ASSESSING STATISTICAL REASONING

JOAN B. GARFIELD
University of Minnesota, USA
jbg@umn.edu

SUMMARY

This paper begins with a discussion of the nature of statistical reasoning, and then 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. Results are presented of a cross-cultural study using the SRA to compare the reasoning of males and females in two countries.

Keywords: Statistics education research; Assessment; Reasoning; Misconceptions

1. INTRODUCTION

Two reform movements have been affecting the teaching and learning of statistics at all educational levels. One reform has focused on content and pedagogy, shifting the focus from computation and procedures to an emphasis on statistical reasoning and thinking (Moore, 1997). A second reform is in the area of student assessment, focusing on better alignment of instruction with important learning goals, and using assessment as a tool to improve student learning (Garfield, 1993, Garfield & Gal, 1999a, Chance & Garfield, 2002). Some of the statements identified with the assessment reform include:

- Traditional forms of assessment (e.g., multiple-choice exams) are too narrow to provide sufficient information about student learning.
- Providing single numbers or letters to students to represent their learning is inadequate and does not promote successful learning.
- Alternative types of assessment are needed, used in combination, and aligned with instructional and curricular goals.

While elementary and secondary schools have embraced new assessment methods such as portfolios and performance tasks, college courses still rely primarily on more traditional tests and exams (Garfield, Hogg, Schau & Whittinghill, 2002). Traditional assessments of statistical knowledge typically look like textbook problems that either rely heavily on numerical calculations or on the ability to recall isolated pieces of information. Although this type of assessment may succeed in providing instructors with a method for determining letter grades, these types of assessment rarely reveal information about how students understand

and reason with statistical ideas or apply their knowledge to solve statistical problems. Garfield and Chance (2001) describe many alternatives to traditional assessment methods. One encouraging trend is the increased number of instructors who are using performance assessments in the form of student projects. However, many instructors of large classes have been hesitant about introducing this type of alternative assessment.

New forms of assessment are needed not only to provide information to students and instructors, but also to use in research on teaching and learning statistics, to evaluate the effectiveness of different curricula or pedagogical approaches, and to explore the development of statistical reasoning. Although student outcomes may best be assessed in personal interviews or via in-depth student work such as projects, there is 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.

2. THE NATURE OF STATISTICAL REASONING

Statistical reasoning may be defined as the way people reason with statistical ideas and make sense of statistical information (Garfield & Chance, 2000). 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, centre, spread, association, uncertainty, randomness, and sampling.

Many people think of mathematics and statistics as the same thing, and therefore, confuse statistical reasoning with mathematical reasoning (Garfield & Gal, 1999b). Today's leading statistical educators view these disciplines and types of reasoning as quite distinct. Gal and Garfield (1997) distinguish between the two disciplines in the following ways:

- In statistics, data are viewed as numbers with a context. The context motivates procedures and is the source of meaning and basis for interpretation of results of such activities.
- The indeterminacy or "messiness" of data distinguishes statistical investigations from the more precise, finite nature characterizing mathematical explorations.
- Mathematical concepts and procedures are used as part of the solution of statistical problems. However, the need for accurate application of computations is rapidly being replaced by the need for selective, thoughtful, and accurate use of technological tools and increasingly more sophisticated software programs.
- Many statistical problems do not have a single mathematical solution, but instead, start with a question and result in an opinion supported by certain findings and assumptions. These answers need to be evaluated in terms of quality of reasoning, adequacy of methods employed, and nature of data and evidence used.

In recent years there has been an appropriate shift from traditional views of teaching statistics as a mathematical topic (with an emphasis on computations, formulas, and procedures) to the current view that distinguishes between mathematics and statistics as separate disciplines. As Moore (1992) argues, statistics is a mathematical science but is not a branch of mathematics, and has clearly emerged as a discipline in its own right, with characteristic modes of thinking that are more fundamental than either specific methods or mathematical theory.

As statistics is being distinguished from mathematics, so is statistical reasoning being distinguished from mathematical reasoning. DelMas (in press) provides an analysis and comparison of these two types of reasoning. He argues that while mathematical and statistical reasoning can be distinguished with respect to the content of reasoning (abstract versus contextual), certain statistical content has an abstract nature that proves difficult for students. He reviews the relevant research on reasoning and uses these findings to identify areas of statistical reasoning that students find most challenging. At the current time, studies on particular aspects of statistical reasoning are still in the early stages of identifying the development of different types of reasoning and how reasoning may be affected by particular teaching activities and technological tools (for a collection of these studies, see BenZvi &Garfield, in press and the STRL-2 summaries published in SERJ 1(1)). Despite a substantial body of knowledge on how to effectively promote statistical reasoning in students, 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 being written or revised to help students comprehend and deal with uncertainty, variability, and statistical information in the world around them. This emphasis in statistical education on developing statistical reasoning illustrates the need for good methods to assess students' statistical reasoning.

2.1. 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 & 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.

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 also 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.

2.2. 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. These topics seemed appropriate for the purpose of the instrument, which was to be used with secondary school students who had been learning the basic techniques of data analysis as part of the ChancePlus Project described above.

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, centre and spread.

These two types of reasoning (about data and representation) were not linked to specific items in the SRA but are related to many of the items assessing reasoning. The following types of reasoning are linked to particular items in the SRA as shown in Table 1.

Reasoning about statistical measures: Understanding what measures of centre, 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 centre as well as a measure of spread and that summaries of centre 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 scatter plot when considering a bivariate relationship; knowing that a strong correlation between two variables does not mean that one causes the other.

3. 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

analysing 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 & 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, used to develop the SRA, 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), 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 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, & 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).

4. 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 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 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. A copy of the instrument is attached to this paper.

Table 1. Correct Reasoning Skills and Misconceptions Measured by the SRA and the Corresponding Items and Alternatives for Measuring Each Conception and Misconception

Correct Reasoning Skills	Items and Alternatives
1. Correctly interprets probabilities	2d, 3d
2. Understands how to select an appropriate average	1d, 4ab, 17c
3. Correctly computes probability	
a. Understands probabilities as ratios	8c
b. Uses combinatorial reasoning	13a, 18b, 19a, 20b
4. Understands independence	9e, 10df, 11e
5. Understands sampling variability	14b, 15d
6. Distinguishes between correlation and causation	16c
7. Correctly interprets two-way tables	5,1d
8. Understands importance of large samples	6b, 12b
Misconceptions	
1. Misconceptions involving averages	
a. Averages are the most common number	1a, 17e
b. Fails to take outliers into consideration when computing the mean	1c 15bf
c. Compares groups based on their averages	17a
d. Confuses mean with median	
2. Outcome orientation misconception	2e, 3ab, 11abd, 12c, 13b
3. Good samples have to represent a high percentage of the population	7bc, 16ad
4. Law of small numbers	12a, 14c
5. Representativeness misconception	9abd, 10e, 11c
6. Correlation implies causation	16be
7. Equiprobability bias	13c, 18a, 19d, 20d
8. Groups can only be compared if they are the same size	6a

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 statistical reasoning and misconceptions are unrelated to students' performance in a first statistics course.

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 (see Table 1). One item (number 7) was not included in a correct reasoning scale because it seemed to be primarily effective in assessing misconceptions. 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 new set of data was gathered from a group of 32 students enrolled in an assessment course for preservice teachers. These students took the SRA and one week later took the same test again. A test-retest reliability analysis yielded a reliability of .70 for the correct total score and .75 for the incorrect reasoning scores (Liu, 1998).

5. CROSS CULTURAL STUDY

Once an appropriate scoring method was developed for the SRA, the instrument was used in cross-cultural study. Liu (1998) used the SRA to determine if gender differences exist in large samples of college students in the USA and in Taiwan. In this study, the SRA was administered to 267 subjects at the University of Iowa, and a translated, Chinese, version of the SRA was administered to 144 subjects at Cheng-Chi University and 101 subjects at Feng-Chia University in Taiwan. All students were tested at the end of an introductory course in business statistics. The students were of comparable ages and had no prior instruction in statistics. However, in the two Taiwan samples there were higher percentages of females (60% and 74%) than in the Iowa sample (43%). The first analyses compared scale scores for students in each country, as shown in Table 2.

Table 2. Comparison of scaled scores on SRA for Taiwan and Iowa Students (scale = 0 to 2 points)

	Taiwan (n=245)	Iowa (n=267)
Correct Reasoning Scales		
1. Correctly interprets probabilities	1.36	1.35
2. Understands how to select an appropriate average	1.19	1.22
3. Correctly computes probability	.90	.91
4. Understands independence	1.47	1.25
5. Understands sampling variability	.46	.44
6. Distinguishes between correlation and causation	1.30	1.31
7. Correctly interprets two-way tables	1.57	1.30
8. Understands importance of large samples	1.52	1.35
Misconceptions Scales		
1. Misconceptions involving averages	.43	.59
2. Outcome orientation (misconception)	.43	.45
3. Good samples have to represent a high percentage of the population	.18	.18
4. Law of small numbers	.68	.58
5. Representativeness misconception	.21	.34
6. Correlation implies causation	.20	.20
7. Equiprobability bias	1.12	1.12
8. Groups can only be compared if they are the same size	.78	1.20

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. These scaled scores suggested that there

are strong similarities in reasoning for the two samples of students in spite of cultural differences. 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).

The second set of analyses investigated gender differences, using total correct reasoning scores and total misconception scores. Table 3 presents two-way analysis of variance results for the total correct reasoning scores by country and gender, which indicates that the country effect is highly significant. As shown in Figure 1, students in Taiwan have higher correct reasoning scores than the students in the United States. While the male sample in Taiwan has higher scores than the females, in the United States males and females have more similar scores. Both the gender effect and the interaction effect between country and gender are not significant.

Table 3. Analysis of Variance Results for Total Correct Reasoning Scores by Country and Sex

Source of Variation	DF	Mean Square	F	F-prob
Country	1	307.268	13.945	<.001 **
Sex	1	69.051	3.134	.077
Country x Sex	1	73.028	3.314	.069
Error	508	22.035		

** significant at the alpha = .01 level

Table 4. Cell Means and Standard Deviations of Total Correct Reasoning Scores for Males and Females in Each Country

	Taiwan	United States
Male	22.90 (4.83)	20.55 (4.59)
Female	21.38 (4.80)	20.57 (4.58)

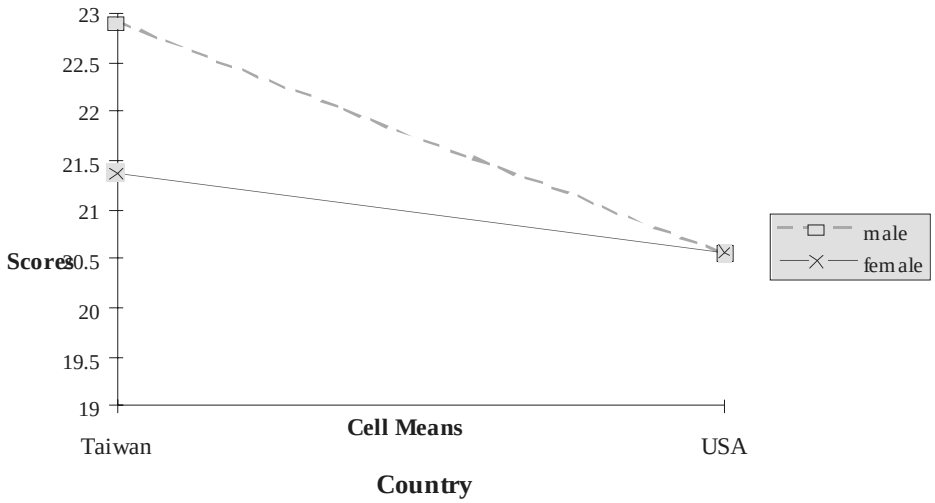


Figure 1: Cell Means of Total Correct Reasoning Scores for Each Sex by Two Countries

The ANOVA results for the total misconception scores by country and sex is presented in Table 5. Both country and sex effects are significant. Students in Taiwan have significantly lower misconception scores than students in the United States. As also shown in Figure 2, the female samples have significantly higher misconception scores than their male counterparts.

Table 5. Analysis of Variance Results for Misconception Total Scores by Country and Sex

Source	of Variation	DF	Mean Square	F	F-prob
Country	1	145.424	9.126	.003 **	
Sex	1	129.701	8.139	.005 **	
Country x Sex		1	31.258	1.962	.162
Error	508	15.935			

** significant at the alpha = .01 level

Table 6. Cell Means of Total Misconception Scores for Males and Females in Each Country

	Taiwan	United States
Male	11.28 (4.42)	12.87 (4.05)
Female	12.81 (3.53)	13.39 (4.11)

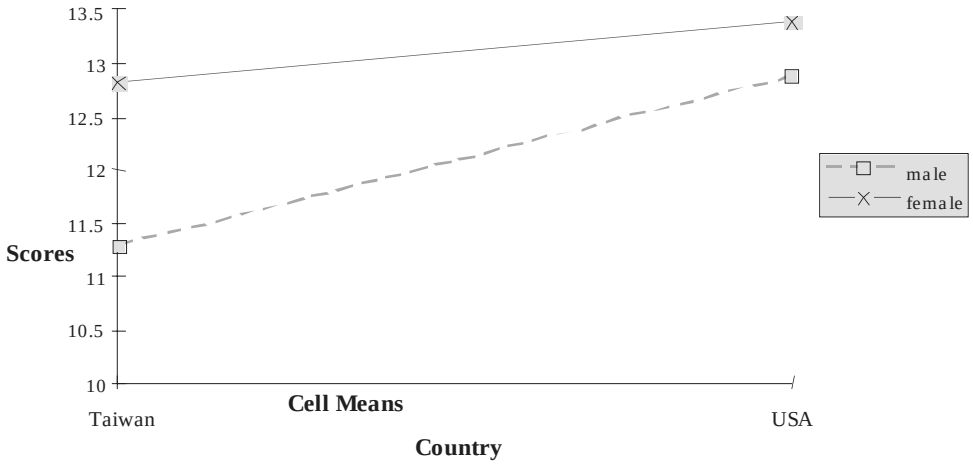


Figure 2: Cell Means of Total Misconception Scores for Each Sex by Two Countries

Liu 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 is interesting to see that despite the seemingly similar scale scores for the students in the two countries, that there are actually striking differences when comparing the male and female groups. However, it is important to note that these non-random samples are not equivalent, and the results should be interpreted with caution. Nevertheless, it will be interesting to see if replications of this study in other countries will yield similar results. The SRA is currently being administered in other countries and similar comparisons will be useful for comparison purposes.

6. SUMMARY

Although there is a growing emphasis on developing students' statistical reasoning, assessing statistical reasoning remains a challenging task, and one that needs more attention in the research literature. Although the SRA is an easy to administer, paper-and-pencil instrument that provides some useful information regarding the reasoning of students, it is nonetheless an imperfect research and evaluation tool. The 16 scales represent only a small subset of reasoning skills and strategies, and attempts to establish the reliability and validity have raised new issues and yielded incomplete results. Konold (2003) is currently working with other researchers to establish an improved set of items to assess students' statistical reasoning that should be available in the near future, and include some of the original SRA items in a revised format. In addition, a new Web-based assessment resource for teachers of statistics is collecting and developing a large item bank for assessing many aspects of statistical reasoning and thinking (Garfield, delMas and Chance, 2003). Despite these two efforts, there is still ample room for more studies that develop new assessments of statistical reasoning, as well as studies that investigate or build on current instruments and items. A set of valid and reliable instruments will be of great use both to teachers and researchers who want to evaluate students' statistical reasoning.

REFERENCES

BenZvi, D. & Garfield, J. (In press). *The Challenge of Developing Statistical Reasoning, Literacy, and Thinking*. Dordrecht: Kluwer.

delMas, R. (In press). A comparison of mathematical and statistical reasoning. In D. BenZvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Reasoning, Literacy, and Thinking*. Dordrecht: Kluwer.

Chance, B. L. & Garfield, J. B. (2002). New approaches to gathering data on student learning for research in statistics education. *Statistics Education Research Journal*, 1(2), 38-44.

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

Gal, I. & 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. (1994). Beyond testing and grading: Using assessment to improve instruction. *Journal of Statistical Education*, 1(2).

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. (2002) The challenge of developing statistical reasoning. *Journal of Statistics Education,* 10(3).

Garfield, J. & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematics Thinking and Learning, 2,* 99-125.

Garfield, J., delMas, R. & Chance, B. (2003). The Web-based ARTIST Project. Paper presented at the *Annual Meeting of the American Educational Research Association*, Chicago.

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

Garfield, J. & Gal, I. (1999b). Teaching and assessing statistical reasoning. In L. Stiff (Ed.), *Developing Mathematical Reasoning in Grades K-12: National Council Teachers of Mathematics 1999 Yearbook* (pp. 207-219). Reston, VA: N.C.T.M.

Garfield, J., Hogg, B., Schau, C. & Whittinghill, D. (2002). First courses in statistical science: the status of educational reform efforts, *Journal of Statistics Education,* 10(2).

Kahneman, D., Slovic, P. & 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.

Konold, C. (2003) The reciprocal relation between learning goals and assessment. Paper presented at the *Annual Meeting of the American Educational Research Association*, Chicago.

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

Liu, H. J. (1998). *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.

Moore, D. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review,* 65(2), 123-137.

Shaughnessy, J. M. (1992). Research in 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.

Joan B. GARFIELD
Department of Educational Psychology
University of Minnesota
Minneapolis, Minnesota 55455
USA

APPENDIX: STATISTICAL REASONING ASSESSMENT (SRA)

Purpose	The purpose of this survey is to indicate how you use statistical information in everyday life.
Take your time	The questions require you to read and think carefully about various situations. If you are unsure of what you are being asked to do, please raise your hand for assistance.

The following pages consist of multiple-choice questions about probability and statistics. Read the question carefully before selecting an answer.

1. A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student are shown below.

6.2 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.2

The students want to determine as accurately as they can the actual weight of this object. Of the following methods, which would you recommend they use?

___ a. Use the most common number, which is 6.2.

___ b. Use the 6.15 since it is the most accurate weighing.

___ c. Add up the 9 numbers and divide by 9.

___ d. Throw out the 15.3, add up the other 8 numbers and divide by 8.

2. The following message is printed on a bottle of prescription medication:

WARNING: For applications to skin areas there is a 15% chance of developing a rash. If a rash develops, consult your physician.

Which of the following is the best interpretation of this warning?

___ a. Don't use the medication on your skin, there's a good chance of developing a rash.

___ b. For application to the skin, apply only 15% of the recommended dose.

___ c. If a rash develops, it will probably involve only 15% of the skin.

___ d. About 15 of 100 people who use this medication develop a rash.

___ e. There is hardly a chance of getting a rash using this medication.

3. The Springfield Meteorological Center wanted to determine the accuracy of their weather forecasts. They searched their records for those days when the forecaster had reported a 70% chance of rain. They compared these forecasts to records of whether or not it actually rained on those particular days.

The forecast of 70% chance of rain can be considered very accurate if it rained on:

___ a. 95% - 100% of those days.

___ b. 85% - 94% of those days.

___ c. 75% - 84% of those days.

___ d. 65% - 74% of those days.

___ e. 55% - 64% of those days.

4. A teacher wants to change the seating arrangement in her class in the hope that it will increase the number of comments her students make. She first decides to see how many comments students make with the current seating arrangement. A record of the number of comments made by her 8 students during one class period is shown below.

Student Initials								
	A.A.	R.F.	A.G.	J.G.	C.K.	N.K.	J.L.	A.W.
Number of comments	0	5	2	22	3	2	1	2

She wants to summarize this data by computing the typical number of comments made that day. Of the following methods, which would you recommend she use?

___ a. Use the most common number, which is 2.

___ b. Add up the 8 numbers and divide by 8.

___ c. Throw out the 22, add up the other 7 numbers and divide by 7.

___ d. Throw out the 0, add up the other 7 numbers and divide by 7.

5. A new medication is being tested to determine its effectiveness in the treatment of eczema, an inflammatory condition of the skin. Thirty patients with eczema were selected to participate in the study. The patients were randomly divided into two groups. Twenty patients in an experimental group received the medication, while ten patients in a control group received no medication. The results after two months are shown below.

	Experimental group (Medication)	Control group (No Medication)
Improved	8	2
No Improvement	12	8

Based on the data, I think the medication was:

___ 1. somewhat effective

___ 2. basically ineffective

If you chose option 1, select the one explanation below that best describes your reasoning.

___ a. 40% of the people (8/20) in the experimental group improved.

___ b. 8 people improved in the experimental group while only 2 improved in the control group.

___ c. In the experimental group, the number of people who improved is only 4 less than the number who didn't improve (12-8), while in the control group the difference is 6 (8-2).

___ d. 40% of the patients in the experimental group improved (8/20), while only 20% improved in the control group (2/10).

If you chose option 2, select the one explanation below that best describes your reasoning.

___ a. In the control group, 2 people improved even without the medication.

___ b. In the experimental group, more people didn't get better than did (12 vs 8).

___ c. The difference between the numbers who improved and didn't improve is about the same in each group (4 vs 6).

___ d. In the experimental group, only 40% of the patients improved (8/20).

6. Listed below are several possible reasons one might question the results of the experiment described above. Place a check by every reason you agree with.

___ a. It's not legitimate to compare the two groups because there are different numbers of patients in each group.

___ b. The sample of 30 is too small to permit drawing conclusions.

___ c. The patients should not have been randomly put into groups, because the most severe cases may have just by chance ended up in one of the groups.

___ d. I'm not given enough information about how doctors decided whether or not patients improved. Doctors may have been biased in their judgments.

___ e. I don't agree with any of these statements.

7. A marketing research company was asked to determine how much money teenagers (ages 13 - 19) spend on recorded music (cassette tapes, CDs and records). The company randomly selected 80 malls located around the country. A field researcher stood in a central location in the mall and asked passers-by who appeared to be the appropriate age to fill out a questionnaire. A total of 2,050 questionnaires were completed by teenagers. On the basis of this survey, the research company reported that the average teenager in this country spends $155 each year on recorded music.

Listed below are several statements concerning this survey. Place a check by every statement that you agree with.

___ a. The average is based on teenagers' estimates of what they spend and therefore could be quite different from what teenagers actually spend.

___ b. They should have done the survey at more than 80 malls if they wanted an average based on teenagers throughout the country.

___ c. The sample of 2,050 teenagers is too small to permit drawing conclusions about the entire country.

___ d. They should have asked teenagers coming out of music stores.

___ e. The average could be a poor estimate of the spending of all teenagers given that teenagers were not randomly chosen to fill out the questionnaire.

___ f. The average could be a poor estimate of the spending of all teenagers given that only teenagers in malls were sampled.

___ g. Calculating an average in this case is inappropriate since there is a lot of variation in how much teenagers spend.

___ h. I don't agree with any of these statements.

8. Two containers, labeled A and B, are filled with red and blue marbles in the following quantities:

Container	Red	Blue
A	6	4
B	60	40

Each container is shaken vigorously. After choosing one of the containers, you will reach in and, without looking, draw out a marble. If the marble is blue, you win $50. Which container gives you the best chance of drawing a blue marble?

___ a. Container A (with 6 red and 4 blue)

___ b. Container B (with 60 red and 40 blue)

___ c. Equal chances from each container

9. Which of the following sequences is most likely to result from flipping a fair coin 5 times?

___ a. H H H T T

___ b. T H H T H

___ c. T H T T T

___ d. H T H T H

___ e. All four sequences are equally likely

10. Select one or more explanations for the answer you gave for the item above.

___ a. Since the coin is fair, you ought to get roughly equal numbers of heads and tails.

___ b. Since coin flipping is random, the coin ought to alternate frequently between landing heads and tails.

___ c. Any of the sequences could occur.

___ d. If you repeatedly flipped a coin five times, each of these sequences would occur about as often as any other sequence.

___ e. If you get a couple of heads in a row, the probability of a tails on the next flip increases.

___ f. Every sequence of five flips has exactly the same probability of occurring.

11. Listed below are the same sequences of Hs and Ts that were listed in Item 8. Which of the sequences is <u>least</u> likely to result from flipping a fair coin 5 times?

___ a. H H H T T

___ b. T H H T H

___ c. T H T T T

___ d. H T H T H

___ e. All four sequences are equally unlikely

12. The Caldwells want to buy a new car, and they have narrowed their choices to a Buick or a Oldsmobile. They first consulted an issue of <u>Consumer Reports,</u> which compared rates of repairs for various cars. Records of repairs done on 400 cars of each type showed somewhat fewer mechanical problems with the Buick than with the Oldsmobile.

The Caldwells then talked to three friends, two Oldsmobile owners, and one former Buick owner. Both Oldsmobile owners reported having a few mechanical problems, but nothing major. The Buick owner, however, exploded when asked how he liked his car:

First, the fuel injection went out — $250 bucks. Next, I started having trouble with the rear end and had to replace it. I finally decided to sell it after the transmission went. I'd never buy another Buick.

The Caldwells want to buy the car that is less likely to require major repair work. Given what they currently know, which car would you recommend that they buy?

___a. I would recommend that they buy the Oldsmobile, primarily because of all the trouble their friend had with his Buick. Since they haven't heard similar horror stories about the Oldsmobile, they should go with it.

___ b. I would recommend that they buy the Buick in spite of their friend's bad experience. That is just one case, while the information reported in <u>Consumer Reports</u> is based on many cases. And according to that data, the Buick is somewhat less likely to require repairs.

___ c. I would tell them that it didn't matter which car they bought. Even though one of the models might be more likely than the other to require repairs, they could still, just by chance, get stuck with a particular car that would need a lot of repairs. They may as well toss a coin to decide.

13. Five faces of a fair die are painted black, and one face is painted white. The die is rolled six times. Which of the following results is more likely?

___ a. Black side up on five of the rolls; white side up on the other roll

___ b. Black side up on all six rolls

___ c. a and b are equally likely

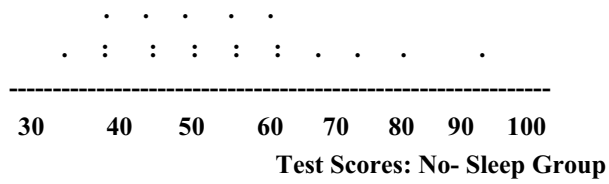
14. Half of all newborns are girls and half are boys. Hospital A records an average of 50 births a day. Hospital B records an average of 10 births a day. On a particular day, which hospital is more likely to record 80% or more female births?

___ a. Hospital A (with 50 births a day)

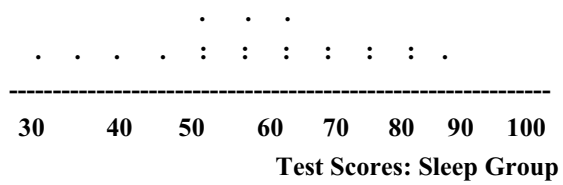
___ b. Hospital B (with 10 births a day)

___ c. The two hospitals are equally likely to record such an event.

15. Forty college students participated in a study of the effect of sleep on test scores. Twenty of the students volunteered to stay up all night studying the night before the test (no-sleep group). The other 20 students (the control group) went to bed by 11:00 p.m. on the evening before the test. The test scores for each group are shown in the graphs below. Each dot on the graph represents a particular student's score. For example, the two dots above the 80 in the bottom graph indicate that two students in the sleep group scored 80 on the test.



Test Scores: No- Sleep Group



Test Scores: Sleep Group

Examine the two graphs carefully. Then choose from the 6 possible conclusions listed below the one you most agree with.

___ a. The no-sleep group did better because none of these students scored below 40 and the highest score was achieved by a student in this group.

___ b. The no-sleep group did better because its average appears to be a little higher than the average of the sleep group.

___ c. There is no difference between the two groups because there is considerable overlap in the scores of the two groups.

___ d. There is no difference between the two groups because the difference between their averages is small compared to the amount of variation in the scores.

___ e. The sleep group did better because more students in this group scored 80 or above.

___ f. The sleep group did better because its average appears to be a little higher than the average of the no-sleep group.

16. For one month, 500 elementary students kept a daily record of the hours they spent watching television. The average number of hours per week spent watching television was 28. The researchers conducting the study also obtained report cards for each of the students. They found that the students who did well in school spent less time watching television than those students who did poorly. Listed below are several possible statements concerning the results of this research. Place a check by every statement that you agree with.

___ a. The sample of 500 is too small to permit drawing conclusions.

___ b. If a student decreased the amount of time spent watching television, his or her performance in school would improve.

___ c. Even though students who did well watched less television, this doesn't necessarily mean that watching television hurts school performance.

___ d. One month is not a long enough period of time to estimate how many hours the students really spend watching television.

___ e. The research demonstrates that watching television causes poorer performance in school.

___ f. I don't agree with any of these statements.

17. The school committee of a small town wanted to determine the average number of children per household in their town. They divided the total number of children in the town by 50, the total number of households. Which of the following statements must be true if the average children per household is 2.2?

___ a. Half the households in the town have more than 2 children.

___ b. More households in the town have 3 children than have 2 children.

___ c. There are a total of 110 children in the town.

___ d. There are 2.2 children in the town for every adult.

___ e. The most common number of children in a household is 2.

___ f. None of the above.

18. When two dice are simultaneously thrown it is possible that one of the following two results occurs: *Result 1:* A 5 and a 6 are obtained. *Result 2:* A 5 is obtained twice.

Select the response that you agree with the most:

___ a. The chances of obtaining each of these results is equal

___ b. There is more chance of obtaining result 1.

___ c. There is more chance of obtaining result 2.

___ d. It is imposible to give an answer. (Please explain why)

19. When three dice are simultaneously thrown, which of the following results is MOST LIKELY to be obtained?

___ a. *Result 1*: "A 5, a 3 and a 6"

___ b. *Result 2*: "A 5 three times"

___ c. *Result 3*: "A 5 twice and a 3"

___ d. All three results are equally likely

20. When three dice are simultaneously thrown, which of these three results is LEAST LIKELY to be obtained?

___ a. *Result 1*: "A 5, a 3 and a 6"

___ b. *Result 2:* "A 5 three times"

___ c. *Result 3*: "A 5 twice and a 3"

___ d. All three results are equally unlikely

IASE 2004 Research Round Table on Curricular Development in Statistics Education, Lund, Sweden, June 28 - July 3, 2004

The Round Table dates coordinate with those of the Tenth International Congress on Mathematical Education, which takes place in Copenhagen, Denmark 4-11 July 2004. Lena Zetterqvist (lena@maths.lth.se) and Ulla Holt will be local organisers. **Those interested** can contact Gail Burrill, Division of Science and Mathematics Education, College of Natural Science, Michigan State University, 116 North Kedzie, East Lansing MI 48824, USA, E-mail: (burrill@msu.edu).

IASE Activities at the 55th Session of the ISI, Sydney, Australia, April 5-12, 2005

Chris Wild is the IASE representative at the ISI Programme Co-ordinating Committee for ISI-55th Session, to be held in Sydney, Australia, April 5-12, 2005. As such he also is Chair of the IASE Programme Committee, which is in charge of preparing a list of Invited Paper Meetings to be organised by the IASE alone or in co-operation with other ISI Sections, Committees and sister societies. The committee will pay special attention to new topics that have been not discussed at the previous ISI Session. There is still time for you to propose a session theme for the IASE sessions for ISI55 in Sydney in 2005. Sessions that are of joint interest to IASE and another ISI section are also sought. Suggestions should normally include the name of the session organiser, a short description of the theme and an indicative list of possible speakers. Please email your proposals to Chris Wild at c.wild@auckland.ac.nz.

ICOTS-7, Working Cooperatively in Statistics Education, Brazil, 2006

We are also glad to announce that the IASE Executive accepted the proposal made by the Brazilian Statistical Association to hold ICOTS-7 in 2006 in Brazil. The proposal is also supported by the statistical associations in Argentina and Chile. Pedro Morettin <pam@ime.usp.br> is the Chair of the Local Organising Committee and Lisbeth Cordani <lisbeth@maua.br> is acting as a link between the IASE Executive and the local organisers. Scientific Committee IPC: Carmen Batanero (Chair), Susan Starkings (Chair Scientific Programme), John Harraway (Scientific Secretary), Allan Rossman and Beth Chance (Editors of Proceedings). More information from Carmen Batanero (batanero@ugr.es).