generating Data from Sample Populations

TIME MANAGEMENT

One class period to introduce concepts and conduct class survey.

One night of homework for family survey.

One class period to compile family data.

One night of homework to create data tables and answer questions on student sheet.

MATERIALS

Student Sheet, pages 9-12

BACKGROUND

This activity will help develop students' ability to compare data columns, to extrapolate data to a larger set, and to make decisions based upon their interpretation of that information. In this activity, students survey their classmates and families about their favorite television shows, although you might adapt the surveying techniques to any subject.

Because this may be the first time your students have ever conducted a survey, take some time to discuss how surveys are conducted. This discussion might be a good opportunity to introduce the concept of "margin of error." Surveys almost always take into account the potential for inaccuracy, or sampling error.

Sampling error is very different from measurement error. Measurement errors are problems with the actual surveys or measurements that are taken, rather than with the sampling set.

Gallup polls use a sample size of 1,329, which may sound pretty small. It's only possible to generalize from such a small sample to 150,000,000 people because Gallup has developed very sophisticated models and very precise ways of getting a sample that has the proper demographics.

Another way to avoid sampling error is to survey the entire set. If you need to know the reading level of all Americans, and you give every American the same reading test, you are testing the whole population. Whatever results you obtain are real. As you get closer and closer to testing a whole set, you get fewer and fewer chances of being wrong.

Another method is multiple sampling. Because no single method is foolproof, combining sampling methods helps surveyors obtain more accurate data. For example, students have probably heard a news reporter make an announcement such as, "about one million people attended the rally." Estimates like this often use a combination of aerial photography, counts from people on the ground, and statistical extrapolation.

Sometimes more unusual techniques are called for: in 1990, the U.S. Census Bureau sent thousands of workers into the streets to count the homeless people who would not otherwise show up in that year's census—which is usually conducted by mail. Ask students why it is important that a survey identify its sampling errors.

Can your students identify ways to decrease the sampling error in the survey they are about to conduct?

PROCEDURE

1. Teacher Preparation

a. Find out how many people live in your community—the information is available through your town's local library, municipal office, or town hall. If you have time to do further research, locate data about your community in the U.S. Census Bureau's *State and Metropolitan Area Data Book*, or *County and City Data Book*. These books feature demographic, economic, and

ACTIVITY SUMMARY

Students survey their class and family to generate a set of data, extrapolate their findings to their community, and identify trends within the data.

DATABASE SKILLS

Comparing categories of data

Identifying sampling error

Identifying trends in data

Extrapolating data

TEACHING TIPS

You may wish to have students gather more information about census-taking, population studies, and extrapolation techniques. It could prove interesting to include faculty from your school's Social Studies department in these and other research exercises. Other resources include:

- **Nielsen Media Research's Communications Department, which can by reached by calling 212-708-7500, through their website (http://www.nielsenmedia.com), or by writing the company at: 299 Park Avenue, New York, NY 10017.**
- **U.S. Census information, which may be accessed through the U.S. Department of Commerce, Bureau of the Census home page at http://www.census.gov/.**
- **Zero Population Growth, an advocacy group which maintains a platform linking to groups and organizations around the world engaged in population, demographic, and census studies. Their Internet address is http://www.igc.apc.org/zpg/popinfo/.**

governmental data for the purpose of multi-county comparisons or single county profiles. You might also visit the Census Bureau's home page at http://www.census.gov/.

b. If the only information available is population size, approximate the percentages of males and females in each age bracket.

c. If you can locate specific information about your community's demographic structure, incorporate those categories into the television survey. For example, if you have data on your community showing what percent has degrees from high school, two-year colleges, four-year colleges, etc., students can record the education levels, in addition to gender and age, of those they survey. The more data categories you use, the more opportunity students have to compare categories.

d. Select 10 current television shows that the survey should rate. Include shows you think might appeal to varying age and gender groups. Also, try to select shows from various time slots—include shows that are on during the morning, during the day, during the evening, and "late-night" shows. (Be sure to select shows all students can watch, rather than those limited to cable.)

2. Reproduce the table on the Student Sheet, page 9, so that each student has two copies.

3. Have students label one table "Table A," and list 10 current television shows in the boxes on the left of the table. Select 10 students to rate them. The students should give each show a rating from 0 to 100 according to the scale below the table. Each student should work independently, so that nobody knows the ratings until everyone is finished. Rather than writing the students' names, however, have students describe themselves by their gender and age bracket. (Instead of writing "Susan," describe her as "female, age 11-19.")

4. Students can then label the second table "Table B," and list the same television shows in the left column of table. Ask them to bring Table B home and survey their family. As in the class survey, interviewees should be identified by their gender and age bracket. (If you have chosen to have students work with additional data categories, ask them to also identify interviewees by that category. For example, if you have added the category of education level, "Dad," might be identified as "Male, age 35-50, Bachelors degree.")

5. The next time class meets, have students compile their family survey results into one set of data.

6. Have students conduct the "Opinion Poll Mini-Activity" on their Student Sheets. (Explain that they will be using the same method for finding out what their community is watching on television.) As students conduct the pre-activity, reproduce the survey results so that each student has a copy.

7. Give students the demographic information about your community's population, and have them extrapolate their findings.

8. For homework, ask students to create three tables: one displaying class results, one displaying family results, and one projecting the community's favorite shows. The tables can take any format students prefer, as long as they clearly display the data. Students might wish to provide legends to help clarify their tables.

ANSWERS TO STUDENT SHEET

Opinion Poll Mini-Activity

1. Extremely concerned: 43 percent, fairly concerned: 28.5 percent; not at all concerned: 28.5 percent.

2. If 3 out of 7 were extremely concerned, then 3 = X% of 7, or 43%, which can then be extrapolated to 43% of 49 = number of people extremely concerned.

3. Twenty-one people were extremely concerned, 14 people were fairly concerned, and 14 people were not at all concerned.

Questions

1. Sampling error is very different from measurement error. Measurement errors are problems with the actual measurements that are taken, rather than with the sampling set.

2. Answer depends on data.

3. Answer depends on data. Students should identify the time slot with the majority of popular shows.

4. Answer depends on data. Students should identify the show that appeals most to men of all age groups.

5. Answer depends on data. Students should identify the show that appeals most to young children or parents with young children.

6. Answers will vary. Students may write that the data could have been more accurate had they surveyed a larger sample of the entire community.

7. Answers will vary.

TV Show (name and time)	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10

THE RATING SYSTEM:

90-100 excellent
70-89 good
50-69 fair
30-49 poor
0-29 terrible.

AGE BRACKETS:

1-10
11-19
20-29
30-39
40-49
50-59
60+

Generating Data from Sample Populations

OPINION POLL MINI-ACTIVITY

In a recent opinion poll, seven people out of a 49-person group were asked, "How concerned are you about global warming?" Three people said "extremely concerned," two people said "fairly concerned," and two people said "not concerned at all."

1. What percent of the respondents were extremely concerned? Fairly concerned? Not at all concerned?

2. Assume that these seven people were a good representation of the entire group. Devise a method to extrapolate your data to the entire group.

3. Use this method to find out how many people of the group were extremely concerned, fairly concerned, and not concerned.

QUESTIONS

1. Explain the difference between sampling error and measurement error.

2. After completing the survey activity, use your data tables to answer the following questions:

a. Which shows were most popular?

b. Who watched these shows? Calculate percentages for each show, based on age and gender. (Use the same method you developed in the mini-activity.)

c. Which age bracket tends to have the most varying tastes in television shows?

d. Which show is ranked as "excellent" by the most disparate age brackets?

e. Which age bracket tends to have the most different preferences from the other age brackets?

f. Describe how your class data differs from the family data.

g. What trends can you identify in your data? Please describe the trend and provide a possible reason.

3. Based on your data, when is the most expensive air time that advertisers can buy? Support your answer with data from your survey.

4. If a men's clothing company wanted to advertise in your community, during which show should it advertise? Why?

5. If a children's toy company wanted to advertise in your community, during which show should it advertise? Why?

6. Do you think your class survey gave an accurate representation of the entire community? Please explain your answer, and describe how the data could have been more accurate.

7. Even the best-planned surveys may have flaws. For example everybody in your survey had at least one thing in common—a child in the same grade! Discuss the problems with using information gathered by a single class to describe the viewing patterns of an entire neighborhood. How might the shows watched by your families differ from those watched by families with younger children, adults without children, or senior citizens?

Detectability and Toxicity

TIME MANAGEMENT

This activity can be completed in two class periods.

MATERIALS

Strips of white paper cut to fit around a beaker
tape
index cards
lemon juice
beakers
distilled water (tap water can be substituted)
non-toxic red food coloring
Student Sheet, page 16

BACKGROUND

Sir Arthur Conan Doyle (1859–1930) entertained the world with his famous character Sherlock Holmes. Books such as *The Hound of the Baskervilles* showcased Holmes' awesome abilities of detection. In this activity, students will get a chance to do some sleuth work of their own. They will first use indicators to track the presence of a substance. Help them draw the connection between these lab indicators and those that might be measured in the "real world"—for example, particulate levels, visibility, and ozone can indicate how clean the air is in a given area.

The concepts explored in this activity—dilution, very large vs. very small quantities, indicators, and experimental bias—are crucial to students' understanding of databases. Before working with any set of data, students should understand the process by which it is gathered; have a good sense of perspective regarding very large and small numbers; and be aware that different indicators are better for gathering different types of data.

The specific concepts—toxicity, ppm, dilutions, etc.—addressed by this activity may only be appropriate for some databases. They are intended to introduce important concepts that will likely be utilized in numerous future volumes in this series.

PROCEDURE A: INDICATORS, PERCEPTION, AND MEASUREMENTS

1. Tell students they will be tracking the presence of two substances—lemon juice and red food coloring—in water. How do your students think they will be able to detect the concentration of these substances? (Color, taste, and odor are three examples.) Have students list several methods for detecting these substances, then discuss them as a class. Choose three detection methods to use later in this activity.

2. Divide students into groups of three or four. How you organize the rest of the activity is somewhat flexible, and depends upon student skill level and how much time you wish to spend on this activity. Two basic options are to have each group perform serial dilutions for both lemon juice and red food coloring, or to have half of the class work with lemon juice and the other half with food coloring.

3. Students should begin with 1 ml of lemon juice in 9 ml of distilled water. From this 10 percent solution (1 part per 10), have students progress through a standard serial dilution. As Sherlock Holmes claims in *The Sign of Four*, "Detection is, or ought to be, an exact science." Exact measurements are crucial when performing a serial dilution; remind students to label each container with both the percent of the solution, and with how many parts lemon juice each contains.

4. Students should continue the dilution until they reach a .00001 percent solution. It can be measured as 1 part per 1,000,000, or 1 ppm.

ACTIVITY SUMMARY

Students perform two serial dilutions, then use color and taste as indicators of concentration. Students then learn to define toxicity and relate toxicity to concentration.

DATABASE SKILLS

Understanding dilutions, solutes, and solvents

Understanding toxicity

Working with parts per million

Working with concentrations

Working with data indicators

Understanding very large and very small numbers

Detection: Understanding Very Large and Very Small Numbers

1. Did each group get the same results? Why or why not?

2. For each of the three detection methods used, answer the following questions:

a. Which substance (lemon juice or food coloring) was detectable at lesser concentrations?

b. Using the volunteer's examinations of the beakers, was one type of indicator more accurate? Did the accuracy of the indicator depend upon the substance (for example, was visual detection more accurate for one substance than for the other substance)? Please refer to your data table as you answer this question, and provide a possible explanation for your answer.

c. An experimental bias occurs when a researcher affects his or her own results. What are some ways this might occur?

3. Would the detection levels have been different if the volunteer had known the actual concentration? Please explain.

4. Consider two hypothetical substances released into the same river. Why might 10 parts per million (ppm) of Substance A be a threat, while 10 ppm of Substance B be considered reasonable?

Comparing Measurements

TIME MANAGEMENT

This activity can be completed in one class period.

MATERIALS

objects of various mass, such as:

- pencils
- erasers
- paper clips
- folded sheets of paper

scale
two-pan balance
Student Sheet, page 19

BACKGROUND

There are two ways of comparing numbers: absolutely and relatively. As numbers increase in size, it becomes easier and more important to understand them in relative terms than in absolute terms. For example, we may know that our dog, Queenie, has a mass of 20 kg. This is an absolute measurement. If we are not sure what 20 kg feels like, we can pick up the dog and find out. Relative measurements deal with comparisons. Rather than describe Queenie as having a mass of 20 kg, a relative measurement might describe the dog as having a mass three times that of the neighbor's cat.

Which type of description is more convenient? It depends how the numbers are being used. The following activity will help students develop an understanding of the difference between absolute and relative measurements, and improve their ability to describe numbers in relation to other numbers.

PROCEDURE

1. Have students use a balance scale to determine the mass of at least six different objects.

2. They should record these data in Table 5.1 on the Student Sheet in the column labeled "Absolute Measurements." (See Sample Table 5.1 on page 18 for an example.)

3. In the real world—talking with friends, reading the newspaper, watching television—we often encounter data that are described in relative terms. It is important to identify such descriptions, so that we understand exactly what is being conveyed. An example of an relative description might be, "The river is twice as high yesterday as it was last year on the same day." In the third column of Table 5.1, have students write relative measurements for the six different objects. (See Sample Table 5.1 on page 18 for an example.)

4. Actual databases often contain data that are in the millionths as well as in the millions. One way to address such very small and very large numbers is to develop a context that is more manageable. This can be done by comparing them to familiar objects. The first step is often creating a scale. Table 5.2 on the Student Sheet provides students with several events and a scale. Ask students to fill in the last blank using the scale provided. (The question is based on geologic time; you might create a different table to coincide with your class' subject matter.)

5. (Optional) If you have the time, students might create visual depictions of Table 5.2. This project might be done as homework or as part of an assessment on scale and measurement. For example, students might:

- mark off the distances on a meter stick as if it were a timeline
- stack reams of paper on top of each other. Letting each sheet of paper represent a certain number of years, they could use arrows to denote when each event occurred.
- Use an inflatable beach ball to rep-

ACTIVITY SUMMARY

Students describe various objects using absolute and relative measurements.

DATABASE SKILLS

Drawing comparisons in absolute and relative terms

Comparing different types of data

Understanding and quantifying mass

Understanding very large and very small numbers

SAMPLE TABLE 5.1

Sample Absolute and Relative Measurements

OBJECT	ABSOLUTE	RELATIVE MEASUREMENT
1. Key	5 grams	Three times as much mass as a pencil
2. Mug	50 grams	Ten times as much mass as a 10-ml beaker
3. Paper clip	1 g	1/100th the mass of a pair of forceps

SAMPLE TABLE 5.2

a. Age when dinosaurs thrived = 100–200 million years ago = 425 km = *(example: about the distance from San Francisco to Santa Barbara.)*

b. Coldest part of the ice age = 20 thousand years ago = 60 meters = shorter than the length of a football field and roughly the height of a fifteen-story building.

c. First theoretical paper on the greenhouse effect = 100 years ago = 30 cm = a little longer than an average sheet of notebook paper.

d. Time in which 100 million more humans will live on Earth = About one year = 3 mm =about the width of the E in the word Earth.

(Adapted from *Forecasting the Future: Exploring Evidence for Global Climate Change*, by the Education Department of the Stephen Birch Aquarium-Museum. 1996. Arlington, VA: National Science Teachers Association.)

resent the Earth, and mark distances directly on the ball.

6. A second way to informally describe data are with words that are "descriptors." Examples of descriptors include half, twice, and equal to. Have students look at Table 5.3 on the Student Sheet. Ask them to write five one-sentence descriptions of the data, using these types of descriptors. One example might be: "The total capacity of the High Plains Aquifer is $2.6x10^4$ L more than its typical annual recharge."

7. Lead a class discussion about the benefits and drawbacks of absolute and relative measurements. Remind students that absolute measurements are more exact and leave less room for interpretation. Exactitude is important because a single database might be used by different people with broadly ranging experiences and methods for interpreting data.

ANSWER TO STUDENT SHEET

Potential answers to the Student Sheet charts are provided in Sample Tables 5.1 and 5.2 on the left. For Table 5.3, student descriptions should meet the requirements set out in Step 6, by writing one-sentence, informal descriptions of the information.

Comparing Measurements

TABLE 5.1

Absolute and Relative Measurements

OBJECT	ABSOLUTE MEASUREMENT	RELATIVE MEASUREMENT
1.		
2.		
3.		
4.		
5.		
6.		

TABLE 5.2

Because the average person lives fewer than 100 years, it is difficult to think about periods of time as long as 5,000,000,000 years. The table below will help place some important events in geologic time into perspective. The first step is to devise a scale. If the age of the Earth, 4.6 billion years, is compared with the diameter of Earth, 13 million meters, you have a scale you can use to come up with comparisons for the events below. Use the scale to fill in the blanks with measurements that are a bit more common than those provided. (An example is provided for the first comparison.)

a. Age when dinosaurs thrived = 100–200 million years ago = 425 km = ____________________
(example: about the distance from San Francisco to Santa Barbara.)

b. Coldest part of the ice age = 20 thousand years ago = 60 meters = ____________________

c. First theoretical paper on the greenhouse effect = 100 years ago = 30 cm = ____________________

d. Time in which 100 million more humans will live on Earth = About one year = 3 mm = ____________________

TABLE 5.3

Balance Sheet Estimates for the High Plains Aquifer

DIMENSIONS	SCIENTIFIC NOTATION	COMMON NOTATION
maximum width	5x102 km	500 km wide
maximum length	1.2x103 km	1150 km long
CAPACITY		
Total Capacity	$3.9\text{x}10^{15}$ L	3.9 million trillion liters
Typical Annual Discharge	$2.7\text{x}10^{13}$ L	27 thousand trillion liters
Typical Annual Recharge	$1.3\text{x}10^{11}$ L	13 trillion liters

(Source: William Ashworth, ed. 1991. *The Encyclopedia of Environmental Sciences.* New York, NY: Facts on File.)

Describing Data

ACTIVITY SUMMARY

A discussion of basic statistical measurements, and how to find them, is followed by individual or group work to practice using these methods.

DATABASE SKILLS

Describing data using statistical measurements: median, mode, and average

Identifying trends in data

TEACHING NOTES

- **Graphing calculators can greatly reduce the work involved in organizing sets of numbers and calculating their central tendencies. See Appendix A, page 47, for the steps to sort numbers and calculate mean, median, and other statistics on a TI-82 calculator.**
- **Neither the median nor the mean can tell us as much about the data as a plot showing all the values, such as a number line plot, a dot plot, or a stem-and-leaf plot. Activity 7 on page 24 provides suggestions for introducing these methods to your students.**
- **Many of the statistical interpretation methods discussed in this activity are also discussed in Activity 7, pages 24-33. Activity 7 provides several advanced methods for looking at data sets, whereas this activity is a good introduction to the basic techniques. You may wish to conduct one activity by itself, or adapt the most relevant parts of both activities to fit your needs.**

TIME MANAGEMENT

This activity can be completed in one class period.

MATERIALS

Chalkboard or overhead projector
Student Sheets, pages 22-23

BACKGROUND

Databases are created to help people answer questions. (What is the pattern of stream flow in the Mississippi River? How have interest rates varied over the past 10 years?) However, the answers don't usually leap out from a set of data. The first step towards answering the larger questions is interpreting the data. Being able to describe a set of data will give you a better understanding of a set of numbers.

Averages are a common method for describing data. Two types of averages—median and mean—summarize the data by giving a measure of the center of the data values. For example, if Peter gets scores of 80, 84, 95, and 90 on five science tests, then his average is:

$$\frac{80 + 96 + 84 + 95 + 90}{5}$$

$$= \frac{445}{5}$$

$$= 89.$$

This way of computing an average is called the mean. Thus, the mean of Peter's scores is 89. Another type of average is the median. To find the median of Peter's test scores, first put them in order from smallest to largest:

80 84 90 96 96

The middle score, 90, is the median. Half of the test scores are lower or equal to the median; half are greater than or equal to the median. If there is an even number of scores, the median is halfway between the two scores in the middle:

25 80 84 90 96 96

The median of the above number set is 87, which is halfway between 84 and 90.

Which type of average—mean or median—you use depends on your purpose in finding a measure of center. When there are no outliers (values that are widely separated from the rest of the data), there generally will not be much difference between the median and the mean, and which method you choose won't matter much.

A third method for describing data is the mode. The mode of a set of data are the responses that occur most frequently. The mode of the number sets above, for example, is 96. Another method is the range of data, which is the difference between the smallest number and the largest number in a number set. The range of Peter's last set of scores is 71 points.

PROCEDURE

1. Distribute a copy of the Student Sheet to each student.
2. Show students how to find the mean, median, mode, and range of a set of data.
3. Divide the class into groups of three or four students, and ask them to answer the questions on the Student Sheets.
4. (Optional) Another way to teach mean, median, and mode is with a simple popcorn activity. Put out a bowl of popcorn, and have each student take three handfuls. Each student counts and records the number of pieces in each handful. They then report the number of pieces in each of their handfuls, and the class creates a combined data table on the board. As

the students eat their popcorn, discuss patterns in the data. Calculate the mean, median and mode of the handfuls. Are these numbers similar? What does this indicate about the data? If the numbers are similar, the data likely are evenly distributed, whereas if the numbers are significantly different there are some handfuls that are much bigger or smaller than most that skew the average.

ANSWERS TO STUDENT SHEET

1. $6.34

2. $5.24

3. $4.74

4. $4.77

5. The median gives a better indication of the prices, because measurements of median are not as affected by data points that are dramatically different from the rest of the data.

6. Ice cream is generally more expensive.

7. $9.35

8. $5.00

9. Characteristics may include:

a. mode—describes the most typical number

b. median—not so affected by outliers, usually lower than the mean

c. mean—somewhat affected by outliers, usually higher than the median

d. range—shows how spread out a set of data are

Describing Data

TABLE 6.1

The prices for the different kinds of ice cream and ice milk/ dairy desserts sold at a Los Angeles supermarket.

ICE CREAM	PRICE FOR HALF A GALLON	ICE MILK/DAIRY DESSERTS	PRICE FOR HALF A GALLON
Ben and Jerry's	$11.80	Breyer's Natural Light	$4.85
Breyer's	4.65	Dreyer's Grand Light	4.79
Dreyer's Grand	4.79	Jerseymaid Lightly Maid	2.85
Frusen Gladje	10.76	Jerseymaid Light	3.85
Haagen Dazs	9.98	Lean Cuisine	4.75
Jerseymaid	2.65	Weight Watchers	10.36
Jerseymaid Natural	2.50		
Jerseymaid Old Fashioned	2.50		
McConnell's	9.96		
Mrs. Field's	4.69		
Royal Request	9.38		
Westwood	2.45		

(note: brands not sold in gallons were calculated based on what the price would be for a gallon.)

1. What is the mean price of a half gallon of ice cream?

2. What is the mean price of ice milk/dairy desserts?

3. What is the median price of a half gallon of ice cream?

4. What is the median price of ice milk/dairy desserts?

5. Which method, median or mode, gives a better indication of the typical price of ice cream? of ice milk/dairy dessert? Explain.

6. In general, which type of product is more expensive?

7. Combine both lists and find the range in prices.

8. Round each price to the nearest dollar. Using the combined list, what is the mode of the prices?

9. What are some characteristics of each of these different types of measurements?

a. mode

b. median

c. mean

d. range

Analyzing Data

ACTIVITY SUMMARY

This activity presents students with several options for analyzing sets of data. Each step introduces a new technique by posing a "Sample Question," which is then answered with a brief explanation of the technique and a teacher-led demonstration.

DATABASE SKILLS

Using the following data analysis techniques:

a. range

b. average (mean)

c. standard deviation

d. median score

e. quartiles

f. examining distribution with stem-and-leaf plots

g. examining distribution with box-and-whiskers plots

h. identifying outliers

i. establishing relationships between two variables

j. examining distribution with scatter plots

TIME MANAGEMENT

This activity can be completed in one or two class periods. To use less class time, assign the questions on the Student Sheet as homework.

MATERIALS

graphing paper
pencils
calculator
Student Sheets, pages 32-33

BACKGROUND

Calculations such as mean, median, and mode describe the center of a set of data. They are therefore called measures of central tendency. While central tendency measurements describe a set of data, dispersion measurements help complete the picture. This activity gives students several options for analyzing data using measurements of both dispersion and central tendency.

The following set of data (Table 7.1, page 25) shows the scores a class produced on two end-of-unit exams and a final exam. While the individual score is important to each student, many additional interpretations can be made by looking at the data as a set. The two most common ways to analyze these data are by finding each student's average score, and to find an average score for each exam. However, some fairly basic methods can help identify even more information from these scores.

Before beginning the activity, provide each student with a copy of the Student Sheet. Try posing the sample questions to students to see how many of them are familiar with each technique. If possible, have a student demonstrate how to perform each calculation.

As each calculation is discussed, have students record definitions, formulas, and procedures on a separate sheet of paper. They can then apply each technique (individually or in groups) by responding to the practice questions on the Student Sheet (page 32). When finished, students will have a record of several data analysis techniques which they can use when working with any number set.

PROCEDURE

1. RANGE

Sample Question: What is the range of scores for exam 1?

The range of scores includes the highest value and the lowest value in a set of data. It is also important to consider the possible range of scores when analyzing data.

Maximum = 100
Minimum = 43
Possible Maximum = 100
Possible Minimum = 0

2. AVERAGE (MEAN)

Sample Question: What is the average (mean) score of exam 1 and what is its standard deviation?

The standard deviation is a measure of how far away the scores in a set are from the average. The standard deviation is often reported as an absolute error (same units as the measurement) in a measurement.

To find average (mean):

$$\bar{X} = \text{Average} = \frac{\Sigma \bar{X}}{n} = \frac{1742}{22} = 79$$

To find standard deviation (where n = total number of data points):

$$\delta = \text{Standard Deviation} = \sqrt{\frac{\Sigma(\bar{X} - X)^2}{n-1}} = 12$$

3. MEDIAN

Sample Question: What is the median score for exam 1 and what are the quartiles?

Yet another way of showing dispersion is to show the proportion of people scoring in the various quartiles. The median and quartiles of a number set are easiest to find when the data are sorted by number value. For this set of data, the data could be sorted by score rather than by name. The median score is the center score. Each half of the data can be split in half to find the upper and lower quartiles.

100, 93, 91, 91, 90, 88, 88, 87, 85, 83, 82,|78, 76, 76, 76, 75, 74, 74, 73, 68, 55, 43
median

$\text{median} = \frac{82 + 78}{2}$ upper quartile = 88 lower quartile = 74

4. DISTRIBUTION PLOTS (STEM-AND-LEAF, BOX-AND-WHISKERS)

Sample Question: How are the scores on exam 1 distributed?

There are two effective techniques for examining the distribution of a data set. The first is a stem-and-leaf plot. To create a stem-and-leaf plot, the first step is to identify the possible range of values. For the exam, the possible scores are from 0 to 100. The first digits of a number are the stem, and the last digit, no matter what decimal place, is the leaf.

The following examples show data from Table 7.1.

The possible stems are first placed on an axis like this:

0 |
1 |
2 |
3 |
4 |
5 |
6 |
7 |
8 |
9 |
10 |

TABLE 7.1

Class 1

NAME	SCORE EXAM 1	SCORE EXAM 2	SCORE FINAL EXAM
John	93	88	90
Mary	74	90	78
Sally	43	70	62
Tisha	88	84	88
Albert	76	68	72
Anton	68	55	45
Anna	100	91	95
James	87	86	86
Sue	76	89	88
Daniel	55	72	70
Lisa	90	89	92
Jennifer	91	90	89
Sophe	74	86	80
Susan	78	77	77
Kevin	83	84	90
Peter	76	45	62
Ben	75	76	80
Stacey	82	81	83
Wesley	72	77	76
Carlene	91	88	91
Jessica	88	72	72
Amy	85	80	81

EMPHASIS ON THE MEDIAN

You may have noticed that this book focuses more on median than on mode. There are two main reasons for this emphasis. First, the median is a simpler idea, requiring students to perform less computation. Second, the median is not affected by a few extremely large or small values (as is the mean).

The mode is generally not useful for interpreting and summarizing data. One reason is that many sets of data, such as 22, 25, 28, and 29, have no mode. In contrast, median and mean are always defined. Second, the mode is unstable. For example, the mode of 1,1,3,5,8, and 9 is 1. However, by slightly changing two values—to 1,2,3,5,9, and 9—the mode becomes 9. Finally, as shown in the last example, the mode does not necessarily indicate the center of the data. For these reasons, statisticians tend to use median-based statistics when analyzing data.

The next step is to add the leaves, which are the last digits of a number.

0	
1	
2	
3	
4	3
5	5
6	8
7	4 6 7 4 8 6 5 3
8	8 7 3 2 8 5
9	3 5 0 1 1
10	0

The final step is to arrange the leaves in order, numerically. If you previously sorted the data, this step may be already done. The final stem-and-leaf plot looks like this:

0	
1	
2	
3	
4	3
5	5
6	8
7	3 4 4 5 6 6 6 8
8	2 3 5 7 8 8
9	0 1 1 3 5
10	0

Types of Plot Distributions

There are four major types of distributions shown on stem-and-leaf plots: J-shaped, bell-shaped, U-shaped, and rectangular. Tables 7.2 to 7.5 show examples of these distribution types (they do not refer to specific data).

TABLE 7.2

A J-shaped leaf plot

0	0 1 1 2 4 6 7
1	1 3 3 4 5
2	2 4 6
3	2 8
4	0
5	1

TABLE 7.4

A U-shaped leaf plot

0	0 0 1 3 5 8
1	0 1 6 9
2	4 7
3	6
4	0 0 4
5	1 1 3 3 9 6

TABLE 7.3

A bell-shaped leaf plot

0	1 2
1	4 6 7
2	0 0 1 4 5 8
3	1 1 3 6
4	2 8
5	1

TABLE 7.5

A rectangular leaf plot

0	0 1 1 3 6
1	2 4 9 8
2	2 2 6 5 4
3	4 5 5 8 8
4	2 6 7 7
5	4 4 5 6 6

THE EVOLUTION OF MAKING PLOTS

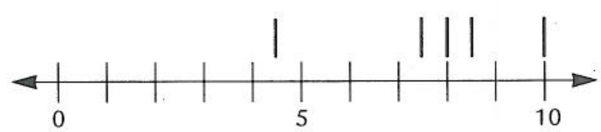
Many of the plotting techniques taught in this activity—such as stem-and-leaf plots, box plots, and fitting a line—were developed during the 1960s and 1970s. They are so new that their names have not been universally agreed upon: fitted lines, for example, are also called Tukey lines, median-fit lines, resistant lines, and robust lines. The techniques themselves are not even set in stone. For example, some people make box plots horizontally; some make them vertically. Some people put a dent in their box plot at the median; others draw a line across the box at the median. Students like to be told that the techniques of making these plots are still evolving. They might be the ones to invent a better technique. Encourage experimentation. Because there is no "correct" way to make these plots, there's no reason to get caught up in the technicalities of making plots exactly the way this book does.

Stem-and-leaf plots also clearly show clusters, gaps, and outliers in the data.

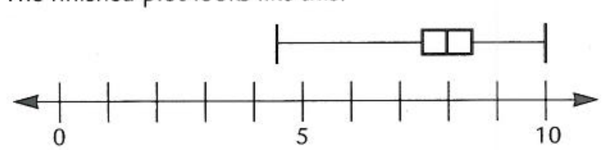
0		
1	2	*outlier*
2		
3		
4		
5		
6	3 4 5	
7		*gap*
8	0 1 2 8 9	*cluster*
9	2 2 4 6 7 8	*cluster*
10	0 0 1 3	*cluster*

The second type of graphical description of a data set is the box-and-whiskers plot. For the box-and-whiskers plot, the first step is to draw an axis with the range of data like this:

Then the median value, upper and lower quartiles, and maximum and minimum values are placed on the plot:



The finished plot looks like this:



Each section of a box-and-whiskers plot represents 25 percent of the values examined. This method makes it easy to determine if the numbers are evenly or unevenly distributed.

5. OUTLIERS

Sample Question: What are the outliers for a set of data?

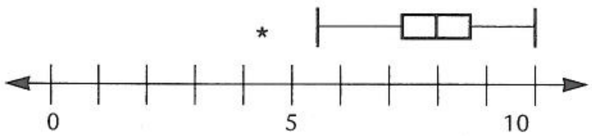
Outliers, points which lie a long way from the bulk of the data, are easy to identify on box-and-whiskers plots. If either the high whisker or the low whisker is much longer than the range of data in the center, one should check for outlier points.

The interquartile range is the upper quartile minus the lower quartile. If any point is more than 1.5 times the interquartile range above the upper quartile, or below the lower quartile, that point may be considered an outlier. Usually the whisker is ended at the next lower point and the outlier marked with a point.

Interquartile range (IQR) = 88 – 74 = 14

14 x 1.5 = 21

Therefore, any point > 88 + 21 or any point < 74 – 21 can be considered an outlier. On exam 1 the point 43 can be considered an outlier.



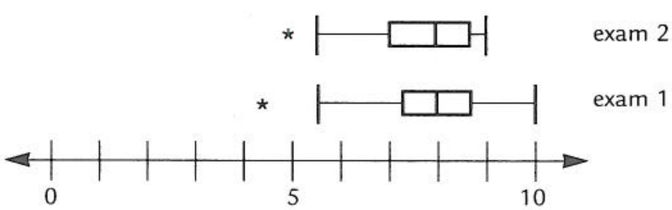
6. COMPARING SETS OF DATA

Sample Question: Is there any difference in the class performance on exam 1 and exam 2?

This question compares two sets of data. The most effective methods for this comparison are back-to-back, stem-and-leaf plots, or stacked box-and-whiskers plots. A back-to-back, stem-and-leaf plot looks like:

	0	
	1	
	2	
	3	
3	4	5
5	5	5
8	6	8
8 6 6 6 5 4 4 3	7	0 2 2 6 7 7
8 8 7 5 3 2	8	0 1 4 4 6 6 8 8 9 9
5 3 1 1 0 9 0	0	1
0	10	

Stacked box-and-whiskers plots look like:



7. RELATIONSHIPS BETWEEN TWO VARIABLES

Sample Question: Is there any relationship between student performance on exam 1 and the final exam?

This question seeks to establish a relationships between two variables. The first step is to create a scatter plot. The scores for each student are graphed with exam 1 as the independent variable and the final exam as the dependent variable. Tables 7.6 and 7.7 on page 30 shows a scatter plot of these data.

The scatter plot can be described as having a positive association, a negative association, or no association. Tables 7.8, 7.9, and 7.10 on page 31 show these types of associations.

NOTE: Several advanced techniques for curve fitting are discussed in Appendix B on page 48, and Appendix C on page 50. These techniques require the use of a graphing calculator or data analysis program on a computer.

ANSWERS TO STUDENT SHEET

Answers should follow formats taught in the Teacher Section's Sample Questions.

TABLE 7.6

Scatter Plot

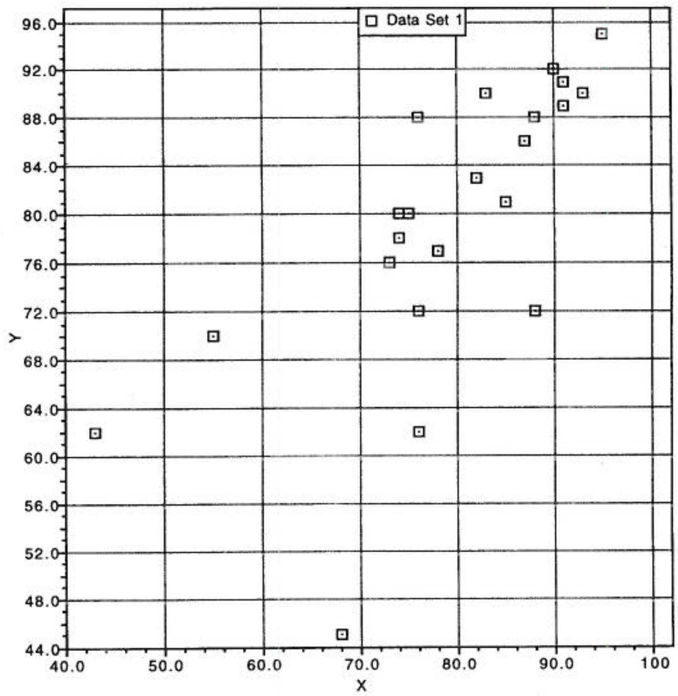


TABLE 7.7

Scatter Plot with Best Fit Line

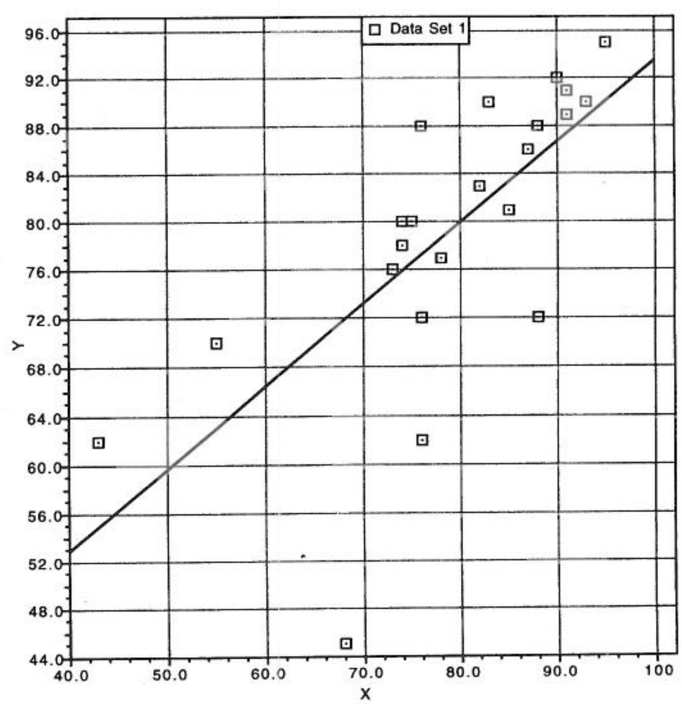


TABLE 7.8

Positive Association

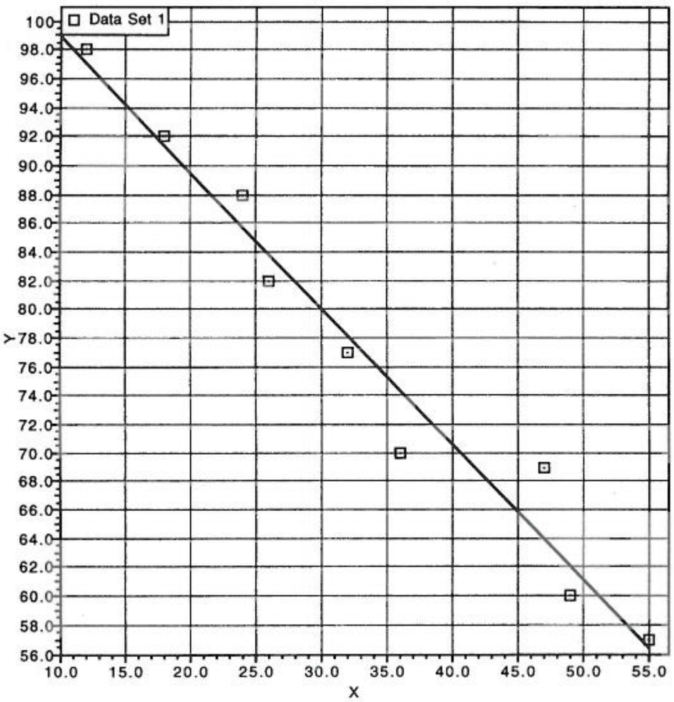


TABLE 7.10

No Association

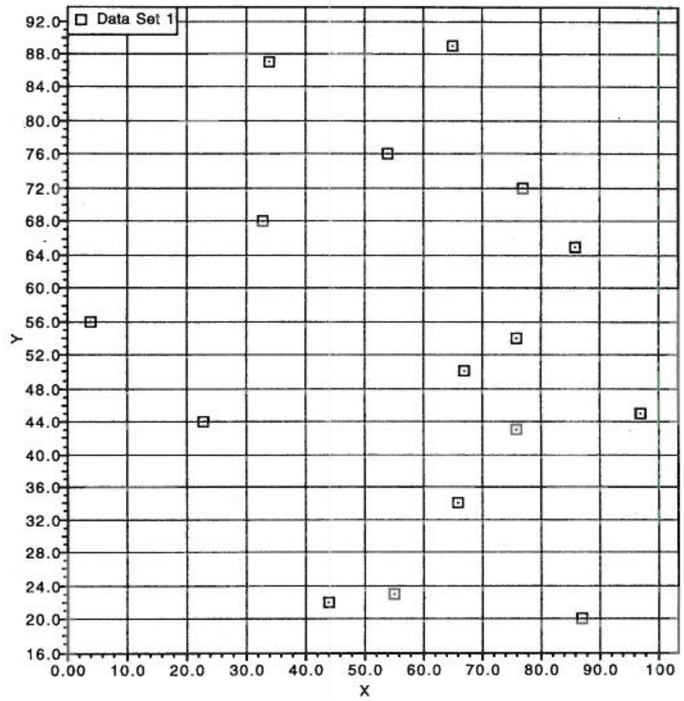
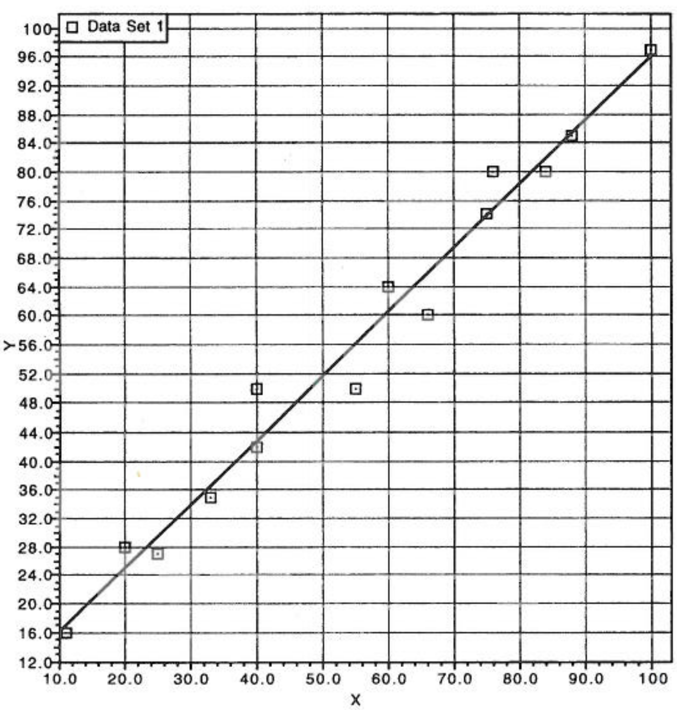


TABLE 7.9

Negative Association



Analyzing Data

1. RANGE

Sample Question: What is the range of scores for exam 1?

Practice Questions:

a. What is the range of scores for exam 2?

b. What is the range of scores for the final exam?

2. AVERAGE (MEAN)

Sample Question: What is the average (mean) score of exam 1 and what is its standard deviation?

Practice Question:

What is the average score of exam 2 and its standard deviation?

3. MEDIAN

Sample Question: What is the median score for exam 1 and what are the quartiles?

Practice Question:

What are the median score and the quartiles for exam 2?

4. DISTRIBUTION PLOTS (STEM-AND-LEAF, BOX-AND-WHISKERS)

Sample Question: How are the scores on exam 1 distributed?

Practice Questions:

a. Use a separate sheet of paper to draw a stem-and-leaf plot for exam 2.

b. Use a separate sheet of paper to draw a box-and-whiskers plot for exam 2.

5. OUTLIERS

Sample Question: What are the outliers for this set of data?

Practice Questions

a. Find the outliers for exam 2.

b. On a separate sheet of paper, draw a revised box-and-whiskers plot, with the whisker ended at the next lower point and the outlier marked with a point.

6. COMPARING TWO SETS OF DATA

Sample Question: Is there any difference in the class performance on exam 1 and exam 2?

Practice Questions:

a. Describe the difference between the class' performances on exam 1, exam 2, and the final exam.

b. Describe the difference in the performance of males and females on exam 1.

7. RELATIONSHIPS BETWEEN TWO VARIABLES

Sample Question: Is there any relationship between student performance on exam 1 and the final exam?

Practice Question:

Is there any relationship between student performance on exam 2 and the final exam?

Displaying Data with Graphics

ACTIVITY SUMMARY

Students examine the way published data are presented graphically to guide their own work in presenting data.

DATABASE SKILLS

Identifying appropriate methods to graphically display data

Communicating scientific research

Explaining data using graphics

TEACHING TIPS

- **If students are unable to locate every type of data graphic they originally listed, suggest they create their own examples based on real data.**
- **An alternative procedure is to assign student groups to a particular category of graphic. The group would then provide examples in that category and develop that section of the class portfolio.**

TIME MANAGEMENT

One class period to introduce and discuss the activity.

One week for out-of-class research.

One day for students to discuss the graphical methods they have researched.

MATERIALS

Pre-Activity Tables, page 37
Data Graphics Guidelines, page 36
Optional materials might include:

- computer on which students can design graphs and tables
- graph paper
- cardboard
- markers

BACKGROUND

Working with statistics and presenting their results provides a good opportunity to increase students' communications abilities. In fact, communicating one's work is a crucial part of any scientific endeavor. This activity focuses on graphically communicating data and statistical information.

According to one graphic designer, "...of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful." (Source: Tufte, Edward E. 1983. *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.) Modern graphics play a large part in communicating any set of information. In fact, data graphics do a lot more than substitute for tables—they are often the most effective way to describe, explore, and summarize data.

Encourage students to discuss how graphics might have supplemented reports they have written for science classes. The addition of well-designed data graphics to clearly-written text can make all the difference for scientists trying to communicate research to their colleagues—or for students handing in a term paper.

Charts, timelines, histograms, and tables are just a few ways to present information. In this activity, students will explore types of data graphics, using a standard format that will allow their collected work to be used as a class portfolio. They will be able to use this portfolio as a reference for conveying data they generate or collect throughout the semester.

PRE-ACTIVITY

1. Distribute copies of Tables 8.1 and 8.2 on page 37.

2. Tell students that the lower two lines on Table 8.2 show precisely the same information as the two lines in Table 8.1. Have students compare these two tables.

3. What differences can students identify in these two graphs? Talking points include:

a. Table 8.2 shows the combined score in addition to the individual trends.

b. To show the combined scores, the Table 8.1 plots data on a 600-point scale that runs from 200 to 800. Table 8.2 shows the SAT score decline plotted on a 100-point scale running from 400 to 500 points.

c. The graph in Table 8.1 might lead people to use words like "plunge" and "plummet." The graph in Table 8.2 looks much less dire.

d. Graphs with more restricted parts of the scale represent a common and often unintentional way of misleading people with a data graphic.

e. It is important to examine graphs to make certain that the scale is properly represented and that the data actually say what is claimed for them.

PROCEDURE

1. Bring several copies of *USA Today* into class. Give each student a different section of the newspaper. Ask them to discuss how the paper displays information—are the graphics interesting? How much information do these graphics present? What is the purpose of these graphics? What is the audience looking for in these graphics?

2. Ask the class to generate a list of techniques used in the paper that might be useful when working with scientific data. Have a volunteer write this list on the chalkboard or overhead projector.

3. When students have finished identifying the newspaper's various data graphics, ask them to look at their list. Can they think of other data graphic styles that might be useful for presenting data? Is there data that students are familiar with that can't be easily presented by the methods they've just reviewed?

4. For homework, have students collect data graphics from other printed or on-line (and printable) materials. Ask students to find examples of different types of data graphics. Sources of graphics include:

- newspapers
- brochures
- scientific and other journals
- almanacs
- textbooks

5. Once students have gathered their data graphics, they should paste or print each graphic onto the top of a blank sheet of paper. Below the graphic, they should write a brief description of each data graphic and explain the data that method is illustrating. Is it a pie chart? A timeline? Does it show one variable? Two variables?

6. After students have collected their examples of data graphics and defined them, assemble the students into groups of three or four to further analyze and critique the data graphics. Hand out the Data Graphics Guidelines Sheet on page 36 to guide their critique.

7. Below their data graphic descriptions, students should record their group's critique. (Students may want to take notes on the group discussion, then rewrite them onto the bottom of the graphic sheet once the discussion is complete.) What is the purpose and quality of the graphic? Are there advantages to using that type of graphic with that data set? Are there disadvantages? How would they improve on the graphic? You may wish to have each group nominate a best and worst graphic for class discussion.

NOTE: The procedure uses *USA Today* because of its particular strength in using data graphics. Other newspapers, journals, or technical books might be used as substitutes.

Data Graphics Guidelines Sheet

Graphics are intended to reveal data, to make them more accessible to your eye and mind than raw numbers. Graphics put complex ideas into a format that is comfortable to look at, and therefore comfortable to understand. Data graphics can also distort the perception of the data. Sometimes this distortion is on purpose; sometimes the graphics are just poorly done.

Data graphics can be designed for a specific audience, or for a broad audience. Consider the audience for which the data graphics you collected were designed. To entertain a general newspaper readership? To show elementary school students a relationship between hours of sunlight and growth of bean plants?

Data graphics can also be designed for a specific purpose. Are the data being presented to support an economics news analysis? To show the interaction of several weather variables? Graphics should reflect, in both tone and content, the purpose for which they were created. They may be descriptive, they may be decorative, they may encourage exploration of a concept behind the data.

GUIDELINES FOR THE PRESENTATION OF DATA IN GRAPHICAL FORM

A successful data graphic

- ▶ shows the data
- ▶ labels the variables and the scales on which they are presented
- ▶ guides the eye so the reader can compare different pieces of data
- ▶ presents many numbers in a small space
- ▶ provides ways to look at several layers of detail, from an overview of the data to the data's fine structure
- ▶ does not distort the meaning of the data
- ▶ induces the reader to think about the substance of the data rather than about the graphic design or the design technology

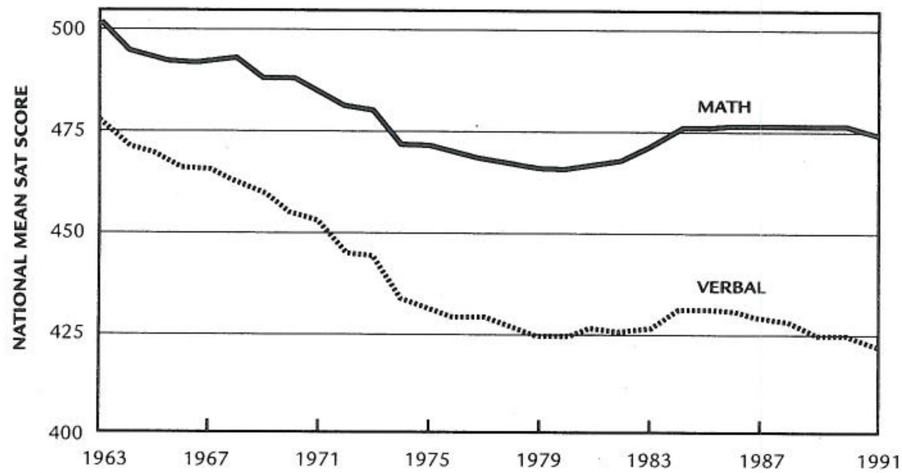
An unsuccessful data graphic

- ▶ does not label the variables or their scales
- ▶ distorts the relationships among data through faulty use of perspective or "artistic" license
- ▶ connects data points with lines that have no meaning
- ▶ distorts the relationships among data by falsely implying relationships, e.g. showing incidence of some event by county on a map, when geography has nothing to do with the events
- ▶ shows relationships or causes that do not exist, e.g. the amount of solar radiation compared with the prices on the New York Stock Exchange
- ▶ neglects to show all the data
- ▶ uses so much color or decoration that the data are obscured

Pre-Activity Tables

TABLE 8.1

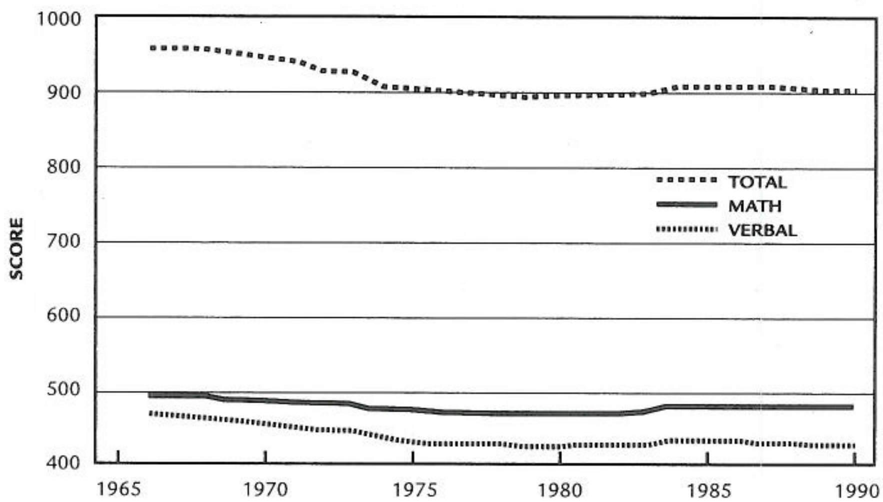
Decline in SAT Scores, Graphed on a 100-Point Axis



(Source: Charles Murray and R.J. Herrnstein. 1992. "What's Really Behind the SAT-Score Decline?" *The Public Interest* No. 106 (Winter 1992), p. 33.)

TABLE 8.2

Decline in SAT Scores, Graphed on a 600-Point Axis



(Source: C.C. Carson, R.M. Huelskamp, and T.D. Woodall. 1992. *Perspectives on Education in America.* Albuquerque, NM: Strategic Studies Center, Sandia National Laboratories, p. 39.)

Using Statistics to Persuade

ACTIVITY SUMMARY

Students research data on an athlete, then use statistics to justify their athlete is the "Most Valuable Athlete."

DATABASE SKILLS

Identifying and researching the most relevant data

Preparing statistical summaries from a set of data

Understanding how statistics can be used for multiple purposes

Using statistics to present an argument

Making oral presentations based on database information

TIME MANAGEMENT

One class period to introduce the activity and concepts.

One week (out of class) for research and presentation preparation.

One class period for student presentations.

MATERIALS

Student Sheet, page 40
Optional materials might include:

- computer on which students can design graphs and tables
- graph paper
- cardboard
- markers
- daily newspaper with sports pages
- sports almanacs (see Potential Resources sidebar on page 39.)

BACKGROUND

Even if you don't read the newspaper sports pages every morning, you've probably heard people arguing that their home team is having its strongest season ever, or that a certain player's career seems to be on the upswing. People talking about sports often use statistics to help prove their argument. Examples of sports statistics include batting averages, horse racing odds, and pass attempts vs. pass completions. Any way you look at it, statistics are an integral component of sports.

Because these same techniques can be used for discussions about nearly everything, it's important to be able to feel comfortable using statistics. This activity will help students learn how to stress different numbers to make different points. They will even learn to make contrasting arguments using the same data. As your class will discover, a successful argument has a lot to do with how the data are presented. Before students prepare their arguments, discuss various methods for using the same data to present different sides of an argument.

PROVIDING ALL THE INFORMATION

One way to present different sides of an argument is to leave out pertinent information. And stressing different numbers can present a very different picture of the same set of data. Presenting the complement of a fact is a similar technique. For example, rather than say that a player got a hit at 33 percent of his "at bats" (a very good average in baseball), report that he was "out" 67 percent of the time.

RANK VS. PERFORMANCE

A fair representation of a study shows all pertinent data. But the fairest representations also show actual performance. Remind students to always keep in mind that ranks obscure performance. For example, in the Atlanta Olympic's 200-meter dash, Michael Marsh of Houston came in last. But his performance (his actual running time) still left him the 8th fastest man in the world at 200 meters. And if he had improved his performance by a mere 7 percent, he would have gotten the gold. Another example: Among people who have hit over 700 home runs, Babe Ruth ranks dead last. Keep rank vs. performance in mind when describing data.

RATES VS. NUMBERS

It is equally important to be aware of whether you are dealing with rates or numbers. For example, until 1997, black students dropped out of school at a higher rate than white students, and Hispanic students had a much higher dropout rate than either of the two other groups. However, the typical dropout is a white, male suburbanite, because whites make up almost 70 percent of the K-12 population. Their dropout rates are lower, but their dropout numbers are higher.

The Hispanic dropout rate provides a good example of how a single statistic can mask important information. Hispanic students originally from other countries drop out at about three times the rate of Hispanic students born in the United States. For that matter, the Hispanic dropout rate varies immensely by country of origin. To address the "high dropout rate among Hispanic students," one would need to know more than just the "Hispanic dropout rate."

Because this activity leaves much room for independent research and decisions, it might be used as a final assessment of students' mastery of statistics. When all the students have had a chance to nominate their athlete, have the class vote for "Most Valuable Athlete."

PROCEDURE

1. Provide the class with several sports almanacs which cover the same season or seasons. (Some useful sports almanacs are listed under "Potential Resources" at right.) You may wish to assign the sport that will be followed, or take a class vote to increase student ownership of this activity. If you have access to statistics for your school's athletics teams, you may want to use these data instead. Another option is to have students research sports legends, such as Babe Ruth or Joe Namath; this method would teach students to compare very different data sets.

2. Have each student select a different athlete to research.

3. Ask students to use the statistics provided in these almanacs to prove that their athlete should be the "Most Valuable Athlete." Explain that they will be asked to argue their athlete's merits based on statistics. Over the course of the next few days, students should use these almanacs to research their athlete's performance. Students following a baseball player, for example, would track hits, runs, on-base percentages, and errors.

4. Students should also learn about other athletes chosen in class so they can make comparisons.

5. To help make their argument, students should provide their data in multiple formats—for example, in addition to verbally discussing their athlete's strengths, students might use charts, tables, and histograms.

6. Devote a class period for presentations of students' data analyses. Give each student a chance to nominate their player for "Most Valuable Athlete," and to defend his or her choice.

7. (Optional) Using the data from this activity, have the class create "advertisements" promoting each athlete's strengths. Students can present their advertisements to the class and explain how and why they emphasized and de-emphasized particular data. In addition to improving students' communication skills, this optional step provides an opportunity for assessing their understanding and facility with these skills.

8. You might assign the Student Sheet on page 40 as homework, or use it as the basis for assessment.

ANSWERS TO STUDENT SHEET

Answers will vary, but should display a clear understanding of how statistics can be used to support arguments. Students should be able to refer to their research and class presentations when answering these questions.

POTENTIAL RESOURCES

School and local libraries contain a variety of almanacs which will contain data for this activity. Examples of sports almanacs include:

Mass Market Paperback, ed. 1996. *1996 Baseball Almanac.* New York, NY: Signet.

Marty Strasen. 1996. *1996–1997 Basketball Almanac.* New York, NY: Signet.

Hughes, M., *et. al.* 1996. *1996–1997 Hockey Almanac.* New American Library.

Hassan, John, ed. 1997. *The 1998 ESPN Information Please Sports Almanac.* New York, NY: Warner Books.

NOTE: The Activity 8 pre-activity, on misleading data graphics, might also be used before conducting this activity.

Using Statistics to Persuade

After hearing all the nominations, answer the following questions on a separate sheet of paper.

1. Write a few paragraphs discussing how different methods of presentation have different benefits.

2. It's important to understand how statistics can be used to prove opposite sides of an argument. Write two paragraphs: one supporting your athlete, the other against your athlete.

4. How did you select the statistics for your presentation?

5. How did you use those statistics to prove a point?

6. When working with databases, you will occasionally need to make predictions. To do this, you will have to look for patterns in past data. Using data discussed during class, project which athlete will be most successful during the following season.

The Steps of Decision Making

TIME MANAGEMENT

One to two class periods to complete the decision charts. (Students might need to perform research outside the classroom between these two periods.)

One class period to make a final decision and discuss the process.

MATERIALS

Student Sheets, pages 43-45
Blank Decision Chart, page 46

BACKGROUND

This activity introduces decision analysis, a method through which students can organize information in order to reach a decision. This method can be applied to almost any problem, whether the decision is personal or part of a class discussion on a particular science subject.

Databases are ideal tools for making decisions. It's a two-way process: skills learned through decision analysis can also help students gauge the importance of different data sets. A good decision maker, however, can assess hard data along with intangible factors when making decisions. Intangible factors may include ethical concerns, economic aspects, and public attitudes or perceptions. Decision analysis gives students a way to compare both data and perception as separate—yet important—factors in the decision-making process.

For students who learn the logic of decision making, studies reported in the news, labs they conduct in class, symbols, and vocabulary words are no longer disconnected pieces but, when combined, comprise the entirety of the scientific process.

PROCEDURE

1. In this activity, students will use a decision chart to decide whether to raise the speed limit. The decision chart on the Student Sheet has been prepared in advance for this particular activity. The format is flexible; more columns and rows can be added depending on your needs.

2. Give each student a copy of the Student Sheets, pages 43-45. Have a volunteer read the scenario (page 44) aloud.

3. Now take a few minutes to have students brainstorm. What goals would they like to achieve? Examples include: increase revenue, reduce traffic fatalities, appease various special-interest groups, etc. Have students list about five or six goals.

4. Some goals are not as important as others. "Appeasing special interest groups" may not be quite as important as "Reducing traffic fatalities." Taking importance into account can be an effective way to organize your thoughts and simplify your decision. Ask students (on their own, rather than as a class) to rank the goals in terms of importance.

5. Have students write these goals in their decision chart according to how they rated their importance. The more important the goal, the higher it should go in the chart. Table 10.1, page 42, provides a sample chart to which you can refer while explaining these steps and how to fit them into the chart.

6. Now what options are available to accomplish your goals? Have students write their options across the top of their decision chart. (Two probable options for this scenario are: raise the speed limit to 65 mph, or keep the speed limit at 55 mph. If your students would like to add additional options, they can list them in the extra columns provided.)

7. The next step is to determine the outcomes of each option. Under each

ACTIVITY SUMMARY

Students create a "decision chart," a decision-analysis method which organizes multiple variables.

DATABASE SKILLS

Using a scientific organization method to reach a logical decision

Learning how to weigh data when forming a decision

Identifying gaps in information

Making decisions without complete data

Understanding how public perception of risk may affect a decision

TEACHING TIPS

- **Before beginning a database unit, you might pose a question to students that relates to a particular set of information. Have them write a paragraph on their initial decision. At the end of the unit, repeat the same question. At this point, students should be able develop a decision chart using information gleaned from the database, their own lab investigations, and/or further research. Their revised paragraphs should reflect their ability to draw upon all these factors in order to reach a final decision.**
- **The Blank Decision Chart, page 46, can be reproduced to use in conjunction with future activities.**
- **Further information about decision analysis is provided in *Decisions, Based on Science.* See the Bibliography and Resources section in Appendix E, page 52, for more information.**

TABLE 10.1

Sample Decision Chart

	OPTION 1: Raise speed limit	OPTION 2: Don't raise speed limit
Goal: Reduce accidents and fatalities	*Accidents that occur are likely to be more severe; Police could be freed from speed traps to concentrate on reducing drunk driving.*	*Town will lose ticket revenues; Accidents contribute business to local car repair shops.*
Goal: Reduce costs to town	*Costs may increase if accidents increase, but it is not clear that they will.*	*Costly accidents may still occur for other reasons; Police will be needed to enforce speed limit which takes them from other policing duties.*
Goal: Reduce emissions	*Emissions are higher at increased speeds, but many drivers are already driving at the higher speed.*	*If the lower speed limit is strictly enforced, emissions will be slightly lower.*
Goal: Increase town revenues	*Town will lose ticket revenues; Accidents contribute business to local car repair shops.*	*Speeding tickets generate a lot of town revenue.*

option, have students write a brief description of how the options would meet each goal. Write these outcomes in the chart's remaining boxes (two outcomes have been provided as an example.)

Students should be as specific as possible; use facts, figures, and numbers wherever possible. If there are several ways an option will affect a particular goal, they should neatly list each result within that outcome box. Many of the outcomes are discussed in the Background section (page 44) of the Student Sheet. If students are unsure about any outcome, they should leave that box blank.

8. When students have filled in all the outcomes of which they are fairly certain, give them time to research the outcomes they have questions about. Sources of information include your school library (encyclopedias, *Facts on File*, and the *Reader's Guide to Periodical Literature* might be especially useful), the Internet, back issues of the newspaper, or groups devoted to that particular subject. Students may need several days to complete this research. Remind students that decision makers rarely have all the information they'd like to have; if they can not find information for several outcome boxes, they should just use their best guess.

9. When you make a decision, you consider which goals are most important, and which results will help you meet those goals. Yet not all results are "sure things." Have students think about how much confidence they have in each of the outcomes that they listed. With 10 representing "very probable" and 1 representing "not probable," have them record their confidence levels in each outcome box. Remind students that few outcomes are as certain as people might wish them to be.

10. We say that two options have the same expected value if the choices are likely to lead to equally good results. The expected value of an outcome is the probability (P) it will happen multiplied by the value (V) of the option. (Expected Value = P x V.) Have students calculate the expected value of each option, then compare them.

11. Now have students make their decision. They've listed their goals by order of importance; they've selected the options (actions) that have to be taken to achieve their goals; they've noted the probability that each goal will be reached by each option. The option that best achieves the most important goals is the decision they choose.

12. After completing the activity, ask students if they feel their final decision is one they would really make. If not, have them examine the thinking that went into their decision. Have they missed any important goals in their analysis? Do they think any facts or estimates they used to set the importance of their goals might be flawed? Add these new ideas, then see if the decision changes as a result.

This cyclic method of reconsidering is central to making decisions. It enables students to logically examine their thinking—and will encourage them to notice that they can provide reasons for what they may have thought were "feelings," or "just a reaction." One purpose of decision analysis is to enable people to recognize inconsistencies—and consistencies—in their thinking and planning.

ANSWERS TO STUDENT SHEET

N/A

The Steps of Decision Making

A decision resolves a divergence of paths; each path will have benefits and risks. A good decision maker identifies each path's benefits and risks, uses evidence to weight them logically, and then decides. The terms used in this activity are: decision (divergence of paths), options (each path), and outcomes (a path's benefits and risks; how the choice of each option will affect each outcome). There are several decision-making techniques, but decision-making processes generally follow four steps.

STEP ONE
What's the Decision?

A. Before you do any analysis, make a tentative decision and record it.

B. Determine whose decision it is to make. Determine at what level (personal, local, national, international) the decision is being made, and make sure all goals and options are stated at this level.

C. Make a list of goals for the decision.

STEP TWO
What Should Happen?

A. In the goals boxes of the Blank Decision Chart, write the three main goals from the Step One goals list.

B. Deliberate on and select two to three options. Write those options in the options boxes of the Decision Chart.

C. How will selecting those options affect the goals? Enter obvious or initial guesses in the outcomes boxes in the Decision Chart.

D. Identify initial research needs from outcome boxes that are missing information.

STEP THREE
What Do We Know?

A. Research outcomes.

B. Complete Decision Chart with information from research.

STEP FOUR
What's the Answer?

A. Assign numerical probability estimates for each outcome in the chart.

B. Assign numerical importance levels for each goal in the chart.

C. We say that two options have the same expected value if the choices are likely to lead to equally good results. The expected value of an outcome is the probability (P) it will happen multiplied by the value (V) of the option. Calculate the expected value of each option.

D. Select which option is most likely to meet the most important goals.

The Steps of Decision Making

Background Information

SCENARIO

You are a member of the town council of your suburban town. A certain busy stretch of road in town carries about 6,000 cars per day. The road runs through a business area of the town. Some citizens would like the speed limit raised because they say it's too low for road conditions. Your popularity might increase if you raised the speed limit. But another group of citizens wants even more traffic controls to control drunk and reckless driving.

The current speed limit is 30 mph. The actual speed traveled by most cars on this two-lane road is 45 mph. About 12 speeding tickets are issued per day on that stretch of road. The average cost of the ticket is $75.00. The accidents tend to occur when people are slowing and turning into businesses along the road. Drunk drivers also cause accidents, especially at night.

As a member of the town council, you must decide whether to change the speed limit on this section of road. Your traffic administrator has given you some facts about national traffic safety. Read the background, then make a decision.

EFFECTIVENESS OF SPEED LIMITS

According to the National Highway Traffic Safety Administration (NHTSA), 77 percent of fatal accidents occur on two lane roads. And 88 percent of speeding related fatalities occur on roads that were not interstate highways. The effectiveness of speed limits to control traffic speed is somewhat debatable because several studies contradict each other. A federal study completed in June 1996, "Effects of Raising and Lowering Speed Limits on Selected Roadway Sections," indicated that most vehicles travel at the same speed along a given stretch of road—and that speed has little to do with the posted speed limit. The speed at which most cars travel is related to driver perception of a prudent speed to travel. When speed limits are raised, drivers do not significantly speed up. There was no evidence that lowering the speed limit reduced the accident rate or that raising the speed limit increased the occurrence of accidents.

Other studies, based on recent increases in state speed limits on interstates, have been mixed. In some state studies, the average speed did increase when speed limits were raised. According to these studies, people seem to pick a speed slightly higher than the speed limit no matter how fast the limit. But in other studies, the average speed on interstates did not significantly increase when the speed limit was changed to any "reasonable and prudent" speed.

SAFETY AND SPEED LIMITS

Reducing injuries and deaths in crashes is an important goal. Although actual numbers of fatalities change from year to year, comparisons of fatalities per vehicle miles traveled are the lowest in the history of the automobile. Although our society travels more miles per year by motor vehicle, we are less likely to die per mile traveled. Some reasons for the low fatality rate per mile traveled are improvements in car and road design, improved vehicle safety features, and more control over drunk driving.

Speed is an important factor in injury prevention because the higher the speed the more severe the crash. NHTSA cites "speed-related factors" as contributing to one-third of all crashes. Other causes include drinking alcohol, driving too fast for road conditions, failing to yield, and careless driving. However, even if the speed limit is 55 mph on a certain road, on wet days a prudent driver would reduce speed. Therefore, posted speed limits are not the only factors in speed-related accidents.

COSTS OF ACCIDENTS

The costs of crashes can be categorized as: property damage, productivity loss, medical, insurance administration, legal, funeral, and emergency service costs. The most costly crashes to society are those in which someone is injured or killed. But the most common crashes are property damage only (86 percent of all crashes).

Alcohol-involved crashes tend to be more severe and cause up to 30 percent of all crash costs. Exceeding the speed limit or driving too fast for conditions were also big factors in costs. However, as noted above, driving too fast for conditions is not necessarily related to the speed limit.

Using safety devices reduces the cost of accidents because they reduce injuries. Used properly, seat belts and airbags do reduce the overall injury and death rate in accidents. Again, these are less effective at high speeds.

Although accidents can be tragic and cost money to society, they contribute income to other parts of society: car repair shops, new car dealers, and medical practitioners. Insurance companies make money, not through the actual accidents, but from people who pay insurance premiums yet don't end up needing insurance money.

POLLUTION AND SPEED

Lower speeds reduce emissions—up to a point. A car that drives smoothly and slowly will pollute slightly less that a car that drives smoothly and quickly. However, a car that stops and starts in traffic will also emit a good deal of pollution.

If one assumes that cars are already traveling faster than the speed limit, raising the limit may not necessarily raise emissions. But if the average speed does increase with the raised speed limit, then emissions will increase.

A DECISION CHART: HOW THE PIECES FIT TOGETHER

	OPTION 1	OPTION 2	OPTION 3
GOAL 1	*Outcome* *Probability* *Expected Value*	*Outcome* *Probability* *Expected Value*	*Outcome* *Probability* *Expected Value*
GOAL 2	*Outcome* *Probability* *Expected Value*	*Outcome* *Probability* *Expected Value*	*Outcome* *Probability* *Expected Value*
GOAL 3	*Outcome* *Probability* *Expected Value*	*Outcome* *Probability* *Expected Value*	*Outcome* *Probability* *Expected Value*

The Environmental Protection Agency (EPA) has found that 10 percent of the cars on the roads emit 50 percent of the pollution from cars. These "gross emitters" are cars that are not tuned correctly or do not have catalytic converters. The majority of cars built since 1990 emit small amounts of pollutants at any speed.

REVENUES FROM SPEEDING TICKETS

Many towns rely on revenues from motor vehicle tickets to pay for essential government functions. When speed limits are slightly lower than prevailing traffic, the police have more opportunities to issue tickets. Raising speed limits might reduce the number of tickets collected.

Motor vehicle crashes cost society an estimated $4,800 per second. The total economic cost of crashes in 1994 was estimated at $150.5 billion. The 1996 costs of speeding-related crashes were estimated to be $28.8 billion—or $913 per second.

Other ways to collect ticket revenues could be to increase ticketing for other offenses—such as broken lights or expired tags—rather than relying on speeding tickets.

Blank Decision Chart

	OPTION 1	OPTION 2	OPTION 3	OPTION 4
GOAL A:				
GOAL B:				
GOAL C:				

Using Graphing Calculators

The following steps are based on the TI-82 and TI-83 calculators. Other calculators have similar features, but the steps may not be exactly the same.

NOTE: Bold print indicates an actual button, such as the up arrow, on the calculator.
Italics indicate words which appear on the screen.

ENTERING DATA

1. Press **stat.**
2. Choose *1:Edit*.
3. Space for three lists should appear on the screen, labeled *L1*, *L2*, and *L3*. You can access additional lists using the right arrow to move the cursor past the edge of the screen.
4. Enter data by typing the number and pressing **enter** or the **down arrow** to scroll down the list.

CLEARING DATA

There are several ways to clear data that have already been entered in the lists. The easiest is to use the **up arrow** to position the cursor over the *L1* (or whatever list is to be cleared) and press **clear** followed by **enter**.

SORTING DATA

The options for sorting data can be found on the stat menu.

1. Press **stat.**
2. Choose between *2:SortA* (ascending order) and *3:SortD* (descending order).
3. You will be returned to the main screen and the operation you chose will appear. If anything else appears on the same line with the operation you chose, use the arrows to move the cursor over it and press **del** (delete).
4. Tell the calculator which list is to be sorted by choosing the appropriate *L_* from above the keys numbered 1 to 6. Be sure to press **2nd** to access the list desired.
5. Press **enter.** (Users of the TI-83 will need to close the automatic parentheses first).
6. The word *done* appears on the screen. The list is now sorted. Press **stat** and choose *1:Edit* from the menu to see the sorted list.

PLOTTING DATA

Both the TI calculators offer several types of graphs: scatter plot, line plot, histogram, and box-and-whiskers plot. This graphing feature is separate from the graphs of functions. For best viewing, turn off or erase any graphs stored under **y=**.

TO GRAPH STATISTICAL DATA

1. Press **2nd** followed by **y=** (stat plot)
2. A list of stat plots appears showing whether they are on and which type of graph they display. Choose a plot. Most likely all other plots should be turned off.
3. Use the **arrows** to move the cursor over *On* and press **enter.**
4. Using the **arrows,** choose a plot to be drawn. Choose a list to be the *Xlist*. Choose either the *Ylist* or the *frequency*, depending on which type of graph was selected. Note that the *frequency* and the *Ylist* can be an entered list or a constant number.
5. Press **graph.** The window may need adjustment. This can be done either by pressing **window** and entering the ranges desired or by pressing **zoom** and choosing one of the calculator-generated sizes.
6. Try using **trace** to move around the graph and observe its values. When used with the box-and-whiskers plot, **trace** will indicate the minimum, maximum, quartiles and median.

CALCULATING STATISTICS

The calculator will compute basic statistics on any data set entered in a stat list.

1. Press **stat.**
2. Use the **right arrow** to move the cursor to the second menu and press *calc.*
3. *1-Var Stats* is as far as most high school courses investigate. Choosing this feature will return the cursor to the main screen and write the words
1-Var Stat.
4. By default, the calculator will use L1. If another list is to be used, enter the list from above the 1 to 6 keys.
5. Press **enter**. The results of the calculations will appear. the results do not fit on a single screen so you will need to use the **down arrow** to scroll to see those which did not fit.

FINDING LINE OF BEST FIT

The calculator offers several regression analyses. The TI-82 provides a coefficient of correlation so students can see how exact the fit is, whereas the TI-83 does not. For both calculators the steps to calculate regression are the same.

1. Press **stat.**
2. Move cursor to *calc* menu.
3. Select the regression desired. This will return the cursor to the main screen.
4. Press **enter.**

Good Graphing Techniques

TABLE B.2

Data with point protectors

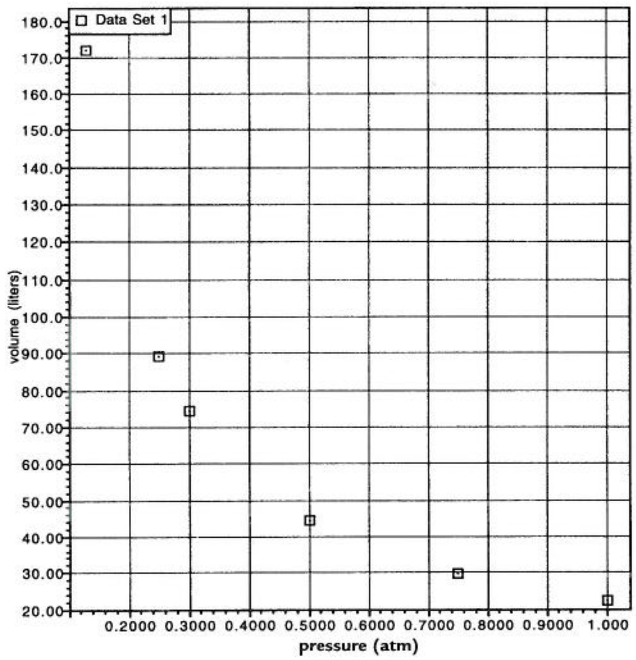
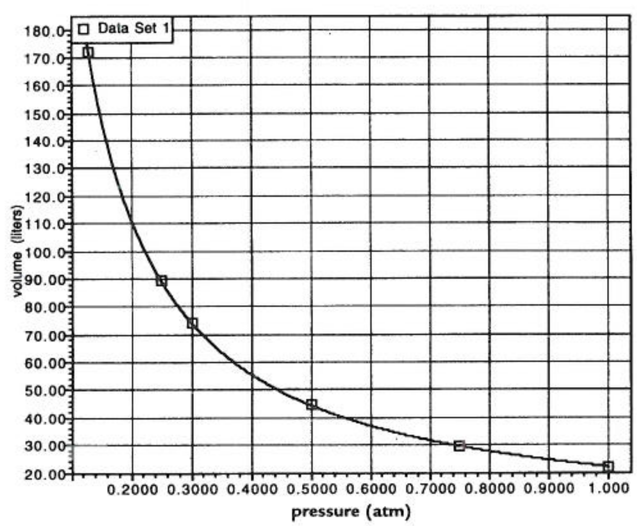


TABLE B.3

Data with fitted curve



$$f(x) = 22.12\ x^{-1.006}$$
$$R = 1.00$$

Good graphing techniques are extremely important to understanding data. Consider the data in Table B.1, which represent the relationship between the pressure (P) of a gas and the volume (V) the gas occupies. This principle is known as Boyle's Law.

TABLE B.1

PRESSURE ATM	ERROR IN P... +-ATM	VOLUME LITERS	ERROR IN V... +-LITERS
0.1300	0.02000	172.1	5.000
0.2500	0.02000	89.28	5.000
0.3000	0.02000	74.35	5.000
0.5000	0.02000	44.49	5.000
0.7500	0.02000	29.55	5.000
1.000	0.0200	22.08	5.000

Several things should be noted about the way the data are recorded. Each column of data are labeled with the quantity represented and the unit in which it was measured. The error in each measurement is also recorded. (Significant figures are used consistently.)

The first step in graphing these data are to decide which variable is the independent (placed on the X-axis) and dependent (placed on the Y-axis.) Boyle's Law is usually represented with pressure as the independent variable.

The second step is to establish a scale for each of the axes. The scale should evenly cover the entire range of data. However, it is not necessary to start at a "0" origin, or to show quadrants that do not have any data in them. This graph has a domain (X-axis) of 0.1300 to 1.000 atm and a range (Y-axis) of 22.8 to 172.1 liters. Each axis should be labeled with the quantity it represents and the unit in which it is measured. Also, each graph should have a title.

Notice the purpose of point protectors, the boxes around the data points. Point protectors allow you to see where the original points were located when additional information is added to the graph. (See Table B.2.)

One piece of information that can be added to a graph is a best fit curve. Students should be able to describe the data with a smoothly-drawn curve through the points. It is important for students to be able to decide when the data are best described by a line, connecting point to point, or a smooth curve. (See Table B.3.)

Error bars can be used to represent the error in each measurement. This error is an absolute error, and is expressed in the same units as the measurement. The size of the error bar should be to scale on the axis. Error bars are most commonly used on the dependent variable; however, they may also be used on the independent variable. (See Tables B.4 and B.5.)

TABLE B.4

Data with error bars in independent variables

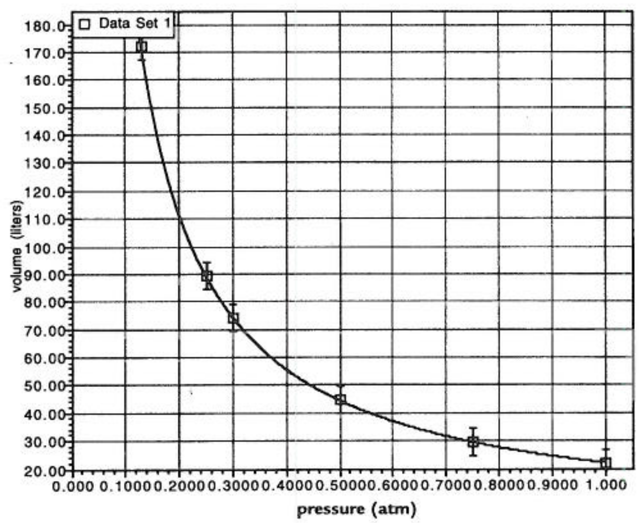
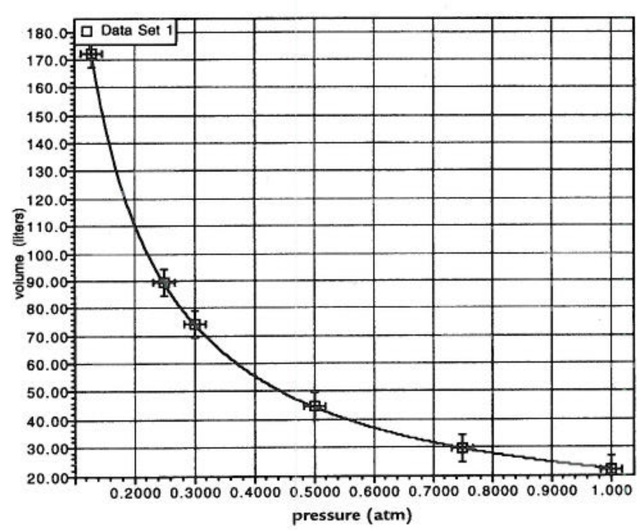


TABLE B.5

Data with error bars in both variables



Representing Data with Equations

The term regression refers to determining the mathematical equation that best represents the data. Traditionally, this was a tedious process requiring extensive algebra. However, new graphing calculators and computer programs now perform this task easily. (See Appendix A.)

If we consider the graph of Boyle's Law (as shown in Appendix B), we can describe the data as an inverse relationship. The formula for Boyle's Law, where P is pressure, V is volume, and *k* is the constant of proportionality, can be represented by:

$$V \propto 1/P$$

or as an equality with k the constant of proportionality:

$$PV = k \text{ or } V = k/P$$

where V is on the Y-axis and P is on the X-axis. There are two effective ways to analyze these data. The first method uses a regression with the power function. This returns an equation in the form y = A*x ^B. Boyle's Law can then be described as:

$$Y = 22.12X^{-1.006}$$

This can be transformed into a recognizable form of Boyle's Law by the following:

$$Y = \text{Volume} = V \quad X = \text{pressure} = P$$

$$\text{substituting } C = 22.12P^{-1.006}$$

$$-1.006 \approx -1$$

$$V = 22.12P^{-1}$$

$$V = \frac{22.12}{P}$$

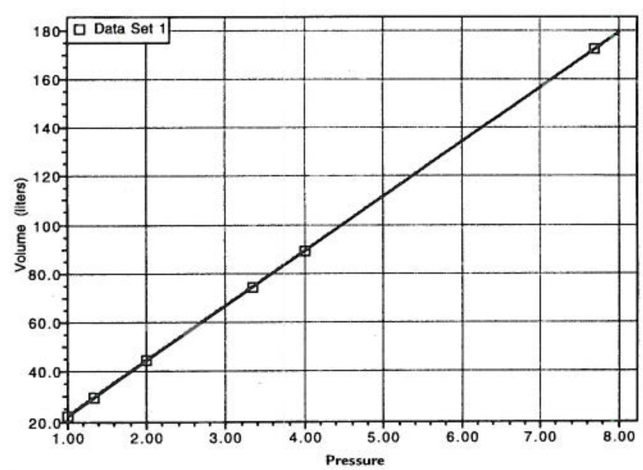
$$PV = 22.12 \text{ or } PV = k$$

Another way of describing these data would be to substitute the measurement X = 1/P. This will put the equation in the form V = *k*X, which is linear. The slope of the best fit line represents the value *k*.

Most calculators include a value that represents how closely the equation fits the data. This value, called the coefficient of regression or correlation coefficient, is represented by R or R^2. R will have the range of -1 to 1, depending on whether the data shows positive or negative associations. The closer the value is to -1 or 1 the better the regression curve fits the data. The value R^2 will have a range of 0 to 1. The closer the value is to 1 the better the regression curve fits the data. Because of the ease of trial and error, many people recommend simply trying different regressions and looking for the best correlation coefficient. However, a much greater understanding of the data can be found when the student understands the shapes created by different regression equations and can choose the regression that provides the most meaningful interpretation of the data.

TABLE C.1

Graph of 1/P



Algebraic Method of Determining Outliers

The word outliers indicates values that are widely separated from the rest of the data. If you think you have spotted an outlier, it is worth some special thought about why it is different from the rest. Trying to make sense out of the outliers can be an important part of interpreting data. It is not reasonable, however, to automatically call the upper and lower extremes outliers. Any data set has extremes, and we don't want to put extra energy into trying to interpret them unless they are separated from the rest of the data. It is helpful to have a test to aid in making the decision, especially when there are a moderate to large number of observations (say 25 or more.) The Q test is a simple algebraic test for determining if a value should be included in an average or if it is an outlier. Given the set of measurements determined experimentally:

SAMPLE	PERCENT IRON
A	36.04
B	36.09
C	35.12
D	36.07
E	36.10

The average (mean) of these data set is calculated to be: $\overline{x} = 35.88$, and the standard deviation is: $\delta = .428$.

Looking at the data, all of the values except C seem to closely agree. The Q test determines mathematically whether the trial C can be discarded from the average. The Q test is applied to the value that is furthest from the average—the deviant result.

$$Q \text{ is defined as: } \frac{|\text{deviant result} - \text{nearest neighbor}|}{\text{max value} - \text{min value}}$$

$$\text{For the data given: } \frac{|35.12 - 36.04|}{36.10 - 35.12} = \frac{.92}{.98} = .939$$

The calculated Q value is compared to a list of tabulated Q values which determine the confidence with which the value can be discarded. For typical data analysis, a confidence of 90 percent is usually used. If the calculated Q value is larger than the tabulated value, the trial should be discarded as an outlier. The values for Q - 90 are:

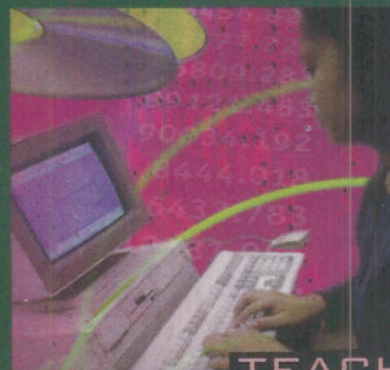
NUMBER OF MEASUREMENTS	Q - 90
3	.94
4	.76
5	.64
6	.56
7	.51
8	.47
9	.44
10	.41

For our example, we have 5 trials. Because the calculated Q is larger than .64, trial C can be discarded as an outlier. The mean and standard deviation can then be recalculated:

New average (mean) without trial C: $\overline{x} = 36.08$

New standard deviation: $\delta = .026$

Databases can bring real-world, up-to-the-minute science to your classroom. But how to use them?



TEACH WITH
DATABASES

Database Basics leads students through the process of gathering data, organizing it, describing it, identifying trends, and then using data to make decisions.

Concepts include:

- Understanding database organization
- Identifying survey and measurement error
- Working with data indicators
- Plotting data
- Describing data with statistics
- Preparing statistical summaries from a set of data
- Explaining data with graphics
- Weighing data when making a decision

Each activity includes teaching tips, data collection methods, sample data sets, procedures, student worksheets, and assessment suggestions. Extensive background information provides suggestions for incorporating these data skills into the science curriculum, and discusses how these skills fit the *National Science Education Standards*. If you've considered teaching with a database, *Database Basics* will help you use these important resources to bring your students a new appreciation for the process of science.

NSTA
National Science
Teachers Association
1840 Wilson Blvd.
Arlington, VA 22201-3000
www.nsta.org

Stock # PB142X ISBN 0-87355-170-2