Assessment in Statistics Education: Issues and Challenges

Joan Garfield, University of Minnesota
Beth Chance, Cal Poly - San Luis Obispo

1. Overview

There have been many changes in educational assessment in recent years, both within the fields of measurement and evaluation as well as in specific disciplines. This paper summarizes current assessment practices in statistics education, distinguishing between assessment at different educational levels and for different purposes. In order to provide a context for assessment of statistical learning, we first describe current learning goals for students, followed by some innovative alternative assessment methods currently utilized to better attain these goals. We will then highlight how these methods have been 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 in tertiary schools will be given. We will focus on three unique examples of effective uses of assessment. The paper concludes with a description of some current assessment challenges.

2. Learning Goals for Students

One of the forces driving change in the assessment of student learning of statistics has been educational reform within mathematics education (e.g., the NCTM Curriculum and Evaluation Standards). This reform has established probability and statistics as integral topics within the pre-college mathematics curriculum and defined new learning goals for students. These calls for reform have also infiltrated tertiary teaching (e.g. Cobb, 1992, Moore, 1997). Gal and Garfield (1997) provide 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, and the “big ideas” that underlie approaches to data-based inquiries. These ideas include the existence of variation, the need to describe populations by collecting data, 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 and how we make inferences from samples to populations.

2. Understand the process of statistical investigations

As Batanero argued, students (and practitioners) often fail to fully understand the logic of statistical inference or its role in experimental research. Students should understand the nature of and processes involved in a statistical investigation and considerations affecting the design of a plan for data collection. They should recognize how, when, and why existing statistical tools can be used to aid an investigative process. They should be familiar with the 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.

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, construct and display tables and different representations of data.

4. Understand probability and chance

Students need to develop an understanding of concepts and words related to probability, and understand probability as a measure of uncertainty. They need to know
how to develop and use models to simulate events, 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 and reflective questions about numerical argument, data reported in the media, and project reports from their classroom peers (e.g., How reliable are the measurements used? How representative was the sample? 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. Underlying this reasoning is a conceptual understanding of important ideas, such as distribution, center, spread, association, uncertainty, randomness, and sampling.
Although the depth or breadth of the seven goals described above may differ according to educational level, they describe the main goals for all students who learn basic statistics.

3. Assessment for individual student evaluation

Traditionally, assessment has been primarily used to assign grades to give periodic feedback on student learning. Assessment methods used to assign grades include quizzes and exams, homework exercises, and often computer lab activities. As statistics instruction begins to change in response to calls for reform, newer methods of assessment are being used to measure students' understanding of probability and statistics and their ability to achieve goals such as being able to explore data and to think critically using statistical reasoning. These alternative assessment approaches, such as projects, portfolios, and journals, better capture how students think, reason, and apply their learning, rather than merely having 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 (e.g., to provide a grade, to provide individual feedback, to evaluate statistical reasoning, to stimulate student reflection), the instructor needs to select an appropriate assessment method for a particular purpose (Garfield, 1994; Radke Sharpe, 1997). Some of the different possible methods include:

- quizzes (including calculations, graphs, and/or essay questions)
- minute papers (e.g., 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

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 this task or 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.

We now give more detail on some of the newer methods of assessment below. More detail is give by Garfield and Gal (1999b).

**Individual or group projects.** These projects typically involve posing a problem, designing an experiment or taking a sample, collecting and analyzing data, and interpreting the results (e.g., Mackisack (1994) or Starkings (1997) at the tertiary level; Hill and Walsh (1997) or Mastromatteo (1993) with secondary students). The project may be written up 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.

**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) provide examples of authentic tasks that integrate assessment with activities appropriate to a student’s life outside of school for primary students. For example, they describe an
assessment situation where students are given data on the weights of some grizzly and black bears living in Montana. Students are asked to solve problems about the bear cubs that require the students to organize, describe, and reason about the weights of the two types of bears. A scoring rubric is used to assign 0 to 3 points to the student responses. Another type of authentic task at the primary level is described by Lesh et al., (1977) who provide students with a detailed problem based on a real context, and use students’ strategies and interpretations as they solve the problem to construct a model of students reasoning. More extensive studies and tasks can be given at higher grade levels, e.g. Konold (1987), Sommers (1985) junior high students, Peck, Haugh, and Goodman (1998), Chatterjee, Handcock, and Simonoff (1996) college students, and Beck (1986) graduate students.

A portfolio 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 both the student and teacher 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) describes 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.

**looking for ref at other grades

Concept maps. Schau and Mattern (1997) describe 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 and/or relationship words are omitted. Students fill in these blanks either by generating the words or by
selecting them from a list which may or may not include distracters. For examples of this type of assessment see Schau and Mattern (1997).

Critiques of statistical ideas or issues in the news. Students may be asked to read and critique a newspaper article responding to particular questions as such as: (1) What do you think is the purpose of the research study described in this article? (2) What method or methods were used to answer the research question? (3) What questions would you like to ask the investigators in order to better understand the study? (4) Are there any aspects of the study that might make you question the conclusions presented in the article?

For examples of assessment items that involve brief articles or graphs presented in the media see Watson (1997) and Gelman, et. al. (1998).

Minute papers are brief, anonymously written remarks provided by students, sometimes on an index card or half-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 understand 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 above may be used in combination with each other as well in combination with traditional quizzes and exams. Chance (1997) provides details on her model for combining different assessment components.

4. Assessment for large group evaluation and study

Assessing statistical (including probabilistic) reasoning using a group procedure, such as a paper-and-pencil test, is a particularly challenging task that has not received much attention in the research literature. In fact, too often standardized exams have served as
examples of poor statistics and probability questions, and a misleading reflection of what we want our students to know. For example, multiple choice questions such as:

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 greater than 50?

focus too often on the calculation in an artificial setting, with no explanation or interpretation required of the students. While these exams are often a poor measure of students statistical reasoning ability, they are even less informative as a research tool.

Statistical reasoning has long been of interest to researchers in cognitive and social psychology, who have conducted studies of how people reason under uncertainty. A primary goal of statistics education today is to enable students to produce reasoned descriptions, judgments, inferences, and opinions about data. Therefore, many statistics instructors are interested in determining how well their students are able to use correct statistical reasoning during or after coursework in statistics.

One approach is developing objective-format questions to assess higher level thinking. Traditional exam questions may be adapted to assess deeper levels of conceptual understanding. 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) offers 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, prior record) by the same dependent variable (whether or not 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 judgment. A second principle is to involve two or more modes of statistical thinking (e.g., both 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 (1998) argues 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 suggests 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.

Most assessment instruments used in research studies of statistical reasoning and understanding consist of items given to students or adults individually as part of clinical interviews or in small groups that are closely observed. Although statistical reasoning may best be assessed through one-to-one communication with students (e.g., interviews or observations) or by examining a sample of detailed, in-depth student work (e.g., a statistical project), carefully designed paper-and-pencil instruments can be used to gather some limited indicators of students reasoning. One such instrument, The Statistical Reasoning Assessment (SRA), is highlighted in section 7. The AP Statistics Exam, highlighted in section 6, is another recent example of an assessment tool to be administered in large groups, that hopes to focus on reasoning as well as calculation.

5. Assessment as a research tool

Now that more students are first encountering statistics in elementary school mathematics classes, much attention is being paid to how these young children learn and understand statistical ideas, in addition to continued research for older students. More and more assessment is used as a method for gaining insight students’ understanding of statistical concepts and for modeling student reasoning, often as part of exploratory research. Furthermore, informative assessment tools are also needed to gauge the effectiveness of new technologies and teaching strategies. While there has been much agreement of the potential of these tools, there has not been significant research quantifying the impact on student learning.
In addition to the case studies presented by Ben-Zvi, Friel et al. (1997) present items for assessing primary students’ knowledge and interpretation of particular graphical representations of data. They describe 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 suggests 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) provide a framework for assessing students’ statistical problem solving skills in this context.

Jones et al. (1997) developed a framework used to assess children’s thinking about probability. This framework included 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 used to assess students’ thinking. Metz (1997) also provides tasks that can be used to describe children’s understanding of basic concepts related to probability. She analyzes her assessment of understanding along three dimensions: cognitive, epistemological, and cultural.

**Something at another grade level?**

In addition to these research studies, there many articles and materials available on assessing mathematical performance of students in primary grades that are relevant for assessing students’ statistical knowledge (e.g., Webb & Coxford, 1993, NCTM, 1995). 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 better understand and diagnose students' statistical reasoning skills, and to gain insight into how statistical reasoning develops.


An influential model for assessment in secondary schools in the United States is the Advanced Placement Statistics examination. This exam was offered by the College Board for the first time in 1997 and was taken by approximately 7600 high school
students. In 1999, more than 25,000 exams were given. The structure of the exam has been 35 multiple-choice questions and six free-response questions, including one investigative task. The exam covers four main topics (The College Board, 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 the future exams.

The free response questions are scored holistically by statistics teachers from colleges and high schools, using a detailed scoring 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?

** could include the general scoring guide from the teacher's guide?

Thus statistical knowledge and communication of statistical ideas are both 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 enough information so that the line of reasoning can be followed, e.g. why the method was used, 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
Assessment in statistics education

the first administrations of the AP exams. Therefore, the College Board, 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 rubrics provide the means for consistent evaluation of students' open responses and critical thinking (e.g. Olsen, 1998, Facione and Facione, 1994). The key to this approach is inter-rater reliability. The graders, or readers, of the AP Statistics exam typically spend 30-60 minutes being trained on how to interpret the rubric and examining sample students responses. Each reader is paired with another reader (typically college & high school instructors are paired) who serves as a first resource if the reader needs assistance interpreting a student answer. Additional questions can be referred to the table leader or the question leader. Table leaders "backread" samples of readers' grading and clarify the rubric as necessary.

The 1999 free response questions are posted at www.collegeboard.org/ap/statistics/faq99/.
The third question is reprinted here:

3. The dentists in a dental clinic would like to determine if there is a difference between the number of new cavities in people who eat an apple a day and in people who eat less than one apple a week. They are going to conduct a study with 50 people in each group.

Fifty clinic patients who report that they routinely eat an apple a day and 50 clinic patients who report that they eat less than one apple a week will be identified. The dentists will examine the patients and their records to determine the number of new cavities the patients have had over the past two years. They will then compare the number of new cavities in the two groups.

a. Why is this an observational study and not an experiment?

b. Explain the concept of confounding in the context of this study. Include an example of a possible confounding variable.

c. If the mean number of new cavities for those who ate an apple a day was statistically significantly smaller than the mean number of new cavities for those who ate less than one apple a week, could one conclude that the lower number of new cavities can be attributed to eating an apple a day? Explain.

This question illustrates the many skills required of the students: 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 Banatenro
urges, students need to realize the limitations of what conclusions can be drawn from a statistically significant result. The rubric (also available at the College Board website) clearly focuses on the students' ability to clearly explain their reasoning.

In grading this problem holistically, readers needed to see if students sufficiently explained

- why this was not an experiment
- the differential effect of the confounding variable on the response
- why a cause and effect conclusion could not be drawn

anywhere in their answer. Thus, the answer to question (a) did not have to appear strictly in the space for question (a) as the problem is graded as a whole.

One difficulty with this particular problem, is that it requires students to define a specific term, "confounding". While students can receive full credit if they explain the concept and clearly identify the concept as confounding, many students confused confounding variables with lurking variables. This focus on terminology surprised some instructors. Further, there is not universal acceptance in current textbooks on the definition of experiments or confounding, if the latter term is explicitly defined at all.

The 1999 Investigative Task concerned data on subjects' ability to correctly predict the outcome of a coin toss:

6. Researchers want to see whether training increases the capability of people to correctly predict outcomes of coin tosses. Each of twenty people is asked to predict the outcome (heads or tails) of 100 independent tosses of a fair coin. After training, they are retested with a new set of 100 tosses. (All 40 sets of 100 tosses are independently generated.) Since the coin is fair, the probability of a correct guess by chance is 0.5 on each toss. The numbers correct for each of the 20 people were as follows.

Students were presented with the raw data and then asked the following questions:

a. Do the data suggest that after training people can correctly predict coin toss outcomes better than the 50 percent expected by chance guessing alone? Give appropriate statistical evidence to support your conclusion.

b. Does the statistical test that you completed in part (a) provide evidence that this training is effective in improving a person's ability to predict coin toss outcomes?

If yes, justify your answer. If no, conduct an appropriate analysis that would allow you to determine whether or not the training is effective.
c. Would knowing a person's score before training be helpful in predicting his or her score after training?
Justify your answer.

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 testing (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.

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

<< I'd stop there, but another good one is A plot of the number of defective items produced during 20 consecutive days at a factory is shown below.

a. Draw a histogram that shows the frequencies of the number of defective items.

b. Give one fact that is obvious from the histogram but is not obvious from the scatterplot.

c. Give one fact that is obvious from the scatterplot but is not obvious from the histogram.
Showing how students have to combine and apply knowledge from different parts of the course.

7. Example 2: The Statistical Reasoning Assessment

The Statistical Reasoning Assessment (SRA) was developed and validated as part of the NSF-funded ChancePlus Project (Konold, 1990; Garfield, 1991), for evaluating 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 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. 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. Items from this instrument have been adapted and used in research projects in other English-speaking countries such as Australia and the United Kingdom and have been translated into Spanish, French and Chinese for studies in Spain, France, and Taiwan.

The following applications of reasoning were used to direct development and selection of SRA items (Garfield and Gal, 1999a).

*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; 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, likelihood to make judgments about uncertain events, knowing that not all outcomes are equally likely, 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 a computer program).

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; being cautious when making inferences made on small or biased samples.

Reasoning about association: knowing how to judge and interpret a relationship between two variables, knowing how to examine and interpret a two way table or scatterplot when considering a bivariate relationship, 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 (e.g., 1982) are well-known for their substantial body of research that 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, 1991, 1995; Lecoutre, 1992). Several of the identified misconceptions or errors in reasoning used to develop the SRA are described below:
Assessment in statistics education

**Misconceptions involving averages:** 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:** 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 (Konold, 1989). For example: 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 didn't do such a good job, because it should have rained on all days on 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 would mean it will not rain.

**Good samples have to represent a high percentage of the population:** 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:** Small samples should resemble the populations from which they are sampled, so small samples are used as a basis for inference and generalizations (Kahneman, Slovic, & Tversky, 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, Slovic, & Tversky, 1982).

**The Equiprobability bias:** Events tend to be viewed as equally likely. Therefore, the chances of getting different outcomes (e.g., three fives or one five on three rolls of a
dice) are incorrectly viewed as equally likely events (Lecoutre, 1992). Similarly, the probability 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.

In order 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 where 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 eight categories of incorrect reasoning 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. A test-retest reliability analysis yielded a reliability of .70 for the correct total score and .75 for the incorrect reasoning scores (Liu, 1998).

Table 1
Correct Reasoning Skills and Misconceptions Measured by the SRA

<table>
<thead>
<tr>
<th>Correct Reasoning Skills</th>
<th>Misconceptions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Correctly interprets probabilities</td>
<td>1. Misconceptions involving averages</td>
<td></td>
</tr>
<tr>
<td>2. Understands how to select an appropriate average</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Correctly computes probability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. understands probabilities as ratios</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. uses combinatorial reasoning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Understands independence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Understands sampling variability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Distinguishes between correlation and causation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Correctly interprets two-way tables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Understands importance of large samples</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Liu, 1998
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

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., sampling variability, 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, males have higher total correct reasoning scores and lower total misconception scores than their female counterparts. Results were more striking in the Taiwan sample than the US 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 problem, it is nonetheless problematic as a research and 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.

**8. 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, 1999). Based on the belief that students develop a shallow and isolated understanding of important foundational concepts in a traditional statistics classes, 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 accompany 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. Ten test banks or instructor guides for introductory statistics texts were examined, and revealed 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 type of question 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 are inadequate. Students might be able to get correct answers or good 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, 1986), Garfield, delMas and Chance (1999) developed instructional units for concepts such as sampling
distribution, confidence interval, and p-values, that involve the use of simulation software, and using them in different institutional settings. Each instructional unit includes a pretest and a post-test. The pretest is a diagnostic test of prerequisite knowledge for the unit, while the posttest evaluates students’ understanding of the important concepts and their ability to apply these concepts in solving problems in a new context.

**Example Pre-test Question:** Students are given two histograms and 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

This question allows instructors to clear up misconceptions about variability before students proceed with the learning activity, as well as to identify the reasoning students are using in their selection.

**Example Post-test Question:** In addition to graphic-based items, we also want to assess if students can apply their knowledge of the central limit theorem's implications. An example scenario is the post-office problem: 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.). Every day for one year, about 5 men registered at a small post office and about 50 men registered at a large post office. At the end of each day, a clerk at each post office computed and recorded the average height of the men who registered 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 inc)?
- (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 71 inches would be greater for the large post office than for the small post office.
(c) There is no basis for predicting which post office would have the greater number of days.

Students can also be asked to sketch the appropriate distribution.

Such tools allow instructors to monitor student understanding and how it is influenced by different instructional tasks. They also allow researchers to compare student performance in different settings and after a time-delay to measure retention.

In creating new instructional models, we utilize the following assessment framework:

a. Identify the desired knowledge and skills to be developed by students related to each concept. Distinguish what students should understand and what students should be able to do with their knowledge.
b. Identify the specific prerequisite concepts and skills that are necessary in order to learn and understand the new concept.
c. Identify typical students misunderstandings and misconceptions related to the concept
d. Develop appropriate assessment items for diagnosing errors or gaps in students’ prerequisite knowledge.
e. Develop appropriate assessment items to reflect students’ understanding of the concept
f. Develop appropriate assessment items to evaluate students’ ability to apply their knowledge in solving problems with a specified context.

As instruments were developed, pilot tested and reviewed, these lists of prerequisite skills and misconceptions were reviewed and revised as well. All materials, including the software, are available on a web site (www.gen.umn.edu/faculty_staff/delmas/stat_tools/index.htm) for easy access by statistics instructors.

9. Current assessment challenges
As more educators adopt alternative methods of assessing students learning, new questions arise that need to be addressed. Gal and Garfield (1997) summarize some of these challenges.

1. What are appropriate methods for assessing students who learn statistics using new technological tools?

2. What are appropriate ways to assess the application or transfer of student learning to interpretive or functional tasks such as those encountered in media or outside the classroom?

3. What are appropriate ways to assess students’ intuitions and reasoning involving probability concepts and processes in classroom settings?

4. What are appropriate ways to assess and grade collaborative statistical project work?

5. What are appropriate methods for using assessment to evaluate the relative utility of new instructional materials and methods?

6. What are appropriate and consistent methods for evaluating students' open ended responses and explanations of statistical concepts?

10. Conclusion

This paper summarized current learning goals for students of statistics. 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 revising their courses in light of reform suggestions will increasingly incorporate newer assessment methods and combinations of assessment methods to provide detailed information to
Assessment in statistics education

students as well as to inform their teaching, and our models of students' reasoning process. We have focused on several examples at different grade levels of various assessment methods. These methods illustrate the different roles that assessment can take beyond assigning student grades. By using techniques which inform the teacher, the student, and the researcher, and which concentrate on students' reasoning and performance in more authentic tasks, assessment can lead to the most important educational goal, improving students’ learning of statistics.

11. References


Hill, N. & Walsh, S. (1997). Summer is here, and statistics means sunflowers!
The writers describe how they taught statistics to year 8
students using a class project that involved growing
sunflowers. Each student grew a sunflower in the biology
greenhouse during a school term. Over this period, students
collected growth data and calculated the mean, median, mode,
and range. Measurements of the sunflowers were used to remind
students of different units of length. Students used the data
they had gathered to create their own database on the
sunflowers and to develop graphs and pie charts


