In This Section

Level A

Children are surrounded by data. They may think of data as a tally of students' preferences, such as favorite type of music, or as measurements, such as students' arm spans and number of books in school bags.

It is in Level A that children need to develop data sense—an understanding that data are more than just numbers. Statistics changes numbers into information.

Students should learn that data are generated with respect to particular contexts or situations and can be used to answer questions about the context or situation.

Opportunities should be provided for students to generate questions about a particular context (such as their classroom) and determine what data might be collected to answer these questions.

Students also should learn how to use basic statistical tools to analyze the data and make informal inferences in answering the posed questions.

Finally, students should develop basic ideas of probability in order to support their later use of probability in drawing inferences at Levels B and C.

It is preferable that students actually collect data, but not necessary in every case. Teachers should take advantage of naturally occurring situations in which students notice a pattern about some data and begin to raise questions. For example, when taking daily attendance one morning, students might note that many students are absent. The teacher could capitalize on this opportunity to have the students formulate questions that could be answered with attendance data.

Specifically, Level A recommendations in the Investigative Process include:

I. Formulate the Question

→ Teachers help pose questions (questions in contexts of interest to the student).

→ Students distinguish between statistical solution and fixed answer.

II. Collect Data to Answer the Question

→ Students conduct a census of the classroom.

→ Students understand individual-to-individual natural variability.

→ Students conduct simple experiments with nonrandom assignment of treatments.

→ Students understand induced variability attributable to an experimental condition.

III. Analyze the Data

→ Students compare individual to individual.

→ Students compare individual to a group.

→ Students become aware of group to group comparison.

→ Students understand the idea of a distribution.

→ Students describe a distribution.

→ Students observe association between two variables.

→ Students use tools for exploring distributions and association, including:

- Bar Graph
- Dotplot
- Stem and Leaf Plot
- Scatterplot
- Tables (using counts)
- Mean, Median, Mode, Range
- Modal Category

IV. Interpret Results

→ Students infer to the classroom.

→ Students acknowledge that results may be different in another class or group.

→ Students recognize the limitation of scope of inference to the classroom.

Example 1: Choosing the Band for the End of the Year Party—Conducting a Survey

Children at Level A may be interested in the favorite type of music among students at a certain grade level. An end of the year party is being planned and there is only enough money to hire one musical group. The class might investigate the question: *What type of music is most popular among students?*

This question attempts to measure a characteristic in the population of children at the grade level that will have the party. The characteristic, favorite music type, is a categorical variable—each child in that grade would be placed in a particular non-numerical category based on his or her favorite music type. The resulting data often are called *categorical data*.

The Level A class would most likely conduct a census of the students in a particular classroom to gauge what the favorite music type might be for the whole grade. At Level A, we want students to recognize that there will be individual-to-individual variability.

For example, a survey of 24 students in one of the classrooms at a particular grade level is taken. The data are summarized in the frequency table below. This *frequency table* is a *tabular representation* that takes Level A students to a summative level for categorical data. Students might first use *tally marks* to record the measurements of categorical data before finding frequencies (counts) for each category.

Table 2: Frequency Count Table

Favorite	Frequency or Count
Country	8
Rap	12
Rock	4

A Level A student might first use a *picture graph* to represent the tallies for each category. A picture graph uses a picture of some sort (such as a type of musical band) to represent each individual. Thus, each child who favors a particular music type would put a cut-out of that type of band directly onto the graph the teacher has created on the board. Instead of a picture of a band, another representation—such as a picture of a guitar, an X, or a colored square—can be used to represent each individual preference. A child who prefers "country" would go to the board and place a guitar, dot, X, or color in a square above the column labeled "country." In both cases, there is a deliberate recording of each data value, one at a time.

Note that a picture graph refers to a graph where an object, such as a construction paper cut-out, is used to represent one individual on the graph. (A cut-out of a tooth might be used to record how many teeth were lost by children in a kindergarten class each month.) The term *pictograph* often is used to refer to a graph in which a picture or symbol is used to represent several items that belong in the same category. For example, on a graph showing the distribution of car riders, walkers, and bus riders in a class, a cut-out of a school bus might be used to represent five bus riders. Thus, if the class had 13 bus riders, there would be approximately 2.5 busses on the graph.

This type of graph requires a basic understanding of proportional or multiplicative reasoning, and for this reason we do not advocate its use at Level A. Similarly, circle graphs require an understanding of proportional reasoning, so we do not advocate their use at Level A.

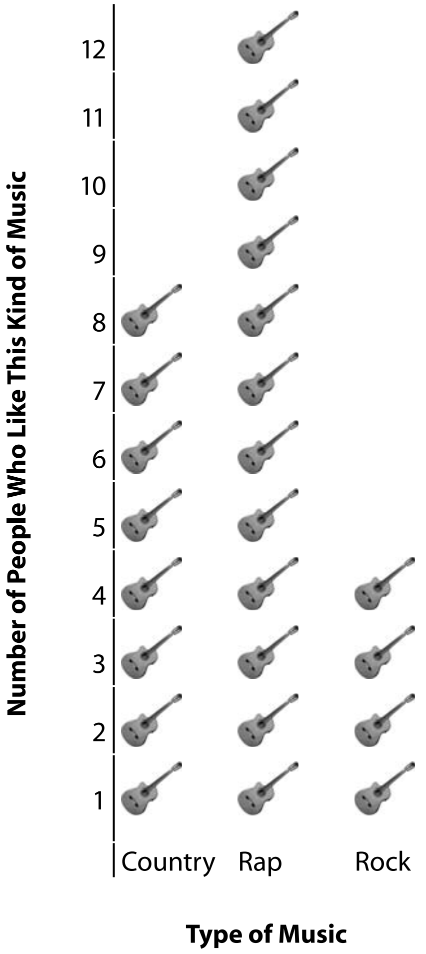

Figure 1: Picture graph of music preferences

A bar graph takes the student to the summative level with the data summarized from some other representation,

“ Students should understand that the mode is the category that contains the most data points, often referred to as the modal category. ”

such as a picture graph or a frequency count table. The bar on a bar graph is drawn as a rectangle, reaching up to the desired number on the *y*-axis.

A bar graph of students’ music preferences is displayed below for the census taken of the classroom represented in the above frequency count table and picture graph.

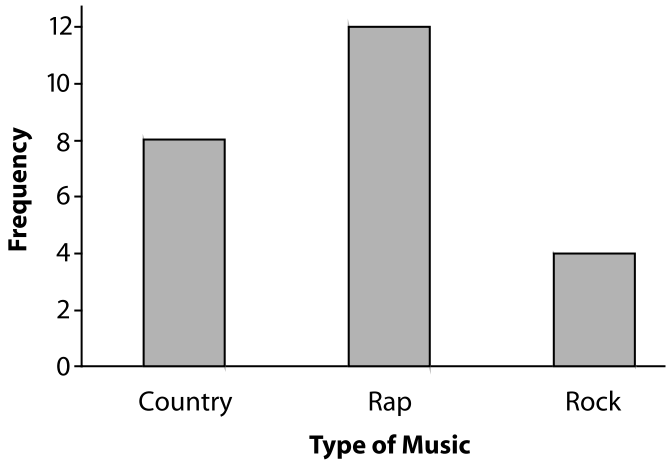

Figure 2: Bar graph of music preferences

Students at Level A should recognize the *mode* as a way to describe a “representative” or “typical” value for the distribution.

The mode is most useful for categorical data. Students should understand that the mode is the category that contains the most data points, often referred to as the *modal category*. In our favorite music example, rap music was preferred by more children, thus the mode or modal category of the data set is rap music. Students could use this information to help the teachers in seeking a musical group for the end of the year party that specializes in rap music.

The vertical axis on the bar graph in Figure 2 could be scaled in terms of the proportion or percent of the sample for each category. As this involves proportional reasoning, converting frequencies to proportions (or percentages) will be developed in Level B.

Because most of the data collected at Level A will involve a census of the students’ classroom, the first stage is for students to learn to read and interpret at a simple level what the data show about their own class. Reading and interpreting comes before inference. It is important to consider the question:

What might have caused the data to look like this?

It is also important for children to think about if and how their findings would *“scale up” to a larger group*, such as the entire grade level, the whole school, all children in the school system, all children in the state, or all people in the nation. They should note variables (such as age or geographic location) that might affect the data in the larger set. In the music example above, students might speculate that if they collected data on music preference from their teachers, the teachers might prefer a different type of music. Or, what would happen if they collected music preference from middle-school students in their school system? Level A students should begin recognizing the limitations of the scope of inference to a specific classroom.

Comparing Groups

Students at Level A may be interested in comparing two distinct groups with respect to some characteristic of those groups. For example, is there a difference between two groups—boys and girls—with respect to student participation in sports? The characteristic "participation in sports" is categorical (yes or no). The resulting categorical data for each gender may be analyzed using a frequency count table or bar graph. Another question Level A students might ask is whether there is a difference between boys and girls with respect to the distance they can jump, an example of taking measurements on a *numerical* variable. Data on numerical variables are obtained from situations that involve taking measurements, such as heights or temperatures, or situations in which objects are counted (e.g., determining the number of letters in your first name, the number of pockets on clothing worn by children in the class, or the number of siblings each child has). Such data often are called *numerical data.*

Returning to the question of comparing boys and girls with respect to jumping distance, students may measure the jumping distance for all of their classmates. Once the numerical data are gathered, the children might compare the lengths of girls' and boys' jumps using a back-to-back ordered *stem and leaf plot,* such as the one below.

Girls		Boys
	8	
	7	
	6	1
	5	2 6 9
9 7 2	4	1 3 5 5 5
5 5 3 3 3 2 1	3	1 1 2 5 6 7
9 8 7 7 6 4 4 3 2	2	2 3 4 6
	1	

Inches Jumped in the Standing Broad Jump

Figure 3: Stem and leaf plot of jumping distances

From the stem and leaf plot, students can get a sense of shape—more symmetric for the boys than for the girls—and of the fact that boys tend to have longer jumps. Looking ahead to Level C, the previous examples of data collection design will be more formally discussed as examples of observational studies. The researcher has no control over which students go into the boy and girl groups (the pre-existing condition of gender defines the groups). The researcher then merely observes and collects measurements on characteristics within each group.

The Simple Experiment

Another type of design for collecting data appropriate at Level A is a *simple experiment,* which consists of taking measurements on a particular condition or group. Level A students may be interested in timing the swing of a pendulum or seeing how far a toy car runs off the end of a slope from a fixed starting position (future Pinewood Derby participants?) Also, measuring the same thing several times and finding a mean helps to lay the foundation for the fact that the mean has less variability as an estimate of the true mean value than does a single reading. This idea will be developed more fully at Level C.

Example 2: Growing Beans—A Simple Comparative Experiment

A *simple comparative experiment* is like a science experiment in which children compare the results of two or more conditions. For example, children might plant dried beans in soil and let them sprout, and then compare which one grows fastest—the one in the light or the one in the dark. The children decide which beans will be exposed to a particular type of lighting. The conditions to be compared here are the two types of lighting environments—light and dark. The type of lighting environment is an example of a categorical variable. Measurements of the plants' heights can be taken at the end of a specified time period to answer the question of whether one lighting environment is better for growing beans. The collected heights are an example of numerical data. In Level C, the concept of an experiment (where conditions are imposed by the researcher) will be more fully developed.

Another appropriate *graphical representation for numerical data on one variable* (in addition to the stem and leaf plot) at Level A is a *dotplot.* Both the dotplot and stem and leaf plot can be used to easily compare two or more similar sets of numerical data. In creating a dotplot, the *x*-axis should be labeled with a range of values that the numerical variable can assume. The *x*-axis for any one-variable graph conventionally is the axis representing the values of the variable under study. For example, in the bean growth experiment, children might record in a dotplot the height of beans (in centimeters) that were grown in the dark (labeled D) and in the light (labeled L) using a dotplot.

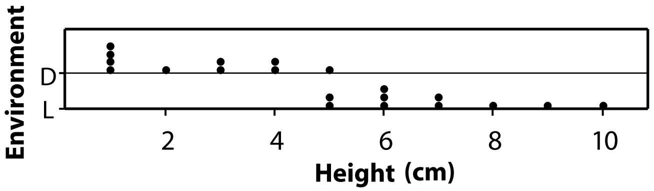

Figure 4: Dotplot of environment vs. height

It is obvious from the dotplot that the plants in the light environment tend to have greater heights than the plants in the dark environment.

Looking for clusters and gaps in the distribution helps students identify the *shape* of the distribution. Students should develop a sense of why a distribution takes on a particular shape for the context of the variable being considered.

- → Does the distribution have one main cluster (or mound) with smaller groups of similar size on each side of the cluster? If so, the distribution might be described as *symmetric*.
- → Does the distribution have one main cluster with smaller groups on each side that are not the same size? Students may classify this as "lopsided," or may use the term asymmetrical.
- → Why does the distribution take this shape? Using the dotplot from above, students will recognize both groups have distributions that are "lopsided," with the main cluster on the lower end of the distributions and a few values to the right of the main mound.

Making Use of Available Data

Most children love to eat hot dogs, but are aware that too much sodium is not necessarily healthy. Is there a difference in the sodium content of beef hot dogs (labeled B in Figure 5) and poultry hot dogs (labeled P in Figure 5)? To investigate this question, students can make use of available data. Using data from the June 1993 issue of *Consumer Reports* magazine, parallel dotplots can be constructed.

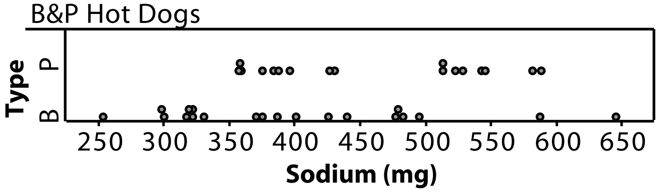

Figure 5: Parallel dotplot of sodium content

Students will notice that the distribution of the poultry hot dogs has two distinct clusters. What might explain the gap and two clusters? It could be another variable, such as the price of the poultry hot dogs, with more expensive hot dogs having less sodium. It can also be observed that the beef sodium amounts are more spread out (or vary more) than the poultry hot dogs. In addition, it appears the center of the distribution for the poultry hot dogs is higher than the center for the beef hot dogs.

As students advance to Level B, considering the shape of a distribution will lead to an understanding of what measures are appropriate for describing center and spread.

> "As students advance to Level B, considering the shape of a distribution will lead to an understanding of what measures are appropriate for describing center and spread."

Describing Center and Spread

Students should understand that the *median* describes the center of a numerical data set in terms of how many data points are above and below it. The same number of data points (approximately half) lie to the left of the median and to the right of the median. Children can create a human graph to show how many letters are in their first names. All the children with two-letter names can stand in a line, with all of the children having three-letter names standing in a parallel line. Once all children are assembled, the teacher can ask one child from each end of the graph to sit down, repeating this procedure until one child is left standing, representing the median. With Level A students, we advocate using an odd number of data points so the median is clear until students have mastered the idea of a midpoint.

Students should understand the *mean as a fair share* measure of center at Level A. In the name length example, the mean would be interpreted as "How long would our names be if they were all the same length?" This can be illustrated in small groups by having children take one snap cube for each letter in their name. In small groups, have students put all the cubes in the center of the table and redistribute them one at a time so each child has the same number. Depending on the children's experiences with fractions, they may say the mean name length is 4 R 2 or 4 1/2 or 4.5. Another example would be for the teacher to collect eight pencils of varying lengths from children and lay them end-to-end on the chalk rail. Finding the mean will answer the question "How long would each pencil be if they were all the same length?" That is, if we could glue all the pencils together and cut them into eight equal sections, how long would each section be? This can be modeled using adding machine tape (or string), by tearing off a piece of tape that is the same length as all eight pencils laid end-to-end. Then, fold the tape in half three times to get eighths, showing the length of one pencil out of eight pencils of equal length. Both of these demonstrations can be mapped directly onto the algorithm for finding the mean: combine all data values (put all cubes in the middle, lay all pencils end-to-end and measure, add all values) and share fairly (distribute the cubes, fold the tape, and divide by the number of data values). Level A students should master the computation (by hand or using appropriate technology) of the mean so more sophisticated interpretations of the mean can be developed at Levels B and C.

The mean and median are *measures of location* for describing the center of a numerical data set. Determining the maximum and minimum values of a numerical data set assists children in describing the position of the smallest and largest value in a data set. In addition to describing the center of a data set, it is useful to know how the data vary or how spread out the data are.

One *measure of spread* for a distribution is the *range*, which is the difference between the maximum and minimum values. Measures of spread only make sense with numerical data.

In looking at the stem and leaf plot formed for the jumping distances (Figure 3), the range differs for boys (range = 39 inches) and girls (range = 27 inches). Girls are more consistent in their jumping distances than boys.

Looking for an Association

Students should be able to look at the possible *association of a numerical variable and a categorical variable* by comparing dotplots of a numerical variable disaggregated by a categorical variable. For example, using the parallel dotplots showing the growth habits of beans in the light and dark, students should look for similarities within each category and differences between the categories. As mentioned earlier, students should readily recognize from the dotplot that the beans grown in the light environment have grown taller overall, and therefore reason that it is best for beans to have a light environment. Measures of center and spread also can be compared. For example, students could calculate or make a visual estimate of the mean height of the beans grown in the light and the beans grown in the dark to substantiate their claim that light conditions are better for beans. They also might note that the range for plants grown in the dark is 4 cm, and 5 cm for plants grown in the light. Putting that information together with the mean should enable students to further solidify their conclusions about the advantages of growing beans in the light.

Considering the hot dog data, one general impression from the dotplot is that there is more variation in the sodium content for beef hot dogs. For beef hot dogs, the sodium content is between 250 mg and 650 mg, while for poultry hot dogs, the sodium content is between 350 mg and 600 mg. Neither the centers nor the shapes for the distributions are obvious from the dotplots. It is interesting to note the two apparent clusters of data for poultry hot dogs. Nine of the 17 poultry hot dogs have sodium content between 350 mg and 450 mg, while eight of the 17 poultry hot dogs have sodium content between 500 mg and 600 mg. A possible explanation for this division is that some poultry hot dogs are made from chicken, while others are made from turkey.

Example 3: Purchasing Sweat Suits—The Role of Height and Arm Span

What about the association between two numerical variables? Parent-teacher organizations at elementary schools have for a popular fund raiser "spirit wear," such as sweatshirts and sweatpants with the school name and mascot. The organizers need to have some guidelines about how many of each size garment to order. Should they offer the shirt and pants separately, or offer the sweatshirt and sweatpants as one outfit? Are the heights and arm spans of elementary students closely related, or do they differ considerably due to individual growing patterns of children? Thus, some useful questions to answer are:

Is there an association between height and arm span?
How strong is the association between height and arm span?

A *scatterplot* can be used to graphically represent data when values of two numerical variables are obtained from the same individual or object. Can we use height

“With the use of a scatterplot, Level A students can visually look for trends and patterns.”

to predict a person’s arm span? Students can measure each other’s heights and arm spans, and then construct a scatterplot to look for a relationship between these two numerical variables. Data on height and arm span are measured (in centimeters) for 26 students. The data presented below are for college students and are included for illustrative purposes.

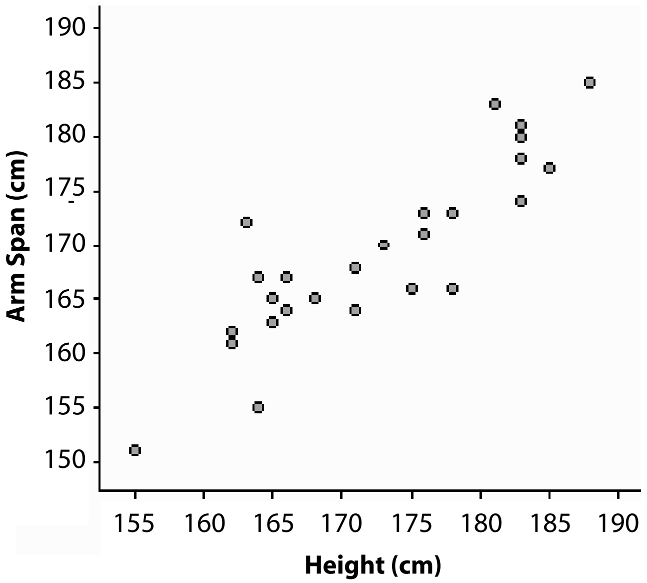

Figure 6: Scatterplot of arm span vs. height

With the use of a scatterplot, Level A students can visually *look for trends and patterns*.

For example, in the arm span versus height scatterplot above, students should be able to identify the consistent relationship between the two variables: generally as one gets larger, so does the other. Based on these data, the organizers might feel comfortable ordering some complete outfits of sweatshirt and sweatpants based on sizes. However, some students may need to order the sweatshirt and sweatpants separately based on sizes. Another important question the organizers will need to ask is whether this sample is representative of all the students in the school. How was the sample chosen?

Students at Level A also can use a scatterplot to graphically look at the values of a numerical variable change over time, referred to as a *time plot*. For example, children might chart the outside temperature at various times during the day by recording the values themselves or by using data from a newspaper or the internet.

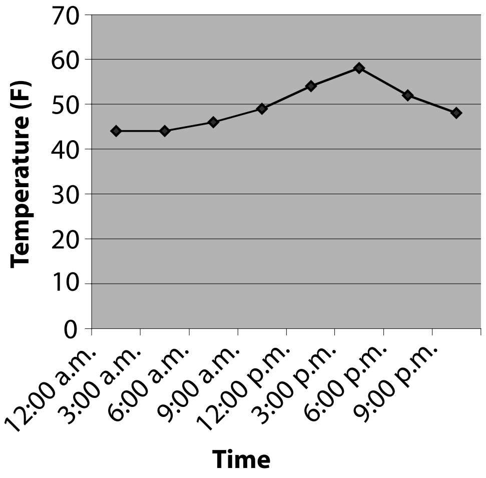

Figure 7: Timeplot of temperature vs. time

When students advance to Level B, they will quantify these trends and patterns with measures of association.

Understanding Variability

Students should explore possible reasons data look the way they do and *differentiate between variation and error.* For example, in graphing the colors of candies in a small packet, children might expect the colors to be evenly distributed (or they may know from prior experience that they are not). Children could speculate about why certain colors appear more or less frequently due to variation (e.g., cost of dyes, market research on people's preferences, etc.). Children also could identify possible places where errors could have occurred in their handling of the data/candies (e.g., dropped candies, candies stuck in bag, eaten candies, candies given away to others, colors not recorded because they don't match personal preference, miscounting). Teachers should capitalize on *naturally occurring "errors"* that happen when collecting data in the classroom and help students speculate about the *impact of these errors* on the final results. For example, when asking students to vote for their favorite food, it is common for students to vote twice, to forget to vote, to record their vote in the wrong spot, to misunderstand what is being asked, to change their mind, or to want to vote for an option that is not listed. Counting errors are also common among young children, which can lead to incorrect tallies of data points in categories. Teachers can help students think about how these events might affect the final outcome if only one person did this, if several people did it, or if many people did it. Students can generate additional examples of ways errors might occur in a particular data-gathering situation.

The notions of error and variability should be used to explain the outliers, clusters, and gaps students observe in the graphical representations of the data. An understanding of error versus natural variability will help students interpret whether an outlier is a legitimate data value that is unusual or whether the outlier is due to a recording error.

At Level A, it is imperative that students begin to understand the concept of variability. As students move from Level A to Level B to Level C, it is important to always keep at the forefront that *understanding variability is the essence of developing data sense.*

The Role of Probability

Level A students need to develop basic ideas of probability in order to support their later use of probability in drawing inferences at Levels B and C.

At Level A, students should understand that *probability is a measure of the chance that something will happen. It is a measure of certainty or uncertainty.* Events should be seen as lying on a continuum from impossible to certain, with less likely, equally likely, and more likely lying in between. Students learn to informally assign numbers to the likelihood that something will occur.

An example of assigning numbers on a number line is given below:

0	¼	½	¾	1
Impossible	Unlikely or less likely	Equally likely to occur and not occur	Likely or more likely	Certain

Students should have experiences *estimating probabilities using empirical data*. Through experimentation (or simulation), students should develop an explicit understanding of the notion that the more times you repeat a random phenomenon, the closer the results will be to the expected mathematical model. At Level A, we are considering only simple models based on equally likely outcomes or, at the most, something based on this, such as the sum of the faces on two number cubes. For example, very young children can state that a penny should land on heads half the time and on tails half the time when flipped. The student has given the expected model and probability for tossing a head or tail, assuming that the coin is "fair."

If a child flips a penny 10 times to obtain empirical data, it is quite possible he or she will not get five heads and five tails. However, if the child flips the coin hundreds of times, we would expect to see that results will begin *stabilizing* to the expected probabilities of .5 for heads and .5 for tails. This is known as the *Law of Large Numbers*. Thus, at Level A, probability experiments should focus on obtaining empirical data to develop relative frequency interpretations that children can easily translate to models with known and understandable "mathematical" probabilities. The classic flipping coins, spinning simple spinners, and tossing number cubes are reliable tools to use in helping Level A students develop an understanding of probability. The concept of relative frequency interpretations will be important at Level B when the student works with proportional reasoning—going from counts or frequencies to proportions or percentages.

As students work with results from repeating random phenomena, they can develop an understanding for the concept of *randomness*. They will see that when flipping a coin 10 times, although we would expect five heads and five tails, the actual results will vary from one student to the next. They also will see that if a head results on one toss, that doesn't mean the next flip will result in a tail. Because coin tossing is a random experiment, there is always uncertainty as to how the coin will land from one toss to the next. However, at Level A, students can begin to develop the notion that although we have uncertainty and variability in our results, by examining what happens to the random process in the *long run*, we can quantify the uncertainty and variability with probabilities—giving a predictive number for the likelihood of an outcome in the long run. At Level B, students will see the role probability plays in the development of the concept of the simple random sample and the role probability plays with randomness.

Misuses of Statistics

The Level A student should learn that proper use of statistical terminology is as important as the proper use of statistical tools. In particular, the proper use of the mean and median should be emphasized. These numerical summaries are appropriate for describing numerical variables, not categorical variables. For example, when collecting categorical data on favorite type of music, the *number* of children in the sample who prefer each type of music is summarized as a frequency. It is easy to confuse categorical and numerical data in this case and try to find the mean or median of the frequencies for favorite type of music. However, one cannot use the frequency counts to compute a mean or median for a categorical variable. The frequency counts *are* the numerical summary for the categorical variable.

Another common mistake for the Level A student is the inappropriate use of a bar graph with numerical data. A bar graph is used to summarize categorical data. If a variable is numerical, the appropriate graphical display with bars is called a *histogram*, which is introduced in Level B. At Level A, appropriate graphical displays for numerical data are the dotplot and the stem and leaf plot.

Summary of Level A

If students become comfortable with the ideas and concepts described above, they will be prepared to further develop and enhance their understanding of the key concepts for data sense at Level B.

It is also important to recognize that helping students develop data sense at Level A allows mathematics instruction to be driven by data. The traditional mathematics strands of algebra, functions, geometry, and measurement all can be developed with the use of data. Making sense of data should be an integrated part of the mathematics curriculum, starting in pre-kindergarten.