MATHEMATICAL THINKING AND LEARNING, *2*(1&2), 99–125

Assessment in Statistics Education: Issues and Challenges

Joan Garfield
Department of Educational Psychology
University of Minnesota

Beth Chance
Department of Statistics
California Polytechnic State University

There have been many changes in educational assessment in recent years, both within the fields of measurement and evaluation and in specific disciplines. In this article, we summarize current assessment practices in statistics education, distinguishing between assessment for different purposes and assessment at different educational levels. To provide a context for assessment of statistical learning, we first describe current learning goals for students. We then highlight recent assessment methods being used for different purposes: individual student evaluation, large-scale group evaluation, and as a research tool. Examples of assessment used in teaching statistics in primary schools, secondary schools, and tertiary schools are given. We then focus on 3 examples of effective uses of assessment and conclude with a description of some current assessment challenges.

One of the forces driving change in the assessment of student learning of statistics has been educational reform within mathematics education (e.g., The Curriculum and Evaluation Standards for School Mathematics produced by the National Council of Teachers of Mathematics, 1989). This reform has established probability and statistics as integral topics within the precollege mathematics curriculum and defined new learning goals for students. Similar calls for reform have also led to major changes in tertiary teaching (e.g., Cobb, 1992; Moore, 1997). Gal and Garfield

Requests for reprints should be sent to Joan Garfield, Department of Educational Psychology, University of Minnesota, 315 Burton Hall, 178 Pillsbury Drive SE, Minneapolis, MN 55455. E-mail: jbg@tc.umn.edu

(1997) provided a summary of currently accepted learning goals for students learning statistics across most grade levels. These goals include:

1. *Understand the purpose and logic of statistical investigations:* Students should understand why statistical investigations are conducted as well as the big ideas that underlie statistical inquiries. These ideas include the omnipresent nature of variation and the use of numerical summaries and visual displays of data. Students also need to understand the nature of sampling: why we study samples instead of populations, how we make inferences from samples to populations, and why designed experiments are needed to establish causation.
2. *Understand the process of statistical investigations:* Students should understand the nature of, and processes involved in, a statistical investigation and the considerations affecting the design of data collection plans. They should be familiar with all specific phases of a statistical inquiry, which include formulating a question, planning a study, collecting, organizing, analyzing and displaying data, interpreting and presenting findings, and discussing conclusions and implications of the study. They should recognize how, when, and why existing inferential tools can be used to aid an investigative process. As Batanero (2000/this issue) argues, students (and practitioners) often fail to fully understand the logic of statistical inference or its role in experimental research.
3. *Learn statistical skills:* Students need to learn important skills that may be used in the process of a statistical investigation. These skills include being able to organize data, compute summary measures, and construct and display tables and different representations of data.
4. *Understand probability and chance:* Students need to develop an understanding of concepts and terminology related to probability and to understand probability as a measure of uncertainty. They need to know how to develop and use models to simulate random phenomena and how to produce data to estimate probabilities.
5. *Develop statistical literacy:* Students need to learn what is involved in interpreting results from a statistical investigation. This includes how to pose critical, reflective questions about numerical arguments, data reported in the media, and project reports from their classroom peers. For example: (a) How reliable are the measurements used? (b) How representative was the sample? and (c) Are the claims being made sensible in light of the data and sample?
6. *Develop useful statistical dispositions:* Students should develop an appreciation for the role of chance and randomness in the world and for statistical methods and planned experiments as useful scientific tools and as powerful means for making personal, social, and business-related decisions in the face of uncertainty. Students should learn to use critical reasoning when faced with an argument that purports to be based on data. This includes reports or conclusions from a statistical investigation, survey, or empirical research.

7. *Develop statistical reasoning:* Statistical reasoning may be defined as the way people reason with statistical ideas and make sense of statistical information. This involves making interpretations based on sets of data, representations of data, or statistical summaries of data. Students need to be able to combine ideas about data and chance, which leads to making inferences and interpreting statistical results.

Although the depth or breadth of these seven goals may differ according to educational level, they describe the main goals for all students who learn basic statistics.

CURRENT ASSESSMENT PRACTICES

Traditionally, assessment has been used primarily to assign grades and give periodic feedback on student learning. More recently, assessment has come to include practices that better inform the instructor of students' understanding and reasoning processes, develop students' learning skills, and improve instructional practices. Figure 1 highlights the various assessment dimensions as described by Ben-Zvi (1999).

Assessment for Individual Student Evaluation

Traditional assessment methods used to assign grades include quizzes, exams, homework exercises, and often laboratory activities. These methods have been

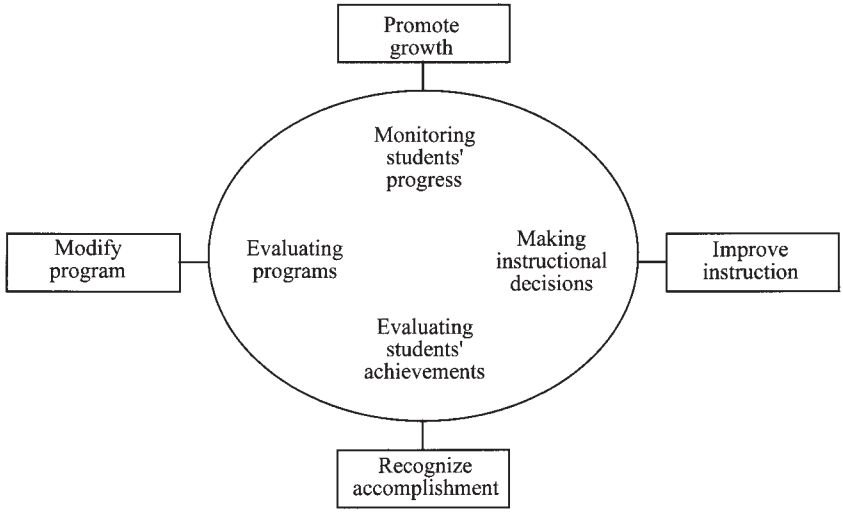


FIGURE 1 Dimensions of assessment. From "Alternative Assessment in Statistics Education," by D. Ben-Zvi, 1999, in *Proceedings of the 52nd Session of the International Statistical Institute* (Vol. 3, pp. 175–176), Helsinki, Finland: Edita Ltd. on behalf of the International Statistical Institute. Used with permission.

used both to monitor student growth and to evaluate student achievement. However, as statistics instruction changes in response to calls for reform, instructors have become increasingly interested in measuring (a) students' ability to produce reasoned descriptions, judgments, inferences, and opinions about data; (b) their understanding of probability and statistics; and (c) their ability to think critically using statistical reasoning. This has led to the use of alternative assessment approaches, such as projects, portfolios, and journals. These newer methods of assessment better capture, during or after course work in statistics, how students think, reason, and apply their learning, rather than merely have students tell the teacher what they have remembered or show that they can perform calculations or carry out procedures correctly.

Given the different purposes for administering a student assessment, the instructor needs to select an appropriate assessment method for a particular purpose (Garfield, 1994; Radke Sharpe, 1998). Some of the different possible methods include the following:

- Quizzes (including calculations, graphs, short-answer questions, and essay questions).
- Minute papers (e.g., short, in-class anonymous essays on what students have learned in a particular class, what they found to be most confusing, or current perceptions of the course).
- Individual or group projects.
- Case studies or authentic tasks.
- Reflective journals (e.g., written reflections on the students' learning and understanding, reports of in-class activities, or analyses of articles in the news that report statistical information).
- Portfolios (including a selection of different materials).
- Exams (covering a broad range of material).
- Attitude surveys (often rating scales about the course, content, or view of statistics).
- Write-ups on in-class activities or computer lab activities.
- Open-ended questions or problems to solve.
- Concept maps.
- Critiques of statistical ideas or issues in the media.

Although these assessment methods may be used to assign a grade, they may also be used to help students learn how to improve their performance, either on the current task or on future ones. The data gathered from these assessments also inform the instructor about what the students are learning and about their competencies, areas of weakness, and reactions to the course.

Next, we provide details on some of the newer methods of assessment (see also Garfield & Gal, 1999).

Individual and group projects. Student projects typically involve posing a problem, designing an experiment or taking a sample, collecting and analyzing data, and interpreting the results (e.g., at the tertiary level, see Mackisack, 1994; Starkings, 1997; at the secondary level, see Hill & Walsh, 1997; Mastromatteo, 1993). The project may be written as a report, presented orally in class, or displayed on a poster. Projects may be assessed using a scoring rubric to assign points (such as 0, 1, 2) to different components of the project. More details on projects and practical work may be found in Hawkins, Jolliffe, and Glickman (1992).

Case studies or authentic tasks. Similar to projects, case studies allow students to study and reflect on actual examples from statistical practice. Colvin and Vos (1997) provided examples of authentic tasks for primary students that integrate assessment with activities appropriate to a student's life outside of school. For example, they described an assessment situation in which students were given data on the weights of grizzly and black bears living in Montana. Students were asked to solve problems about the bear cubs, which required the students to organize, describe, and reason about the weights of the two types of bears. A scoring rubric was used to assign 0 to 3 points to the student responses.

Another type of authentic task at the primary level is described by Lesh, Amit, and Schorr (1977), who provided students with a detailed problem based on a real context and used students' strategies and interpretations as they solved the problem to construct a model of students' reasoning. More extensive studies and tasks can be given at higher grade levels (for examples of tasks for secondary students, see Konold, 1987; Sommers, 1992; for case studies to use with college students, see Chatterjee, Handcock, & Simonoff, 1995; Peck, Haugh, & Goodman, 1998; for projects used with graduate nursing students, see Beck, 1986).

Portfolios of students' work. A portfolio consists of a collection of a student's work, often gathered over an entire course. The selection is often done by the student and teacher together and may include a variety of components, such as computer output for data analyses, written interpretations of statistical analyses, and reflections on what has been learned. Keeler (1997) described the use of portfolios in a statistics class for graduate students. Portfolios are also being used to allow students to demonstrate their achievement and how well they are able to integrate their learning throughout an entire program of study (e.g., National Council of Teachers of Mathematics, 1995).

Concept maps. Schau and Mattern (1997) described different uses of concept maps to assess students' understanding of conceptual connections. Concept maps include the concepts (referred to as nodes and often represented visually by ovals or rectangles) and the connections (referred to as links and often represented with arrows) that relate them. Students may be asked to construct their own maps

for a particular statistical topic (e.g., hypothesis testing) or to fill in missing components from a partially constructed concept map. The general process for this second approach involves first constructing a master map. Keeping that map structure intact, some or all of the concept words or relationship words, or both, are omitted. Students fill in these blanks by either generating the words or by selecting them from a list that may or may not include distractors.

Critiques of statistical ideas or issues in the news. Students may be asked to read and critique a newspaper article, responding to particular questions such as (a) What do you think is the purpose of the research study described in this article? (b) What method or methods were used to answer the research question? (c) What questions would you like to ask the investigators to better understand the study? and (d) Are there any aspects of the study that might make you question the conclusions presented in the article? Brief graphs or articles may also be used to assess students' statistical thinking, including their understanding of basic terminology, their ability to interpret concepts in a wider context, and their ability to question claims based on data (Watson, 1997).

For examples of assessment items that involve articles or graphs presented in the media and research literature, see Gal (in press) and Gelman and Nolan (1998).

Minute papers. These are brief, anonymously written remarks provided by students, sometimes on an index card or half of a sheet of paper, during the last few minutes of class (Angelo & Cross, 1993). These remarks can cover a variety of topics, such as a summary of what students do or do not understand on a topic or students' reactions to various aspects of a course (e.g., the use of cooperative groups, the textbook, or the teacher's explanations in class). Some statistics teachers use minute papers to have students describe their understanding of a particular concept or procedure discussed in class that day or to have them respond to the question: "What was the most confusing idea in today's class?"

The different assessment methods described previously may be used in combination with each other as well as in combination with traditional quizzes and exams. Chance (1997) provided details on her model for combining different assessment components.

Assessment for Large-Group Evaluation

Assessing statistical (including probabilistic) reasoning using a group procedure, such as a paper-and-pencil test, is a particularly challenging task that has received little attention in the research literature. The only large-scale study in this area was Green's (1983) survey of probabilistic understanding, which, despite careful pilot testing, experienced difficulties due to problematic wording (Hawkins et al., 1992).

Although this was the only large survey devoted entirely to probabilistic understanding, items assessing probability and statistics learning have been included in standardized tests and national assessments (e.g., Lindquist, 1989).

Unfortunately, too often, standardized exams have served as examples of poor statistics and probability questions and as a misleading reflection of what we want our students to know. For example, multiple-choice questions such as those shown in Figures 2 and 3, taken from two national examinations, focus too much on the calculation in an artificial setting, with no explanation or interpretation required of the students.

Probability items such as these typically ask students to compute probabilities for marbles in urns, rolls of dice, or calculations involving similar random devices. These items tend to show students' ability to compute correctly but do not reveal their probabilistic intuitions or reasoning skills. Indeed, as Konold (1995) has stated, problems such as these lead students to believe that routine skills and memorized formulas are what teachers view as important.

A bag contains 100 balls numbered from 1 to 100. One ball is removed. What is the probability that the number of the ball is even or less than 30?

a. .1

b. .5

c. .6

d. .7

e. .8

FIGURE 2 Typical probability item on a national exam.

The average and SD of a set of 50 scores are 30 and 7, respectively. If each of these scores is increased by 10, then which of the following is true for the new set of scores?

a. The average is 60

b. The average is 40

c. The SD is 17

d. The SD is 7.2

FIGURE 3 Typical statistics item on a national exam.

Statistics items on similar exams typically ask students to compute measures of center for a set of data, such as number of inches of snowfall for 15 years. Although these types of items result in more than a simple list of numbers, Hawkins et al. (1992) cautioned that just adding a context to a set of numbers is not sufficient for a good assessment item, unless there is a meaningful purpose for calculating a statistic or graphing a data set. In a recent analysis of probability and statistics items used in the Seventh National Assessment of Educational Progress, Zawojewski and Shaughnessy (in press) found that, although some items went beyond simple computation of statistical measures by asking students to select an appropriate measure of center (mean or median) for a set of data, they failed to provide a clear purpose or context, which would have allowed a better assessment of students' reasoning about measures of center.

Assessment of deeper levels of statistical reasoning and understanding has traditionally been restricted to research studies in which items are given (a) to students or adults individually as part of clinical interviews, or (b) to small groups that are closely observed. Although statistical reasoning may be best assessed through such one-to-one communication with students or by examining a sample of detailed, in-depth student work (e.g., a statistical project), carefully designed paper-and-pencil instruments may be used to gather some limited indicators of statistical reasoning. One such instrument, the Statistical Reasoning Assessment (SRA), is discussed later in some detail. The Advanced Placement (AP) Statistics Exam, also described later, is another recent example of an assessment tool to be administered in large groups that attempts to focus on reasoning as well as calculation. Computerized testing has been another area of development. For example, Cicchitelli, Bartolucci, and Forcina (1999) explored uses of computerized testing as an alternative to the oral exams used in Italy to assess university students' understanding after completing a statistics course.

In addition, in many cases, traditional exam questions may be adapted to assess deeper levels of conceptual understanding. One approach is to develop and use objective-format questions to assess higher level thinking. For example, enhanced multiple-choice items or items that require students to match concepts or questions with appropriate explanations may be used to capture students' reasoning and measure conceptual understanding. Cobb (1998) offered five principles for designing objective-format questions that assess statistical thinking. One principle is to ask for comparative judgments, not just category matching. For example, a set of two-way tables is presented to students with data representing factors related to the death sentence. Each table displays frequencies for the breakdown of different independent variables (e.g., race of defendant, race of victim, or prior record) by the same dependent variable (e.g., whether a convicted murderer is sentenced to death). Students are asked which factors in the tables are most strongly associated with whether a convicted murderer is sentenced to death. They need to compare the strength of the interaction between variables in each table to make this judg-

ment. A second principle is to involve two or more modes of statistical thinking (e.g., visual and verbal–intuitive thinking). For example, students may be asked to match verbal descriptions to four different plots of data or to match boxplots with normal probability plots or histograms.

Similarly, Hubbard (1997) argued for more variety and less predictability in exam questions, requiring students to apply their knowledge to real problems in new ways and to think beyond the calculations. She suggested techniques that include asking students to construct a setting that accomplishes given criteria (e.g., a research question that can be solved with regression, a data set with the mean larger than the median) and having students link graphical and symbolic representations of a concept.

Assessment as a Research Tool

Now that more students are first encountering statistics in elementary school classes, more attention is being paid to how young children learn and understand statistical ideas. Assessment is increasingly being used as a method for gaining insight into students' understanding of statistical concepts and for modeling student reasoning, often as part of exploratory research. Furthermore, informative assessment tools are needed to gauge the effectiveness of new technologies and teaching strategies. Although there has been much agreement regarding the potential of new pedagogical methods and technological tools, there has been little research measuring their impact on student learning. One reason has been the lack of available assessment methods and instruments; however, that is changing.

In addition to the case studies presented by Ben-Zvi (2000/this issue), Friel, Bright, Frierson, and Kader (1997) presented items for assessing primary students' knowledge and interpretation of particular graphical representations of data. They described a method used to categorize students' responses to these types of tasks. Graphs, particularly those found in the media, are also used by Watson (1997). She suggested ways to use items such as these to assess students' basic statistical literacy and thinking.

Students may also be assessed as they work together in groups to interpret a graph or solve a problem. Curcio and Artzt (1997) provided a framework for assessing students' statistical problem-solving skills in this context.

Jones, Langrall, Thornton, and Mogill (1997) developed a framework used to assess children's thinking about probability. This framework includes four constructs (sample space, probability of an event, probability comparisons, and conditional probability) and four levels of thinking within each construct. This framework was used to generate probability tasks to assess students' thinking. Metz (1997) also provided tasks that can be used to describe children's understanding of basic concepts related to probability. She analyzed her assessment of

understanding along three dimensions: cognitive, epistemological, and cultural. Konold, Pollatsek, Well, and Gagnon (1997) used assessment items to probe secondary students' reasoning about data as they used a statistical software package.

In addition to these research studies, there are many articles and materials available on assessing mathematical performance of students in primary grades, which are relevant for assessing students' statistical knowledge (e.g., National Council of Teachers of Mathematics, 1995; Webb & Coxford, 1993).

EXAMPLES

We now describe three examples of assessment tools in more detail. These tools have been effectively used to evaluate statistical understanding in large groups, to understand better and to diagnose students' statistical reasoning skills, and to gain insight into how statistical reasoning develops.

Example 1: The AP Statistics Exam

An influential model for assessment in secondary schools in the United States is the AP Statistics Examination. This exam was offered by the College Board for the first time in 1997 and was taken by approximately 7,600 high school students; in 1999, the number had risen to more than 25,000 students. The structure of the exam has been 35 multiple-choice questions and 6 free-response questions, including 1 investigative task. The exam covers four main topics (College Entrance Examination Board and Educational Testing Service, 1998):

- *Exploring data:* Observing patterns and departures from patterns.
- *Planning a study:* Deciding what and how to measure.
- *Anticipating patterns:* Producing models and using probability and simulation.
- *Statistical inference:* Confirming models.

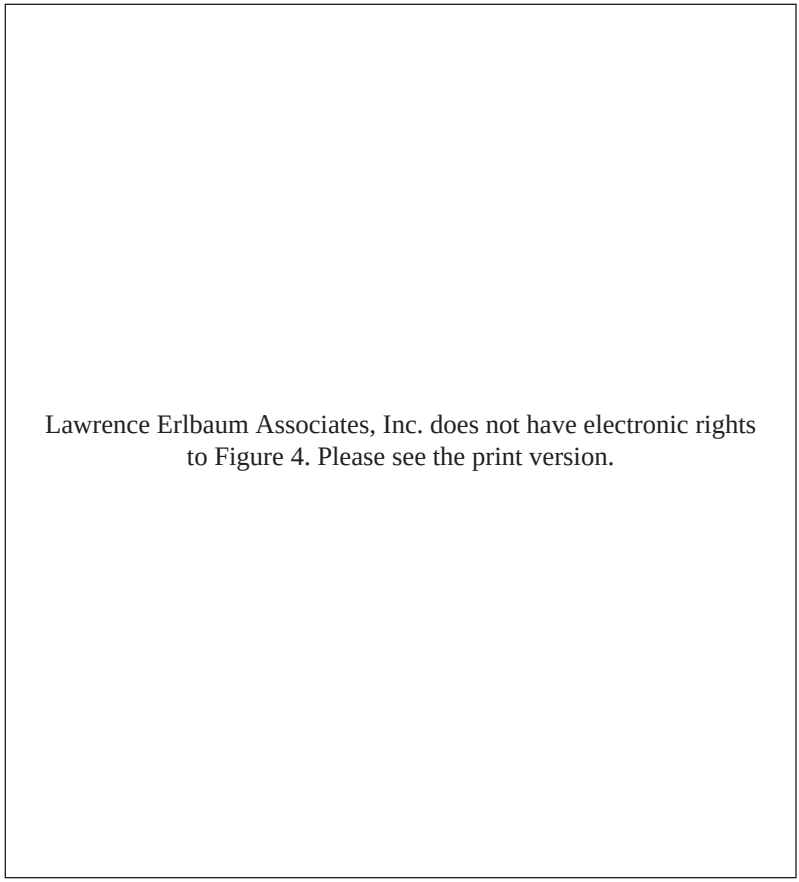
The 1997 multiple-choice questions, which count for half of the student's total score, have now been released for students and teachers to use in preparing for future exams.

The free-response questions are scored holistically by statistics teachers from colleges and high schools using a detailed scoring rubric (see Figure 4 for the general framework of the AP Statistics rubric). Students' responses to each of these statistical problems are evaluated using the following criteria:

- Did the student demonstrate knowledge of the statistical concepts involved?

- Did the student communicate a clear explanation of what was done in the analysis and why?
- Did the student express a clear statement of the conclusions drawn?

Thus, both statistical knowledge and communication of statistical ideas are weighted heavily, and, generally, more weight is given to clear communication of the correct idea than to correct computations (Scheaffer, 1999). The rubrics give credit for any correct method used in the solution, but the student must present



(continued)

FIGURE 4 Scoring rubric for the AP Statistics free-response questions. Reprinted by permission of Educational Testing Service and the College Entrance Examination Board, the copyright owners. For limited use by Cal Poly San Luis Obispo.

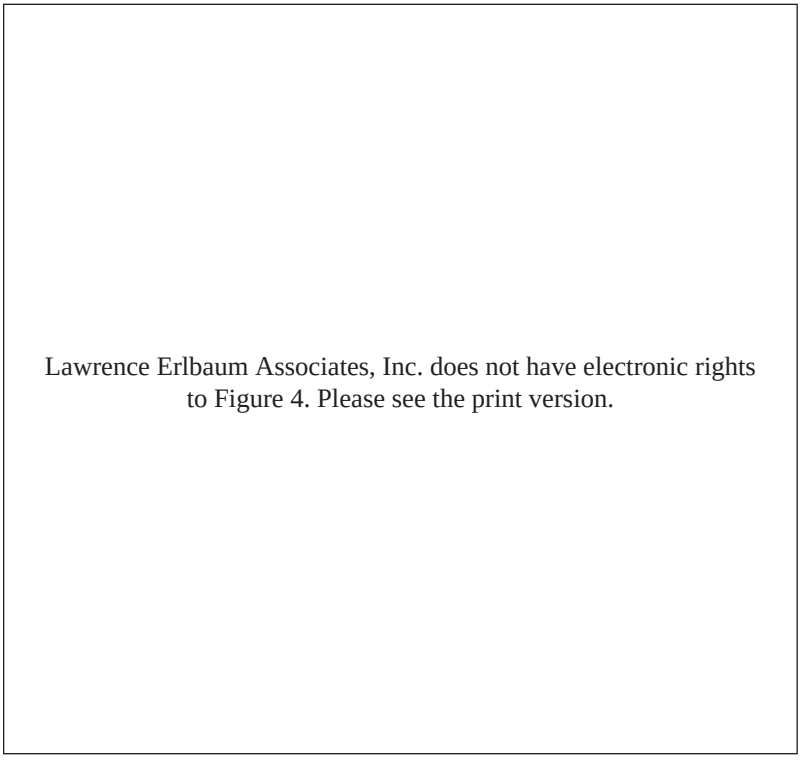


FIGURE 4 Scoring rubric for the AP Statistics free-response questions. Reprinted by permission of Educational Testing Service and the College Entrance Examination Board, the copyright owners. For limited use by Cal Poly San Luis Obispo.

enough information so that the line of reasoning can be followed—for example, why the method was chosen, the assumptions of the method, verification of the validity of the method, and a final conclusion in the context of the original problem. Solutions that lack these explanations resulted in lower scores on the first administrations of the AP exams. Therefore, the College Entrance Examination Board, the Educational Testing Service, and the Test Development Committee are distributing specific information regarding what is expected in students' answers, highlighting the need for clear and detailed explanations. In this way, students and educators are being shown that calculations alone are insufficient for a complete statistical analysis.

The scoring rubrics provide the means for consistent evaluation of students' open responses and critical thinking (in general, see Facione & Facione, 1994; regarding the AP Statistics Exam, see Olsen, 1998). The key to this approach is interrater reliability. The graders, or readers, of the AP Statistics Exam typically

spend 30 to 60 min being trained on how to interpret the rubric and examine sample student responses. Each reader is paired with another reader (typically college and high school instructors are paired), who serves as a first resource if the reader needs assistance interpreting a student answer. Additional grading questions can be referred to the table leader or the question leader. Table leaders also backread samples of readers' grading and clarify the rubric as necessary.

The 1999 free-response questions are posted at www.collegeboard.org/ap/statistics/frq99/. Figures 5 through 7 contain three questions from recent exams.

FIGURE 5 Free-response question for the AP Statistics Exam. Reprinted by permission of Educational Testing Service and the College Entrance Examination Board, the copyright owners. For limited use by Cal Poly San Luis Obispo.

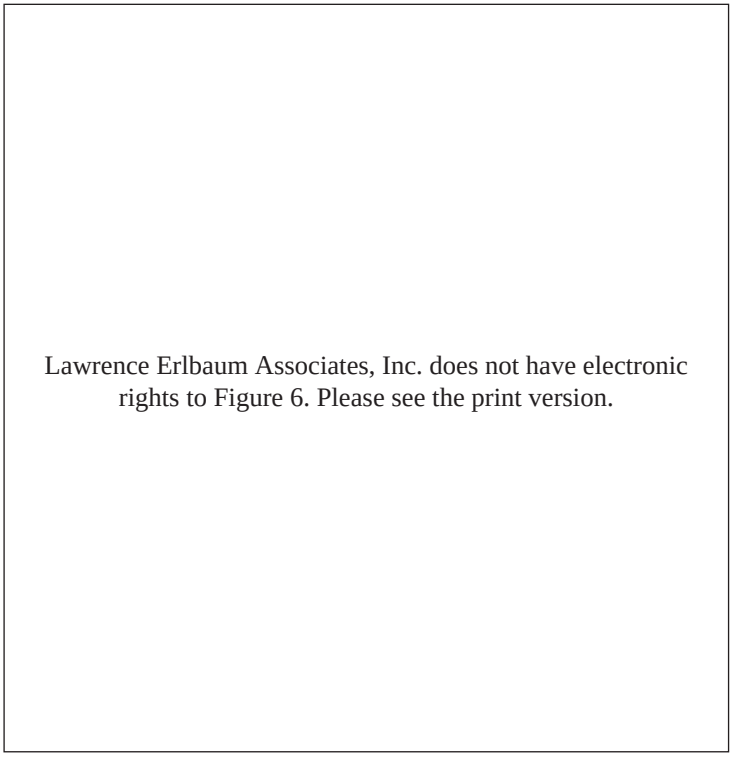


FIGURE 6 The AP Statistics investigative task. Reprinted by permission of Educational Testing Service and the College Entrance Examination Board, the copyright owners. For limited use by Cal Poly San Luis Obispo.

Figure 5 is the third question from the 1999 exam. This question illustrates how, often, many skills are required of the students in the same question. Students needed to demonstrate understanding of types of study and the proper conclusions that can be drawn from different studies as well as the ability to explain their conclusions and provide examples of statistical concepts in context. As Batanero (2000/this issue) urges, students need to realize the limitations of what conclusions can be drawn from a statistically significant result. The rubric for this question (also available at the College Board Web site) clearly focuses on the students' ability to clearly explain their reasoning.

In grading this problem holistically, readers needed to see if students sufficiently explained anywhere in their answer the following aspects:

- Why this was not an experiment.
- The differential effect of the confounding variable on the response.

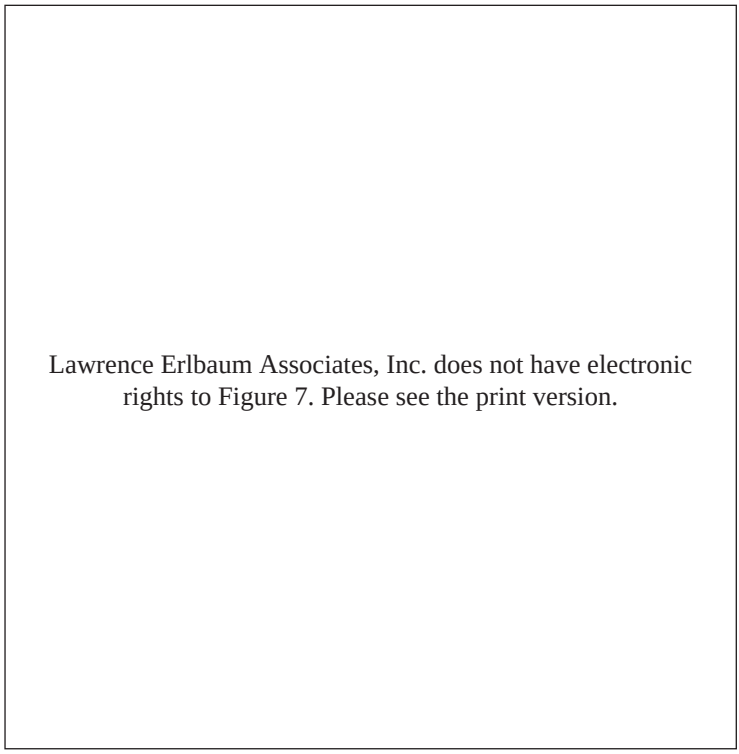


FIGURE 7 Integrative item from the AP Statistics Exam. Reprinted by permission of Educational Testing Service and the College Entrance Examination Board, the copyright owners. For limited use by Cal Poly San Luis Obispo.

- Why a cause and effect conclusion could not be drawn.

Thus, the answer to Question A in Figure 5 did not have to appear strictly in the space for Question A because the problem is graded as a whole.

The 1999 investigative task concerned data on people's ability to correctly predict the outcomes of coin tosses, as shown in Figure 6. This question was designed to test students' ability to choose the appropriate statistical procedure based on the question posed. Students also had to carry out the procedures in detail, including stating and checking (e.g., by producing and examining a graph) the technical conditions required for the validity of the procedure. A complete answer also included a final statement of their conclusion in the context of the problem.

A question from the 1998 exam, as shown in Figure 7, also targets the need for students to be able to synthesize ideas from different sections of the course and to understand the primary information contained in different graphs.

Although these questions highlight some of the strengths of the AP Statistics Exam, there are a few cautions. One difficulty with the problem shown in Figure 5 is that it requires students to define a specific term, *confounding.* Although students could receive full credit if they explained the concept and clearly identified the concept as confounding, many students confused confounding variables with lurking variables. This focus on terminology surprised some instructors. Furthermore, there is not universal acceptance in current textbooks on the definitions of *experiments* and *confounding*, if the latter term is explicitly defined at all.

One difficulty with the item as shown in Figure 6 is the context. Many students focused on the randomness of the outcome of the coin toss, rather than on a person's prediction. Although students need to be able to relate the calculations to the context, it is also important that they remember to base their objective conclusions on the information in the data and not be overwhelmed by the context.

The AP Statistics Exam is an excellent model of an exam that can be administered and graded en masse although still focusing on students' reasoning ability, conceptual knowledge, and communication skills. Still, it is important that such exams do not rely too heavily on terminology in which usage is not widespread or on potentially misleading contexts.

Example 2: The SRA Instrument

The SRA was developed and validated as part of the ChancePlus Project (Garfield, 1998; Konold, 1989) funded by the National Science Foundation to evaluate the effectiveness of a new secondary-level statistics curriculum in achieving its learning goals. At that time no other instrument existed that would assess high school students' ability to understand statistical concepts and apply statistical reasoning. The SRA has been used not only with the ChancePlus Project but with other high school and college students, in a variety of statistics courses, to evaluate the effectiveness of curricular materials and approaches as well as to describe the level of students' statistical reasoning.

The SRA is a multiple-choice test consisting of 20 items. Each item describes a statistics or probability problem and offers several choices of responses, both correct and incorrect. Most responses include a statement of reasoning explaining the rationale for a particular choice. Students are instructed to select the response that best matches their own thinking about each problem. Items from this instrument have been adapted and used in research projects in other English-speaking countries, such as Australia and the United Kingdom, as well as for studies in Spain, France, and Taiwan.

The following applications of reasoning were used to direct development and selection of the SRA:

• *Reasoning about data:* Recognizing or categorizing data as quantitative or qualitative, discrete or continuous; and knowing how the type of data leads to a particular type of table, graph, or statistical measure.

• *Reasoning about graphical representations of data:* Understanding the way in which a plot is meant to represent a data set; understanding how to read and interpret a graph; knowing how to modify a graph to better represent a data set; and being able to identify the overall pattern, center, and spread in a distribution.

• *Reasoning about statistical measures:* Understanding what measures of center, spread, and position tell about a data set; knowing which are best to use under different conditions and how they do or do not represent a data set; knowing that using summaries for predictions will be more accurate for large samples than for small samples; and knowing that a good summary of data includes a measure of center as well as a measure of spread and that summaries of center and spread can be useful for comparing data sets.

• *Reasoning about uncertainty:* Understanding and using ideas of randomness, chance, and likelihood to make judgments about uncertain events; knowing that not all outcomes are equally likely; and knowing how to determine the likelihood of different events using an appropriate method (such as a probability tree diagram or a simulation using coins or technology).

• *Reasoning about samples:* Knowing how samples are related to a population and what may be inferred from a sample; knowing that a larger, well-chosen sample will more accurately represent a population and that there are ways of choosing a sample that make it unrepresentative of the population; and being cautious when making inferences made on small or biased samples.

• *Reasoning about association:* Knowing how to judge and interpret a relation between two variables; knowing how to examine and interpret a two-way table or scatterplot when considering a bivariate relation; and knowing that a strong correlation between two variables does not mean that one causes the other.

In addition to determining what types of reasoning skills students should develop, it was also important to identify the types of incorrect reasoning students should not use when analyzing statistical information. Kahneman, Slovic, and Tversky (1982) are well known for their substantial body of research, which reveals some prevalent ways of thinking about statistics that are inconsistent with a technical understanding. More recent research suggests that even people who can correctly compute probabilities tend to apply faulty reasoning when asked to make an inference or judgment about an uncertain event, relying on incorrect intuitions (Garfield & Ahlgren, 1988; Shaughnessy, 1992). Other researchers have discovered additional misconceptions or errors of reasoning when examining students in classroom settings (e.g., Konold, 1989, 1995; Lecoutre, 1992). Several of the identified misconceptions or errors in reasoning used to develop the SRA are described here:

• *Misconceptions involving averages:* Examples include the following: Averages are the most common number; to find an average one must always add up all the numbers and divide by the number of data values (regardless of outliers); a mean is the same thing as a median; and one should always compare groups by focusing exclusively on the difference in their averages.

• *The outcome orientation:* Konold (1989) postulated an intuitive model of probability that leads students to make yes or no decisions about single events rather than looking at the series of events. For example, consider this item: "A weather forecaster predicts the chance of rain to be 70% for 10 days. On 7 of those 10 days it actually rained. How good were his forecasts?" Many students will say that the forecaster did not do such a good job because it should have rained on all days for which he gave a 70% chance of rain. They appear to focus on outcomes of single events rather than being able to look at the series of events, believing that a 70% chance of rain means that it should rain. Similarly, a forecast of 30% rain means it will not rain.

• *Good samples have to represent a high percentage of the population:* According to this belief, it does not matter how large a sample is or how well it was chosen—it must represent a large percentage of a population to be a good sample.

• *The law of small numbers:* According to this "law," small samples should resemble the populations from which they are sampled, so small samples are used as a basis for inference and generalizations (Kahneman et al., 1982).

• *The representativeness misconception:* People estimate the likelihood of a sample based on how closely it resembles the population. Therefore, a sample of coin tosses that has an even mix of heads and tails is judged more likely than a sample with more heads and fewer tails (Kahneman et al., 1982).

• *Equiprobability bias:* Events tend to be viewed as equally likely. Therefore, the chances of getting different outcomes (e.g., three 5s or one 5 on three rolls of a dice) are incorrectly viewed as equally likely events (Lecoutre, 1992). In general, the probabilities of any two outcomes happening are automatically judged to be equal.

Once the items had been written, borrowed, or adapted to represent areas of correct and incorrect reasoning, all items went through a long revision process. The first step of this process was to distribute items to experts for content validation, to determine if each item was measuring the specified concept or reasoning skills, and to elicit suggestions for revisions or addition of new items. A second step was to administer items to groups of students and to investigate their responses to open-ended questions. These responses were used to phrase justifications for different responses to use in a subsequent multiple-choice format in the instrument. After several pilot tests of the SRA, administration of the instrument in different settings, and many subsequent revisions, the current version was created.

An attempt was made to determine criterion-related validity by administering the SRA to students at the end of an introductory statistics course and correlating

their scores with different course outcomes (e.g., final score, project score, quiz total, etc.). The resulting correlations were all extremely low, suggesting that students' statistical reasoning and misconceptions are unrelated to their performance in a first statistics course. One plausible explanation is that industrious students with good study skills may be able to do well in a course, regardless of whether their statistical reasoning has been changed.

To determine the reliability of the SRA, different reliability coefficients were examined. An analysis of internal consistency reliability coefficients indicated that the intercorrelations between items were quite low and that items did not appear to be measuring one trait or ability. A test–retest reliability coefficient appeared to be a more appropriate method to use, but first a new scoring method was needed.

Although individual items could be scored as correct or incorrect, and total correct scores could be obtained, this single numerical summary seemed uninformative and did not adequately identify students' reasoning abilities. Therefore, a method was created in which each response to an item was viewed as identifying a correct or incorrect type of reasoning. Eight categories, or scales, of correct reasoning were created, and 8 categories of incorrect reasoning (or misconceptions) were also developed (see Table 1). Scores for each scale range from 2 to 8, depending on how many responses contribute to that scale. In addition to the 16 scale scores, total scores for correct and incorrect reasoning may be calculated by adding the 8 scale scores that pertain to them. A test–retest reliability analysis yielded a reliability of .70 for the correct reasoning total score and .75 for the incorrect reasoning total score (Liu, 1998).

The SRA was administered to two large groups of college students, and scale scores were compared (Garfield, 1998). Because each scale could have a different number of points, all scales were divided by the number of items to yield scores on a scale of 0 to 2. An analysis of these scaled scores suggested that there were strong similarities in reasoning for the two samples of students. These scores also show the types of reasoning that are most difficult for students (e.g., relating to sampling variability and probability) and the misconceptions that are most prevalent (e.g., equiprobability bias).

In a cross-cultural comparison of American and Taiwanese college students to identify possible gender differences on the SRA, Liu (1998) found seemingly similar scale scores for students in the two countries but striking differences when comparing the male and female groups. She concluded that, based on her samples, men have higher total correct reasoning scores and lower total misconception scores than women. Results were more striking in the Taiwan sample than the United States sample. It will be interesting to see if replications of this study in other countries will yield similar results.

Although the SRA is an easy-to-administer paper-and-pencil instrument that provides some useful information regarding the thinking and reasoning of students as they solve statistical problems, it is nonetheless problematic as a research and

TABLE 1
Correct Reasoning Skills and Misconceptions Measured by the Statistical Reasoning Assessment

Correct reasoning skills	1. Correctly interprets probabilities 2. Understands how to select an appropriate average 3. Correctly computes probability a. Understands probabilities as ratios b. Uses combinatorial reasoning 4. Understands independence 5. Understands sampling variability 6. Distinguishes between correlation and causality 7. Correctly interprets two-way tables 8. Understands importance of large samples
Misconceptions	1. Misconceptions involving averages a. Average is the most common number b. Fails to take outliers into consideration when computing the mean c. Compares groups based on their averages d. Confuses mean with median 2. Outcome orientation misconception 3. Good samples have to represent a high percentage of the population 4. Law of small numbers 5. Representativeness misconception 6. Correlation implies causation 7. Equiprobability bias 8. Groups can only be compared if they are the same size

evaluation tool. The 16 scales represent only a small subset of reasoning skills and strategies. It is important to recognize that attempts to establish the reliability and validity of this instrument have yielded less than impressive results. More work needs to be done in developing other assessments of statistical reasoning and in finding appropriate ways to determine their reliability and validity so that better tools may be utilized in future research and evaluation studies.

Example 3: Tools for Teaching and Assessing Statistical Inference

Another example of assessment of statistical reasoning is a project that examines the learning of concepts related to statistical inference (Garfield, delMas, & Chance, 1999b). Based on the belief that students in traditional statistics classes develop a shallow and isolated understanding of important foundational concepts and do not develop the deep understanding needed to integrate these concepts and use them in their reasoning, the Tools for Teaching and Assessing Statistical Inference Project has developed a set of innovative teaching and assessment materials to ac-

company instructional software. The goal of these materials and software is to help students develop better statistical reasoning by first developing a rich conceptual understanding of foundational concepts.

Before designing new materials, a study was done of how students' understanding of sampling distributions is assessed. Test banks or instructor guides for introductory statistics texts were examined, revealing that none of the items included figures or graphs; most based their assessment of understanding on student selection of the correct definition; and several asked specific questions about the shape, center, and spread of sampling distributions, according to the central limit theorem. Another commonly used question was one that asked students to apply the central limit theorem by calculating standard errors for a specified sample size or determining probabilities using the standard normal table.

An analysis of these assessment items indicated that typical evaluations of conceptual understanding of sampling distributions are inadequate. Students might be able to produce correct answers and even receive good test grades, yet still not understand the ideas and maintain misconceptions (e.g., believing that a sampling distribution should always resemble the shape of the population). This led to the development of new assessment instruments for evaluating students' understanding of sampling distributions as well as other key concepts related to statistical inference.

Using a classroom research approach (Cross & Steadman, 1996), Garfield, delMas, and Chance (1999a) developed instructional units for concepts such as sampling distribution, confidence interval, and *p* values, which involve the use of simulation software, and used them in different institutional settings. In creating each instructional unit, they utilized the following assessment framework:

- A pretest that measures prerequisite knowledge and intuitions that may affect students' interactions with the activity. The results of the pretest are used as the basis for discussions with students to clear up misconceptions that might interfere with the learning activity.
- A list of assessment goals that specifies desired learning outcomes and is used to develop the learning activity.
- Assessment items that are embedded in the learning activity, which students use to make and evaluate predictions and promote conceptual change.
- A posttest of desired outcomes that assess correct types of conceptual reasoning or misconceptions, including items parallel to those used in the instructional activity.
- A delayed posttest that consists of items that could be included in an end-of-course final exam to measure long-term retention.

As instruments were developed, pilot tested, and reviewed, lists of prerequisite skills and misconceptions were reviewed and revised as well. All materials, in-

cluding the software, are available on a Web site (www.gen.umn.edu/faculty_staff/delmas/stat_tools/index.htm).

Each unit consists of a pretest and a posttest. These pretests of prerequisite knowledge consist of sets of items that may be used for diagnostic purposes. The instructor may administer all or some of these items before beginning an instructional activity to determine if students have misunderstandings that may be corrected before beginning the activity. For example, many students reveal a misunderstanding of the term *variability*, thinking that it refers to the bumpiness of a distribution rather than to the spread of the distribution. This led to the item shown in Figure 8, from a pretest for sampling distributions. By giving this item as part of a diagnostic pretest, instructors are better able to diagnose and remedy misconceptions about variability before students proceed with a learning activity that involves visually assessing variability of sampling distributions.

The posttests are designed to evaluate students' understanding of key concepts and their ability to use these concepts in solving statistical problems. Each test consists of a variety of items to assess students' learning after completing the unit. Most items are multiple choice or matching format for easier scoring. The use of these items varies according to the research purpose, the instructional purpose, or both. Some or all of these items may be included in a posttest at the end of the activity, unit quiz, or final exam. The item shown in Figure 9 was adapted from one used in studies by Konold, Well, Lohmeier, and Pollatsek (1993) to be used here on a posttest concerning sampling distributions.

Items such as these allow instructors to monitor student understanding as it is influenced by different instructional tasks. They also allow researchers to compare

Students are given two histograms: one is bumpy with a narrow spread, the other resembles a normal distribution and has a wider spread. Students are asked to identify which graph exhibits more variability and to check whatever statements explain their reasoning from the following list:

(a) Because it's bumpier

(b) Because it's more spread out

(c) Because it has a larger number of different scores

(d) Because the values differ more from the center

(e) Other

FIGURE 8 Pretest question for sampling distributions unit.

1. American males must register at a local post office when they turn 18. In addition to other information, the height of each male is obtained. The national average height for 18-year-old males is 69 inches (5 ft. 9 in.). At the small, local post office, about 5 men register each day. At the large, city post office, about 50 men register each day. At the end of each day, the clerk at each post office computes the average height of the men who registered there that day

Which of the following predictions would you make regarding the number of days on which the average height for the day was more than 71 inches (5 Ft. 11 in)?

a. The number of days with average heights over 71 inches would be greater for the small post office than for the large post office.

b. The number of days with average heights over 71 inches would be greater for the large post office than for the small post office.

c. The number of days with average heights over 71 inches would be the same for large and small post offices.

d. There is no basis for predicting which post office would have the greater number of days.

FIGURE 9 Posttest item for sampling distributions.

student performance in different settings and, after a time delay, to measure retention of learning.

CURRENT ASSESSMENT CHALLENGES

As more educators adopt alternative methods of assessing student learning, new questions arise that need to be addressed. Garfield and Gal (1999) summarized some of these challenges:

1. *Assessment of students using new technological tools:* We need more effective ways to assess what students can do and how they reason when they use computers or other technological aids (Ben-Zvi, 2000/this issue). We also need to find ways to assess the nature of, and limitations on, inferences that can be drawn from assessments when students learn with computers but are tested without computers.
2. *Assessment of statistical literacy:* We need ways to assess the application or transfer of student learning to interpretive or functional tasks such as those en-

countered in media or outside the classroom. The challenge in assessment of statistical literacy is that it should involve examining not only what students think when asked to reflect on a graph or report in the media but also their tendency to do so without being cued.

3. *Assessment of students' understanding of big ideas:* We need ways to assess students understanding of the important ideas in statistics, such as variation, error, bias, sampling, or representativeness. Often their meaning may depend on the context of the problem. Assessment items or tasks are needed that can evaluate students' understanding of, and sensitivity to, the prevalence and importance of such "big ideas" in different contexts.

4. *Assessment of students' intuitions and reasoning involving probability concepts and processes:* We need ways to transfer and adapt promising assessment methods and instruments used by researchers to formats that are reasonably acceptable and accessible to teachers and that can be used for routine classroom use.

5. *Assessment of outcomes of group work:* There is a need for good methods of assessing and grading students' work when it is done in groups, which is important as more teachers include cooperative group activities and projects in their statistics classes. These methods need to be fair to students and should also motivate them to participate equally.

6. *Developing models to use in evaluating and comparing curricula:* As new curricula, innovative textbooks, and instructional software replace traditional approaches to teaching statistics, there is an increasing need for reliable, valid, practical, and accessible assessment instruments to use in evaluating the relative utility of these materials and methods. As long as statistics items used in large-scale or standardized assessments remain focused on computations (as opposed to statistical reasoning) and provide little context or meaningless contexts, the relative effectiveness of statistics courses or units will remain difficult to ascertain.

CONCLUSIONS

As educational reforms lead to additional changes in statistics education, assessment of students who are learning statistics will continue to be a challenging endeavor. Instructors at all educational levels need to incorporate newer assessment methods and combinations of assessment methods to provide detailed information to students as well as to inform their teaching. We have provided several examples, at different educational levels, of various assessment methods. These methods illustrate the different roles that assessment can take beyond the traditional role of assigning student grades. By using techniques that inform the teacher, student, and researcher and that focus on students' reasoning and performance in more authentic tasks, assessment can contribute to the most important educational goal: improving students' learning of statistics.

ACKNOWLEDGMENTS

We thank Allan Rossman, Carmen Batanero, and Brian Greer, whose valuable comments greatly improved this article.

REFERENCES

Angelo, T., & Cross, K. P. (1993). *Classroom assessment techniques.* San Francisco: Jossey-Bass.

Batanero, C. (2000/this issue). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2,* 75–97.

Beck, C. T. (1986). Use of nonparametric correlation analysis in graduate students' research projects. *Journal of Nursing Education, 25,* 41–42.

Ben-Zvi, D. (1999). Alternative assessment in statistics education. In *Proceedings of the 52nd Session of the International Statistical Institute* (Vol. 3, pp. 175–176). Helsinki, Finland: Edita, on behalf of the International Statistical Institute.

Ben-Zvi, D. (2000/this issue). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning, 2,* 127–155.

Chance, B. (1997). Experiences with authentic assessment techniques in an introductory statistics course. *Journal of Statistics Education* [Online], *5*(3). Retrieved from the World Wide Web: http://www.amstat.org/publications/jse/v5n3/chance.html

Chatterjee, S., Handcock, M. S., & Simonoff, J. S. (1995). *A casebook to accompany a first course in data analysis.* New York: Wiley.

Cicchitelli, G., Bartolucci, F., & Forcina, A. (1999, August). *Assessment in statistics using the personal computer.* Paper presented at the 52nd International Statistical Institute Session, Helsinki, Finland.

Cobb, G. W. (1992). Teaching statistics. In L. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action* (pp. 3–43). Washington, DC: Mathematical Association of America.

Cobb, G. W. (1998, April). *The objective-format question in statistics: Dead horse, old bath water, or overlooked baby?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

College Entrance Examination Board and Educational Testing Service. (1998). *Advanced placement course description.* Princeton, NJ: Author.

Colvin, S., & Vos, K. (1997). Authentic assessment models for statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 27–36). Amsterdam: IOS Press.

Cross, K. P., & Steadman, M. H. (1996). *Classroom research: Implementing the scholarship of teaching.* San Francisco: Jossey-Bass.

Curcio, F., & Artzt, A. (1997). Assessing students' statistical problem-solving behaviors in a small-group setting. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 123–138). Amsterdam: IOS Press.

Facione, N. C., & Facione, P. A. (1994). *Holistic critical thinking scoring rubric.* Millbrae: California Academic Press.

Friel, S., Bright, G., Frierson, D., & Kader, G. (1997). A framework for assessing knowledge and learning in statistics (K–8). In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 35–63). Amsterdam: IOS Press.

Gal, I. (in press). Statistical literacy: Conceptual and instructional issues. In D. Coben, J. O'Donoghue, & G. FitzSimons (Eds.), *Adults learning mathematics: Research and practice.* London: Kluwer.

Gal, I., & Garfield, J. (Eds.). (1997). *The assessment challenge in statistics education.* Amsterdam: IOS Press.

Garfield, J. (1994). Beyond testing and grading: Using assessment to improve students' learning. *Journal of Statistics Education* [Online], *2*(1). Retrieved from the World Wide Web: http://www.amstat.org/publications/jse/v2n1/garfield.html

Garfield, J. (1998, April). *Challenges in assessing statistical reasoning.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in statistics: Implications for research. *Journal for Research in Mathematics Education, 19,* 44–63.

Garfield, J., delMas, R., & Chance, B. (1999a, April). *The role of assessment in research on teaching and learning statistics.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Garfield, J., delMas, R., & Chance, B. (1999b, January). *Using technology to improve statistical reasoning.* Paper presented at the annual Joint Mathematics Meetings, San Antonio, TX.

Garfield, J., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review, 67,* 1–12.

Gelman, A., & Nolan, D. (with Men, A., Warmerdam, S., & Bautista, M.). (1998). Student projects on statistical literacy and the media. *The American Statistician, 52,* 160–166.

Green, D. R. (1983). A survey of probability concepts in 3000 students aged 11–16 years. In D. R. Grey, P. Holmes, V. Barnett, & G. M. Constable (Eds.), *Proceedings of the 1st International Conference on Teaching Statistics* (pp. 766–783). Sheffield, England: Teaching Statistics Trust.

Hawkins, A., Jolliffe, F., & Glickman, L. (1992). *Teaching statistical concepts.* London: Longman.

Hill, N., & Walsh, S. (1997). Summer is here, and statistics means sunflowers! *Mathematics Teaching, 161,* 42–43.

Hubbard, R. (1997). Assessment and the process of learning statistics. *Journal of Statistics Education* [Online], *5*(1). Retrieved from the World Wide Web: http://www.amstat.org/publications/jse/v5n1/hubbard.html

Jones, G., Langrall, C., Thornton, C., & Mogill, A. (1997). A framework for assessing and nurturing young children's thinking in probability. *Educational Studies in Mathematics, 32,* 101–125.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press.

Keeler, C. (1997). Portfolio assessment in graduate level statistics courses. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 165–178). Amsterdam: IOS Press.

Konold, C. (1987). Teaching probability through modeling real life problems. *The Mathematics Teacher, 80,* 232–235.

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

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education* [Online], *3*(1). Retrieved from the World Wide Web: http://www.amstat.org/publications/jse/v3n1/konold.html

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 151–167). Voorburg, The Netherlands: International Statistical Institute.

Konold, C., Well, A., Lohmeier, J., & Pollatsek, A. (1993, September). *Understanding the law of large numbers.* Paper presented at the 15th Annual Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education, Pacific Grove, CA.

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23,* 557–568.

Lesh, R., Amit, M., & Schorr, R. (1997). Using "real-life" problems to prompt students to construct conceptual models for statistical reasoning. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 65–83). Amsterdam: IOS Press.

Lindquist, M. M. (Ed.). (1989). *Results from the Fourth Mathematics Assessment of the National Assessment of Educational Progress.* Reston, VA: National Council of Teachers of Mathematics.

Liu, H. J. (1998). *A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Mackisack, M. (1994). What is the use of experiments conducted by statistics students? *Journal of Statistics Education* [Online], *2*(1). Retrieved from the World Wide Web: http://www.amstat.org/publications/jse/v2n1/mackisack.html

Mastromatteo, M. (1993). Assessment of statistical analysis in eighth grade. In N. Webb (Ed.), *Assessment in the mathematics classroom* (1993 Yearbook of the National Council of Teachers of Mathematics, pp. 159–166). Reston, VA: National Council of Teachers of Mathematics.

Metz, K. (1997). Dimensions in the assessment of students' understanding and application of chance. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 223–238). Amsterdam: IOS Press.

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65,* 123–137.

National Council of Teachers of Mathematics. (1989). *The curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1995). *The assessment standards for school mathematics.* Reston, VA: Author.

Olsen, C. (1998, August). *What's a rubric? Scoring the open-ended questions.* Paper presented at the Joint Statistical Meetings, Dallas, TX.

Peck, R., Haugh, L. D., & Goodman, A. (Eds.). (1998). *Statistical case studies: A collaboration between academe and industry.* Philadelphia: American Statistical Association/Society of Industrial and Applied Mathematics.

Radke Sharpe, N. (1998, August). *Assessment issues in introductory and advanced statistics courses.* Roundtable paper presented at the Joint Statistical Meetings, Dallas, TX.

Schau, C., & Mattern, N. (1997). Assessing students' connected understanding of statistical relationships. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 91–104). Amsterdam: IOS Press.

Scheaffer, R. L. (1999, Spring). Making the grade—Again: AP statistics, 1998. *Stats: The Magazine for Students of Statistics, 25,* 3–5.

Shaughnessy, J. M. (1992). Research on probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465–494). New York: Macmillan.

Sommers, J. (1992). Statistics in the classroom: Written projects portraying real-world situations. *Mathematics Teacher, 85,* 310–313.

Starkings, S. (1997). Assessing student projects. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 139–151). Amsterdam: IOS Press.

Watson, J. (1997). Assessing statistical thinking using the media. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121). Amsterdam: IOS Press.

Webb, N. L., & Coxford, A. (Eds.). (1993). *Assessment in the mathematics classroom.* Reston, VA: National Council of Teachers of Mathematics.

Zawojewski, J., & Shaughnessy, J. M. (in press). Data and chance. In E. Silver & P. Kenney (Eds.), *Results from the Seventh Mathematics Assessment of the National Assessment of Educational Progress.* Reston, VA: National Council of Teachers of Mathematics.