THE STATISTICAL REASONING ASSESSMENT: DEVELOPMENT AND VALIDATION OF A RESEARCH TOOL

Joan B. Garfield, University of Minnesota, USA

This paper describes the development and validation of the Statistical Reasoning Assessment (SRA), an instrument consisting of 20 multiple-choice items involving probability and statistics concepts. Each item offers several choices of responses, both correct and incorrect, which include statements 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 provides 16 scores which indicate the level of students' correct reasoning in eight different areas and the extent of their incorrect reasoning in eight related areas. Although the 16 scales represent only a small subset of reasoning skills and strategies, they provide useful information regarding the thinking and reasoning of students when solving statistical problems.

THE NATURE OF 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. Much of statistical reasoning combines 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.

A primary goal of statistics education is to enable students to produce reasoned descriptions, judgments, inferences, and opinions about data. Current mathematics curricula for students in elementary and secondary schools are designed to help students to comprehend and deal with uncertainty, variability, and statistical information in the world around them, and to participate effectively in an information-laden society. (Gal and Garfield, 1997). This requires students to develop good statistical reasoning skills.

ASSESSING STATISTICAL REASONING

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 which are closely observed. Most paper and pencil assessment instruments focus on computational skills or problem solving, rather than on reasoning and understanding.

Traditional test questions involving statistical content often lack appropriate context and tend to focus on accuracy of statistical computations, correct application of formulas, or correctness of graphs and charts. Questions and task formats that culminate in simple "right or wrong" answers do not adequately reflect the nature of students' thinking and problem solving, and therefore provide only limited information about students' statistical reasoning processes and their ability to construct or interpret statistical arguments (Gal and Garfield, 1997).

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 is *The Statistical Reasoning Assessment* (SRA).

The SRA was developed and validated as part of the NSF-funded ChancePlus Project (Konold, 1990; Garfield, 1991), to use in evaluating the effectiveness of a new statistics curriculum for high school students 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. There was a practical need to have an easily scorable instrument that captures students' thinking, reasoning, and application of knowledge, rather than a test where students "tell" the teacher what they have remembered or show that they can perform calculations or carry out procedures correctly.

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. This instrument has been translated into French, Spanish, and Chinese versions.

STATISTICAL REASONING GOALS FOR STUDENTS

The first step in developing or considering an assessment of statistical reasoning is to clarify the types of reasoning skills students should develop. The following types of reasoning were used to develop and select items to use in 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 representations of data: understanding the way in which a plot is meant to represent a sample, understanding how to read and interpret a graph and knowing how to modify a graph to better represent a data set; being able to see beyond random artifacts in a distribution to recognize general characteristics such as shape, center and spread.

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.

INCORRECT STATISTICAL REASONING

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 that reveals some prevalent ways of thinking about statistics that are inconsistent with a technical understanding. Their 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 and Ahlgren, 1988, Shaughnessy, 1992). Other researchers have discovered additional misconceptions or errors of reasoning when examining students in classroom settings (e.g., Konold, 1989; Lecoutre, 1992). Several of the identified misconceptions or errors in reasoning were used to develop the SRA, which are described below:

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), 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 series of events-70% chance of rain means that it should rain. Similarly, a forecast of 30% rain would mean it won't 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, and 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, and 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).

VALIDITY AND RELIABILITY ANALYSES

Once 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 revision 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 of selected responses to use in a subsequent multiple-choice format in the instrument. After several pilot tests of the SRA followed by administration of the instrument in different settings, and after many subsequent revisions, the current version was created.

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 number summary seemed uninformative and did not adequately reflect 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. Scores for each scale ranged from 2 to 8, depending on how many responses contributed to that scale. In addition to the 16 scale scores, total scores for correct and incorrect reasoning could be obtained. A test retest reliability analysis yielded a reliability of .70 for the correct total score and .75 for the incorrect reasoning scores (Liu, 1998).

CURRENT WORK

Now that an appropriate scoring method has been developed for the SRA, the instrument is being used in crosscultural studies. Liu (1998) used the SRA to determine if gender differences exist in large samples of college students in the USA and in Taiwan. Comparisons of college students in the two countries yielded striking similarities in reasoning scale scores. Plans are currently being made to administer the SRA in France and in Spain for similar analyses.

REFERENCES

Gal, I. (1995). Statistical tools and statistical literacy: The case of the average. *Teaching Statistics, 17 (3)*, 97-99.

Gal, I. and Garfield, J. (Eds.) (1997).*The Assessment Challenge in Statistics Education.* Amsterdam: IOS Press.

Garfield, J. (1991). Evaluating Students' Understanding of Statistics: Development of the Statistical Reasoning Assessment. In *Proceedings of the Thirteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Volume 2* (pp. 1-7). Blacksburg, VA.

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

Kahneman, D., Slovic, P. and Tversky, A. (1982) *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

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

Konold, C. (1990). *ChancePlus: A Computer-Based Curriculum for Probability and Statistics*. Final Report to the National Science Foundation. Scientific Reasoning Research Institute, University of Massachusetts, Amherst.

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

Liu, C. A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States. Doctoral dissertation, University of Minnesota, Minneapolis.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. (Ed.) *Handbook of Research on Mathematics Teaching and Learning*, edited by D. A. Grouws. New York: Macmillan, 465-494.