EFFECTS OF TEACHING STATISTICAL LAWS ON REASONING ABOUT PROBLEMS

by

Pertti Olavi Kosonen
B.A., Simon Fraser University, 1971
M.A., Simon Fraser University, 1979

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the Faculty
of
Education

© Pertti Olavi Kosonen 1992
SIMON FRASER UNIVERSITY
March 1992

All rights reserved. This work may not be reproduced in whole or in part, by photocopy or other means, without permission of the author.
APPROVAL

Name: Pertti Olavi (Peter) Kosonen
Degree: Doctor of Philosophy
Title of Thesis: Effects of Teaching Statistical Laws on Reasoning About Problems

Examinining Committee:
Chair: Jack Martin

Philip H. Winne
Senior Supervisor

Evelyn Ng
Assistant Professor

Bernice Wong
Professor

Richard Mayer
Professor, Department of Psychology
University of California
Santa Barbara, CA 93106-9660 U.S.A.
External Examiner

Date Approved March 20, 1992
PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of Thesis/Project/Extended Essay

EFFECTS OF TEACHING STATISTICAL LAWS

ON REASONING ABOUT PROBLEMS

Author: ____________________________

(signature)

Pertti Olayi (Peter) Kosonen

(name)

March 20, 1992

(date)
Contemporary psychology, in the tradition of Thorndike and James, has largely been characterized by the belief that instructionally useful inferential rules are not abstract or general. In contrast, recent research supports a view that instruction in abstract rule systems can improve reasoning about ill-defined problems that abound in real life. More specifically, the research shows that instruction in statistics can influence the way people reason about events involving uncertainty in everyday life and can produce knowledge that readily transfers to types of problems outside those focused on in instruction.

Building on a study by Fong, Krantz, and Nisbett (1986), who found that training people to use the heuristic of the law of large numbers substantially enhanced their reasoning about everyday statistical problems, three experiments, that investigated the effects of instructing students about this statistical heuristic, were performed. The experiments were carried out with a total of 315 participants in university as well as secondary and elementary schools.

The results (a) indicate that students learned a good deal about how to reason statistically as a consequence of instruction; (b) reveal no problem format-specificity of instructional effects; and, (c) suggest that typical reasoning errors were reduced by instruction in the statistical heuristic. These results stand in contrast to antiformalist views that stress domain-specificity in problem solving and lend support to the formalist view that teaching people to apply formal rules of inference can help them to reason more accurately about a variety of probabilistic events. In general, the findings bolster an optimistic view about the potential of rather limited instruction to foster valid reasoning about problems for students in university as well as in secondary and elementary school.
ACKNOWLEDGEMENTS

This work is dedicated to Robin, Amanda, and Elisa who make it all worthwhile.

To the following individuals, I extend my heartfelt appreciation: To Phil Winne, my senior supervisor, whose deep knowledge, energy, humour, and thoughtful leadership inspired me; to Evelyn Ng, whose caring and insight contributed important revisions; to Bob Prosser and Tim Seifert, whose statistical knowledge helped immeasurably; and, to fellow graduate students whose friendship and encouragement meant a great deal.

In addition, I extend warmest thanks to the elementary and secondary students, teachers, and principals in School District 43 (Coquitlam) and to the undergraduate students who participated.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPROVAL</td>
<td>ii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER ONE: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>A Normative View of Teaching Reasoning</td>
<td>3</td>
</tr>
<tr>
<td>Formal Rules of Inference</td>
<td>5</td>
</tr>
<tr>
<td>Pragmatic Inferential Rules</td>
<td>7</td>
</tr>
<tr>
<td>General Research Questions</td>
<td>10</td>
</tr>
<tr>
<td>Overview</td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER TWO: REVIEW OF LITERATURE</td>
<td>11</td>
</tr>
<tr>
<td>Problem Solving</td>
<td>12</td>
</tr>
<tr>
<td>Problem Space</td>
<td>13</td>
</tr>
<tr>
<td>Problem Representation</td>
<td>14</td>
</tr>
<tr>
<td>Inferential Rules</td>
<td>17</td>
</tr>
<tr>
<td>Syntactic versus Pragmatic Rules</td>
<td>18</td>
</tr>
<tr>
<td>Empirical Rules versus Inferential Rules</td>
<td>20</td>
</tr>
<tr>
<td>Implications for Instruction</td>
<td>21</td>
</tr>
<tr>
<td>Studies of Statistical Reasoning</td>
<td>25</td>
</tr>
<tr>
<td>Failure to Consider Statistical Information</td>
<td>33</td>
</tr>
<tr>
<td>Summary</td>
<td>35</td>
</tr>
<tr>
<td>Research Questions</td>
<td>36</td>
</tr>
<tr>
<td>CHAPTER THREE: EXPERIMENT 1</td>
<td>37</td>
</tr>
<tr>
<td>Method</td>
<td>39</td>
</tr>
<tr>
<td>Sample</td>
<td>39</td>
</tr>
<tr>
<td>Treatments and Controls</td>
<td>39</td>
</tr>
<tr>
<td>Measures</td>
<td>42</td>
</tr>
<tr>
<td>Procedure</td>
<td>42</td>
</tr>
<tr>
<td>Coding System</td>
<td>44</td>
</tr>
<tr>
<td>Interrater Agreement</td>
<td>47</td>
</tr>
<tr>
<td>Scoring</td>
<td>48</td>
</tr>
<tr>
<td>Results</td>
<td>48</td>
</tr>
<tr>
<td>Effects of Instruction on Statistical Reasoning</td>
<td>48</td>
</tr>
<tr>
<td>Relationship between Instructional Domain and Statistical Reasoning</td>
<td>49</td>
</tr>
</tbody>
</table>
APPENDICES

A. Example Problems and Analyses for Experiment 1 ........................................ 132
B. Test Problems for Experiment 1 with Directions ............................................. 146
C. Written Explanation of the Law of Large Numbers for Experiments 1 and 2 .... 151
D. Verbal Presentation of the Law of Large Numbers for Experiments 1 and 2 ..... 153
E. Example Problems and Analyses for the Rule Condition and the Rule-Plus-
   Error Condition: Experiment 2 ........................................................................ 157
F. Training Materials for Errors in the Rule-Plus-Error Condition: Experiment 2 .. 159
G. Pretest Problems and Directions: Experiment 2 .......................................... 162
H. Posttest Problems and Directions: Experiment 2 .......................................... 167
I. Probes of Conditional Reasoning, Domain Knowledge, and Sampling:
   Experiment 2 ................................................................................................. 170
J. Law of Good Sampling: Demonstration Script .................................................. 173
K. Examples and Analyses: Experiment 3 ............................................................ 177
L. Rule Group Teacher Instructions ...................................................................... 179
M. Law of Good Sampling Review ....................................................................... 181
N. Probe of Sampling Knowledge: Experiment 3 .................................................. 187
O. Error Materials: Experiment 3 .......................................................................... 192
P. Pretest and Posttest Problems and Directions: Experiment 3 ....................... 199
Q. Letter of Permission and Consent Form ........................................................... 214
LIST OF TABLES

Table 1  Descriptive Statistics on Quality of Statistical Reasoning ........................................ 50
Table 2  Results of A priori t-tests Comparing Quality of Statistical Reasoning by Instructional Condition Contrasted to Control Condition ........................................ 52
Table 3  Mean Number of Reasoning Errors Committed by Students in Problem Formats ........................................ 54
Table 4  Results of A priori t-tests Comparing Errors in Statistical Reasoning by Instructional Condition Contrasted to Control Condition ........................................ 57
Table 5  Correlation Matrix for Variables in Undergraduate Multiple Regression Analysis ........................................ 58
Table 6  Results of Multiple Regression Predicting Undergraduates' Quality of Statistical Reasoning on All Problems by Achievement, Major, and Courses ........................................ 59
Table 7  Pretest Problems by Format and Structure ........................................ 65
Table 8  Posttest Problems by Format and Structure ........................................ 66
Table 9  Descriptive Statistics on Quality of Undergraduate Subjects' Statistical Reasoning in Treatment Groups and Problem Formats ........................................ 71
Table 10  Descriptive Statistics on Undergraduates' Reasoning Errors in Treatment Groups and Problem Formats ........................................ 73
Table 11  University Undergraduates: Results of Repeated Measures MANOVA Comparing Treatment Conditions on Statistical Reasoning and Reasoning Errors ........................................ 76
Table 12  Correlation Matrix for Variables in Undergraduate Multiple Regression Analysis ........................................ 78
Table 13  Results of Stepwise Multiple Regression Predicting Undergraduate Posttest Scores in Statistical Reasoning ........................................ 79
Table 14  Pretest and Posttest Problems by Format and Structure ........................................ 87
Table 15  Descriptive Statistics on Grade 10 Students' Statistical Reasoning in Treatment Groups and Problem Formats ........................................ 91
Table 16  Descriptive Statistics on Grade 7 Students' Statistical Reasoning in Treatment Groups and Problem Formats ........................................ 92
Table 17  Descriptive Statistics on Grade 10 Students' Reasoning Errors in Treatment Groups and Problem Formats ................................................................. 95

Table 18  Descriptive Statistics on Grade 7 Students' Reasoning Errors in Treatment Groups and Problem Formats ................................................................. 96

Table 19  Grade 7 and Grade 10: Results of Repeated Measures MANOVA Comparing Treatment Conditions on Statistical Reasoning and Reasoning Errors ........................................................................................................ 100

Table 20  Correlation Matrix for Variables in Grade 10 and Grade 7 Multiple Regression Analyses ........................................................................................................ 104

Table 21  Results of Stepwise Multiple Regression Predicting Posttest Scores in Statistical Reasoning for Grade 10 Students ............................................................................ 105

Table 22  Results of Stepwise Multiple Regression Predicting Posttest Scores in Statistical Reasoning for Grade 7 Students ............................................................................ 106
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Mean Statistical Reasoning Score as a Function of Instructional Condition and Problem Format</td>
<td>51</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Mean Number of Reasoning Errors as a Function of Instructional Condition and Problem Format</td>
<td>55</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Plotted Means for Undergraduates' Statistical Reasoning as a Function of Instructional Condition and Problem Format</td>
<td>72</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Plotted Means for Undergraduates' Reasoning Errors as a Function of Instructional Condition and Problem Format</td>
<td>74</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Error Types as Presented to Grade 7 and 10 Students</td>
<td>86</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Plotted Means for Grade 10 Students' Statistical Reasoning as a Function of Instructional Condition and Problem Format</td>
<td>93</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Plotted Means for Grade 7 Students' Statistical Reasoning as a Function of Instructional Condition and Problem Format</td>
<td>94</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Plotted Means for Reasoning Errors in Grade 10 Students as a Function of Instructional Condition and Problem Format</td>
<td>97</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Plotted Means for Reasoning Errors in Grade 7 Students as a Function of Instructional Condition and Problem Format</td>
<td>98</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Plotted Means and Interaction Effects for Statistical Reasoning in All Problems as a Function of Instructional Condition and Level of Schooling</td>
<td>108</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Plotted Means and Interaction Effects for Reasoning Errors in All Problems as a Function of Instructional Condition and Level of Schooling</td>
<td>109</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

A student surveys her lecture notes to determine where to focus her efforts during the limited amount of time she has to study before the major exam. A businessman considers advantages and disadvantages of going into partnership with a friend. A national leader weighs the recommendations of advisors on a new foreign policy in the Middle East. A professor reflects on the likelihood of his research being funded. A basketball coach attempts to predict the kind of defense the team's perennial rival will open with in the opening game of the season.

These are all examples of reasoning in the face of uncertainty. The problem solver may have very little familiar, domain-specific information available for reasoning about the problem. Yet, uncertainty does not prevent people from making decisions. In fact, fuzzy, ill-defined sorts of problems abound in real life. According to Lesgold (1988), the essence of problem solving is being able to deal with new situations or problems that one has not been specifically trained to solve. In such everyday situations, people are often compelled to explain and predict events under conditions of limited information about the events. The information available may be data of variable and, sometimes, dubious quality. Nevertheless, decisions are based on such explanations and predictions, either because there is too little time to get additional information or because additional information is unavailable.

A predominant view among psychologists concerned with students' reasoning and problem solving is that (a) reasoning and problem solving requires knowledge of rules about events in particular domains of knowledge; and (b) domain-specific knowledge, not abstract inferential rules or heuristics, acts as the primary path toward improving students' problem solving (Gagne, 1980; Johnson-Laird, Legrenzi, & Sonino-Legrenzi, 1972; Newell, 1980).
Educators, adopting this view and interested in improving students' problem solving in everyday situations, are faced with a dilemma: While remaining pessimistic about the transferability of domain-specific knowledge from one domain to another, they will need to identify, from among numerous domains, specific domain knowledge which will enhance students' problem solving in everyday contexts.

In contrast to the dominant view emphasizing domain-specific knowledge, recent research by Fong, Krantz, and Nisbett (1986), as well as others (e.g., Cheng & Holyoak, 1985; Holland, Holyoak, Nisbett, & Thagard, 1989; Nisbett, Krantz, Jepson, & Kunda, 1983), supports a view that instruction in abstract rule systems can improve reasoning about fuzzy, ill-defined problems that abound in real life. More specifically, the research argues that instruction in statistics can influence the way people reason about events involving uncertainty in everyday life and can produce knowledge that readily transfers to domains outside those focused on in instruction. Since much of the recent research on reasoning under uncertainty involves artificial vignettes of problem situations provided to subjects as the major dependent measure, it is not based on authentic decision making situations encountered by people in the real world. However, this research on people's thinking about written descriptions of problems has shown improved statistical reasoning in a large variety of decision-making contexts after brief instructional sessions as well as after extensive formal training in statistics.

The present investigation builds on this line of research to explore how students in elementary school, secondary school, and university reason under conditions of limited information about probabilistic events and how instruction in formal rules of inference—more specifically, instruction in statistical heuristics—might assist them to reason more evidentially.
A Normative View of Teaching Reasoning

One view of intelligence (Barrow, 1991) maintains that the way we sift, organize, and use information—how we deal with what we know instead of just what we know—is a crucial component of intellectual competence. This view also holds that certain traditions of inquiry, including the scientific tradition, are logically necessary for human reasoning. Barrow argues that if we abandon the traditions of inquiry, we literally do not know how to reason and that schools need to promote these traditions in learners in order to develop their intelligence. The scientific tradition, among other traditions of intellectual inquiry, is seen by Barrow as contributing to a "normative sense of intelligence". According to this sense, intelligent and unintelligent performance must be distinguished from each other according to norms established in the traditions of inquiry. For example, magical explanations and scientific explanations of events in the everyday world cannot be accorded equal value because the former do not satisfy a norm about acceptable evidence. It follows that educational research and practice in the future should employ what we know about the nature of intellectual inquiry in order to develop human reasoning and invoke a normative sense of intelligence when according value to problem solutions. This conclusion about the necessity of reasoned evidence is an important one for this study. Like "bounded rationality", regarded by Herbert Simon (1979) as the normative starting point for research on problem solving, the premise taken here is that traditions of intellectual inquiry can serve to identify the products of reasoning as more or less evidential and accurate in utilizing information given in a problem. In other words, reasoning about fuzzy, ill-defined problems can be evaluated as being more or less valid.

A similar, normative view of human intellectual competence is offered by Nisbett and Ross (1980) in their landmark book, Human Inference: Strategies and Shortcomings of Social
Judgment. Rooted in the scientific tradition framed by Bacon, Descartes, Locke, Hume, Kant, and Mill, these authors examine the potential for improving inferential reasoning. Like the early epistemologists, Nisbett and Ross are concerned with describing processes and structures of human thought in the form of empirically testable propositions as well as normative prescription of more effective thought based on such empirical evidence. Nisbett and Ross cite other investigators whose empirical work is guided by normative comparisons of people's thinking to formal canons of inference: e.g., Herbert Simon, Daniel Kahneman, Amos Tversky, Paul Slovic, and Sarah Lichtenstein. The reasoner, in this view, is an "intuitive scientist" whose inferential shortcomings in attempting to understand, predict, and control events reflect, in part, a "failure to use the normative principles and inferential tools that guide formal scientific inquiry" (Nisbett & Ross, 1980, p. 3).

The scientific tradition provides principles and tools for thinking about new situations or problems that one has not been specifically trained to solve. For instance, systematic methods of observation and hypothesis testing are tools for thinking arising from the scientific tradition. Since education aims to prepare students for life beyond school, where new situations are commonplace, it is important for instructional psychology to research how people reason about problems and whether formal canons of inference arising from the scientific tradition can improve reasoning that gives rise to solutions.

What normative principles and inferential tools provide assistance for the student choosing topics to study for the impending exam; for the businessman deciding about the viability of partnership; for the politician judging foreign policy; for the professor considering an application for research funds; for the basketball coach who needs to predict the rival's opening defense? What tools and principles are useful when people have to make predictions based on data of variable and sometimes dubious quality? One set of normative principles is formal rules of inference.
Formal Rules of Inference

How reasoning in everyday situations is affected by instruction in formal rules of inference has been recently explored by a host of researchers (e.g., Cheng & Holyoak, 1985; Holland, et al., 1989; Nisbett, et al., 1983; Tversky & Kahneman, 1983). Furthermore, a growing number of studies focus on the effects of instructing people to apply statistical heuristics in solving problems (e.g., Fong, et al., 1986; Fong & Nisbett, 1990; Lehman, Lempert, & Nisbett, 1988).

The idea that abstract rule systems enhance reasoning has a long history. Ancient Greeks, including Plato and Aristotle, believed that people inherently understand abstract inferential rules and use them to solve problems. Roman thinkers believed that the study of mathematics and grammar as formal disciplines improved reasoning. Later, during the Middle Ages, scholars added logic to these disciplines. Subsequently, humanists included the study of Latin and Greek as domains believed to improve reasoning. In twentieth century psychology, this formalist view—that people solve inferential problems using general and abstract rules and schemas—is associated with the theories of Piaget and Simon.

The formalist views of Piaget and Simon are not widely reflected in contemporary psychology which has largely been characterized by a belief that instructionally useful inferential rules are not abstract or general. At the turn of the century, William James and E.L. Thorndike, from their studies of transfer in specific learning tasks, called into question the notion of abstract rules. Thorndike's argument was one of domain specificity: Transfer of training was dependent on "identical elements" between an instructional task and a new task. Research by Wason (1966) and Johnson-Laird, Legrenzi, and Sonino-Legrenzi (1972) strengthened the belief, that if people solve a problem correctly, it is because they are familiar with the content and grasp rules that are within the domain of the problem. Additional
support for this antiformalist stance came from findings by Kahneman and Tversky (1973) who showed how people violate laws of formal logic and rules of statistics, even when given formal instruction. For example, subjects persisted in deriving conclusions from small samples, apparently believing in a "law of small numbers" (Tversky & Kahneman, 1971), or subjects continued to disregard important information in favour of stereotypical representations of information. To recapitulate, antiformalist views of reasoning argue that if people solve a problem correctly, it is because they are familiar with the content domain and use domain-specific, not generalizable, rules.

In contrast to the antiformalist stance that emphasized domain specificity in problem solving (e.g., Johnson-Laird, 1972; Wason, 1966; Kahneman & Tversky, 1973), the research of Nisbett, Lehman, Fong, and others may support a view that teaching people about formal rules of inference can assist them to reason more validly about fuzzy and probabilistic events. In addition, the latter studies suggest that the effects of such instruction can transfer to subject areas outside the domain emphasized during instruction, thus reviving a two-thousand-year-old idea that instruction in abstract rule systems may enhance reasoning. In large measure, this recent research is a revision of the Piagetian tradition (e.g. Braine, 1978) which argued in favour of inferential rules similar to statistical rules. However, while Piaget believed in cognitive development as the basis for acquiring abstract inferential rules, the recent findings are more optimistic about the role of instruction in improving people's reasoning abilities.

This recent research re-addresses Kahneman and Tversky's notion of how problem solvers, inappropriately, rely on a "law of small numbers". It concurs with Nickerson's (1988) observation about "the error of psychologists in this century to assume that all rules show equally poor transfer and generalization properties based on a limited amount of research on a relatively small set of rules" (p.15). However, so far, it has not answered
Shulman and Elstein's (1975) call for studies of how typical error patterns in problem solving can be identified and reduced. This latter consideration is an important one for educators. What alternative inferential strategies do people apply when thinking about probabilistic events? Can the alternative approaches be considered errors? If so, teaching people about the errors as well as formal inferential rules may provide educators with a powerful means for improving people's reasoning abilities.

Pragmatic Inferential Rules

Recent work (Nisbett, et al., 1983; Tversky & Kahneman, 1983; Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986; Fong, et al., 1986; Holland, et al., 1989; Nisbett, Fong, Lehman, & Cheng, 1987; Lehman, et al., 1988; Fong & Nisbett, 1990) has identified types of rule systems or schemas labelled "pragmatic inferential rules". These, according to Lehman, et al., are "naturally occurring inferential rules that people use to solve everyday problems" (p. 432). People may induce primitive versions of these rules by virtue of their day-to-day practical experience. Furthermore, the researchers argue that these existing pragmatic inferential rules are improvable by instruction, an undertaking made easier because people already use primitive versions of the rules.

Among the types pragmatic inferential rules is a "family of related schemas or heuristics having to do with the law of large numbers; the rule that sample values resemble population values as a direct function of sample size and as an inverse function of population variability; and the related regression or base rate principles, for example, the rule that extreme values for an object or sample are less likely to be extreme when the object is reexamined or a new sample observed" (Lehman, et al., p. 433). The study by Fong, et al. (1986) found that people's solutions to everyday problems using statistical rules are greatly enhanced by a relatively brief training session in the law of large numbers. Fong, et al. attributed their
instructional effects to the idea that people already possess rudimentary statistical intuitions, thus making it possible to improve the rule system by relatively brief and direct instruction.

While the above mentioned research is optimistic about teaching people to reason more accurately about probabilistic events, other empirical evidence on the question of improving reasoning ability is limited and not very encouraging. In addition to the findings of James, Thorndike, Wason, and Johnson-Laird, cited earlier, the literature on transfer effects, as well as the literature on instruction in general problem solving strategies (e.g. Hayes & Simon, 1977; Gick & Holyoak, 1987; Nickerson, Perkins, & Smith, 1985; Resnick, 1987), offers a pessimistic view of improving reasoning. And, Perkins (1985) has shown how schooling, from elementary to graduate school, seems to do very little to foster critical thinking and valid reasoning about everyday events.

Three arguments, which counter the pessimistic view, give impetus to the general research questions in the present investigation: (a) The fuzzy, ill-defined sorts of problems that occur in real life were not studied by many of the previous "antiformalist" researchers. (b) Rules of logic and other rules not readily transferred to everyday problem solving are emphasized in twentieth-century antiformalist arguments. And, (c) the instructional practices critiqued by antiformalist researchers may not represent the most effective instructional methods for teaching students to think and reason about everyday events.

The present study investigates instructional variations that use the recent, promising pragmatic inferential rules postulated in the research on improving reasoning. The study anchors these rules in the tradition of scientific inquiry as a normative source of principles and tools for enhancing intellectual competence. The study, from a new perspective detailed more specifically in Chapter 2, revisits historical formalist beliefs about using abstract inferential rules to improve reasoning. Newell (1980) considers this old dichotomy of domain-independence versus domain-specificity as unresolved. He states that the issue concerns the
basic nature of problem solving and its teachability: "Our ultimate dichotomy involves not just the nature of general problem solving but whether it can be taught independently of specific subject matter. ... With respect to teaching, ... there seems to be little in the way of science about the issue." (Newell, 1980, pp. 187-188). This dissertation responds to Newell's further call for "experimental-theoretical forays" that "would do wonders for our understanding of this issue" (p. 188).

First, the study focuses on teaching students about statistical heuristics having to do with the law of large numbers. These heuristics are regarded as thinking tools for improving problem solving in everyday situations where uncertainty is prevalent. Second, the study investigates what typical alternative thinking strategies—competing heuristics—people use when solving such problems. The purpose of the second focus is to identify common error patterns that may then be amenable to instructional intervention. Third, the study explores how students' statistical reasoning and error commission relates to other variables. This final aim is descriptive.
General Research Questions

Three experiments were carried out with a total of 315 participants. The first and second experiments were conducted with university undergraduate students; the third involved elementary school students (grade 7) and secondary school students (grade 10). The primary questions addressed are as follows.

1. Will the problem solving ability of students increase after instruction in statistical inferential rules?

2. Will the problem solving ability of students increase when instruction about errors of reasoning is provided?

3. Does use of statistical inferential rules and commission of reasoning errors vary as a function of students' level of schooling, gender, achievement, previous coursework, knowledge about statistical rules, knowledge about domain areas in problems, and the domain of instruction?

Overview

Chapter 2 reviews theoretical literature on problem solving and inferential rules. It overviews studies about teaching problem solving based on statistical rules and concludes with specific hypotheses to be tested in the present research. Chapters 3 and 4 describe the purposes, methods, and results of Experiments 1 and 2 which were conducted with university students. Chapter 5 contains parallel material for Experiment 3 which was conducted with students in secondary and elementary schools. Chapter 6 contrasts grade 7, grade 10, and university results. Chapter 7 contains discussion and a synthesis of the results from the experiments. In addition, it offers conclusions, based on theory and previous research, and recommends directions for future research. Instructional materials and measures are found in the Appendices along with all procedural documents necessary for replication of the experiments.
Education is a matter of the mind, of thinking, and of intellectual development. One of the goals of education is to teach students to reason about problems they may encounter outside school. While many of the problems posed to students in school, such as those in mathematics and science, are well-defined—that is, they include all the necessary information and there exist methods which lead to a correct answer—many problems outside school are ill-defined and are presented with a limited amount of information. Frederiks (1984) suggests that the dichotomy of well- and ill-structured problems is too simple, and offers three categories that might better encompass the variations in problems: (1) "Well structured" problems are ones that are clearly formulated, with a known algorithm, and with criteria available for testing the accuracy of a solution; (2) "structured" problems are similar to well-structured problems but all or some aspect of the problem solving procedure must be generated by the problem solver; and, (3) "ill structured" problems lack clear formulation, a procedure that guarantees the correct solution, and criteria for evaluating solutions. If one purpose of education is to prepare students for life outside a school's perimeter, then it is important to inform instructional practice with research on teaching students to reason about problems that are not well structured.

This chapter overviews current theoretical and empirical literature on problem solving and on rules of inference useful in conceptualizing instruction in everyday problem solving. Concepts and empirical findings from the literature inform the questions addressed and the methodology employed in this study. The first part of the chapter reviews principles derived from the theoretical literature on problem solving. The second part of the chapter describes the nature of inferential rules with an emphasis on the recent theorizing of Holland, Holyoak,
Nisbett, and Thagard (1989). The next part of the chapter describes research studies where subjects engaged in solving everyday problems containing statistical information. The final section in the chapter addresses reasons why people may fail to consider statistical information in addressing everyday problems.

Problem Solving

Problem solving, defined as goal-directed activity in the face of obstacles to achieving a goal, is a well-established notion in cognitive psychology (Bransford, 1984; Rubinstein, 1975; Simon, 1978). Mayer (1983) contends that any definition of "problem" should include "the three ideas that (1) the problem is presently in some state, but (2) it is desired that it be in another state, and (3) there is no direct, obvious way to accomplish the change." (p. 5).

Simon (1978) defined problem solving in this way: "A human being is confronted with a problem when he has accepted a task but does not know how to carry it out ... accepting a task implies having some criterion he can apply to determine when the task has been successfully completed" (p. 272). Simon's theory contends that problem solving is the interaction between the information processing system and the "task environment." The latter can be described as an omniscient observer's way of describing the problem. The "problem space" is the way in which the problem solver views the task environment. More formally, it is a problem representation generated by the problem solver. The three concepts together, according to Simon—information processing system, task environment, and problem space—form the framework for problem solving behaviour.

Simon also postulates four "laws of qualitative structure" for human problem solving:

1. The information processing system is adaptive, capable of modifying its behaviour over time by learning.

2. The system will represent the task environment as a problem space and problem solving will take place in this space.
3. The structure of the task environment determines possible features of the problem space.

4. The structure of the problem space determines strategies that may be applied to problem solving.

Simon's laws underscore the importance of procedural knowledge and executive strategies commonly described in theories of human information processing (e.g., Anderson, 1983).

Problem Space

The concept of problem space and means that develop and modify the problem space are central features in the theory of Newell and Simon (1972). The functionality of these concepts can be viewed primarily in two ways.

First, this view implies the development of constraints that progressively restrict the problem space to include only valid solutions (Sacerdoti, 1977). One might visualize this activity as a bull's eye target in which regions defined by the concentric circles represent solution spaces. Each smaller space represents a progression toward the bull's eye, which is the solution itself. Narrowing the problem space involves selecting the right elements of information that are relevant to the problem at hand.

Since working memory is limited in its capacity to deal with information, chunking information helps the problem solver to retain and process more information in short term memory. The problem solver can profit from tools for conceptualizing information and chunking it, such as diagrams, charts, principles, and other conceptual tools. In this way, the problem space can become progressively more restricted toward a solution.

It should be noted that the problem solver may introduce unnecessary constraints, such as reasoning errors, into this kind of bull's eye targeting process. Rubinstein (1975) identifies categories of unnecessary constraints such as association constraints (where previously learned associations or functions may prevent the problem solver's processing system from
examining more effective or correct patterns for a solution) and world view constraints (where an individual's adopted world view makes it difficult for a problem solver to explore a number of frames of reference before becoming deeply committed to one).

Second, functioning within a problem space implies the application of a large and well-structured knowledge base in an increasingly effective and automatic manner. This requires organized, hierarchical knowledge structures. Procedural knowledge that becomes increasingly automatized is clearly an important feature of successful problem solving. Automatization of procedural knowledge can be viewed on a continuum rather than as an either automatic or controlled processing. Sternberg's (1986) notion of "effortful perception" connects automatization to reasoning:

Many processes will be automatized but only to some degree. The greater the degree of automatization, the less the degree of reasoning involved in process execution ... the extent to which a task measures reasoning is a function of the interaction between the person and the task, rather than merely a function of the task ... reasoning is, in part, effortful perception. (p. 287)

Discussions of the interaction between the person and the problem solving task that are similar to Sternberg's notion of effortful perception are found in Eysenck (1984) and Sinnott (1975).

Problem Representation

One of the major features of Simon's theory is that the relative ease of solving a problem depends on how successful the problem solver has been in representing critical features of the task environment in his problem space. Findings from think-aloud protocols of problem solving activity show that people typically do not search for the most efficient representation of a problem, to make solving easier, but often quickly adopt an obvious or straightforward representation.
Gagné (1985) presents another view of what is involved in solving problems. She defines problem solving as students achieving goals for which they do not have an automatic solution. This supplements Simon's definition of problem solving and complements Sternberg's concept of a continuum of automatic and controlled processing. It also fits Lesgold's (1988) contention that the essence of problem solving is being able to deal with new situations for which one has not been specifically trained. Gagné contends that the problem solver forms propositions and images in short term memory, producing representations and assigning values to variables in the problem representation process. The representation then activates related knowledge, and the activated knowledge is applied to the new situation. This combination of activation and application is referred to as transfer. The notion of transfer, that is, the question of what facilitates application of knowledge learned in one situation to a somewhat different situation, is "at the heart of successful problem solving" (Gagné, 1985, p. 151).

Problem solvers typically tend to overestimate the extent to which they have exhausted an issue because they access only a fraction of the knowledge they have that is relevant to a given problem. They may, in fact, possess a great deal of "inert knowledge" (Bereiter & Scardamalia; 1985) for solving the problem. Gagné sees problem representation as critical to accessing inert knowledge. Problem representation determines what knowledge will be activated in long term memory; and, long term memory is a key to solving ill-defined problems. As a result, transfer is further facilitated by the availability of additional, potentially useful knowledge in new problem situations for which one has not been specifically trained.

In addition, problem representation can be described as an understanding process with two subprocesses (Simon, 1978): The first is an interpreting process; the second is a constructing process. The interpreting process interprets the language of the problem. It "...
reads the sentences of the problem text and extracts information from them, guided by a set of
information extraction rules" (p. 284). The constructing process builds a representation of the
problem space by (a) a situation description—this includes problem elements, relations
among elements, initial states, and goal states; and (b) a set of operators—a production
system where conditions are represented as states of the situation and actions are represented
as processes for making changes in the situation. It is the responsibility of the construction
process to make sure that the subprocesses of describing the situation and of applying
operators are compatible.

As noted previously, there may be a need to delay this constructing process, to
interpret more deeply than the first impulse for expeditious representation might invite. This
need for delay is supported by studies of the Tower of Hanoi problem (Simon, 1978) and by
expert novice studies of problem solving such as those conducted by Chi, Glaser, and Rees
(1982). They found that experts, at least those working in well-defined domains, carried out
more initial qualitative analysis of a problem that led to substantial differences between
experts' and novices' qualities of representation. Novices' node-link networks reflected the
surface structure and were object-oriented while experts' networks were schema-oriented.
Experts' schema-orientation led them to spend time looking at problems in ways that triggered
deeper principles leading to solutions.

Nickerson (1986) characterizes novices' reluctance to spend the initial time in
analyzing problems as a common type of reasoning fallacy. Included among Nickerson's
examples of reasoning fallacies of this type are uncritical acceptance of simple explanations,
hasty closure, and inappropriate persistence. Such typical reasoning errors may provide
another source of understanding about improving learner's representation skills. Students
who are alerted to these and other reasoning errors may improve their problem
representations. According to Nickerson, reasoning fallacies prevent people from reasoning effectively by providing ways in which people "succeed in being unreasonable" (p. 111).

Another way to explain expert novice differences in problem representation is in terms of declarative and procedural knowledge. The expert has potential procedures or patterns of solutions and represents the problem in terms of potential procedures. "It is the procedural knowledge that will ultimately determine how well a person can solve a problem" (Chi, Glaser, & Rees, 1982, p. 54). Perkins, Allen, and Hafner (1983) referred to skilled reasoners as having a "critical epistemology" as opposed to a novice reasoner's "makes-sense epistemology". To successfully reason about problems, the former epistemologist requires (a) a large repertoire of knowledge, (b) efficient knowledge evocation, and (c) deliberative effort to interrogate the knowledge base in order to construct arguments. The latter epistemologist may not have one or more of these requirements. Perkins et al. refer to the makes-sense epistemologist's typical shunning of complicated cognitive activity as a "default condition" while a skilled reasoner may activate a formal repertoire of knowledge that contains "knowledge of logical and heuristic forms that can be applied to the practical business of reasoning" (p. 188). This formal repertoire should include elementary statistical knowledge about sufficient sample sizes. Statistical knowledge, the authors contend, "can help guard against unwarranted inferences and may sanction warranted ones, and such knowledge is applicable in many informal contexts" (p. 188).

Inferential Rules

The major differences between Simon's theory of problem solving and that of Holland et al. (1989) arise from assumptions about how information processing occurs during problem solving. Holland et al. agree with Simon that (a) problems are defined as initial and goal states; (b) problem solving is a search through state space; (c) there is need for a set of
operators to transform initial states to goal states; (d) constraints are required to reach acceptable solutions; (e) some problem solving methods are general and some are specific; and, (f) general methods are not sufficient to account for expert problem solving skill (i.e., human expertise is dependent on specialized methods and knowledge about the relevant domain). Holland et al. state that "at a global level ... problem solving appears to be a process based on domain-specific knowledge, if that is available, or reliance on general methods if it is not" (p. 11). Notably, in the case of ill-defined problems in everyday contexts, domain specific knowledge is not always available.

Holland et al. argue with Simon's model by stating that his model is inadequate for ill-defined problems. Ill-defined problems are characterized by uncertainty: The initial state, the goal state, the required operators, and the applicable constraints may not be clearly defined at the outset of problem solving.

**Syntactic versus Pragmatic Rules**

Holland et al. believe that people construct models of situations, thereby creating a problem space. These models are then "mentally run" to produce expectations about the environment and to make predictions about potential solutions. The most basic building block of these mental models is the "if-then", condition-action rule. Sets of condition-action rules provide structures that help to run mental models. These rules are not "imperatives" in the natural language sense, but rather guides. How these guides, or suggestions, compete and interact with other guiding rules is what is important. In other words, inferential rules create a network of interacting, competing, not necessarily consistent hypotheses. Rule systems can create new ideas from combinations of rules and disconfirm old predictions. Holland et al.'s theory of rule systems augments problem solving theory advanced by Simon by postulating processing mechanisms for seeking additional, inert knowledge stored in memory that may
help represent and clarify ill-defined problems. Such systems can represent environments where there is novelty, like the contexts of everyday problems. The mental models described by Holland et al. do not rely on what the authors regard as more inflexible structures, such as schemata.

An important distinction is made by Holland et al.'s theory between inductive inferential rules, which are pragmatic, and rules of formal logic, which are syntactic. They describe this difference as follows. "Though people possess inductive inferential rules for statistical and causal analysis at a high level of abstraction, they possess few, if any, abstract rules corresponding to those of formal logic" (p. 7). Such inductive inferential rules for statistical and causal analysis are based on how induction in specific realistic contexts becomes constrained. According to Holland et al., the general nature of the human information processing system is characterized by a pursuit of goals in a complex environment and the utilization of feedback about the results of that pursuit. While the processing constraints that help the system reach goals are specified by current knowledge in the system, the current knowledge also becomes modified through use. And, because particular situations activate particular prior knowledge, induction is highly context dependent. Deduction, on the other hand, is distinguished from induction by the fact that the truth of premises may guarantee the truth of inferences based on the premises. While deductions may be formally valid, they may be useless in real-world contexts. Holland et al. contend that most research on induction in psychology has been inappropriately dominated by formal syntactic approaches arising from the tenets of logical positivism:

Because of its emphasis on the role of the system's goals and the context in which induction takes place, we characterize the theory proposed here as pragmatic. In contrast, most treatments of the topic have looked at purely syntactic structure of the knowledge to be expanded and leaving the pragmatic aspects, those concerned with goals and problem-solving contexts, to look out for themselves" (p. 5).
Empirical Rules versus Inferential Rules

Holland et al. believe that statistical reasoning involves the use of one of four types of pragmatic inferential rules which have the function of generating other rules that generate a more immediate representation of a problem. In describing pragmatic inferential rules, Holland et al. differentiate between "empirical rules" which act to model the world and "inferential rules" which generate better empirical rules. Empirical rules describe the environment and its likely next states. They are the "bread and butter performance rules of the system" (p. 41). Inferential rules are seen as "providing relatively domain-independent procedures for altering the general knowledge base" (p. 41). Inferential rules, therefore, have the function of generating empirical rules and adding these to the permanent base of general procedural knowledge. While the function of empirical rules is to model the world, the function of inferential rules is to produce more accurate empirical rules. In this theoretical scheme, inferential rules are "relatively independent of specific domain of events ... such rules can sometimes be applied to domain of events for which the person has never used them before" (p. 41).

There are four types of inferential rules according to Holland et al.:

Specialization Rules  If a prediction based on a strong rule fails, then create a more specialized rule that includes a novel property associated with the failure in its condition and the observed unexpected outcome as its action.

Unusualness Rules  If a situation has an unexpected property, then be prepared to use that property in the condition of a new rule if rule generation is triggered in close proximity to the occurrence of the situation.

Law of Large Numbers Heuristics  If S is a sample providing an estimate of the distribution of property P over some population, then create a rule stating that the entire population has that distribution, with the strength of the rule varying with the size of S.

Regulation Schemas  If you have the rule "If you want to do X, then you must first do Y," then create the rule "If you do not do Y, then you cannot do X." (p. 43)
What makes the third type of inferential rule, Law of Large Numbers Heuristics, so receptive to instructional intervention is the recent research establishing that many aspects of inductive reasoning are statistical at base. Students appear to already possess statistical rules in highly abstract form and can apply these rules to ordinary events. It seems that learner's errors in statistical reasoning are better understood as an inability to code events in terms that make contact with statistical rules rather than as ignorance of the statistical rules themselves. In fact, Holland et al. state that the strongest evidence for the highly abstract existence of these rules is "the fact that the teaching of statistical rules in an abstract way, as is normally done in statistics courses, for example, has an effect on the way people reason about problems in everyday life" (p. 256).

Implications for Instruction

The theory about inferential rules outlined in this chapter suggests implications for injecting new rules into human information processing systems in school settings. Essentially, the type of inferential rule Holland et al. call Law of Large Numbers Heuristics is part of the formal array of scientific reasoning tools employed by statisticians and scientists. Chapter 1 advanced a normative premise for teaching reasoning and improving human intellectual competence through principles and tools for thinking arising from the scientific tradition of inquiry. The formal array of scientific reasoning tools is embedded in that scientific tradition and, in large measure, is already part of school curricula. However, the present view that such tools may also include inferential rules that are relatively independent of specific domains of knowledge and are applicable in everyday contexts is not a commonly held view in school settings. The promise of teaching abstract inferential rules to improve problem solving is highlighted by Holland et al. as follows:
We have spoken about systems as if they had to go it alone, learning everything by the sweat of their brows or the hum of their central processors. Fortunately, that is not the case. New rules can be suggested to systems by external agents such as teachers and programmers... The most important thing to say about inserted rules is that they will tend to act in some important aspects like induced rules. That is to say, they will enter into competition for the right to represent the environment along with the rest of the rules in the system's possession. (p. 96).

Since statistical reasoning involves abstract, domain-independent inferential rules, it may be possible to improve students' abilities to reason about a wide range of events by formally teaching the abstract statistical rules themselves. This is, in some ways, made easier because students may already possess intuitive versions of some statistical rules that they have induced by reasoning about problems where uncertainty is prevalent. They may have acquired the means for encoding information in problems in terms of statistical rules. However, learners' abilities to encode and apply such abstract rules vary. Application can be improved by instruction in encoding the information contained in problems in appropriate ways so that it makes contact with the rules.

There are two main issues to be considered in planning instruction. First, newly-introduced rules compete with old rules. This issue, highly relevant to instruction, is often missed by teachers who teach as if students were tabula rasa. According to Holland et al.:

*education should not be thought of as replacing the rules that people use for understanding the world but rather as introducing new rules that enter into competition with the old ones. People reliably distort the new rules in the direction of the old ones, or ignore them altogether except in the highly specific domains in which they were taught.* (p. 206)

Therefore, it is not appropriate to assume that any learner's mind is a blank slate.

Instructional techniques must take students' existing beliefs (Clement, 1982) and prior knowledge, including related heuristics, into account. For example, students commonly learn one set of rules for school science and another for the real world (West & Pines, 1985).
Second, the newly acquired rules will not compete successfully with older induced rules if the application requires encoding that the system has not yet mastered. Knowledge about how students of various ages and capabilities encode statistical information in problems from various domains is useful information for instructional planning. Nisbett et al. (1983) have argued that people call on statistical heuristics in addressing everyday problems according to (a) the degree to which features of a problem can be encoded in terms of statistical rules (i.e., when sample space and sampling processes are clear, when the events can be coded in common units); (b) the presence of chance factors or random variation in the problem; and (c) the presence of events in the problem that are "culturally" associated with random variations (e.g., gambling). Based on this analysis, three problem "domains" of statistical reasoning were identified by Fong et al.: (a) Probabilistic problems are those in which random variation in the sample data was made clear in the problem. (b) Objective problems are those that describe sample data in some objective metric (e.g., pounds as opposed to “fat”), but there is no explicit cue about randomness of the sample. And, (c) Subjective problems are those in which data describing a sample is not reliably measurable (e.g., “funny”). Fong et al. argue that these three are domains of everyday problems where statistical heuristics apply.

The two instructionally significant issues—newly-introduced rules compete with old rules and newly acquired rules will not compete successfully with older induced rules if the application requires encoding that the system has not yet mastered—need to be considered in teaching abstract inferential rules such as statistical heuristics. If students readily use statistical rules only in a domain where they can easily encode the variability of events such as the Probabilistic domain, they need to be shown how to encode events in terms that make contact with the rules when events are not so easily encoded. This might be done with examples from the Objective and Subjective domains. Second, if students have strong prior
intuitions—pre-established, erroneous statistical or other heuristics that they apply to problems where uncertainty is prevalent—instructional approaches that incorporate demonstrations may overcome and defeat those prior intuitions (see Siegler, 1983; McCloskey & Kaiser, 1984). Holland et al. are specific in their suggestions for the types of demonstrations that teachers should undertake to defeat pre-existing errors in reasoning:

What is needed is demonstrations where students are required to make predictions about the likely outcomes of events, then to articulate the bases of their predictions, then to explore the possibility that other rules might exist that would do a better job of prediction. (p. 228)

Teachers' knowledge of the other, competing heuristics or typical reasoning errors and intuitions is critical. In addition teachers need to become informed about common "default conditions" (Perkins et al.) that are often triggered in the cognitions of makes-sense epistemologists who encounter problem solving activities.

Knowledge of students' existing rules and misconceptions is encompassed in a recent model of teaching called the cognitive-mediational model where student cognitions are seen as the cause for learning. Winne and Marx (1987) refer to two models of teaching that correspond to process-product and cognitive-mediational paradigms. The first of these is the performance-based model of teaching effectiveness. It was predicated on the view that a simple cause-and-effect chain exists between what teachers do and how well students learn. It meant that teaching could be analyzed as a set of discrete skills or behaviours. Specific skills/behaviours used by teachers would increase student learning, and a linking together of skills/behaviours would result in effective teaching. The authors argue that the most critical problem with the performance-based model is that "it does not include a formal description of how students actually learn from teaching" (Winne & Marx, 1987, p. 269). The missing causal link between teaching and learning is addressed by the second of the two models, the cognitive mediational model. It adds two dimensions to the performance-based model. One
is a cognitive dimension focusing on the kinds of cognitions which students use in trying to learn. The other is a mediational dimension focusing on the need for students to interpret teacher behaviours in the ambiguous context of classrooms. The cognitive-mediational model shifts the cause-effect relationship for learning from teacher behaviours as causes to students' cognitions as causes. It follows that knowing how learners reason about problems becomes prerequisite understanding for improving such reasoning.

Studies of Statistical Reasoning

The following eight studies represent the corpus of research on how the inferential rule called the law of large numbers heuristics is applied in everyday problem solving.

How Do Learners Encode Statistical Information?

To investigate whether people normally reason more statistically about objective, easy-to-code events than about subjective, hard-to-code events, Jepson et al. (1983) tested subjects with two types of problems. In the first type of problem, the behaviour to be considered by the problem solver was an achievement or some other outcome, such as number of friends, that could be assessed by objective means. It was hypothesized that the objective nature of the information in the problem would allow easier identification of statistical properties such as variability and randomness. In the second type of problem, the behaviour to be considered was only subjectively assessable, for example, friendliness. Subjects were observed to be much more likely to use the law of large numbers and the regression principle with the first type of problem. The investigators concluded that inductive reasoning will be guided by statistical inferential rules when people are solving problems in domains where events are readily evaluated with respect to their variability and with respect to the randomness of the
occurrence of the events. When people solve problems in domains where variability and randomness are difficult features to detect, statistical principles will be used less often.

**Does Domain-Specific Knowledge Increase Statistical Reasoning?**

To determine if experts (with domain-specific expertise) reason more statistically about problems in their area of expertise, Nisbett *et al.* (1983) presented subjects with problems from two domains, sports and acting. Subjects' level of expertise was based on self-reports of experience in those areas. The researchers believed that experts would have greater awareness of the variability of events and the role of chance in producing events in their area of expertise and would avoid over-generalizing from a small sample.

Nisbett *et al.* found that expertise had a strong positive correlation with statistical reasoning: 56% of experts in sports versus 35% non-experts preferred statistical explanations for problems; 59% of experts in acting versus 29% non-experts preferred statistical explanations. The investigators concluded from these results that expertise in a content domain can influence statistical reasoning.

**Do Learners Use Abstract Inferential Rules and Does Training Help?**

In a series of four studies, Fong *et al.* (1986) sought to establish support for the view that people possess an abstract inferential rule system that acts like an intuitive version of the Law of Large Numbers and that this rule system can be improved by formal instruction. The researchers argued for a formalist theory of reasoning: learners reason using very abstract rules and, because the rule system is not dependent on any particular content domain, formal instruction could improve the system. In addition, it was hypothesized that subjects would be able to apply such rules over a wide range of problems from differing content areas. It was argued that the formalist position would be weakened if instruction did not assist people in
solving problems, especially if subjects had actually learned the rules. The formalist position would be undermined further if subjects improved in only the more obvious, probabilistic types of problems.

In Experiment 1, Fong et al. worked with 347 subjects—229 adults and 118 high school students. The adults varied widely in age and education and almost all were females who had participated previously in psychology experiments at Bell Laboratories. The subjects were paid to participate and were instructed in groups of 4 to 6 with training condition determined randomly. The five conditions were: (a) rule training, which included reading materials and a demonstration about the law of large numbers; (b) examples training, which included examples of problems and analyses of problems using concepts from the law of large numbers; (c) full training, which included both of the foregoing; (d) demand training, which included a one-sentence statement about the law of large numbers; and (e) control, which involved no training.

Adults and high school students showed the same pattern of results and, therefore, were combined in the analysis presented by the authors. The results of Experiment 1 show that both quality and frequency of statistical reasoning was enhanced for all problems ranging from more obviously probabilistic ones to more difficult-to-assess subjective ones. According to the researchers, a three-level ordering of the conditions resulted. At the lowest level, the control and demand conditions were least likely to employ statistical reasoning (42% and 44%, respectively). At the highest level, subjects in the full training condition—rule and examples training combined—were most likely to use statistical reasoning (64%). In addition, the frequency of probabilistic answers by training condition and problem type—labelled probabilistic, objective, and subjective—were examined. No interaction between training and problem type was found. Strong support for the formalist position is claimed.
Because of an absence of interaction between training and problem type in the results of the first experiment, training effects were judged to be domain-independent. This result was believed to be very important because of the typically strong domain-specificity displayed in subjects' problem solving when they engaged in statistical reasoning without training. This strong domain specificity was similar to the results of Jepson et al.'s (1983) study. As a result of the findings from Experiment 1, the second experiment by Fong et al. was designed to explore further the lack of domain-specificity in training effects. Subjects were 166 University of Michigan undergraduates enrolled in introductory educational psychology classes. Training was conducted in small groups assigned to one of three training groups or a control group. The three training groups received the full training in rules plus examples described in Experiment 1. However, only one type of problem and analysis—probabilistic, objective, or subjective—was studied in each condition. Subjects were tested immediately after training.

As in the first experiment, it was found that training in the law of large numbers enhanced the frequency of statistical responses when compared to the control group. Subjects in the control conditions responded with 53% statistical responses; subjects in the training groups had 72%, 81%, and 79% statistical responses in the probabilistic, objective, and subjective training groups, respectively. These percentages do not account for quality of statistical responses. No interaction between domain of example problems and domain of test problems was found. Again, this finding was taken to support the formalist position.

In Experiment 3, Fong et al. examined the effect of differing amounts of formal statistical training on the way subjects reasoned about two different versions of a problem from everyday life. One version of the problem included a randomness cue while the other version had no randomness cue. Subjects were 42 undergraduates with no college level statistics courses; 56 undergraduates with an introductory level statistics course; 72 graduate
students in psychology who were attending the first session of a course in statistical methods and who had taken one or more statistics courses; and, 33, mostly Ph.D.-level scientists with many statistics courses who were attending a colloquium on probabilistic reasoning.

The researchers found that almost none of the subjects without statistics courses used statistical reasoning when presented the problem without the randomness cue. In contrast, 80% of the Ph.D.-level scientists used statistical reasoning in this instance. For the problem containing a randomness cue, 50% of subjects without statistics courses used statistical reasoning. The overall data suggests that a broad range of statistical expertise correlates positively with increased statistical reasoning about everyday problems. Subjects with several years of training in statistics displayed greater ability to analyze problems with reference to the law of large numbers than subjects without or with little formal statistical training.

Fong, Krantz, and Nisbett's fourth experiment continued to examine effects of formal statistical training on reasoning about everyday life. Sources of confounding identified in Experiment 3 were removed by incorporating a testing context where subjects would not be cued into using statistical rules. In this experiment, the researchers examined the effects of formal statistical training in a setting completely outside the context of training. Subjects were 193 randomly selected male students enrolled in an introductory statistics course. A telephone survey, unconnected to the course, was conducted to survey students' opinions of sports. Some of the questions in the survey could be analyzed with reference to statistical concepts. One half of the subjects were randomly selected to be contacted and surveyed during the first week of the semester. The other half of the subjects were surveyed during the last week of the semester. Of those contacted at the beginning of the semester, 16% gave statistical answers to a regression question about a major league baseball player who wins Rookie of the Year but does not perform as well in the subsequent year. Of those interviewed at the end of the term, 37% responded with statistical thinking. The results suggest that statistical
education provided in the introductory-level statistics course enhanced the use of statistical rules in reasoning about everyday life. Similar results were obtained on a second question. Two other questions did not elicit enhanced statistical answers, and the researchers report having no explanation for this.

To summarize their four experiments, Fong et al. concluded:

These studies suggest very strongly that people make use of abstract inferential rules in the form of statistical heuristics. We also know this because training on the purely formal aspects of the law of large numbers improves statistical thinking over a broad range of content, and because showing subjects how to use the rule in a given content domain generalizes completely to quite different content domains. We are aware of no more convincing evidence, in fact, for the existence of abstract rules of reasoning than the present work.

What is the origin of abstract inferential rules about the law of large numbers? Why do people develop such high-level representations of the law of large numbers? We suspect the answer comes, in large measure, from the ubiquity of the principle. The basic notion that large samples are more reliable than small samples underlies concept formation and generalization. It can be argued that during cognitive development, the child learns, through repeated exposure to the law of large numbers across many domains, a highly abstract representation of the principle. (p. 282)

Does Professional Training Affect Statistical Reasoning?

In a cross-sectional study by Lehman, Lempert, and Nisbett (1988), it was hypothesized that training in probabilistic sciences (psychology and medicine) would affect statistical reasoning by sensitizing people to errors such as making inferences from small samples. University graduate-level subjects in two probabilistic sciences (psychology and medicine), in a nonprobabilistic or deterministic science (chemistry), and in a non-science (law) were tested for statistical reasoning. The test items measured subjects' abilities to apply the law of large numbers and regression principles to both everyday life problems and scientific problems. The results were "dramatic", to use the researchers' description. Psychology students changed in statistical reasoning from first to third year. This result,
using Rosenthal's r-index, translates into an effect size of .56. Medical students, showed no significant improvement in statistical reasoning, according to the authors. The disciplines where formal statistics were taught appeared to produce increases in students' statistical reasoning about problems from everyday life.

A replication of this cross-sectional study was conducted by the same authors at a different university, this time involving only psychology and chemistry graduate students. Twenty-seven first-year and 27 third-year psychology students plus 35 first-year and 28 third-year chemistry students were the subjects. After preliminary analyses revealed smaller training effects than those for the previous study, the psychology students were further subdivided into two groups, social science psychologists and natural science psychologists. The subjects who were natural science psychologists—students in physiological learning and behaviour and students in experimental psychology—showed little change in reasoning scores. Chemistry students showed no increase. However, statistical reasoning in the social science psychologists—students in fields of personality, social, developmental, and clinical psychology—amounted to a 59% improvement over two years of study. This represents an effect size of .86, again applying the r-index.

Are Trained Subjects Using Analogy or Statistical Laws?

Fong and Nisbett (1990) were concerned that the "domain independence" suggested by their earlier experiments might be due to subjects' ability to work by analogy from example problems. As a result they conducted an additional study with 231 University of Michigan undergraduates. Subjects were trained in a given content domain (sports or ability testing) and then tested either in that domain or in the other domain immediately after training or after a two week delay. The results showed strong domain independence when testing was immediate. After two weeks there was no decrease in statistical reasoning in the trained
domain. However, there was significant decrease in the untrained domain, although subjects still performed better than control group subjects. Using measures of memory, the researchers found that maintenance of training effects was due to memory for the rule system rather than memory for content of the example problems. This was felt to be contrary to what would be expected if subjects employed direct analogies and, again, was argued to be supportive of a strong formalist position.

These eight studies provide evidence for the existence of statistical inferential rules among adult and young adult problem solvers. The evidence suggests that the rules are abstract, apparently lacking in domain-specificity, and can be transferred through formal instruction.

Recently, two projects have undertaken research on statistical reasoning in children. The first, entitled STARC, at the University of Pennsylvania, has begun to examine if children who acquire statistical skill through the context of classroom instruction in mathematics actually use the skill outside the immediate classroom environment. The researchers working in this project (Gal, Rothschild, & Wagner, 1990) report a dearth of studies in the area. They interviewed 122 children in grades 3, 6, and 9 from three middle-class private schools to elicit the children's knowledge of averages and their usage. They found that while the subjects were, in many instances, able to compute averages, they did not necessarily understand the meaning or applications of the concept in everyday contexts. The second of the two projects, called ELASTIC, sponsored by the U.S National Science Foundation, has begun to address why students generally find basic concepts of sampling and statistical inference so difficult to grasp. Rubin, Bruce, and Tenney (1990) report a study exploring the underlying conceptions and heuristics students bring to the study of statistics. Their findings, after interviewing 12 senior high school students, include errors in students' understanding of the relationship between samples and populations as well as fluctuating
usage of sample variability and sample representativeness notions. The work initiated in projects such as STARC and ELASTIC provide additional perspectives in synthesizing the results of the three experiments reported in this dissertation.

Failure to Consider Statistical Information

How might learners reason when they do not pay attention to statistical information in a problem? One simple judgmental heuristic that people seem to rely on when solving problems is termed the representativeness heuristic (Kahneman, Slovic, & Tversky, 1982). The general premise of the representativeness heuristic is the idea that "the perceiver is an active interpreter, one who resolves ambiguities, makes educated guesses about events that cannot be observed directly, and forms inferences about associations and causal relations" (Nisbett & Ross, 1980, p. 17). In making a judgment, this heuristic involves applying resemblance criteria to categorization. The problem solver assesses the degree to which salient features of the problem are representative of or similar to other problems he has encountered.

There are a number of ways that a problem solver can be misled by applying the representativeness heuristic. Detailed examples of misapplications of judgmental heuristics are found in Nisbett and Ross (1980). Despite this propensity to err, the value of the heuristic may be in its explanatory power. It may provide a robust explanation of why people process probabilistic information much less effectively than the ideal probabilistic reasoner exemplified by Bayes theorem. To illustrate how the representativeness heuristic makes people insensitive to statistical information such as sample size, consider the results of a study conducted by Tversky and Kahneman (1971). Subjects were informed that on a given day 60% of the babies born in a hospital were boys although 50% was the expected result. They were asked if the 60% result was more likely in a large hospital were 50 babies were born per
day or in a small hospital where 15 babies were born per day. Subjects should have indicated, that based on the small sample size, the small hospital was more likely to produce the unusual result. However, they did not favour either answer. Both hospitals were judged to be equally representative of the result. Subjects in this study appeared insensitive to statistical information.

One form of the representativeness heuristic is the "fundamental attribution error" (Nisbett & Ross, 1980; Ross & Anderson, 1980). This error is described as a human tendency to attribute behaviour to an individual's dispositions and to ignore strong situational determinants of behaviour such as statistical base rate information. Nisbett and Ross (1980) characterize the fundamental attribution error in this way: "...in large measure the error, we suspect, lies in a very broad proposition about human conduct, to wit, that people behave as they do because of a general disposition to behave in the way that they do" (p. 31). The authors regard such theories of the intuitive scientist as major contributors to the meaning people extract from social situations. Since everyday problems often involve social interactions and behaviour, theories that people use when they act as intuitive scientists in judging causes of human behaviour are an important source of inferential rules that learners might have in solving such problems.

The strength of these theories of the intuitive scientist are supported by the research showing how people fail to consider statistical information in problems and instead represent problems according to the representativeness heuristic. However, there is very little research on how learners typically represent problems when they ignore or underuse base rate information available in the problem. For example, is the fundamental attribution error a typical nonstatistical response to everyday problems containing statistical information? If students do not use law of large numbers heuristics, for example, then what other, possibly
erroneous, heuristics do they use that may compete with more accurate representations in inferential rule systems?

Summary

Emerging from the literature on problem solving, inferential rules, and statistical heuristics is a model to guide research questions and methodology in the present series of experiments. Problem solvers engage in goal-directed activity and work toward representing problems as accurately as they can. Constraints that progressively restrict the problem space may also be unnecessary or misrepresentative. Yet, how a problem is represented is critical because it determines what knowledge will be activated. Statistical reasoning is dependent on domain specific knowledge that allows problem solvers to assess the expected variability and randomness of events in the problem in an accurate way based on their understanding of the likely variability and randomness of the particular events in question. That domain knowledge may remain inert if competing rules and heuristics overcome potentially useful statistical rules. In this way, abstract, pragmatic heuristics might be regarded as intermediate inferential rules that inform domain-specific procedural knowledge. While general problem solving strategies are weak methods, these intermediate rules may be directly improvable so that students can become more accurate in their problem solving skills.

There is still a paucity of research in how such rule systems may operate in human problem solving. More research that explores the theory of pragmatic inferential rules and investigates the theory's explanatory potential for instruction in everyday problem solving is needed. The research conducted to date has: (a) concentrated on adult and young adult learners; (b) been laboratory, not classroom, oriented; (c) not examined competing rules or heuristics or how instruction might reduce the frequency with which common reasoning errors invade reasoning in these situations; (d) not advanced convincing evidence about
domain-independence; and, (e) not investigated interactions with many other variables that may inform instruction such as gender, general reasoning ability, prior statistical reasoning ability, and achievement. For example, none of the previous studies describe measures of learning the law of large numbers concepts taken after training to address how learning predicts statistical reasoning in everyday problems. More particularly, is declarative knowledge about statistical laws a predictor of statistical reasoning? These and other issues have not been explored.

Research Questions

1. Will students receiving instruction in statistical inferential rules make more use of such rules in reasoning about everyday events when compared with students not receiving instruction? (eg. Fong et al)

2. Will students receiving instruction in statistical inferential rules make fewer errors in reasoning about everyday events when compared with students not receiving instruction?

3. Does the use of statistical inferential rules vary as a function of the problem domain emphasized during instruction? (eg. Fong et al)

4. Does the use of statistical inferential rules vary as a function of students' achievement, coursework (eg. Lehman et al), gender, general reasoning ability, statistical knowledge, knowledge about reasoning errors, and domain knowledge (eg. Nisbett et al. 1983)?

5. Will students receiving both instruction in statistical inferential rules and instruction in reasoning errors make more and better use of such rules and commit fewer reasoning errors when compared with students receiving only instruction in statistical inferential rules?

6. What developmental trends in statistical reasoning exist among elementary, secondary, and university-aged students in the absence of instruction as well as after training, and how do these groups differ in statistical reasoning and reasoning errors?
CHAPTER 3

EXPERIMENT 1

The ubiquity of probabilistic thinking is underscored by Halpern's (1984) observation that probability plays a crucial role in almost all aspects of our lives and that good thinking will require an understanding of probability, as well as by Kahneman and Tversky's (1972) contention that "the decisions we make, the conclusions we reach, and the explanations we offer are usually based on our judgments of the likelihood of uncertain events" (p. 430). The fallibility of probabilistic thinking is underscored by Nickerson, et al. (1985) who state that, while people are constantly in the business of generating and evaluating inductive hypotheses, people are very likely to manifest errors and biases in everyday inductive situations because they: (a) fail to sample in an unbiased way; (b) are conservative in their use of probabilistic information; or, (c) fail to consider abstract statistical information. The notion that instruction in statistical laws might improve this situation for people of varying ages and abilities is investigated in the three experiments of this study.

The first experiment in the present study replicated, extended, and redesigned Fong et al.'s (1986) Experiment 2. The purpose of that earlier work was to explore whether formal statistical training about what the researchers termed the law of large numbers—basic information about population and sampling—can be used to teach people to reason more statistically about everyday inferential problems and whether training effects vary as a function of training domain. Their intent was to investigate domain-specific effects of training. The view supporting domain-specificity in problem solving suggests that subjects would show more improvement for problems in the domain where they were instructed than for other problems. The formalist view supporting abstract inferential rules in problem solving suggests that there would be no interaction between instructional domain and testing
domain. Since Fong, et al. found no reliable domain-specificity in their study, they argued for the formalist position.

The present investigation departs from the previous experiment in five main ways. First, Fong et al. trained and tested subjects in a single session. In response to their recommendation to separate the time and context of training from the time and context of testing, to reduce "the artifactual possibility of salience" (p.276), delayed main effects plus delayed transfer effects of statistical instruction were explored by separating instructional and testing sessions by a week. Second, the directions given to the control group were revised. Fong et al.'s directions referred to the law of large numbers. If subjects are unfamiliar with this label, but know very well what the concept of this principle is, they may underperform. A brief explanatory phrase was substituted for the label to redress this problem. Third, the training and testing procedures were modified slightly. The training procedures involved two subject volunteers to assist with a demonstration, and the testing procedures were implemented in a 45 min period of time with subjects receiving encouragement to maintain brevity in their written answers due to the time constraint. Fourth, interactions of statistical reasoning with achievement, previous coursework, and undergraduate major were explored. The small number of male subjects precluded gender analysis. And, fifth, the data was analyzed for error patterns in statistical reasoning that may be typical. These patterns can guide the planning of training programs in judgment and decision making as well as suggest foci for additional research.

In addition to these five modifications of the research, the present study changes important terminology used in the previous experiment. Fong et al.'s reference to "domains" of problem solving is revised to "formats" of problem solving. Since Fong et al's use of "domain" is quite different from the commonly understood notion of content domain found, for example, in expert-novice studies, this difference needs to be acknowledged.
Method

Sample

Participants in this research were students (N = 105) in an undergraduate course in educational psychology, 86 females and 19 males. They were assigned to one of four conditions: instruction based on one of the three formats (see Treatments) and a control group receiving no instruction. Group sizes for instruction in the course's tutorials ranged from eight to 12 people. Of the 105 subjects, 15 were omitted from the data used in statistical analysis due to incomplete protocols.

Subjects could not be randomly assigned to the four conditions because instruction was provided to intact tutorial groups in the educational psychology course. Timetables and lack of space precluded the luxury of random assignment. Since subjects could not be randomly assigned to the four conditions, equivalency of the condition groups was examined by using subjects' final marks in the educational psychology course as a proxy for achievement. One way ANOVA's indicated that none of the 4 condition groups and none of the 10 individual tutorial groups were significantly different at the .05 level with respect to achievement.

Treatments and Controls

As described in Chapter 2, Nisbett et al. (1983) have argued that people call on statistical heuristics in addressing everyday problems according to: (a) the degree to which features of a problem can be encoded in terms of statistical rules (i.e., when sample space and sampling processes are clear and when the events can be coded in common units); (b) the presence of chance factors or random variation in the problem; and (c) the presence of the events in the problem that are "culturally" associated with random variation (e.g., gambling). Based on this analysis, three problem formats of statistical reasoning were identified by Fong et al.: (a)
**Probabilistic** problems are those in which random variation in the sample data was made clear in the problem. An example of this type of problem follows.

Bert H. has a job checking the results of an X-ray scanner of pipeline welds in a pipe factory. Overall, the X-ray scanner shows that the welding machine makes a perfect weld about 80% of the time. Of 900 welds each day, usually about 680 to 740 welds are perfect. Bert has noticed that on some days, all of the first 10 welds were perfect. However, Bert has also noticed that on such days, the overall number of perfect welds is usually not much better for the day as a whole than on days when the first 10 welds show some imperfections.

Why do you suppose the number of perfect welds is usually not much better on days where the first batch of welds was perfect than on other days?

(b) **Objective** problems are those that describe sample data in some objective metric (e.g., pounds as opposed to "fat"), but there is no explicit cue about randomness of the sample. An example of this type of problem follows.

The superintendent of schools was urging the school board to make an expensive curriculum shift to a "back-to-basics" stress on fundamental learning skills and away from the electives and intensive immersion in specialized arts and social studies topics that had recently characterized the secondary schools in the district. He cited a study of 120 school systems that had recently begun to emphasize the basics and 120 school systems that had a curriculum similar to the district's current one. The "back-to-basics" school systems, he said, were producing students who scored half-a-year ahead of the students in the other systems on objective tests of reading, mathematics, and science. Of the 120 "back-to-basics" school systems, 85 had shown improved skills for students in the system vs. only 40 with improved skills in the 120 systems which had not changed. One of the school board members took the floor to argue against the change. In her opinion, she said, there was no compelling reason to attribute the improved student skills in the "back-to-basics" systems to the specific curriculum change, for two reasons: (1) School systems that make curriculum changes probably have more energetic, adventurous administrators and faculty and thus the students would learn more in those school systems no matter what the curriculum was. (2) Any change in curriculum could be expected to produce improvement in student performance because of increased faculty interest and commitment.

Comment on the reasoning of both the superintendent and the board member. On the basis of the evidence and arguments offered, do you think it is likely that the "back-to-basics" curriculum is intrinsically superior to the district's current curriculum?
(c) Subjective problems are those in which data describing a sample is not reliably measurable (e.g., "funny"). An example follows.

Gerald M. had a 3-year-old son, Timmy. He told a friend: "You know, I've never been much for sports, and I think Timmy will turn out the same. A couple of weeks ago, an older neighbor boy was tossing a ball to him, and he could catch it and throw it all right, but he just didn't seem interested in it. Then the other day, some kids his age were kicking a little soccer ball around. Timmy could do it as well as the others, but he lost interest very quickly and started playing with some toy cars while the other kids went on kicking the ball around for another 20 or 30 min."

Do you agree with Gerald's reasoning that Timmy is likely not to care much for sports? Why or why not?

One treatment group (N=21) studied examples of problems drawn from the probabilistic format plus analyses concerning the validity of conclusions about this format of problems. Another treatment group (N=23) studied objective problems and corresponding analyses in this format. The third treatment group (N=23) studied subjective problems plus analyses in this format (See Appendix A for the three sets of example problems and analyses). The fourth group (N=23) received no training.

Fong et al. devised underlying problem structures to systematize the kinds of problems presented to subjects. Their structures varied according to types of samples, types of decisions, and types of competing information. Three of these structures were used in this study: The first structure is drawing conclusions about a population from a single small sample. The problem about Gerald and his son is an illustration of this structure. The second structure is explaining why an outcome initially selected because of its extreme deviation was not maintained in a subsequent sample. The problem about Bert H. and the welding machine is an example of this structure. The third structure is pitting a large sample against a plausible
theory not founded on data. This structure is illustrated by the problem about the school superintendent and the board member.

**Measures**

Students were tested for applications of statistical reasoning by presenting them with 11 of the 18 problems used by Fong *et al.* (See Appendix B for the 11 test problems and directions to treatment and control groups). For each problem, students were asked to "think carefully ... and then write down answers that are sensible to you" as they explained the validity or invalidity of a conclusion drawn by a character described in the problem situation.

Nine of the 11 problems were instances to which the law of large numbers could be applied and two were instances for which the law of large numbers was not relevant (because conclusions were drawn from a sample that was large but biased). The 9 test problems followed a 3 x 3 design, with problem format (*probabilistic, objective, and subjective*) crossed with problem structure. The order of the test problems was randomized with the constraint that no two problems with the same structure appeared in succession.

In addition to the test items, subjects completed a brief questionnaire indicating their major area of undergraduate study and the number of courses completed in statistics, mathematics, and psychology.

**Procedure**

The control group received no instruction prior to testing. The three instructional groups participated in one 45 min lesson one week prior to testing. This lesson began with students reading a text that introduced concepts associated with sampling and the law of large numbers. (See Appendix C for the written explanation read by the students.) This was followed by a live demonstration where varying sizes of samples were drawn from an urn
filled with coloured gumballs to teach the concepts of population, sampling, and regression to the mean. Larger and larger random samples were drawn from the urn, and information about the proportions of the coloured gumballs in each successively larger sample was recorded on a blackboard. Using this data, the law of large numbers and its associated concepts was explained. A script was used by the instructor to standardize the demonstration across the instructional groups (See Appendix D for the script and blackboard chart format). Finally, students read the written examples of everyday problems and studied analyses of conclusions drawn that used the statistical concepts just presented (Appendix A). Treatments for the three instructional conditions were identical except that, in each group, the example problems and associated analyses were drawn from only one of the three formats of everyday problems: probabilistic, objective, or subjective.

The following brief examples illustrate what the example problems and analyses were designed to do. Subjects in the group instructed using probabilistic examples studied a decision-making situation faced by a character playing a Monopoly-like board game. In the problem, random variation of sample data was made clear in the information available to the player. The analysis accompanying the description of the problem critiqued the character's decision by using concepts such as population and sampling distribution to evaluate the information available. To illustrate, the problem and analysis read by subjects for the item about the Monopoly-like board game is included here.

Patricia was playing a new Monopoly-like board game called "Margin," for the first time. One of the main features of this game is that each player has the option to buy "disaster insurance." Unfortunately, this insurance is extremely costly and tends to cripple one's "investment" program. But not buying it is very risky—one can be wiped out by a "disaster."

"Disasters" come about when picking from a very large pack of cards called "fate cards." Some of these cards have neutral or favorable results, others have minor or major disasters. Each time a player has to draw a "fate card," the card is returned afterward to the pack and the pack is shuffled.
Patricia didn't have any detailed knowledge of the kinds of "fate cards" and so at first she decided to buy insurance as soon as she got near the first place on the board where she might have to draw a "fate card," with no major disaster. (One got a neutral result and one a minor disaster.) Patricia decided on that basis to continue for a while with no insurance.

Comment on the thinking that led Patricia to this decision. Is it basically sound? Does it have weaknesses?

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.

Patricia is trying to draw a conclusion about a certain population. We can think of the members of this population as fate cards. If we divide the members of this population into two categories, "neutral or minor trouble" and "major disaster," we can think of the population distribution as the % in each category. Patricia has concluded that the percentage in the "neutral or minor trouble" category is high enough to justify ignoring the risk and skipping insurance. This conclusion was based on observing a sample of size = 2, in which the sample distribution was 100% "neutral or minor trouble," 0% "major disaster."

Apart from any other considerations, however, the sample distribution for size 2 is apt to be quite different from the population distribution: The latter could be only 60% or 50% or even as low as 40% "neutral or minor trouble" and a 2-0 sample split would not be so unusual; just as one would not be at all amazed to draw 2 out of 2 silver gumballs from an urn with only 40% silvers. So Patricia's attitude is quite unwarranted: a larger sample is needed.

Subjects in the objective examples studied a problem describing the rationale of a partner in a law firm for hiring graduates from specific law schools. While sample data was measurable in this problem, there was no cue about randomness. As in the probabilistic problems, the analysis that followed described weaknesses in the reasoner's thinking based on concepts associated with the law of large numbers. A similar approach was used in the examples and analyses provided for subjects in the subjective treatment group. Unlike the probabilistic and objective examples, the problems drawn from the subjective format described a sample that was not reliably measurable.

Coding System

Codes were recorded for two measures in this study. Coding for the use of statistical reasoning distinguishes each response on the basis of whether a statistical concept was used
and whether it was a "good" statistical response. The 3-point coding system developed by Fong, et al. was used. Each response to test problems where the law of large numbers is applicable (9 problems X 90 subjects = 810 responses) was classified into one of three categories:

1 = an entirely deterministic response. The subject makes no use of statistical concepts. In a response of this type, there is no mention of sample size, randomness, or variance.

2 = a poor statistical response. Responses contain some mention of statistical concepts, but are incomplete or incorrect. These responses contain one or more of the following characteristics: (a) the subject uses both deterministic and statistical reasoning, but the deterministic reasoning is judged by the coder to dominate the subject's analysis; (b) the subject uses incorrect statistical reasoning, such as the gambler's fallacy; (c) the subject refers to luck or chance but is not explicit how the statistical concept is relevant.

3 = a good statistical response. The subject makes correct use of a statistical concept. Some form of the law of large numbers is used and the sampling elements are correctly identified. If the subject uses both good statistical and deterministic reasoning, the statistical reasoning is judged by the coder as preferred in this case.

To study the use of deterministic reasoning in the responses, categories of reasoning "errors" were generated by a qualitative analysis using "comparative analysis" (see Glaser & Strauss, 1967) applied to all control group responses (N=30) plus a random sample of five test booklets from each of the three treatment conditions (15 x 9 = 135 responses). The main question asked in this analysis was, In what types of reasoning do subjects engage? The purpose of the qualitative analysis was to ground categories in the data.

Three sources of literature were used to search for categories of reasoning errors that might converge with those found in the data: Nisbett and Ross (1980), Nickerson (1986), and Holland, et al. (1989). Types of errors from the data and from the literature were compared.
Four categories of reasoning errors that occurred most frequently in both sources (the protocols and the literature) were defined. These categories are similar to reasoning fallacies enumerated by Nickerson (1986) in one important sense: Like Nickerson's term "fallacy", the term "error" denotes an approach to a problem. What may be an inappropriate approach to a problem (a fallacy or an error) in one context may be quite appropriate in another. The nature of the problem requires that the notion of "error" be considered conditionally. In terms of Holland et al.'s theory, these categories represent competing heuristics in a problem solver's rule system.

The four hypothesized categories of errors are labelled and defined below. Examples are drawn from the protocols collected in Experiment 1.

Error Category A = Attribution to Disposition. The subject discounts the problem situation and offers an explanation based on a character's disposition. For example, "The talent scout is biased about the way players improve." "Howard tends to stereotype kids." "The superintendent is just ambitious." "Gerald has a secret wish to compensate for his own inability in sports." "The boss is in an administrative position where one is required to believe in what he believes."

Error Category B = Egocentric Bias. The subject offers an explanation based on a personal opinion or experience. For example, "The economist is right; because, if I won a lottery, I would start a business to fight boredom." "I don't think back-to-basics is is superior; because, from my own experience, the current system is more intrinsically motivating." "Personally, I'm not shy so I don't know if this would affect taste in a restaurant." "No, my brother and father had a similar scenario; he felt threatened by dad everytime he did sports."

Error Category C = Appeal to Truism. The subject relies on a deterministic principle, maxim, or law. For example, "Everybody makes mistakes." "Most kids are looking for
attention; that's why Howard's approach works. "The superintendent should follow the rule, 'Don't fix it if it isn't broken'." "Food is food; the taste doesn't change." "The boss is correct because of human nature." This error category is similar to Rubinstein's "world view" constraint.

**Error Category D = Speculation.** The subject adds extra evidence or information *not* provided in the problem statement or hypothesizes without any apparent evidence. This category is somewhat difficult to delineate since subjects' responses may be based on beliefs, theories, or schemata that, for the subject, are supported by information in the problem. However, to the coder, the explanations are speculative guesses based on unidentified data or some unidentified law. There is nothing in the given information to support the statement. For example, "There would be no incentive for people to take "crummy" jobs." "The player is pressured by the scout's presence." "She's probably lonely." "The welding machine is warmed up properly for the first batch." "The nurses are trying to impress their bosses in the first few days."

**Interrater Agreement**

Interrater agreement was tested by having two coders code a sample of 11 test booklets (99 law of large numbers problems). The responses were retyped and presented as separate items to mask within-subject consistency. The unit of analysis in establishing agreement was a subject's complete response to each problem. The percentage of agreement was established by checking whether the coders agreed or disagreed on each code and dividing the number of agreements by the total number of agreements and disagreements. Interrater agreement of 93% in statistical reasoning codes and 85% in deterministic reasoning codes was established. Finally, the primary coder coded all the responses, blind to conditions.
Scoring

Responses to all nine law of large numbers problems for each subject were coded on a 3-point scale and assigned a total score out of 27. Since each of the three problem formats—probabilistic, objective, and subjective—was tested with three problems, the total possible for each format is 9. The problems act as repeated measure to develop a continuous variable representing each subject's quality of statistical reasoning.

Subjects' errors were coded 1 if one or more instances of a specific error category occurred in a response to a single problem, and 0 if the error did not occur. For example, a subject might obtain a score of 4 in the error category labelled Attribution to Disposition by committing this error 7 times in 4 of the responses to the 9 law of large numbers problems. Scores for each of the four error categories were created.

A reliability analysis was conducted to examine the scales measuring subjects' statistical reasoning and error commission. The analysis of the scale for statistical reasoning reveals strong internal consistency of the nine items (alpha = .76). The analysis of the reasoning errors scale provides a moderate indication that errors are internally coherent when aggregated across problems and types of errors (alpha = .56).

Results

Mean scores in statistical reasoning and mean scores in deterministic reasoning errors were used for analysis of variance. Descriptive statistics were generated as well. Stepwise multiple regression was employed to explore relationships among variables.

Effects of Instruction on Statistical Reasoning

The first purpose of this experiment was to investigate the effects of instruction in a statistical heuristic on students' use of statistical reasoning in thinking about everyday
inferential problems. Table 1 and Figure 1 represent descriptive statistics for the four conditions by problem format.

To test whether instruction in the law of large numbers significantly enhanced the quality of statistical reasoning in probabilistic, objective, and subjective types of problems, all types of problems were pooled to make one scale and a t-test was computed by pooling all three instructed groups and comparing them to the no instruction group. The t-value was 5.68 ($p < .001$). The effect size was 1.48. This shows that subjects in the instructional conditions learned a good deal about how to reason statistically as a consequence of instruction.

The results from the test of the first hypothesis are similar to the results of Fong, et al's experiments. It was found that a minimal amount of instruction (45 minutes) can have a marked effect on the way university undergraduates reason about everyday problems that involve a statistical concept such as the law of large numbers.

Relationship between Instructional Domain and Statistical Reasoning

The second purpose of the experiment was to examine whether improvements in statistical reasoning vary as a function of the problem format emphasized during instruction. To explore this question, the data were entered into oneway ANOVAs with a priori contrasts that compared each instructional condition to the no instruction group separately on each problem format. The results of these analyses are presented in Table 2.
Table 1

Descriptive Statistics on Quality of Statistical Reasoning

<table>
<thead>
<tr>
<th>Problem Format</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probabilistic</td>
</tr>
<tr>
<td></td>
<td>Examples</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>7.95</td>
</tr>
<tr>
<td></td>
<td>1.24</td>
</tr>
<tr>
<td>Objective</td>
<td>6.38</td>
</tr>
<tr>
<td></td>
<td>1.28</td>
</tr>
<tr>
<td>Subjective</td>
<td>5.76</td>
</tr>
<tr>
<td></td>
<td>1.89</td>
</tr>
<tr>
<td>All Problems</td>
<td>20.10</td>
</tr>
<tr>
<td></td>
<td>3.35</td>
</tr>
</tbody>
</table>

Note. Upper numbers are means; lower numbers are standard deviations.
Figure 1

Mean Statistical Reasoning Score as a Function of Instructional Condition and Problem Format

Note. Closed points (•) = performance when instruction included probabilistic examples; open points (○) = performance when instruction included objective examples; squares (□) = performance when instruction included subjective examples; triangles (△) = performance of the No Instruction group.
Table 2

Results of A priori t-tests Comparing Quality of Statistical Reasoning by Instructional Condition Contrasted to Control Condition

<table>
<thead>
<tr>
<th>Group</th>
<th>Problem Format</th>
<th>Probabilistic Examples</th>
<th>Objective Examples</th>
<th>Subjective Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probabilistic vs. Control</td>
<td>1.88</td>
<td>2.05</td>
<td>2.27</td>
</tr>
<tr>
<td></td>
<td>.07</td>
<td>.05</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.52</td>
<td>.63</td>
<td>.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Objective vs. Control</td>
<td>3.59</td>
<td>3.32</td>
<td>2.88</td>
</tr>
<tr>
<td></td>
<td>.001</td>
<td>.002</td>
<td>.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.19</td>
<td>1.17</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subjective vs. Control</td>
<td>3.43</td>
<td>4.75</td>
<td>5.07</td>
</tr>
<tr>
<td></td>
<td>.001</td>
<td>&lt; .001</td>
<td>&lt; .001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.08</td>
<td>1.75</td>
<td>1.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All Problems vs. Control</td>
<td>3.83</td>
<td>4.53</td>
<td>4.54</td>
</tr>
<tr>
<td></td>
<td>&lt; .001</td>
<td>&lt; .001</td>
<td>&lt; .001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.28</td>
<td>1.64</td>
<td>1.51</td>
<td></td>
</tr>
</tbody>
</table>

Note. Top numbers are t-values; middle numbers are probabilities; bottom numbers are effect sizes. Separate variance estimates are used. No Instruction group's standard deviations are used to compute effect sizes.
These results show that each instructional condition reliably enhanced statistical reasoning in each of the three problem formats when contrasted with the control condition. As in the Fong et al. study, the present experiment did not detect domain specificity of instructional effects based on these formats of problems. Generally, subjects taught with examples in one format could solve problems about equally as well in other formats.

Effects of Instruction on Reasoning Errors

An additional purpose of this study—a question not addressed by the Fong et al. work—was to look for typical reasoning errors committed by subjects, especially when they underutilized or did not utilize the base rate information in a problem situation. In other words, what nonstatistical, competing heuristics did subjects use that led them astray in dealing with problems? Four types of errors were identified. Three of the four were examined in the present experiment to determine if instruction in the law of large numbers reduced the number of these reasoning errors. These three are Attribution to Disposition, Egocentric Bias, and Appeal to Truism. The fourth category, Speculation, was judged to be less useful as a category of reasoning errors, because this category was more difficult to delineate from the limited amount of information in subjects' protocols. Descriptive statistics are presented in Table 3 and Figure 2.

Instruction decreased the number of errors in probabilistic, objective, and subjective types of problems when these types of problems were considered together. All types were pooled to make one scale and a t-test was computed by pooling all three instructed groups and comparing them to the no instruction group. The t-value was -3.80 ($p < .001$). The effect size was 1.15.
Table 3

Mean Number of Reasoning Errors Committed by Students in Problem Formats

<table>
<thead>
<tr>
<th>Problem Format</th>
<th>Probabilistic Examples</th>
<th>Objective Examples</th>
<th>Subjective Examples</th>
<th>No Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic</td>
<td>.24</td>
<td>.39</td>
<td>.45</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>.44</td>
<td>.72</td>
<td>.60</td>
<td>.80</td>
</tr>
<tr>
<td>Objective</td>
<td>.71</td>
<td>.43</td>
<td>.68</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>.90</td>
<td>.59</td>
<td>.72</td>
<td>1.16</td>
</tr>
<tr>
<td>Subjective</td>
<td>1.05</td>
<td>.78</td>
<td>.91</td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>1.24</td>
<td>.67</td>
<td>1.11</td>
<td>.99</td>
</tr>
<tr>
<td>All Problems</td>
<td>2.00</td>
<td>1.61</td>
<td>2.05</td>
<td>3.78</td>
</tr>
<tr>
<td></td>
<td>1.95</td>
<td>1.31</td>
<td>1.73</td>
<td>2.19</td>
</tr>
</tbody>
</table>

Note. Upper numbers are means; lower numbers are standard deviations. Three types of errors are coded for each problem: Attribution to Disposition, Egocentric Bias, and Appeal to Truism.
Figure 2

Mean Number of Reasoning Errors as a Function of Instructional Condition and Problem Format

Note. Closed points (*) = performance when instruction included probabilistic examples; open points (o) = performance when instruction included objective examples; squares (□) = performance when instruction included subjective examples; triangles (Δ) = performance of the No Instruction group.
In addition, pooling all problem types, a Pearson correlation between overall statistical reasoning and error commission shows a strong negative correlation \( r = -0.65; p = 0.01 \).

**Relationship between Instructional Format and Reasoning Errors**

Also examined was the reduction in reasoning errors as a function of the problem format emphasized during instruction. As with statistical reasoning, one-way ANOVAs with *a priori* contrasts that compared each instructional condition to the no instruction group separately on each problem format were computed. The results of this analysis are presented in Table 4.

The results show that, although the effect sizes are not large, each instructional condition reliably reduced reasoning errors in each of the three problem formats when contrasted with the control condition. Reasoning errors in objective and subjective types of problems were reduced to the greatest degree when instruction used objective examples.

**Statistical Reasoning and Other Variables**

To explore relationships between subjects' overall statistical reasoning and five predictor variables, data were entered into a stepwise multiple regression with statistical reasoning on all nine law of large numbers problems as the dependent variable. The correlation matrix for the variables is presented in Table 5. The regression model is presented in Table 6.

Subjects' achievement is their final grade in the educational psychology course. The number of postsecondary mathematics, psychology, and statistics courses subjects reported having taken and subjects' major course of undergraduate study—psychology, humanities, or other—comprise the other predictors.
Table 4

Results of A priori t-tests Comparing Errors in Statistical Reasoning by Instructional Condition Contrasted to Control Condition

<table>
<thead>
<tr>
<th>Problem Format</th>
<th>Probabilistic Examples</th>
<th>Objective Examples</th>
<th>Subjective Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic vs. Control</td>
<td>-2.81</td>
<td>-1.72</td>
<td>-1.33</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>.09</td>
<td>.19</td>
</tr>
<tr>
<td></td>
<td>.66</td>
<td>.49</td>
<td>.41</td>
</tr>
<tr>
<td>Objective vs. Control</td>
<td>-2.44</td>
<td>-3.84</td>
<td>-2.90</td>
</tr>
<tr>
<td></td>
<td>.02</td>
<td>&lt; .001</td>
<td>.006</td>
</tr>
<tr>
<td></td>
<td>.59</td>
<td>.83</td>
<td>.61</td>
</tr>
<tr>
<td>Subjective vs. Control</td>
<td>-1.76</td>
<td>-3.80</td>
<td>-2.57</td>
</tr>
<tr>
<td></td>
<td>.09</td>
<td>&lt; .001</td>
<td>&lt; .013</td>
</tr>
<tr>
<td></td>
<td>.57</td>
<td>.84</td>
<td>.71</td>
</tr>
<tr>
<td>All Problems vs. Control</td>
<td>-2.85</td>
<td>-4.08</td>
<td>-3.00</td>
</tr>
<tr>
<td></td>
<td>.007</td>
<td>&lt; .001</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>.81</td>
<td>1.00</td>
<td>.79</td>
</tr>
</tbody>
</table>

Note. Top numbers are t-values; middle numbers are probabilities; bottom numbers are effect sizes. Separate variance estimates are used. No Instruction group's standard deviations are used to compute effect sizes.
Table 5

Correlation Matrix for Variables in Undergraduate Multiple Regression Analysis

<table>
<thead>
<tr>
<th></th>
<th>stat reasoning</th>
<th>achievement</th>
<th>major</th>
<th>math courses</th>
<th>stat courses</th>
<th>psych courses</th>
</tr>
</thead>
<tbody>
<tr>
<td>stat reasoning</td>
<td>.478</td>
<td>.066</td>
<td>.145</td>
<td>-.011</td>
<td>.115</td>
<td></td>
</tr>
<tr>
<td>achievement</td>
<td>-.002</td>
<td>.151</td>
<td>-.002</td>
<td>-.212</td>
<td>-.504</td>
<td></td>
</tr>
<tr>
<td>major</td>
<td>-.082</td>
<td>-.212</td>
<td>.451</td>
<td>.192</td>
<td>.285</td>
<td></td>
</tr>
<tr>
<td>math courses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stat courses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>psych courses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6

Results of Multiple Regression Predicting Undergraduates' Quality of Statistical Reasoning on All Problems by Achievement, Major, and Courses

<table>
<thead>
<tr>
<th>Predictor</th>
<th>b</th>
<th>β</th>
<th>t</th>
<th>p</th>
<th>inc R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement</td>
<td>.74</td>
<td>.45</td>
<td>4.59</td>
<td>&lt; .001</td>
<td>.23</td>
</tr>
<tr>
<td>Major</td>
<td>.51</td>
<td>.11</td>
<td>1.01</td>
<td>.32</td>
<td>.00</td>
</tr>
<tr>
<td>Mathematics Courses</td>
<td>.36</td>
<td>.09</td>
<td>.86</td>
<td>.39</td>
<td>.01</td>
</tr>
<tr>
<td>Statistics Courses</td>
<td>-.36</td>
<td>-.06</td>
<td>-.50</td>
<td>.62</td>
<td>.00</td>
</tr>
<tr>
<td>Psychology Courses</td>
<td>.16</td>
<td>.10</td>
<td>.85</td>
<td>.40</td>
<td>.01</td>
</tr>
</tbody>
</table>

Predictors are listed in order of entry into the regression model.
Overall R² = .25; F = 5.50; p = < .001
Results show that student achievement is the strongest predictor of statistical reasoning when all 90 subjects are included in the analysis with the other predictors adding very little to the overall $R^2$. However, when "naturally occurring" statistical reasoning—that is, data from only the noninstructed group (N=23)—was used, Pearson correlations between statistical reasoning on all types of problems and the same independent variables were low. These results suggest that, for undergraduates, limited training in probabilistic subject-areas, as represented by undergraduate courses in statistics, psychology, and mathematics, as well as by an undergraduate major in psychology, does not have a strong relationship with statistical reasoning in everyday problems. However, statistical reasoning among instructed subjects correlates strongly with subjects’ achievement.
EXPERIMENT 2

Experiment 2 was designed to extend the work done in Experiment 1 by (a) examining questions about main effects of statistical instruction using a tighter pretest-posttest design; (b) testing whether instruction in reasoning errors derived from Experiment 1 would further improve statistical reasoning and further reduce such errors in university undergraduate students' thinking about everyday problems; and (c) exploring links between learning to reason statistically and other variables including subjects' commission of errors, achievement, general reasoning ability, understanding of sampling, and knowledge of content domains in problems.

Experiment 1 suggested that subjects who reason more statistically also commit fewer reasoning errors. Materials to teach people about such error types were designed for Experiment 2. If students became aware of types of errors, it was felt that further gains in statistical reasoning and reductions in errors would result because the errors, as competing heuristics, would be less likely to interfere with students' "naturally occurring" statistical heuristics. The margin of improvement over simply teaching the statistical laws was of interest.

The two main questions investigated in Experiment 2 are (1) Will students receiving instruction in statistical rules increase their use of such rules and make fewer reasoning errors? (2) Will students receiving instruction in statistical inferential rules as well as additional instruction in reasoning errors increase their use of statistical rules and make fewer errors when compared to students receiving only instruction in statistical inferential rules? The first question addresses the effects of instruction per se. The second question addressed the power of adding instruction about errors to instruction about rules. Finally, the data was
used to explore relationships between statistical reasoning and other variables to describe trends at the university level of schooling that could be compared to trends indicated at the secondary and elementary levels of schooling studied in Experiment 3.

Method

Sample

Subjects in this study were undergraduate students (N = 35), 28 females and 7 males, in an educational psychology course who volunteered to participate in all three phases of the experiment: (a) pretest, (b) instruction, and (c) posttest, plus an additional 24 students who volunteered to participate in the instruction and the posttest, but these subjects did not complete the pretest. Subjects with pretest data were assigned to one of two treatment conditions: (a) Rule Condition (N = 19), instruction in the law of large numbers similar to that used in Experiment 1; and (b) Rule-Plus-Error Condition (N = 16), instruction in the law of large numbers plus additional instruction in how to recognize and avoid reasoning errors.

Training was conducted during intact course tutorials with groups ranging in size from 6 to 12 students and involved all students in 11 tutorial groups. Hence, subjects were not randomly assigned to the two conditions. Oneway ANOVA's indicate that the eleven groups where subjects were trained were not significantly different, at the .05 level, with respect to achievement as measured by course grade.

To examine practice or pretest sensitization effects arising from subjects' completion of the pretest, posttest data from all 35 subjects who did both the pretest and the posttest was contrasted to posttest data from 24 subjects who did not take the pretest. A oneway ANOVA revealed no significant differences on all problems pooled. In separate ANOVA's on problems from each of two formats, no differences (p<.05) were found.
Treatments and Controls

All training was delivered by the experimenter. The Rule Condition: (a) read about the law of large numbers (Appendix C); (b) participated in a live demonstration of the law of large numbers (see Appendix D for the script and chart) that was identical to that used in Experiment 1 except for using a gumball machine instead of an urn and gumballs and a laminated chart instead of the blackboard; (c) studied three example problems and analyses of the information presented in the problems using concepts from the reading and demonstration (Appendix E). The example problems were three subjective problems participants had already read and answered in the pretest. Results of Experiment 1 suggest that students learning from only subjective examples performed as well on probabilistic and objective problems as students who studied probabilistic or objective examples.

The Rule-Plus-Error condition received the same instruction as the Rule condition, as well as further instruction about reasoning errors. Study material about the errors was comprised of written (a) information that a student volunteer read aloud while others listened and followed on printed sheets, (b) examples of errors in hypothetical analyses of the same problems they had studied earlier in the statistical material, and (c) exercises that provided practice identifying errors (Appendix F). A simple acronym, TEDS, was used to help subjects remember the error types. Each letter stands for the name of one of the types: Truism, Egocentric Bias, Attribution to Disposition, and Speculation.

Measures

Ten of the eleven test problems developed by Fong et al. and used in Experiment 1 were used in this experiment. In addition, six problems—three in the probabilistic format and three in the objective format—were developed. The six new problems are isomorphic in both
domain and structure to the six probabilistic and objective problems constructed by Fong et al. For instance, their probabilistic item about Bert H. and the welding machine is paralleled by the following new problem.

Brenda has a summer job at a well-known hamburger restaurant. The restaurant has brand new, automatic food-wrapping machines. One of Brenda's duties is to check how the new packaging equipment is working. She does this for one hour each day. Overall, the machines wrapped hamburgers and put french fries and pies into their containers perfectly about 90 percent of the time. This meant that out of 800 items done during the hour, about 700 to 740 packages are perfectly done. Brenda has noticed something she finds interesting. On some days, all of the first 15 packages done during the hour were perfect. However, Brenda also noticed that for those hours, the overall number of perfectly done packages is about the same as for any other hours when the first fifteen were not perfect.

Why do you think the overall number of perfectly done packages is the same -- even for those hours when the first fifteen are perfect?

The six probabilistic and objective problems in the pretest (Appendix G) and the six isomorphic items in the posttest (Appendix H) are the pre-post measures in this experiment. These items, in each the two tests, follow a 2x3 design, with problem format crossed with problem structure. As in Experiment 1, the order of the test problems was randomized so that no two problems with the same structure appeared in succession. As mentioned previously, the three subjective problems included in the pretest were later used as example problems during the instructional phase. An additional, isomorphic false-alarm problem was included in each of the pretest and posttest booklets. Answers to false alarm problems were not coded and were not included in the present data analysis. The sequence, formats, and structures of the pretest and posttest booklets are presented in Tables 6 and 7.
Table 7: Pretest Problems by Format and Structure

<table>
<thead>
<tr>
<th>Problem Sequence and Format</th>
<th>Problem Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Objective</td>
</tr>
<tr>
<td>2</td>
<td>Objective</td>
</tr>
<tr>
<td>3</td>
<td>Objective</td>
</tr>
<tr>
<td>4</td>
<td>Objective</td>
</tr>
<tr>
<td>5</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>6</td>
<td>Subjective</td>
</tr>
<tr>
<td>7</td>
<td>Subjective</td>
</tr>
<tr>
<td>8</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>9</td>
<td>Subjective</td>
</tr>
<tr>
<td>10</td>
<td>Probabilistic</td>
</tr>
</tbody>
</table>

Note. Structure 1 (conclusion from a single small sample with part 'a' as a false alarm question). Structure 2 (large sample pitted against a plausible theory not founded on data). Structure 3 (regression). Structure 4 (false alarm problem where sample data is large but biased).
Table 8: Posttest Problems by Format and Structure

<table>
<thead>
<tr>
<th>Problem Sequence and Format</th>
<th>Problem Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1   Objective</td>
<td></td>
</tr>
<tr>
<td>2   Objective</td>
<td></td>
</tr>
<tr>
<td>3   Objective</td>
<td></td>
</tr>
<tr>
<td>4   Objective</td>
<td></td>
</tr>
<tr>
<td>5   Probabilistic</td>
<td></td>
</tr>
<tr>
<td>6   Probabilistic</td>
<td></td>
</tr>
<tr>
<td>7   Probabilistic</td>
<td></td>
</tr>
</tbody>
</table>

Note. Structure 1 (conclusion from a single small sample with part 'a' as a false alarm question).
Structure 2 (large sample pitted against a plausible theory not founded on data).
Structure 3 (regression).
Structure 4 (false alarm problem where sample data is large but biased).
Immediately prior to instruction, three quick probes were administered (Appendix I) during a five min period. The first of these, the classic Wason selection task (Margolis, 1987), was intended to reveal subjects' conditional reasoning ability. Four cards were displayed showing an "A", a "B", the number 4, and the number 7. Students were asked to identify all and only the cards that must be turned over to determine if the following rule holds: "If a card has an A on one side, then it has a 4 on the other side." The correct answer is to turn over the cards with A and 7. The second of the probes was intended to measure subjects' domain knowledge about the content domain in each of the posttest problems. This was a self-report instrument. On a five item scale, students rated how familiar they were with topics such as welding, basketball, and university housing. The third device presented five diagrams about sampling from a population and a multiple-choice question about regression to the mean. The diagrams represented populations with samples of various sizes and degrees of randomness. Students were asked to identify samples which gave the best estimate of the population. Aggregated responses were intended to provide an indicator of subjects' prior sensitivity to sampling concepts.

Procedure

Subjects received pretest booklets in a course lecture. The first two of the ten items were administered by the experimenter who read directions and provided subjects with practice in reading and writing about the problems in the allotted time of 40 min. Eight minutes was allocated to the first two problems. Students were asked to do the remaining problems over a 32 min period as homework and to return the completed test booklet to the following week's lecture.

Instruction occurred in 11 tutorial sessions during the two-week period after pretests were collected. Immediately prior to instruction, the set of probes was administered. Next,
students in the Rule condition as well as the Rule-Plus-Error condition read about the law of large numbers. Instruction for both condition groups was identical to that in Experiment 1 except that the three examples and associated analyses were based on the three subjective items already encountered in the pretest. This 30 min instructional phase for the Rule condition was concluded with a verbal encouragement from the experimenter to the students to remember as much as possible about the law of large numbers.

The Rule-Plus-Error condition continued after the 30 min of initial instruction and engaged in an additional 15 min of training in reasoning errors. They silently read definitions and examples of the types of errors as a fellow class member read this material aloud. Then, they completed a written exercise identifying categories of errors. In the exercise, subjects reread the three subjective problems from the pretest and identified what types of errors were committed in 12 written statements by characters in the problems. Finally, subjects in the Rule-Plus-Error condition received feedback from the experimenter about correct answers to the error exercise. Results of students' answers in the exercise were checked by the experimenter. All participants completed the exercises accurately. One week after instruction, each condition group completed the seven posttest problems during a 30-minute period in their tutorials.

**Coding and Scoring**

Answers to the six problems involving statistical reasoning were scored as in Experiment 1. A reliability analysis was conducted to examine the six item scales on statistical reasoning. Analysis of the pretest scale reveals a moderate indication that statistical reasoning is internally coherent across problems (alpha = .51). The posttest analysis suggests a lack of internal consistency (alpha = .21). The same reliability analysis was also conducted
by condition. For the Rule condition, pretest and post test alpha's are .58 and .35 respectively. For the Rule-Plus-Error condition, they are .43 and a negative value.

Results

Experiment 2 is a pretest-posttest, two-treatment, two factor design. Mean scores in statistical reasoning and mean scores in reasoning errors were used for repeated measures MANOVA. The between groups factor is the treatment condition. The within subjects factors are (a) the measures for statistical reasoning and error commission and (b) the pretest-posttest occasions. In addition, variables were entered into multiple regression analyses to explore predictors of statistical reasoning.

Effects of Instruction

The first main question investigated in this experiment, "Will students receiving instruction in statistical rules increase their use of such rules and make fewer reasoning errors?", addresses the overall effects of instruction per se. Table 9 and Figure 3 present descriptive statistics on the quality of undergraduates' statistical reasoning in the two treatment groups and problem formats. Effect sizes comparing pretest to posttest scores for the Rule condition in categories of probabilistic, objective, and all problems are .54, 1.78, and 1.13 respectively. Effect sizes for the Rule-Plus-Error condition are larger: 1.10, 2.63, and 2.05, respectively. Table 10 and Figure 4 present descriptive statistics on the frequency of undergraduates' reasoning errors in the two treatment groups and problem formats. Moderate effects are indicated for both treatment conditions with slightly larger effect sizes for the Rule condition.

To examine relationships between statistical reasoning and error commission at pretest and posttest occasions, reasoning and error rates were correlated. A Pearson correlation between
overall statistical reasoning and error commission at the pretest level shows a near-zero correlation ($r = .08$). However, the same correlation at posttest level reveals a strong negative correlation ($r = -.48; p = .01$). The findings are similar to those in Experiment 1.

To summarize, effect sizes resulting from testing the first question are similar to those from Experiment 1. Even the smaller amount of instruction—30 min in the second experiment versus 45 min in the first experiment—had a substantial effect on the way undergraduates reason about everyday problems that involve statistical concepts from the law of large numbers. Instruction enhanced reasoning about both probabilistic and objective events. Adding 15 min of error instruction resulted in effect sizes in statistical reasoning for the Rule-Plus-Error condition that are larger than those for the Rule Only training. In fact, when the results of Experiment 1 and 2 are compared, Experiment 2 resulted in effect sizes for objective problems are over twice as large as effect sizes for probabilistic problems. This outcome was not revealed in Experiment 1 where the difference between effect sizes for objective and probabilistic problems was marginal. In contrast to the results on statistical reasoning, effect sizes for reductions in reasoning errors are consistently larger for the Rule group.

Effects of Rule versus Rule-Plus-Error Instruction

The second main question investigated in Experiment 2, "Will students who receive instruction in statistical inferential rules as well as additional instruction in reasoning errors
Table 9
Descriptive Statistics on Quality of Undergraduate Subjects' Statistical Reasoning in Treatment Groups and Problem Formats

<table>
<thead>
<tr>
<th>Treatment Problem Format</th>
<th>N</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Rule Condition</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td></td>
<td>6.90</td>
<td>1.76</td>
<td>7.84</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>4.32</td>
<td>1.00</td>
<td>6.11</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td>11.21</td>
<td>2.42</td>
<td>13.95</td>
</tr>
<tr>
<td>Rule Plus Error Condition</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td></td>
<td>6.56</td>
<td>1.17</td>
<td>8.31</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>4.25</td>
<td>1.00</td>
<td>6.88</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td>10.81</td>
<td>2.14</td>
<td>15.19</td>
</tr>
</tbody>
</table>

Note. Pretest standard deviations are used to compute effect sizes.
Figure 3

Plotted Means for Undergraduates' Statistical Reasoning as a Function of Instructional Condition and Problem Format

Pretest       Posttest

Note. Closed points (●) = performance on probabilistic problems; open points (○) = performance on objective problems. Dotted lines represent rule condition change. Solid lines represent rule + error condition change.
Table 10

Descriptive Statistics on Undergraduates' Reasoning Errors in Treatment Groups and Problem Formats

<table>
<thead>
<tr>
<th>Treatment Problem Format</th>
<th>N</th>
<th>Pretest</th>
<th></th>
<th>Posttest</th>
<th></th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Rule Condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td>19</td>
<td>.68</td>
<td>.67</td>
<td>.47</td>
<td>.70</td>
<td>.31</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>.95</td>
<td>.91</td>
<td>.26</td>
<td>.73</td>
<td>.76</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td>1.63</td>
<td>1.07</td>
<td>.74</td>
<td>1.15</td>
<td>.83</td>
</tr>
<tr>
<td>Rule Plus Error Condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td>16</td>
<td>.31</td>
<td>.60</td>
<td>.19</td>
<td>.40</td>
<td>.20</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>.88</td>
<td>1.03</td>
<td>.50</td>
<td>.82</td>
<td>.37</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td>1.19</td>
<td>1.33</td>
<td>.69</td>
<td>.95</td>
<td>.38</td>
</tr>
</tbody>
</table>

Note. Pretest standard deviations are used to compute effect sizes.
Figure 4

Plotted Means for Undergraduates' Reasoning Errors as a Function of Instructional Condition and Problem Format

Note. Closed points (•) = performance on probabilistic problems; open points (○) = performance on objective problems. Dotted lines represent rule condition change. Solid lines represent rule + error condition change.
increase their use of statistical rules and decrease their commission of reasoning errors when compared to students who receive instruction in only statistical inferential rules?" contrasts effects of the two treatment conditions. To test whether additional instruction in reasoning errors significantly enhanced the quality of statistical reasoning and reduced the number of reasoning errors in probabilistic and objective types of problems, data were entered into a repeated measures MANOVA to compare treatment conditions. The results are presented in Table 11. A Box M test for homogeneity was conducted.

A multivariate test of significance ($s = 1, m = 3, n = 12$) for the condition effect yielded a nonsignificant Wilk's Lambda of .649 ($p = .13$). While this result was not significant at the .05 default level, univariate F tests suggested that the groups may differ in statistical reasoning performance on posttest problems in the objective format ($F = 3.90, df 1,33; p = .06$). Further examination using a test which is a generalization of Tukey's HSD post hoc procedure, suggested by Spjotvall and Stoline (1973), revealed that the difference in statistical reasoning between conditions on posttest problems from the objective format is reliable at the .05 level of significance. This suggests that instruction in the Rule-Plus-Error condition improved statistical reasoning on objective problems to a reliably greater degree than instruction in the Rule Only condition. Post hoc tests revealed no reliable differences between treatment groups in statistical reasoning on probabilistic problems or in error rates in either format.

In addition, the MANOVA reveals reliable differences in statistical reasoning between objective and probabilistic types of problems. Subjects were more likely to reason statistically and committed less errors on problems from the probabilistic format than on problems from the objective format.
Table 11

University Undergraduates: Results of Repeated Measures MANOVA Comparing Treatment Conditions on Statistical Reasoning and Reasoning Errors.

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Between</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>.13</td>
<td>1</td>
<td>.13</td>
<td>.10</td>
<td>.76</td>
</tr>
<tr>
<td>Error</td>
<td>44.06</td>
<td>33</td>
<td>1.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Within</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measures</td>
<td>2389.62</td>
<td>1</td>
<td>2389.62</td>
<td>948.79</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Condition by Measures</td>
<td>1.94</td>
<td>1</td>
<td>1.94</td>
<td>.77</td>
<td>.39</td>
</tr>
<tr>
<td>Error</td>
<td>83.11</td>
<td>33</td>
<td>2.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td>55.32</td>
<td>1</td>
<td>55.32</td>
<td>87.23</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Condition by Format</td>
<td>2.09</td>
<td>1</td>
<td>2.09</td>
<td>3.30</td>
<td>.08</td>
</tr>
<tr>
<td>Error</td>
<td>20.93</td>
<td>33</td>
<td>.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occasion</td>
<td>35.49</td>
<td>1</td>
<td>35.49</td>
<td>43.32</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Condition by Occasion</td>
<td>4.49</td>
<td>1</td>
<td>4.49</td>
<td>5.48</td>
<td>.03</td>
</tr>
<tr>
<td>Error</td>
<td>27.03</td>
<td>33</td>
<td>.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measures by Format</td>
<td>87.81</td>
<td>1</td>
<td>87.81</td>
<td>118.18</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Co by Me by Fo</td>
<td>.07</td>
<td>1</td>
<td>.07</td>
<td>.10</td>
<td>.76</td>
</tr>
<tr>
<td>Error</td>
<td>24.52</td>
<td>33</td>
<td>.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measures by Occasion</td>
<td>78.56</td>
<td>1</td>
<td>78.56</td>
<td>91.70</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Co by Me by Oc</td>
<td>1.68</td>
<td>1</td>
<td>1.68</td>
<td>1.96</td>
<td>.17</td>
</tr>
<tr>
<td>Error</td>
<td>28.27</td>
<td>33</td>
<td>.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format by Occasion</td>
<td>1.07</td>
<td>1</td>
<td>1.07</td>
<td>1.77</td>
<td>.19</td>
</tr>
<tr>
<td>Co by Fo by Oc</td>
<td>.07</td>
<td>1</td>
<td>.07</td>
<td>.12</td>
<td>.73</td>
</tr>
<tr>
<td>Error</td>
<td>20.02</td>
<td>33</td>
<td>.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Me by Fo by Oc</td>
<td>6.47</td>
<td>1</td>
<td>6.47</td>
<td>6.37</td>
<td>.02</td>
</tr>
<tr>
<td>Co by Me by Oc</td>
<td>.04</td>
<td>1</td>
<td>.04</td>
<td>.04</td>
<td>.86</td>
</tr>
<tr>
<td>Error</td>
<td>33.48</td>
<td>33</td>
<td>1.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Condition (training in statistical heuristics and training in statistical heuristics plus training in reasoning errors) is the between groups factor. Measures (statistical reasoning and reasoning errors); Format (probabilistic problems and objective problems); and, Occasion (pretest problems and posttest problems) are within subjects factors.
Statistical Reasoning Predicted by Other Variables

To explore factors that might predict statistical thinking about problems, subjects' scores on (a) the Wason selection task, (b) the self-report of domain knowledge, and (c) the probe of sensitivity to sampling concepts, as well as (d) pretest scores (e) final marks in the educational psychology course, and (f) gender, were considered as candidates for entry into stepwise multiple regression analysis with statistical reasoning scores on the posttest serving as the dependent variable.

All predictors except two of these were used. First, because of the very limited number of male subjects (n=7), gender was not entered. And, second, only three subjects achieved the correct answer on the Wason card problem. A review of the literature indicates that this may not to be an unusual result. Holland et al. (1989) and others report that fewer than ten percent of college students can produce the correct answer in the Wason selection task. Table 12 presents the correlation matrix for the variables, and Table 13 shows the results of the multiple regression analysis.

The only reliable predictor of subjects' statistical reasoning on the posttest was the pretest. Undergraduates' statistical reasoning on the posttest was not reliably predicted by self-reports of content domain knowledge. For instance, students who felt they knew a great deal about the domains did not reason more statistically than students who indicated that they did not know about these topics. In addition, statistical reasoning on posttest problems was not reliably predicted by student achievement level measured by grades in the educational psychology course where the experiment took place. Finally, statistical reasoning was not predicted by the probe of sensitivity to sampling concepts.
Table 12
Correlation Matrix for Variables in Undergraduate Multiple Regression Analysis

<table>
<thead>
<tr>
<th></th>
<th>posttest scores</th>
<th>pretest scores</th>
<th>domain knowledge</th>
<th>sampling knowledge</th>
<th>achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td>posttest scores</td>
<td>.354</td>
<td>-.283</td>
<td>.187</td>
<td>.038</td>
<td></td>
</tr>
<tr>
<td>pretest scores</td>
<td>-.053</td>
<td>.250</td>
<td>.423</td>
<td></td>
<td></td>
</tr>
<tr>
<td>domain knowledge</td>
<td></td>
<td></td>
<td>-.227</td>
<td>-.113</td>
<td></td>
</tr>
<tr>
<td>sampling knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>achievement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.046</td>
</tr>
</tbody>
</table>
Table 13

Results of Stepwise Multiple Regression Predicting Undergraduate Posttest Scores in Statistical Reasoning

<table>
<thead>
<tr>
<th>Predictor</th>
<th>b</th>
<th>β</th>
<th>t</th>
<th>p</th>
<th>inc $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest Scores</td>
<td>.33</td>
<td>.40</td>
<td>2.18</td>
<td>.04</td>
<td>.13</td>
</tr>
<tr>
<td>Domain Knowledge</td>
<td>-.13</td>
<td>-.27</td>
<td>-1.64</td>
<td>.11</td>
<td>.07</td>
</tr>
<tr>
<td>Sampling Knowledge</td>
<td>.05</td>
<td>.03</td>
<td>.19</td>
<td>.85</td>
<td>.00</td>
</tr>
<tr>
<td>Achievement</td>
<td>-.16</td>
<td>-.16</td>
<td>-.91</td>
<td>.37</td>
<td>.02</td>
</tr>
</tbody>
</table>

Predictors are listed in order of entry into the regression model. Overall $R^2 = .22$; $F = 2.10$; $p = .10$
Experiments 1 and 2 offer convincing evidence for the power of teaching university students about the law of large numbers and some support for teaching them about reasoning errors. These results bolster arguments of Fong et al. and others (e.g. Cheng & Holyoak, 1985; Holland et al.; Nisbett et al., 1983; Fong & Nisbett, 1990; Lehman, Lempert, & Nisbett, 1988) who contend that teaching statistical inferential rules in a formal way can enhance adults' and young adults' reasoning about everyday problems to which statistical heuristics can be applied.

With the exception of Fong et al. who included grade 11 and 12 students in the subject pool of one of their experiments, previous studies investigated adult populations who, in most cases, had achieved at a high level in secondary school in order to gain entry to university. It may be argued that these higher-achieving subjects learn quickly and efficiently from limited instruction because they are strategic learners. What about less accomplished and younger learners? Is abstract rule instruction effective with such learners?

Along with increased use of good statistical thinking, the previous experiments suggest that reasoning errors are reduced with brief statistical instruction even if there is no direct instruction in errors. In fact, the effect sizes in posttest error rates in Experiment 2, while not significant at the .05 level, were consistently larger for the Rule Only condition. This lack of increased effect for instructing about errors may be due to minimal time spent in instruction about errors—15 min may not create adequate grounds to improve the quality of statistical reasoning and reduce reasoning errors.

The design of Experiment 3 extended the earlier two experiments by (a) taking the research to elementary and secondary schools to include students of all ability levels who
work in the general population of those schools; (b) redesigning and adapting existing materials, tests, and procedures so that they would be appropriate for younger learners and, importantly, so that they would be ecologically valid in school classrooms; and (c) developing additional data that would allow comparisons among learners at the three levels of schooling.

Four main questions are addressed in Experiment 3: (1) Will elementary and secondary students receiving instruction in statistical inferential rules make more use of those rules when reasoning about everyday events? (2) Will students receiving instruction in statistical inferential rules make fewer errors when reasoning about everyday events? (3) Will students receiving instruction in statistical inferential rules, as well as revised instruction in the reasoning errors, make more use of statistical rules and less errors when reasoning about everyday events than students not receiving the instruction in errors? (4) Does use of statistical inferential rules vary as a function of gender, achievement in mathematics, achievement in English, overall achievement in school, amount of knowledge about statistical concepts, and amount of knowledge about reasoning errors.

Method

Sample

Participants in this study were grade 7 students (N = 88) and grade 10 students (N = 63). The grade 7 students, age 12 and 13, were in the final year of their elementary school education. The grade 10 students, age 15 and 16, were in the final year of junior secondary school. Instruction and testing took place in the students' classrooms: four regular grade seven classrooms in two elementary schools; four social studies classrooms in one junior secondary school. It was felt that social studies provided a relatively "neutral" context in the subject-area classes available in secondary school. For example, a mathematics classroom
might have cued students to use statistical reasoning on the problems rather than any other heuristics they might normally use. The schools are in a mainly middle-class, suburban school district. The eight class sizes ranged from 18 to 32 with most class sizes approximating 30 students.

Within each class, students were ranked on achievement based on their most recent report card marks in English for junior secondary and reading and writing grades for elementary. Then, assignment to treatment groups was made by taking successive pairs of subjects from the ranked list and randomly assigning them. Since the experimental tasks involved a considerable amount of reading and writing, it was felt that this procedure would further ensure equivalency of the treatment groups. In addition to the English/reading and writing grade, a grade point average (GPA) was available for each subject in grade 10. This was an overall achievement indicator based on the average of scores on a 7-point scale for each course mark on a student's most recent report card: 7=A, 6=B, 5=C+, 4=C, 3=C-, 2=D, 1=F.

Students participated in conditions similar to those in Experiment 2: (a) Rule condition (grade 7, N = 45; grade 10, N = 30) and (b) Rule-Plus-Error condition (grade 7, N = 43; grade 10, N = 33) except that instruction occurred over two sessions instead of a single session.

Virtually all testing and training was conducted by the experimenter in classrooms over four sessions during a 1 1/2 week period in each school. Classroom teachers assisted in one phase of the experiment by supervising a review lesson. The total number of students in the eight classrooms was 203 with 151 yielding complete data for the present analysis. Some students, particularly at the junior secondary level, missed one or more sessions due to absences for a variety of reasons such as a band trip, student government, school office procedures, and illness.
Treatments and Controls

In the first session (50 min) both the Rule condition and the Rule-Plus-Error condition participated in a live demonstration of the law of good sampling—the label used in previous experiments was changed to more accurately describe the heuristics (See Appendix J for the revised script of the demonstration). A gumball machine was used, and the chart was reproduced on the familiar blackboard in each class. Then, both treatment groups read two examples and analyses (Appendix K), one from the probabilistic format and one from the objective format. To illustrate, the probabilistic item with its analysis is included below.

At a large high school, some students who were members of Student Council used a computer to pick names of other students. The computer randomly chose 450 students from the school. Then the Council members handed out a ballot to the 450 chosen students in homeroom classes. On the ballot, students were asked their opinion about increasing the cost of tickets to school dances by 2 dollars for each ticket. The extra money would help the school radio station buy new equipment and new music. After counting all the ballots, the Student Council members calculated that about 72 percent (72 out of every 100) of the 450 students said yes, they would be in favour of paying the extra money; 28 percent of the students said no.

The President of Student Council agreed that the extra 2 dollars for each ticket would help, but she said, "We have so many students in this school who really can't afford the extra money. They wouldn't go to dances if it cost more. I know that most of the students who were asked said yes, they would be willing to pay, but it's certainly far from sure that most of the students in the whole school would be willing to pay. A lot of people haven't been asked about it yet."

Do you agree with the Student Council President that it's "far from sure" that most of the whole student population would be willing to pay more for the tickets? Explain.

Please think about this problem for a few minutes. After you have considered the problem for a minute or two, turn the page for our analysis.

Analysis:

The members of Student Council are trying to find out about the attitude of the students at the high school toward spending more money on dance tickets. The money would be used to improve the school radio station. To use the words and ideas from The Law of Good Sampling, they are trying to find out the population distribution of high school students' attitudes toward paying more for the tickets. To do this, they randomly picked 450 students -
- a sample size of 450 — and asked the people in the sample if they were in favour of the price increase. 72 percent of the 450 said yes (and 28 percent said no).

The Law of Good Sampling says that the bigger the sample is, the better it is at estimating the population. Well, here there is a lot of evidence to show that most of the students in the whole school are in favour of the idea. Because, if you remember in the gumball demonstration, randomly picked samples of 25 were very good estimates of the population. It meant that the samples were not much different from the whole population. Well, here there is a sample size of 450 randomly picked students! This is an extremely accurate estimate of the population. From this, we know that most of the students in the whole school would pay the extra 2 dollars for dance tickets.

What about the President's argument that most students probably do not agree with the price increase because they can't afford it? Although this argument sounds OK, it is not an accurate one because she ignores the information available. In fact, the president's argument is contradicted by the large sample of 450 students.

In the second session (50 min), the half of the class assigned to the Rule Condition was supervised by their teacher in one classroom (see Appendix L for instructions provided to teachers) while the experimenter instructed the half of the class assigned to the Rule-Plus-Error condition. Subjects in both treatment conditions read a booklet (Appendix M) reviewing the concepts from the law of good sampling that were previously presented in the live demonstration. The two earlier examples with analyses were included in the booklet. A third, new example was also included. Next, both groups completed a Law of Good Sampling Quiz (Appendix N). The quiz presented a problem about a forester who wanted to determine the proportion of birch and oak trees in a forest using concepts from the law of good sampling. Following the problem, fifteen questions probing subjects' knowledge about correct application of the concepts were presented. The quizzes were collected, and the correct answers were read from an answer key to provide immediate feedback to the students. This part of the session was completed in 20 minutes. The Rule condition group spent the next 30 minutes working on regular homework or classroom assignments supervised by their teacher. The Rule-Plus-Error condition studied a booklet entitled Wild Boars and Thinking Mistakes (Appendix O). That material included explanations, examples and exercises—based on Experiment 2 materials—on the error categories identified in Experiment 1, as well
as a 20 item quiz to test students' knowledge of the categories. The example problems in the booklet illustrating reasoning errors were the ones previously studied by the students in the context of statistical rules. An acronym, BOAR, with each letter standing for an error type, was used in the material to help students remember the errors. The error types were referred to as "Wild BOARS". Figure 5 displays the error types as they were presented to subjects.

Measures

In addition to example problems, 12 isomorphic test problems were developed for the younger subjects of this experiment. Problems were pilot-tested in a grade 6 classroom in another elementary school in the same school district. Based on teacher-conducted guided-interviews with the grade six pupils and analysis of the written responses from students, revisions were made to improve readability of the items by reducing the number of propositions in some sentences and by replacing some vocabulary items. These were 6 pretest and 6 posttest problems (Appendix P) identical to the two formats and to three of the four structures used with undergraduates in Experiment 2. The false alarm problem (structure 4) was omitted. The 12 problems were randomly assigned to pretest and posttest occasions, and ordered, as in the previous experiment, so that no two formats or structures followed in succession. The sequence, formats, and structures are listed in Table 14.

In the directions of the pretest and posttest, students were instructed to underline information. Students, in most cases, provided these additional traces of their reasoning by underlining what they regarded as important information in the printed problems as well as in their written answers. These traces were used by the experimenter to supplement
Figure 5: Error Types as Presented to Grade 7 and 10 Students

Wild BOARS

B

B Blaming reasons Problem solvers sometimes blame the people or the person in the problem and say that they have a prejudice to act in a certain way. For example, they might say, "That individual is narrow-minded." or "They're ambitious, and that's why they act that way...."

O

O Opinion reasons Problem solvers sometimes think their own personal opinion is most important in answering the problem. For example, they might say, "I just don't agree with it." or "If that happened to me, I would...."

A

A Always reasons Problem solvers sometimes believe that some things, like a situation in the problem, are always true even when there might be another explanation. For example, they might say, "It's only human nature." or "Everybody makes mistakes."

R

R Reckless reasons Problem solvers sometimes make a wild guess and invent new information that is not really in the problem. This is just reckless, because it can result in a mistake, in this case, a thinking mistake.
Table 14

Pretest and Posttest Problems by Format and Structure

<table>
<thead>
<tr>
<th>Problem Sequence and Format</th>
<th>Problem Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1 Objective</td>
<td>X</td>
</tr>
<tr>
<td>2 Probabilistic</td>
<td></td>
</tr>
<tr>
<td>3 Objective</td>
<td></td>
</tr>
<tr>
<td>4 Probabilistic</td>
<td>X</td>
</tr>
<tr>
<td>5 Probabilistic</td>
<td></td>
</tr>
<tr>
<td>6 Objective</td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Structure 1 (conclusion from a single small sample)  
Structure 2 (large sample pitted against a plausible theory not founded on data)  
Structure 3 (regression)
written answers when a decision was made about a coding category. This proved especially helpful with these younger grade 10 and 7 subjects whose written responses were not as lengthy as those of university students.

Procedure

After meetings with classroom teachers and school principals, letters of permission and consent forms were sent home with students (see Appendix Q). In the first 50 min session, students completed the six pretest problems. In the second 50 min session, students participated in instruction as outlined in the previous section. Three student volunteers assisted with the demonstration by counting and reporting results of samples drawn from the gumball machine. These results were recorded on the blackboard by the experimenter. In the third 50 min session, the two condition groups were separated to engage in the instructional tasks described in the previous section. In the final 50 min session, subjects wrote answers to the posttest problems. The four sessions occurred within a one-week period at the elementary level and within a one and one half week period at the secondary level. This difference was due to timetabling constraints at the secondary level.

Coding and Scoring

The same coding and scoring system used in Experiments 1 and 2 was utilized in this experiment. Reliability analysis of pretest and posttest scales, conducted separately by grade, on statistical reasoning indicated low internal coherence when aggregating across items at the pretest level for grade 7 (alpha=.30) and strong internal consistency of items at the posttest level (alpha=.82). Respective results for grade 10 were alpha's of .52 and .77.
Results

Experiment 3, as Experiment 2, is a pretest-posttest, two-treatment, two factor design. However, unlike either of the previous experiments, subjects were assigned randomly in each classroom to condition groups. Mean scores in statistical reasoning and error commission were used for a repeated measures MANOVA. The two between groups factors are the treatment condition and the grade. The three within subjects factors are the pretest-posttest occasion, problem format, and measure (statistical reasoning and error commission). Multiple regression was conducted as an additional exploratory analysis.

The additional data generated by the present experiment allows tentative contrasts by level of schooling (grade) of "naturally occurring" (pretest) statistical reasoning and error commission as well as after-instruction effects (posttest). Such comparisons are more likely to be accurate between the grade 7 and 10 groups because problems and treatments were identical for those grades; but, with caveats, contrasts with the older, university students are also offered. These grade level comparisons of the results are described in Chapter 6. Some initial information is offered in the present chapter.

Effects of Instruction

The first two of the four main questions investigated in this experiment address the overall effects of instruction per se: (1) Will elementary and secondary students receiving instruction in statistical inferential rules make more use of such rules when reasoning about problems? (2) Will elementary and secondary students receiving instruction in statistical inferential rules make fewer errors when reasoning about problems?

Tables 15 and 16 present descriptive statistics on the quality of students' statistical reasoning in grade 10 and grade 7 respectively. These results are displayed by treatment
groups and problem formats. See Figures 6 and 7 for the plotted means for statistical reasoning as a function of instructional condition and problem format. For comparison, performance by university subjects in Experiment 2 on isomorphic posttest problems is also displayed.

When compared to the university undergraduates in Experiment 2—whose "natural" statistical reasoning, as measured by pretest results, is somewhat higher—the grade 7 and 10 students benefitted approximately as much from instruction as the university students did. Generally, the effect sizes, which reflect increases in secondary and elementary students' statistical reasoning after instruction, are similar to those obtained for the university students studied in the previous experiment. A closer examination of the pretest results by problem format suggests that the three levels of schooling engage in strikingly similar "natural" statistical reasoning on objective problems.

Tables 17 and 18, along with Figures 8 and 9 display descriptive statistics on the frequency of reasoning errors for the two grades. In addition, university students' performance from the previous experiment is pictured. As with the undergraduates, grade 10 and grade 7 students generally committed fewer errors in the posttest. However, two exceptions are apparent. Grade 10 students in the Rule-Plus-Error condition and grade 7 students in the Rule condition display a slight increase in reasoning errors. These Format by occasion interactions were tested with Spjotvall and Stoline's (1973) post hoc procedure and found not reliable.
Table 15

Descriptive Statistics on Grade 10 Students' Statistical Reasoning in Treatment Groups and Problem Formats

<table>
<thead>
<tr>
<th>Treatment Problem Format</th>
<th>N</th>
<th>Pretest M</th>
<th>SD</th>
<th>Posttest M</th>
<th>SD</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule Condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td>30</td>
<td>4.90</td>
<td>1.42</td>
<td>6.53</td>
<td>2.03</td>
<td>1.17</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>4.07</td>
<td>1.23</td>
<td>5.03</td>
<td>1.94</td>
<td>.78</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td>8.97</td>
<td>2.22</td>
<td>11.57</td>
<td>3.62</td>
<td>1.17</td>
</tr>
<tr>
<td>Rule Plus Error Condition</td>
<td>33</td>
<td>5.27</td>
<td>1.31</td>
<td>7.27</td>
<td>1.61</td>
<td>1.53</td>
</tr>
<tr>
<td>Probabilistic</td>
<td></td>
<td>4.30</td>
<td>1.31</td>
<td>5.49</td>
<td>1.56</td>
<td>.90</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>9.58</td>
<td>1.94</td>
<td>12.76</td>
<td>2.85</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Note. Pretest standard deviations are used to compute effect sizes.
Table 16

Descriptive Statistics on Grade 7 Students' Statistical Reasoning in Treatment Groups and Problem Formats

<table>
<thead>
<tr>
<th>Treatment Problem Format</th>
<th>N</th>
<th>Pretest M</th>
<th>Pretest SD</th>
<th>Posttest M</th>
<th>Posttest SD</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule Condition</td>
<td>45</td>
<td>4.51</td>
<td>1.10</td>
<td>6.00</td>
<td>1.98</td>
<td>1.35</td>
</tr>
<tr>
<td>Probabilistic</td>
<td></td>
<td>4.27</td>
<td>1.10</td>
<td>5.36</td>
<td>2.12</td>
<td>0.99</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>8.78</td>
<td>1.70</td>
<td>11.36</td>
<td>3.80</td>
<td>1.51</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule Plus Error Condition</td>
<td>43</td>
<td>4.61</td>
<td>1.14</td>
<td>5.84</td>
<td>1.98</td>
<td>1.08</td>
</tr>
<tr>
<td>Probabilistic</td>
<td></td>
<td>4.37</td>
<td>1.05</td>
<td>4.98</td>
<td>1.68</td>
<td>0.58</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>8.98</td>
<td>1.75</td>
<td>10.81</td>
<td>3.40</td>
<td>1.05</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Pretest standard deviations are used to compute effect sizes.
Figure 6

Plotted Means for Grade 10 Students' Statistical Reasoning as a Function of Instructional Condition and Problem Format

Note. Closed points (•) = performance on probabilistic problems; open points (○) = performance on objective problems. Dotted lines represent rule condition change. Solid lines represent rule + error condition change. Short horizontal lines represent performance of undergraduates.
Figure 7

Plotted Means for Grade 7 Students' Statistical Reasoning as a Function of Instructional Condition and Problem Format

Note. Closed points (●) = performance on probabilistic problems; open points (○) = performance on objective problems. Dotted lines represent rule condition change. Solid lines represent rule + error condition change. Short horizontal lines represent performance by undergraduates.
Table 17

Descriptive Statistics on Grade 10 Students' Reasoning Errors in Treatment Groups and Problem Formats

<table>
<thead>
<tr>
<th>Treatment Problem Format</th>
<th>N</th>
<th>Pretest M</th>
<th>SD</th>
<th>Posttest M</th>
<th>SD</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule Condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td>30</td>
<td>.73</td>
<td>.74</td>
<td>.33</td>
<td>.55</td>
<td>.54</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>.87</td>
<td>1.17</td>
<td>.57</td>
<td>.77</td>
<td>.26</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td>1.60</td>
<td>1.59</td>
<td>.90</td>
<td>1.03</td>
<td>.44</td>
</tr>
<tr>
<td>Rule Plus Error Condition</td>
<td>33</td>
<td>.42</td>
<td>.71</td>
<td>.18</td>
<td>.47</td>
<td>.34</td>
</tr>
<tr>
<td>Probabilistic</td>
<td></td>
<td>.79</td>
<td>.86</td>
<td>.85</td>
<td>.76</td>
<td>-.07</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>1.21</td>
<td>1.34</td>
<td>1.03</td>
<td>.95</td>
<td>.13</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Pretest standard deviations are used to compute effect sizes.
Table 18

Descriptive Statistics on Grade 7 Students' Reasoning Errors in Treatment Groups and Problem Formats

<table>
<thead>
<tr>
<th>Treatment Problem Format</th>
<th>N</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Rule Condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td>45</td>
<td>.58</td>
<td>.81</td>
<td>.31</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>.49</td>
<td>.70</td>
<td>.53</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td>1.07</td>
<td>.99</td>
<td>.84</td>
</tr>
<tr>
<td>Rule Plus Error Condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td>43</td>
<td>.42</td>
<td>.70</td>
<td>.21</td>
</tr>
<tr>
<td>Objective</td>
<td></td>
<td>.81</td>
<td>.79</td>
<td>.37</td>
</tr>
<tr>
<td>All Problems</td>
<td></td>
<td>1.23</td>
<td>1.15</td>
<td>.58</td>
</tr>
</tbody>
</table>

Note. Pretest standard deviations are used to compute effect sizes.
Figure 8

Plotted Means for Reasoning Errors in Grade 10 Students as a Function of Instructional Condition and Problem Format

Note. Closed points (•) = performance on probabilistic problems; open points (○) = performance on objective problems. Dotted lines represent rule condition change. Solid lines represent rule + error condition change. Short horizontal lines represent performance by undergraduates.
Figure 9

Plotted Means for Reasoning Errors in Grade 7 Students as a Function of Instructional Condition and Problem Format

Note. Closed points (•) = performance on probabilistic problems; open points (○) = performance on objective problems. Dotted lines represent rule condition change. Solid lines represent rule + error condition change. Short horizontal lines represent performance by undergraduates.
To examine relationships between statistical reasoning and error commission at pretest and posttest occasions, reasoning and error rates, aggregated over grades 7 and 10, were correlated. A Pearson correlation between overall statistical reasoning and error commission at the pretest level shows a near-zero correlation \((r = .08)\), identical to the pretest \(r\) for undergraduates in Experiment 2. The same correlation at posttest level, similar to undergraduates, reveals a negative correlation \((r = -.37; p = .01)\). These findings are similar to those in both Experiment 1 and 2. Elementary and secondary students, like university students, after instruction in the Law—or after instruction in the Law plus error types—were more likely to engage in statistical reasoning with accompanying reductions in reasoning errors.

**Effects of Rule versus Rule-Plus-Error Instruction**

The third main question investigated in Experiment 3 compares effects of the two treatment conditions: Will elementary and secondary students receiving instruction in statistical inferential rules as well as additional revised instruction in reasoning errors make more use of statistical rules and less errors when reasoning about everyday events than students not receiving the instruction in errors?

To test whether instruction in reasoning errors, when added to instruction in the law of good sampling, reliably improved the quality of statistical reasoning and reduced error commission in subjects' analyses of problems, and whether there were reliable differences in subjects' analyses when problem *formats* and *measures* of statistical reasoning and error commission were considered, a repeated measures MANOVA (see Table 19) on the data from 151 students was performed. A Box M test for homogeneity was conducted.
Table 19

Grade 7 and Grade 10: Results of Repeated Measures MANOVA Comparing Treatment Conditions on Statistical Reasoning and Reasoning Errors.

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Between</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>1.39</td>
<td>1</td>
<td>1.39</td>
<td>.52</td>
<td>.47</td>
</tr>
<tr>
<td>Grade</td>
<td>17.96</td>
<td>1</td>
<td>17.96</td>
<td>6.65</td>
<td>.01</td>
</tr>
<tr>
<td>Condition by Grade</td>
<td>4.50</td>
<td>1</td>
<td>4.50</td>
<td>1.67</td>
<td>.20</td>
</tr>
<tr>
<td>Error</td>
<td>397.15</td>
<td>147</td>
<td>2.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Within</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measures</td>
<td>6328.15</td>
<td>1</td>
<td>6328.15</td>
<td>1669.92</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Condition by Measures</td>
<td>3.76</td>
<td>1</td>
<td>3.76</td>
<td>.99</td>
<td>.32</td>
</tr>
<tr>
<td>Grade by Measures</td>
<td>4.25</td>
<td>1</td>
<td>4.25</td>
<td>1.12</td>
<td>.29</td>
</tr>
<tr>
<td>Co by Gr by Me</td>
<td>6.08</td>
<td>1</td>
<td>6.08</td>
<td>1.60</td>
<td>.21</td>
</tr>
<tr>
<td>Error</td>
<td>557.06</td>
<td>147</td>
<td>3.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td>28.46</td>
<td>1</td>
<td>28.46</td>
<td>30.80</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Condition by Format</td>
<td>.24</td>
<td>1</td>
<td>.24</td>
<td>.26</td>
<td>.61</td>
</tr>
<tr>
<td>Grade by Format</td>
<td>6.62</td>
<td>1</td>
<td>6.62</td>
<td>7.16</td>
<td>.008</td>
</tr>
<tr>
<td>Co by Gr by Fo</td>
<td>.00</td>
<td>1</td>
<td>.00</td>
<td>.00</td>
<td>.98</td>
</tr>
<tr>
<td>Error</td>
<td>135.82</td>
<td>147</td>
<td>.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occasion</td>
<td>81.64</td>
<td>1</td>
<td>81.64</td>
<td>79.77</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Condition by Occasion</td>
<td>.01</td>
<td>1</td>
<td>.01</td>
<td>.01</td>
<td>.94</td>
</tr>
<tr>
<td>Grade by Occasion</td>
<td>2.11</td>
<td>1</td>
<td>2.11</td>
<td>2.07</td>
<td>.15</td>
</tr>
<tr>
<td>Co by Gr by Oc</td>
<td>5.90</td>
<td>1</td>
<td>5.90</td>
<td>5.77</td>
<td>.02</td>
</tr>
<tr>
<td>Error</td>
<td>150.44</td>
<td>147</td>
<td>1.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measures by Format</td>
<td>96.15</td>
<td>1</td>
<td>96.15</td>
<td>91.59</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Co by Me by Fo</td>
<td>3.38</td>
<td>1</td>
<td>3.38</td>
<td>3.22</td>
<td>.08</td>
</tr>
<tr>
<td>Gr by Me by Fo</td>
<td>16.67</td>
<td>1</td>
<td>16.67</td>
<td>15.88</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Co by Gr by Me by Fo</td>
<td>.24</td>
<td>1</td>
<td>.24</td>
<td>.23</td>
<td>.63</td>
</tr>
<tr>
<td>Error</td>
<td>154.32</td>
<td>147</td>
<td>1.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measures by Occasion</td>
<td>163.65</td>
<td>1</td>
<td>163.65</td>
<td>110.87</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Co by Me by Oc</td>
<td>.07</td>
<td>1</td>
<td>.07</td>
<td>.05</td>
<td>.83</td>
</tr>
<tr>
<td>Gr by Me by Oc</td>
<td>2.17</td>
<td>1</td>
<td>2.17</td>
<td>1.47</td>
<td>.23</td>
</tr>
<tr>
<td>Co by Gr by Me by Oc</td>
<td>.16</td>
<td>1</td>
<td>.16</td>
<td>.11</td>
<td>.74</td>
</tr>
<tr>
<td>Error</td>
<td>2.16.99</td>
<td>147</td>
<td>1.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format by Occasion</td>
<td>4.73</td>
<td>1</td>
<td>4.73</td>
<td>9.99</td>
<td>.002</td>
</tr>
<tr>
<td>--------------------</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Co by Fo by Oc</td>
<td>.59</td>
<td>1</td>
<td>.59</td>
<td>1.26</td>
<td>.26</td>
</tr>
<tr>
<td>Gr by Fo by Oc</td>
<td>.02</td>
<td>1</td>
<td>.02</td>
<td>.04</td>
<td>.84</td>
</tr>
<tr>
<td>Co by Gr by Fo by Oc</td>
<td>.78</td>
<td>1</td>
<td>.78</td>
<td>1.64</td>
<td>.20</td>
</tr>
<tr>
<td>Error</td>
<td>69.55</td>
<td>147</td>
<td>.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Me by Fo by Oc</td>
<td>10.27</td>
<td>1</td>
<td>10.27</td>
<td>16.27</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Co by Me by Fo by Oc</td>
<td>.00</td>
<td>1</td>
<td>.00</td>
<td>.00</td>
<td>.96</td>
</tr>
<tr>
<td>Gr by Me by Fo by Oc</td>
<td>.70</td>
<td>1</td>
<td>.70</td>
<td>1.11</td>
<td>.29</td>
</tr>
<tr>
<td>Co by Gr by Me by Fo by Oc</td>
<td>.51</td>
<td>1</td>
<td>.51</td>
<td>.82</td>
<td>.37</td>
</tr>
<tr>
<td>Error</td>
<td>92.80</td>
<td>147</td>
<td>.63</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Condition (training in statistical heuristics and training in statistical heuristics plus training in reasoning errors) and Grade (grade 7 and grade 10) are between groups factors. Measures (statistical reasoning and reasoning errors); Format (probabilistic problems and objective problems); and, Occasion (pretest problems and posttest problems) are within subjects factors.
A multivariate test of significance ($s = 1$, $m = 3$, $n = 69$) for the condition effect yielded a nonsignificant Wilk's Lambda of $0.935$ ($p = .30$). Univariate F-tests performed on condition effects showed no reliable differences in statistical reasoning or error rates. Univariate F-tests performed on condition by grade interaction effects also yielded no reliable differences. As in Experiment 2, the MANOVA results for Experiment 3 reveal reliable differences in statistical reasoning between objective and probabilistic formats of problems. Subjects were more likely to reason statistically about problems from the probabilistic format.

**Statistical Reasoning Predicted by Other Variables**

Experiment 2 explored factors that might predict statistical reasoning by entering scores on a self-report of domain knowledge, a probe of sensitivity to sampling concepts, course marks as a proxy for achievement, and pretest scores into stepwise multiple regression analysis with statistical reasoning scores on the posttest serving as the dependent variable. A similar exploratory procedure was intended for the present experiment to address the fourth main question: Do uses of statistical inferential rules vary as a function of gender, achievement in mathematics, achievement in English, overall school achievement, amount of knowledge about statistical concepts, and amount of knowledge about reasoning errors. The correlation matrix for all variables is presented in Table 20. Grade 10 correlations are in the upper triangle; grade 7 correlations are in the lower triangle. Tables 21 and 22 display the results of the analyses.

Stepwise multiple regression equations predicted posttest scores in statistical reasoning from (a) pretest scores, (b) gender, (c) achievement based on English and math marks, and (d) results on knowledge of the concepts in the law of good sampling. For grade 10's, knowledge of the law of good sampling entered the equation first ($p < .001$), and achievement entered the equation second ($p < .001$). Pretest results entered third ($p < .001$).
This result suggests that the amount of knowledge—which may have resulted from instruction—about the concepts encapsulated in the law of good sampling, as measured by the quiz, is predictive of grade 10 subjects' statistical reasoning on everyday problems in the posttest. Statistical reasoning in the posttest appears to be well explained by this model (overall $R^2 = .65$). For grade 7s, the first predictor to enter the equation was achievement ($p = .001$). The second to enter the equation was pretest results ($p < .001$). The third was knowledge of sampling concepts ($p = .02$). Gender was not a reliable predictor at either grade level.
Table 20
Correlation Matrix for Variables in Grade 10 and 7 Multiple Regression Analyses

<table>
<thead>
<tr>
<th></th>
<th>pretest</th>
<th>posttest</th>
<th>achievement</th>
<th>sampling knowledge</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>pretest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>posttest</td>
<td>.578</td>
<td></td>
<td></td>
<td></td>
<td>.066</td>
</tr>
<tr>
<td>achievement</td>
<td>.505</td>
<td>.638</td>
<td></td>
<td></td>
<td>-.101</td>
</tr>
<tr>
<td>sampling knowledge</td>
<td>.391</td>
<td>.540</td>
<td>.554</td>
<td></td>
<td>.151</td>
</tr>
<tr>
<td>gender</td>
<td>-.092</td>
<td>-.184</td>
<td>-.182</td>
<td>-.041</td>
<td></td>
</tr>
</tbody>
</table>

Note. Grade 10 correlations are in the upper triangle. Grade 7 correlations are in the lower.
Table 21

Results of Stepwise Multiple Regression Predicting Posttest Scores in Statistical Reasoning for Grade 10 Students

<table>
<thead>
<tr>
<th>Predictor</th>
<th>b</th>
<th>β</th>
<th>t</th>
<th>p</th>
<th>inc R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Knowledge</td>
<td>.50</td>
<td>.37</td>
<td>4.10</td>
<td>&lt; .001</td>
<td>.42</td>
</tr>
<tr>
<td>Achievement</td>
<td>.37</td>
<td>.36</td>
<td>4.19</td>
<td>&lt; .001</td>
<td>.14</td>
</tr>
<tr>
<td>Pretest</td>
<td>.52</td>
<td>.33</td>
<td>3.82</td>
<td>&lt; .001</td>
<td>.09</td>
</tr>
<tr>
<td>Gender</td>
<td>.17</td>
<td>.03</td>
<td>.32</td>
<td>.75</td>
<td>.00</td>
</tr>
</tbody>
</table>

Predictors are listed in order of entry into the regression model.
Overall $R^2 = .65; F = 27.10; p = < .001$
### Table 22

Results of Stepwise Multiple Regression Predicting Posttest Scores in Statistical Reasoning for Grade 7 Students

<table>
<thead>
<tr>
<th>Predictor</th>
<th>b</th>
<th>β</th>
<th>t</th>
<th>p</th>
<th>inc R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement</td>
<td>.45</td>
<td>.34</td>
<td>3.41</td>
<td>.001</td>
<td>.41</td>
</tr>
<tr>
<td>Pretest</td>
<td>.65</td>
<td>.31</td>
<td>3.48</td>
<td>&lt;.001</td>
<td>.08</td>
</tr>
<tr>
<td>Sampling Knowledge</td>
<td>.25</td>
<td>.22</td>
<td>2.45</td>
<td>.02</td>
<td>.03</td>
</tr>
<tr>
<td>Gender</td>
<td>-.61</td>
<td>-.08</td>
<td>-1.09</td>
<td>.28</td>
<td>.01</td>
</tr>
</tbody>
</table>

Predictors are listed in order of entry into the regression model. Overall $R^2 = .53; F = 23.18; p = < .001$
CHAPTER 6

DEVELOPMENTAL TRENDS ACROSS THE THREE EXPERIMENTS

The three experiments conducted in this study investigated statistical reasoning and reasoning errors by students at three levels of schooling: university, secondary, and elementary. The final research question of the investigation asked: What developmental trends in statistical reasoning exist among elementary, secondary, and university-aged students before and after instruction about statistical heuristics and reasoning errors. Also of interest was whether students at these three age levels responded differently to similar instruction about heuristics and reasoning errors, and whether the levels of schooling differed in how other variables, such as student achievement and knowledge about sampling, might predict statistical reasoning.

Figures 10 and 11 present plotted means for statistical reasoning and reasoning errors, aggregated across formats, as a function of instructional condition and level of schooling (grade). In these graphs, all three grades display similar degrees of improvement in statistical reasoning. In pretest and posttest problems when the format results are combined, undergraduates in both instructional conditions, as well as grade 10's in the Rule-Plus-Error condition, reason more statistically than grade 10's in the Rule condition and Grade 7's in either condition. In addition, the three grades display similar rates of decrease in reasoning errors with one exception: Grade 10's in the Rule-Plus-Error condition committed fewer errors at the pretest level than grade 10's in the Rule condition but more at the posttest level. A similar result was achieved by Grade 7's in the Rule condition when compared to Grade 7's in the Rule-Plus-Error condition. Furthermore, unlike the results in statistical reasoning, the aggregated rates of error commission did not differentiate university students from grade 10's and grade 7's.
Figure 10

Plotted Means and Interaction Effects for Statistical Reasoning in All Problems as a Function of Instructional Condition and Level of Schooling

Note. Closed points (•) within short lines represent performance by undergraduates; open points (○) show performance by Grade 10's; squares (□) display performance by Gr. 7's. Dotted lines represent rule condition increase. Solid lines represent rule + error condition increase.
Figure 11

Plotted Means and Interaction Effects for Reasoning Errors in All Problems as a Function of Instructional Condition and Level of Schooling

Note. Closed points (•) within short lines represent performance by undergraduates; open points (○) show performance by Grade 10's; squares (□) display performance by Gr. 7's. Dotted lines represent rule condition increase. Solid lines represent rule + error condition increase.
In the preliminary grade level comparisons outlined in Chapter 5, statistical reasoning was more specifically investigated by problem format (Figures 6 and 7) instead of by combining formats. It was observed there that, compared to undergraduates—whose natural statistical reasoning, as measured by pretest results, is somewhat higher—the grade 7’s and 10’s benefitted approximately as much as undergraduates did from instruction. The similarity in improvement is reflected in the combined problem format results displayed in the Figure 10. However, a closer inspection of the previous format-specific outcomes suggests that the "natural" propensity of university students to reason more statistically applies only to problems from the probabilistic format. The three grade levels are similar in their performance on objective problems in both pretests and posttests as shown by Figures 6 and 7.

In the previous chapter, reasoning errors were also plotted by format (Figures 8 and 9). All three grades experienced reductions in errors after instruction, and no systematic differences among grades are apparent in the graphs. Generally, there is a trend toward more errors on objective problems than on probabilistic problems in both pretests and posttests. The same result was observed in Experiment 1.

To further examine grade level differences, mean scores in statistical reasoning and errors were used for a repeated measures MANOVA. The two between group factors are treatment condition and grade. Occasion and format are the two within subjects factors.

Pretest Comparisons

It has been suggested that pretest results may be regarded as "natural" reasoning. Since no instruction took place, students' written responses to isomorphic or identical problems from everyday contexts are samples of "natural" reasoning at each grade level.
A multivariate test of significance \((s=2, m=1/2, n=89)\) for the pretest grade effects yielded a significant Wilk's Lambda of \(.694 (p < .001)\). Univariate F-tests on the grade effects revealed a reliable difference in grades on probabilistic problems \((F=34.53; df^2, 183; p < .001)\). This result confirms the observed plotted differences between university and younger subjects. Post hoc tests revealed that undergraduates engaged in more statistical reasoning on probabilistic problems in the pretest than both grade 10's and 7's but not more statistical reasoning on objective problems. No reliable differences in pretest results in statistical reasoning were found between grade 10's and grade 7's. Finally, there were no reliable differences in reasoning errors on the pretest among the three levels of schooling.

**Posttest Comparisons**

A multivariate test of significance \((s=2, m=1/2, n=89)\) for grade effects on posttest results yielded a significant Wilk's Lambda of \(.751 (p < .001)\). Univariate F-tests indicate reliable differences in statistical reasoning on probabilistic problems \((F=18.89; df^2, 183; p < .001)\); objective problems \((F=7.26; df^2, 183; p = .001)\); and, errors on objective problems \((F=3.19; df^2, 183; p = .04)\). However, subsequent post hoc tests indicated that none of these differences were reliable at the .05 level.

**Comparing Predictors of Statistical Reasoning**

Experiment 2 explored factors that might predict statistical thinking about everyday problems in university students. Factors such as students' knowledge about sampling or students' professed knowledge about the content domains of problems did not reliably predict statistical reasoning in that experiment. Additionally, student achievement did not provide a reliable predictor. For the university undergraduates, only pretest scores entered the multiple regression equation as a reliable predictor of posttest results in statistical reasoning.
In contrast, the results of Experiment 3 suggest that student achievement and knowledge about the law of good sampling are good predictors of success in posttest statistical reasoning.
CHAPTER 7
DISCUSSION

The idea that traditions of inquiry—including the scientific tradition—are logically necessary for human reasoning (Barrow, 1991) implies the existence of abstract inferential tools that are transferable across specific content domains to help learners sift, organize, and use information. One aspect of this broad question concerns how people might improve their thinking about problems containing statistical information. Therefore, it is important to research how people reason about such problems and whether formal canons of inference arising from the scientific tradition can improve reasoning. Statistical heuristics such as the law of large numbers are among these pragmatic inferential rules. Statistical thinking is also part of the scientific tradition. More fundamentally, the law of large numbers—the basic notion that large samples are more reliable than small samples—is thought to underly concept formation and generalization (Fong et al., 1986).

It is particularly important to investigate how young people in schools can learn to reason more accurately because public education's policies increasingly emphasize the teaching of problem solving skills across subject areas. In addition, there is a growing literature, established by psychologists, that describes how reasoning in everyday situations is affected by instruction in formal rules of inference (Lehman et al., 1988; Holland et al., 1989; Fong & Nisbett, 1990). Given the ubiquity of the law of large numbers; its place in the scientific tradition of inquiry; its recently researched role in promoting accurate reasoning about everyday problems; and the recognition by policy-makers that instructing students to solve problems is an increasingly important aim of schools, the present study is a timely one.
Previous Research and Present Findings

The present study confirms many of the findings of previous research with adult populations and extends the investigation of teaching abstract inferential tools, specifically statistical heuristics, to elementary and secondary school students. Jepson et al.'s (1983) work with university students found that inductive reasoning will be guided by statistical inferential rules when people think about problems in which events are readily evaluated with respect to their variability and randomness. This was confirmed here by the results of Experiments 2 and 3 with university, grade 10, and grade 7 students. Pretest results from those experiments reflect students' "natural" statistical reasoning and indicate that they more regularly used statistical inferential rules in probabilistic problems as opposed to objective problems. The latter is a format of everyday problems in which variability and randomness are more difficult to detect. Jepson et al.'s results also are supported by the findings of Experiment 1, where the control group, which received no instruction, applied more statistical reasoning to probabilistic problems than to objective and subjective items.

Nisbett et al.'s (1983) finding that domain-specific expertise has a strong positive correlation with statistical reasoning also was investigated in the present study. It was found that university students' self-ratings of domain knowledge related to problems did not reliably predict statistical reasoning. What remains largely unexplored is the extent to which expertise in a content domain, that is domain-specific knowledge, increases awareness of the variability of events and the role of chance in producing events. Additional research in this area is required.

Fong et al.'s (1986) finding, that training greatly enhanced statistical reasoning in problems ranging from more obviously probabilistic ones to more difficult-to-assess subjective ones, was also confirmed. As in their study, no significant interaction between problem formats seen in training and test problems' formats was found here. That is, training
effects were format-independent ("domain-independent", according to Fong et al.). However, there was a slight trend to suggest that using objective and subjective examples in "guided induction" of the law of large numbers improved statistical reasoning more on test problems from the subjective format. Probabilistic example problems, while resulting in significant improvement for all three formats, did so to a lesser degree in the subjective format. This difference was not reliable at the .05 level. Exploring this trend is recommended for a future study. Overall, Fong et al.'s conclusions about people's use of abstract statistical heuristics in everyday problem solving was replicated. Assertions about the ubiquity and generalizability of the law of large numbers from one format to another are upheld.

Lehman et al. (1988) found that university courses in probabilistic sciences (social science, psychology) sensitized graduate students to errors in everyday statistical reasoning, such as making inferences from small samples. Fong et al. (1986) found that statistical education, provided in an introductory-level statistics course, enhanced the use of statistical rules in everyday situations. In the latter experiment, students enrolled in an undergraduate statistics course were tested at the beginning and were found to have enhanced statistical reasoning at the end of the course. The present findings contradict both of these studies. However, some important differences may account for this variation among prior studies and this one. First, unlike Lehman et al.'s findings with graduate students, the present results apply to courses in probabilistic sciences available for undergraduates. It was found that students who had completed a greater number of these courses did not reason more statistically. Second, unlike Fong et al.'s findings, which were based on measurements immediately adjacent to a statistics course, in the present study, students enumerated courses they had completed in previous semesters, and these were used to predict statistical reasoning on test problems. This approach searched for long-term benefits similar to those studied by
Lehman et al. Again, however, Lehman et al.'s results are based on graduate programs where probabilistic reasoning is studied much more intensively.

Fong and Nisbett's (1990) finding, again with undergraduates, that training effects are due to subjects' memory for the formal rule appears to be supported here. Grade 10 students, who demonstrated a stronger memory for the formal rule on the Law of Good Sampling Quiz, reasoned more statistically on posttest problems than students who could not remember the rule.

In summary, comparing the results of previous studies with those of the present study confirms many, but not all, of the earlier findings. Differences usually can be accounted for, however. What follows is a more detailed discussion of the current results.

General Discussion

While supporting major findings in previous studies, the present work extends earlier research by (1) including younger learners; (2) taking the research into ecologically-valid classroom contexts; (3) examining competing rules or heuristics and how instruction might reduce the frequency with which common reasoning errors invade reasoning in everyday problem solving; and (4) investigating interactions with other variables such as gender, achievement, and knowledge about statistical laws. Six main questions were addressed in this study.

1. Will students receiving instruction in statistical inferential rules make more use of such rules in reasoning about everyday events when compared with students not receiving instruction?

Reliable main effects of instruction for statistical reasoning were obtained in all three experiments. Students in the instructional conditions learned a good deal about how to reason statistically in all three problem domains. Large effect sizes were obtained, and this generates
new evidence for the generalizability of abstract inferential rules. Teaching students of various ages in a formal way about the law of large numbers had a considerable effect on the way they reasoned about problems of everyday life, even when instruction was limited to one or two brief sessions.

These improvements in students' reasoning lived beyond the training session. In Fong, et al.'s study, subjects were tested immediately after training. This created a suspicion, as the authors pointed out, that this "temporal relation ... might be essential in order to produce any effects of training at all."(p.275) The present study tested undergraduates at least one week after training and still found large effects. Future work could investigate how enduring the results are and if statistical reasoning is applied in contexts outside school classrooms.

Pretest and control group data show that people use statistical rules to solve problems prior to instruction. The contention of Nisbett and others that these primitive rules are improvable by instruction may help to explain why only brief formal instruction is necessary for impressive results. The existing rules may be highly improvable because of their primitiveness. Kahneman and Tversky's research attests to people as poor intuitive scientists, so the margin of "improvability" may be great. Thus the three experiments may have resulted in impressive gains in statistical reasoning for students because statistical heuristics were already part of the students' inferential repertoire and were also readily improvable.

Errors of inductive reasoning involving statistical rules may be better understood as students' inability to code events in terms that make contact with statistical rules rather than simply ignorance of the rules themselves. According to Holland et al. (1989), the newly inserted rules will tend to act like induced rules and enter into competition for the right to represent the environment along with other, existing rules in the system's possession. Typically, the newly-introduced rules will not compete successfully with older induced rules if their application requires encoding that the system has not yet mastered. Rubin et al. (1990)
characterize this difficulty as follows: "... as soon as non-determinism enters the classroom—as it does in the study of sampling—confusion becomes more common and more tenacious." (p. 2)

Rubin et al.'s view is contradicted by the results of the present experiments which brought non-determinism to classrooms in a university, a secondary school and two elementary schools. The use of example problems may partly account for this contrary outcome. After formal instruction in the law of large numbers, the examples may have provided learners with coding rules to link modifications in their existing rules to events in the example problems. The scripted instruction about the law of large numbers and examples that elaborate the instruction act very much like the demonstrations, recommended by Holland et al., in which students make predictions about the likely outcomes of events to articulate bases of their predictions, and then explore the possibility that other rules might exist that would do a better job of prediction. In Experiments 1 and 2, students made their original predictions and articulated grounds for them in the pretests. They then explored the possibility that other rules (the law of large numbers) might exist that would do a better job of prediction. Thus, the testing and instructional interventions parallel Holland et al.'s recommendations and may have assisted students to master encoding that could then compete successfully with older, induced, deterministic rules.

2. Will students receiving instruction in statistical inferential rules make fewer errors in reasoning about everyday events when compared with students not receiving instruction?

Reductions in frequency of errors was obtained in all three experiments. Not only did subjects engage in improved statistical reasoning as a result of instruction in the law of large numbers, they made fewer reasoning errors identified from the data of the first experiment. Effect sizes obtained for error reduction are moderate but consistent across grade levels. In
this way, Shulman and Elstein's (1975) call for research on how typical error patterns in problem solving can be identified and reduced is answered.

The issue of whether these particular categories of errors represent the most important competing heuristics at all three levels of schooling is problematic for at least three reasons. First, out of the four error categories identified in the initial investigation, Speculation was not used in statistical analyses because it was too difficult to delineate for coding purposes. This category shows promise but requires further conceptualizing. Second, the other three categories—Attribution to Disposition, Egocentric Bias, and Truism—identified as types of errors committed by undergraduates, were later used in analyzing grade 10 and grade 7 results. Additional error types that might be specific to those lower grades were not explored in the present study. Such exploratory work is recommended. Third, if the errors represent deterministic reasoning that "naturally" competes with statistical heuristics, then a negative correlation with statistical reasoning at pretest level might be expected. This was not the case. Statistical reasoning and errors correlate at a near-zero level for all grades in pretests and negatively in posttests.

3. Does the use of statistical inferential rules vary as a function of the problem format emphasized during instruction?

There were reliable differences in statistical reasoning between problems in the probabilistic and objective formats at the pretest level in both Experiments 2 and 3, as well as in the control group studied in Experiment 1. Subjects were more likely to reason statistically about problems from the probabilistic format than about problems from the objective format. This lends some support to the definitions of the "domains" offered by Nisbett, Fong, and others. There was consistent differentiation among the three problem formats in the way
people normally apply statistical reasoning to everyday problems at pretest and in the way people do so in a posttest after instruction.

Undergraduates performed no better on pretest problems from the *objective* format than students in grade 10 and 7. This implies that, without instruction, applying statistical thinking to problems where there is no explicit cue about randomness of the sample is no more advanced for these older, more educated students. This finding may not be surprising when we consider studies of the impact of schooling on informal reasoning in students (e.g., Perkins, 1985).

Experiment 2 obtained appreciably larger effect sizes on statistical reasoning in *objective* problems than Experiment 1. Comparing the results of the two experiments should be done with care because population and procedures were not identical. Given these caveats, the most notable difference between the two experimental procedures was the administration of a pretest in Experiment 2. As mentioned previously, the pretest may have provided students with opportunity to contrast their own heuristics with the subsequently learned alternative statistical rules although pretest sensitization was tested and not revealed. In school settings, where students often search for what is expected of them, this opportunity to re-evaluate isomorphic problems may partly account for the increases in statistical thinking on *objective* problems. The pretest opportunity may still be a valuable learning experience in its own right, and may contribute to the larger effect sizes in *objective* problems in Experiment 2.

4. *Does the use of statistical inferential rules vary as a function of students' achievement, previous coursework, gender, general reasoning ability, statistical knowledge, knowledge about reasoning errors, and domain knowledge?*

Exploring relationships between students' use of statistical inferential rules and these variables is concerned with generating new descriptive information about what factors might influence the improvement of human thought. This question asks whether formal education,
training in statistics, conditional reasoning ability, declarative knowledge about sampling, awareness of errors in reasoning, and expertise in a content area are predictive of statistical reasoning. In addition, the question explores gender differences in statistical reasoning. Describing such relationships may assist in generating testable propositions and hypotheses about how people of different ages learn to reason more evidentially and accurately. However, it should be emphasized that many of the predictor variables used in this research are limited indicators. For example, the final mark in the educational psychology course was used as a proxy for subjects' achievement in Experiments 1 and 2, when students' overall GPA, had it been available, might have served as a better indicator. In addition, self-ratings were used to indicate knowledge of problem content domains. A more accurate measure to assess subjects' knowledge of these areas may be required. A 5-minute probe of sampling knowledge was administered to undergraduates in Experiment 2. A more comprehensive instrument to measure students' intuitions about sampling may be needed. Furthermore, because of the small number of male subjects available in the university investigations, gender was not pursued as a predictor in the first two experiments. Finally, the low rate of success on the Wason card problem precluded its use as an indicator of conditional reasoning ability. In contrast to these shortcomings, the predictor data in the third experiment are generally more valid and reliable because better measures and indicators were available. With these warnings, results of the multiple regression analyses are briefly discussed here primarily as a source of tentative suggestions for future research.

In Experiment 1, statistical reasoning on everyday problems by university undergraduates was best predicted by student achievement. The number of courses students had completed in statistics, mathematics, or psychology was not a reliable predictor. Also, student major, categorized as psychology, humanities, or other, was not predictive of statistical reasoning performance. In the first experiment, pretest scores were not used as a predictor.
In Experiment 2, pretest scores were entered into the regression model along with achievement, domain knowledge, and sampling knowledge. Pretest scores provided a reliable predictor of statistical reasoning in this case while the other variables were not.

In Experiment 3, four variables were entered into the multiple regression model to predict performance by the younger subjects: pretest results, achievement, sampling knowledge, and gender. For Grade 10's, knowledge of sampling, as measured by the Law of Good Sampling Quiz, provided the best predictor. For Grade 7's, achievement provided the best predictor.

The results are limited to posttest performance in statistical reasoning. Generally, they suggest that higher achieving students, students who already engage in higher levels of statistical reasoning, and students who learn most about the law of large numbers are likely to do well on the posttest.

5. Will students receiving both instruction in statistical inferential rules and instruction in reasoning errors make more and better use of such rules and commit fewer reasoning errors when compared with students receiving only instruction in statistical inferential rules?

This question applies to Experiments 2 and 3 where the Rule and Rule-Plus-Error instructional conditions were used. In Experiment 2, instruction in the Rule-Plus-Error condition improved statistical reasoning on objective problems to a reliably greater degree than the Rule condition. There were no reliable differences in reasoning on probabilistic problems and no reliable differences in error rates.

The expectation that additional training in errors would help to reduce errors in general did not materialize in these results. The additional instruction served to sharpen the statistical reasoning skills of subjects only on objective problems. This result may be due to the nature of the error instruction. Examples used in the instruction about the law were also used in
instruction about error categories. The increased time students spent with these examples in the Rule-Plus-Error condition may account for the difference. However, this result may also support the idea that error instruction taught students about heuristics that do compete with statistical heuristics in objective problems.

In Experiment 3, no reliable differences in condition effects were revealed. Also, no reliable condition by grade interaction effects were unearthed. These findings apply to both statistical reasoning and error rates. One explanation for these results may be that the categories of reasoning errors may not be entirely age appropriate for secondary and elementary school populations. As described previously, the categories were generated from undergraduate protocols from Experiment 1. Students in the younger grades may commit different types of errors when they ignore or underutilize statistical information in problems. There may exist types of deterministic reasoning among younger students that are more prevalent than those identified from the data generated by undergraduates.

It should be noted that, in Experiment 2, effect sizes in statistical reasoning for the Rule-Plus-Error group are consistently higher than for the Rule group, but the Rule group revealed higher effect sizes in error reduction, especially on objective problems. It seems that additional error instruction increases errors in these types of problems. There is no ready explanation for this.

In summary, the expected effects for instruction about reasoning errors did not appear with the exception of the results in objective problems for undergraduates. The lack of treatment group differences may simply be due to the power of learning the law of large numbers and the comparative weakness of additional error instruction to either reduce errors or improve reasoning. Furthermore, work in identifying other competing heuristics at each of the grade levels appears warranted.
6. What developmental trends in statistical reasoning exist among elementary, secondary, and university-aged students in the absence of instruction as well as after training, and how do these groups differ in statistical reasoning and reasoning errors?

The previously mentioned warnings about comparing the results of the three experiments offered under Research Question 3 apply here as well. Apparent overall similarities in the performance of the three levels is one of the most striking findings of this study. It was expected that developmental differences in statistical reasoning would become apparent, especially in pretests results. Such differences were revealed on probabilistic problems, but only undergraduates performed significantly higher in statistical reasoning. There were no reliable differences between grade 10 and 7 students. The most striking similarity is in the statistical reasoning performance of the three grade levels on objective problems in the pretest. Here, the grades were almost identical. In addition, there were no reliable differences among the three grade levels in reasoning errors on the pretests. This seems to indicate that the types of reasoning errors identified in Experiment 1 are applicable across the grades; but, again, the idea requires further research.

After instruction, the three grade levels demonstrated similar rates of increase in statistical reasoning and reduction in reasoning errors as shown in Figures 9 and 10. Post hoc testing posttest results yielded no reliable differences among the grades.

Implications

Over a decade ago, Newell (1980) called for experimental and theoretical research that would deepen our understanding of problem solving. This recommendation was made in the context of a distinction that described general problem solving methods as induced from daily life, as opposed to domain-specific problem solving methods that could be taught. The
present results suggest that research in a third, intermediate approach to teaching generalizable problem solving methods may contribute to deepening our understanding of problem solving.

The idea that problems are can be described as "well-structured" and "ill-structured", but also "structured" (Frederiksen, 1984), undergirds this intermediate approach and supplements Newell's earlier view of problem solving as resting on domain-specific or general methods. The present findings support a notion that people at several levels of schooling possess naive inductive inferential rules, at a high level of abstraction, for statistical reasoning. These rules are similar to formal principles for thinking in the scientific tradition and they underlie many topics taught in schools. Improving solutions to structured problems that call for statistical reasoning requires that the problem solver represent statistical information within a problem format. The findings of the three experiments reported here suggest that people of various ages can be instructed to represent such problems more accurately by enhancing naive statistical heuristics that they already know. This approach may well generalize to attempts to solve problems containing novelty or uncertainty that are encountered in everyday life.

Extending this study's findings to current practices in schools suggests some means for enhancing problem solving with relatively minor shifts in curricular and instructional foci. For instance, teaching probability in the grade six and seven mathematics curriculum might be accompanied by examples of everyday problem solving in which students are guided to use statistical information in their solutions. Teaching scientific methods in secondary grades' science classes might be accompanied by instruction in sampling and its applications to real-life situations. By extension, one may speculate that other naive inferential systems, such as causal and contractual rules, may be amenable to improvement by instruction and may generalize to multiple disciplines currently taught in schools.
What needs to be studied further—within the cognitive-mediational paradigm—are students' cognitions in learning about such heuristics. How do newly introduced rules compete with old rules? What is the nature of students' existing prior knowledge and beliefs about statistics? Finally, additional work to improve our understanding of the mediational dimension is required: What teacher behaviours would best assist students in improving statistical, or other, reasoning about everyday problems?

Conclusion

Contemporary psychology has largely adopted a belief, in the tradition of Thorndike and James, that useful problem solving rules are domain specific and will not transfer to other domains. The results of this series of experiments suggests otherwise—some inferential rules are generalizable and can be taught by fine-tuning statistical heuristics that people have already induced by virtue of daily experiences. Instructing people about such pragmatic inferential rules seems to enhance rule systems that people already possess. Identifying other systems of naive rules and finding out whether they can be enhanced to improve problem solving may contribute to both the psychology of human reasoning and to improving human affairs.
REFERENCES


Barrow, R. (1991, February 12). Educating intelligence. The President's Faculty Lectures, Simon Fraser University, Burnaby, B.C.


APPENDIX A
Example Problems and Analyses for Experiment 1 (dotted lines signify page breaks in the training materials; subheadings labelling problem types were not included)

Probabilistic Example Problems

One reason that the Law of Large Numbers is important to learn is that it applies not only to urns and gumballs. The basic principles involved in the Law of Large Numbers apply whenever you make a generalization or an inference from observing a sample of objects, actions, or behaviors.

To give you an idea of how broad the Law of Large Numbers is, we have, in this packet, presented three situations in which the Law of Large Numbers applies. Each situation is analyzed in terms of the Law of Large Numbers.
Patricia was playing a new Monopoly-like board game called "Margin," for the first time. One of the main features of this game is that each player has the option to buy "disaster insurance." Unfortunately, this insurance is extremely costly and tends to cripple one's "investment" program. But not buying it is very risky--one can be wiped out by a "disaster."

"Disasters" come about when picking from a very large pack of cards called "fate cards." Some of these cards have neutral or favorable results, others have minor or major disasters. Each time a player has to draw a "fate card," the card is returned afterward to the pack and the pack is shuffled.

Patricia didn't have any detailed knowledge of the kinds of "fate cards" and so at first she decided to buy insurance as soon as she got near the first place on the board where she might have to draw a "fate card," with no major disaster. (One got a neutral result and one a minor disaster.) Patricia decided on that basis to continue for a while with no insurance.

Comment on the thinking that led Patricia to this decision. Is it basically sound? Does it have weaknesses?

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.
Patricia is trying to draw a conclusion about a certain population. We can think of the members of this population as fate cards. If we divide the members of this population into two categories, "neutral or minor trouble" and "major disaster," we can think of the population distribution as the % in each category. Patricia has concluded that the percentage in the "neutral or minor trouble" category is high enough to justify ignoring the risk and skipping insurance. This conclusion was based on observing a sample of size = 2, in which the sample distribution was 100% "neutral or minor trouble," 0% "major disaster."

Apart from any other considerations, however, the sample distribution for size 2 is apt to be quite different from the population distribution: The latter could be only 60% or 50% or even as low as 40% "neutral or minor trouble" and a 2-0 sample split would not be so unusual; just as one would not be at all amazed to draw 2 out of 2 silver gumballs from an urn with only 40% silvers. So Patricia's attitude is quite unwarranted: a larger sample is needed.

Brian has lived in Japan for several years, but never has been under pressure to learn the language since, in his business activities, most people speak English quite well. When he vacations or eats out downtown, he takes the totally unreadable menu and blindly lets a pencil fall on it, selecting the entree that is nearest the pencil-point. Sometimes he obtains really excellent meals in this way. When this happens, Brian tries to eat at the excellent restaurant as often as possible. However, he finds that when he does this, he is usually disappointed. Subsequent meals that are chosen by the same pencil-point method rarely seem as good to him as the first meal, and his ultimate opinion of the restaurant is usually merely that it is "good" or "better than average."

Why do you suppose that Brian usually has to revise downward his opinion of restaurants that he initially thought were excellent?

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.
Each restaurant can be thought of as consisting of a population of meal-experiences. Since Brian is interested in true excellence, we can divide the members of the population into two categories: "really excellent" meal-experiences and "non-excellent" meal-experiences ("good," "better than average," and so on). We can think of the population distribution as the percentage in each category. For many restaurants, the population distribution is actually 0% excellent, 100% non-excellent: there is no "really excellent" meal-experiences to be had there. For many others, there is a small or moderate percentage of "really excellent" gumballs in the urn. Recall that Brian randomly selects his meals at restaurants. When he obtains an excellent meal, it is usually because he has stumbled onto one of the small number of restaurants that can offer "really excellent" meal-experiences and has drawn one of the "lucky gumballs." Such an excellent meal is just a sample of size = 1, and therefore, according to the Law of Large Numbers, is not a reliable indicator of the true quality of the restaurant: it does not indicate that the population distribution for that restaurant is 100% "really excellent," or even that it is predominantly "excellent." It is reasonable to think that there are really very few restaurants that have population distributions with a large percentage of "excellent" meals; and so when Brian does get such a meal, the chances are it is just a lucky draw in a restaurant that has some, but not many, "excellent" meal-experiences. Therefore, when he goes back to the same restaurant, and takes a larger sample, most of the additional meal-experiences from such a restaurant are not as good as his first.
At a large urban university, a student organization used the computer to select 1500 student numbers at random, and contacted those selected to ask whether they were willing to have $20 added to their fees for the construction of a new gym. About 72% of the 1500 students questioned said they would favor such a fee in order to have expanded athletic facilities. Although the Dean of Student Services at the university agreed the $20 per student would probably be adequate to pay the annual mortgage on a new gym, he argued that: "Since there are a very large number of students at our university who are from lower and lower-middle class families, a $20 fee would be quite a hardship on them. Thus, it is very unlikely that a majority of students at the university would be willing to have $20 added to their fees for the new gym. While a majority of the people you asked were in favor of the fee, it is far from certain that a majority of the entire student population is in favor of the fee. A great many other people have not been consulted.

Do you agree with the Dean that it is "far from certain" that a majority of the student population favors the fee for the new gym? Explain.

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.

The student organization is trying to assess the attitude of the students at the university toward spending $20 per student to build the new gym. In terms of the Law of Large Numbers, they are trying to find out the population distribution of the university students' attitudes toward the fee increase. To do this, the organization randomly selected 1500 students—a sample of size 1500—and asked them if they were in favor of the fee for the new gym. Of the 1500, 72% were in favor of it (and 28% were against it). According to the Law of Large Numbers, which states that the larger the sample, the better it is in estimating the population, samples of size 25 were very good estimates of the population: samples did not differ much from the population. Extending the argument, samples of size 1500 are extremely accurate estimates of the population. Thus, it can be concluded that a majority of students at the university are in favor of the fee.

What about the Dean's argument that it is unlikely that a majority of the student population favors the fee because they would consider a $20 fee a hardship? Although this argument may have intuitive appeal, it should be discounted because it is not supported by any data, and is in fact contradicted by the large sample of 1500 students.
Objective Example Problems

One reason that the Law of Large Numbers is important to learn is that it applies not only to urns and gumballs. The basic principles involved in the Law of Large Numbers apply whenever you make a generalization or an inference from observing a sample of objects, actions, or behaviors.

To give you an idea of how broad the Law of Large Numbers is, we have, in this packet, presented three situations in which the Law of Large Numbers applies. Each situation is analyzed in terms of the Law of Large Numbers.

A major New York law firm had a history of hiring only graduates of large, prestigious law schools. One of the senior partners decided to try hiring some graduates of smaller, less prestigious law schools. Two such people were hired. Their grades and general record were similar to those of people from the prestigious schools hired by the firm. Although their manners and "style" were not as polished and sophisticated as those of the predominantly Ivy League junior members of the firm, their objective performance was excellent. At the end of three years, both of them were well above average in the number of cases won and in the volume of law business handled. The senior partner who had hired them argued to colleagues in the firm that, "This experience indicates that graduates of less prestigious schools are at least as ambitious and talented as graduates of the major law schools. The chief difference between the two types of graduates is in their social class background, not in their legal ability, which is what counts."

Comment on the thinking that went into this senior partner's conclusion. Is the argument basically sound? Does it have weaknesses? (Disregard your own initial opinion, if you had one, about graduates of non-prestigious law schools, and concentrate on the thinking that the senior partner used.)

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.
The senior partner is trying to draw a conclusion about a certain population. We can think of the members of this population as newly graduated lawyers, from nonprestigious law schools, who otherwise meet the law firm's hiring standards. If we divide the members of this population into two categories, "excellent" and "mediocre or worse," we can think of the population distribution as the % in each category. The senior partner has concluded that the % in the "excellent" category is very high, or anyway, just as high as in another population, involving graduates of prestigious law schools. This conclusion was based on observing a sample of size = 2, in which the sample distribution was 100% "excellent," 0% "mediocre or worse."

Apart from any other considerations, however, the sample distribution for size 2 is apt to be quite different from the population distribution: the latter could be only 60% or 50% or even perhaps as low as 40% "excellent," and a 2-0 sample split would not be so unusual; just as one would not be at all amazed to draw 2 out of 2 silver gumballs from an urn with only 40% silvers. So the senior partner's attitude is quite unwarranted: a larger sample is needed.

Susan is the artistic director for a ballet company. One of her jobs is auditioning and selecting new members of the company. She says the following of her experience: "Every year we hire 10-20 young people on a one-year contract on the basis of their performance at the audition. Usually we're extremely excited about the potential of two or three of these young people—a young woman who does a brilliant series of turns or a young man who does several leaps that make you hold your breath. Unfortunately, most of these young people turn out to be only somewhat better than the rest. I believe many of these extraordinarily talented young people are frightened of success. They get into the company and see the tremendous effort and anxiety involved in becoming a star, and they get cold feet. They'd rather lead a less demanding life as an ordinary member of the corps de ballet."

Comment on Susan's reasoning. Why do you suppose that Susan usually has to revise downward her opinion of dancers that she initially thought were brilliant?

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.
We can analyze this problem using the Law of Large Numbers by thinking of each ballet dancer as possessing a population of ballet movements. Susan is interested in excellence, so we can divide the members of each population into two categories: "brilliant movements" and "non-brilliant, or other movements." We can think of the population distribution as the percentage or proportion in each category. For many dancers, the population distribution is actually 0% brilliant and 100% other: these dancers simply lack the talent to perform a brilliant movement. For many other dancers, there is a small or moderate percentage of "brilliant movement" gumballs in their urn. A true ballet star would therefore have a population distribution with a greater percentage of "brilliant" movements than an ordinary member of the corps de ballet.

By conducting auditions, Susan is observing samples of each dancer's population distribution. An audition, however, is a very small sample of a dancer's movements. We know from the Law of Large Numbers that small samples are very unreliable estimates of the population. When a dancer performs some brilliant moves during an audition, it is often because the dancer has happened to draw a couple of the "lucky gumballs" that day: it does not prove that the population distribution for that dancer consists of a large percentage of "brilliant movements." It is reasonable to think that there are really very few dancers that have population distributions with a large percentage of brilliant movements; and so when Susan sees a dancer performing brilliantly at audition, the chances are it is just a lucky draw from a dancer who is capable or performing some, but not necessarily a great number of "brilliant movements." Therefore, when Susan hires such dancers and evaluates them after seeing a much larger sample of their movements, it is not surprising that she finds that many of these dancers that were brilliant at audition turn out to be only somewhat better than the rest.
Kevin, a graduate student in sociology, decided to do a research project on "factors affecting performance of major league baseball players" in which he gathered a great amount of demographic data on birthplace, education, marital status, etc., to see if any demographic factors were related to the performance of major league baseball players (e.g., batting average, pitching victories). Kevin was unable to use data for all the major league teams because information for some of the players was unavailable, but he was able to obtain data for some 200% players in the major leagues.

One finding that interested Kevin concerned the 110 married players. About 68% of these players improved their performance after getting married, while the remainder had equal or poorer performance. He concluded that marriage is beneficial to a baseball player's performance. At a social hour sponsored by the Office of the Commissioner of Major League Baseball, he happened to mention his finding to a staff member of the office. The staff member listened to Kevin's results and then said, "Your study is interesting but I don't believe it. I'm sure that baseball performance is worse after a marriage because the ball player suddenly has to take on enormous responsibilities: taking care of his spouse and children. Plus the factor of being stressed by having to be on the road so much of the time and therefore away from the family. The player will no longer be able to devote as much time to baseball as before he was married. Because of this he will lose that competitive quality that is necessary for good performance in baseball."

What do you think of the staff member's argument? Is it a sound one or not? Explain your reasoning.

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.
Kevin is trying to find out how performance in major league baseball is affected by being married. To do this, he obtained data for 200 players in the major leagues and discovered that out of the 110 that had gotten married, 68% had improved performance after the wedding (and 32% had equal or poorer performance). According to the Law of Large Numbers, which states that the larger the sample, the better it is in estimating the population, there is substantial evidence that marriage is beneficial to baseball players' performance. Recall that in the gumball demonstration, samples of size 25 were very good estimates of the population: these samples did not differ much from population. Extending the argument, samples of size 110 are extremely accurate estimates of the population. Thus, it can be concluded that, in general, marriage is beneficial to baseball players' performance.

What about the staff member's theory that baseball performance is worse after a marriage because the ball player assumes enormous responsibilities and will no longer be able to devote as much time to baseball as before? Although this argument may have some intuitive appeal, it should be discounted because it is not supported by any data, and is in fact contradicted by Kevin's large sample of 110 players.
Subjective Example Problems

One reason that the Law of Large Numbers is important to learn is that it applies not only to urns and gumballs. The basic principles involved in the Law of Large Numbers apply whenever you make a generalization or an inference from observing a sample of objects, actions, or behaviors.

To give you an idea of how broad the Law of Large Numbers is, we have, in this packet, presented three situations in which the Law of Large Numbers applies. Each situation is analyzed in terms of the Law of Large Numbers.

Joan is a bright, sophisticated, young woman with a keen sense of humor. It is important to her that her friends, of both sexes, share her sense of humor. Joan has been invited to a party by Jeremy, a man whom she's met several times and doesn't know at all well. He's good looking and is said to be intelligent and athletic. Recalling her encounters with Jeremy, Joan remembers seeing him once with a group of guys and overhearing him tell a really corny story. She also recalls a time when he was in line for a movie and made a really weak pun. Joan strongly prefers sophisticated humor and so she decides to turn down Jeremy's invitation, reasoning that Jeremy's sense of humor is mostly corny and unsophisticated.

Comment on Joan's reasoning. Do you think her decision is reasonable? Explain your answer.

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.

Joan is trying to draw a conclusion about a certain population. We can think of the members of this population as Jeremy's attempts at humor. If we divide the members of this population into two categories, "corny" and "sophisticated," we can think of the population distribution as the % in each category. Joan has concluded that the % in the "corny" category is very high. This conclusion was based on observing a sample of size = 2, in which the sample distribution was 100% "corny" and 0% "sophisticated."

Apart from any other considerations, however, the sample distribution for a sample size of 2 is apt to be quite different from the population distribution: the latter could be only 60% or 50% or even as low as 40% "corny," and a 2-0 sample split would not be so unusual, just
a some would not be at all amazed to draw 2 out of 2 red gumballs from an urn with only 40% reds.

Joan's conclusion that Jeremy's sense of humor is mostly corny and sophisticated is therefore a premature one. A sample size of 2 is not a reliable estimate of the population. As for Joan's decision to turn Jeremy down: it may be that any corny humor is enough to make a potential date unattractive to Joan. So, we can't criticize Joan's decision to refuse Jeremy's invitation, but we certainly can criticize her conclusion that Jeremy's sense of humor is mostly corny and unsophisticated.

Catherine is a sales representative for a furniture manufacturer. Her job takes her on lengthy trips all over North America. She goes to a particular city and meets with as many furniture wholesalers and retailers as possible over a period of days or weeks. Although she enjoys her work, she finds that one of the disadvantages is the necessity of eating at restaurants all the time. In a typical visit of two weeks to a particular city, she tries out several restaurants in the first few days. Usually the set of restaurants she picks includes some good, some bad, and a majority that are merely mediocre. Occasionally, though, in the first few days, the set of restaurants includes one that seems really excellent or even superb to her. When that happens, she tries to eat at the excellent restaurant as often as possible. However, she finds that when she does this, she is usually disappointed. Subsequent meals rarely seem as good to her as the first meal, and her ultimate opinion of the restaurant is usually merely that it is "good" or "better than average."

Why do you suppose that Catherine usually has to revise downward her opinion of restaurants that she initially thought were excellent?

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.

Each restaurant can be thought of as consisting of a population of meal-experiences. Since Catherine is interested in true excellence, she divides the members of the population into two categories, "really excellent" meal-experiences, and "non-excellent" meal-experiences ("good," "better than average," and so on). We can think of the population distribution as the percentage in each category. For many restaurants, the population distribution is actually 0% excellent, 100% non-excellent: there is no "really excellent" meal experiences to be had there.
For many others, there is a small or moderate percentage of "really excellent" gumballs in the urn. When Catherine has a really great meal-experience, it is usually because she has stumbled onto one of these and drawn one of the "lucky gumballs." Such a superb meal is just a sample of size = 1 and therefore, according to the Law of Large Numbers, is not a reliable indicator of the true quality of the restaurant: it does not indicate that the population distribution for that restaurant is 100% "really excellent" or even that it is predominantly "really excellent." It is reasonable to think that there are really very few restaurants that have population distributions with a large percentage of "really excellent" meals; and so when Catherine does get such a meal, the chances are it is just a lucky draw in a restaurant that has some, but not many, "really excellent" meal-experiences. Therefore, when she goes back and takes a larger sample, most of the additional meal-experiences from such a restaurant are not as good as her first.

Kelly is not what you would call a last minute Christmas shopper. In fact, choosing presents is one of her favorite activities, and she always starts each summer, making mental lists of friends and relatives and picking up really attractive, welcome gifts at bargain prices. She often spends considerable time on birthday and anniversary presents as well. Finding the right item for a friend is a source of real pleasure, and getting it at a good price is a source of genuine satisfaction.

Kelly's best friend, Margaret, sometimes teases her about this preoccupation. And when Kelly becomes pregnant for the first time, one April, and tells Margaret about it in May, Margaret predicts to a mutual acquaintance of theirs that Kelly's feelings about shopping will soon change. "This summer she will be focusing on her tummy and thinking about names and breast-feeding, and she'll drop her preoccupation with her friends' and relatives' Christmas presents. She will find that buying presents for friends will no longer be pleasurable because they will be like chores to her."

Comment on Margaret's reasoning. Perhaps Margaret is a bit jealous of Kelly, but is her reasoning basically sound? Or does it have weaknesses?

Please consider this problem for a few minutes. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.
The question here is predicting what Kelly's attitude toward shopping will be now that she is pregnant: Will she now think of shopping and finding presents for friends a pleasure or an unpleasurable chore?

We know that before Kelly becomes pregnant, finding presents for friends was a pleasure to her. Although it is not explicitly stated in the problem, we can assume that she has considered present-hunting to be pleasurable for a long time. In terms of the Law of Large Numbers, we can say that a large sample over many years of Kelly's past habits indicates that she considers present-hunting to be pleasurable (or in gumball terms, that in her population distribution of "present-hunting" gumballs, there are many more "pleasurable present-hunting" gumballs than "unpleasurable present-hunting" gumballs).

Margaret has a theory that Kelly will begin to view present-hunting as being unpleasurable because she is now pregnant and will be more occupied with thinking about the baby-to-be. Although this theory is somewhat plausible, it is weak. Being pregnant may change Kelly's attitude slightly, but is very unlikely that being pregnant will change her attitude as drastically as Margaret believes. In addition, it is plausible that the opposite may be true: Kelly might become even more enthusiastic about present-shopping because she is very happy about the expected baby and wants to share her happiness with her friends and relatives by picking out extra-special presents for them. The point here is that a plausible theory that is unsupported by data is not a good basis for prediction, especially when there is a large sample that contradicts the prediction of the theory. Thus, Margaret's reasoning is not sound: We should expect that Kelly will still receive pleasure from present-hunting.
APPENDIX B

Test Problems for Experiment 1 with Directions

Control Group Directions

We are interested in studying how people go about explaining and predicting events under conditions of very limited information about the events. It seems to us to be important to study how people explain and predict under these conditions because they often occur very frequently in the real world. Indeed, we often have to make important decisions based on such explanations and predictions, either because there is too little time to get additional information or because it is simply unavailable.

Experts who study human inference have found that some commonsense principles are helpful in explaining and predicting events, especially under conditions of limited information. One such principle is that whenever you make a generalization or inference from observing a sample of objects, actions, or behaviours pay attention to how large the sample is. For example, a public opinion poll based on a large sample is more likely to be accurate than to be based on a small sample.

On the pages that follow, there are a number of problems that we would like you to consider. As you will see, they represent a wide range of real-life situations. We would like you to think carefully about each problem, and then write down answers that are sensible to you. In many of the problems, you may find that the principle just mentioned is helpful.

Treatment Group Directions

On the pages that follow, there are a number of problems that we would like you to consider. As you will see, they represent a wide range of real-life situations. We would like you to think carefully about each problem, and then write down answers that are sensible to you. In many of the problems, you may find that the Law of Large Numbers is helpful.
1. An economist was arguing in favor of a guaranteed minimum income for everyone. He cited a recent study of several hundred people in the United States with inherited wealth. Nearly 92% of those people, he said, worked at some job that provided earned income sufficient to provide at least a middle-class lifestyle. The study showed, he said, that contrary to popular opinion, people will work in preference to being idle. Thus a guaranteed income policy would result in little or no increase in the number of people unwilling to work.

Comment on the economist's reasoning. Is it basically sound? Does it have weaknesses?

2. A talent scout for a professional basketball team attends two college games with the intention of observing carefully the talent and skill of a particular player. The player looks generally excellent. He repeatedly makes plays worthy of the best professional players. However, in one of the games, with his team behind by 2 points, the player is fouled while shooting and has the opportunity to tie the game by making both free throws. The player misses both free throws and then tries too hard for the rebound from the second one, committing a foul in the process. The other team then makes two free throws, for a 4-point lead, and goes on to win by 2 points.

The scout reports that the player in question "has excellent skills, and should be recruited. He has a tendency to misplay under extreme pressure, but this will probably disappear with more experience and better coaching."

Comment on the thinking embodied in the scout's opinion that the player (a) "has excellent skills" and that the player has (b) "a tendency to misplay under extreme pressure." Does the thinking behind either conclusion have any weaknesses?

3. Howard was a teacher in a junior high school in a community known for truancy and delinquency problems among its youth. Howard says of his experiences: "Usually, in a class of 35 or so kids, 2 or 3 will pull some pretty bad stunts in the first week—they'll skip a day of class, get into a scuffle with another kid, or some such thing. When that kind of thing happens, I play it down and try to avoid calling the class' attention to it. Usually, these kids turn out to be no worse than the others. By the end of the term you'll find they haven't pulled any more stunts than the others have." Howard reasons as follows: "Some of these kids are headed toward a delinquent pattern of behavior. When they find out nobody is very impressed, they tend to settle down."

Comment on Howard's reasoning:

(a) Do you agree that it is likely that the students who pull a "pretty bad stunt in the first week" are "headed toward a delinquent pattern of behavior?"

(b) Do you agree that it is likely that the students who initially pull a "pretty bad stunt" turn out to be no worse than the others because they find no one is impressed with their behavior?
4. Martha was talking to a fellow passenger on an airplane. The fellow passenger was on his way to Hawaii for a month's vacation. "I don't like vacations myself," Martha said. "I've always worked. I put myself through college and law school and now I have a full-time legal practice. Frequently, of course, I've had slow periods when I wasn't working at all, but I never liked those times. For example, there would usually be a week or two between the end of school and the beginning of a summer job and another week or two of enforced idleness at the end of the summer. And there were many occasions when I was getting started in my career when I had no real work to do for fairly long periods. But I never enjoyed the leisure. I know there are some people who talk about using vacations to "recharge" themselves. But I suspect many of these people don't really enjoy their work or don't have a very high energy level. I do have a lot of energy, and I do enjoy my work, and I guess that's why I don't really like vacations."

Analyze Martha's reasoning. Do you think she had good evidence for feeling she doesn't like vacations?

5. The superintendent of schools was urging the school board to make an expensive curriculum shift to a "back-to-basics" stress on fundamental learning skills and away from the electives and intensive immersion in specialized arts and social studies topics that had recently characterized the secondary schools in the district. He cited a study of 120 school systems that had recently begun to emphasize the basics and 120 school systems that had a curriculum similar to the district's current one. The "back-to-basics" school systems, he said, were producing students who scored half-a-year ahead of the students in the other systems on objective tests of reading, mathematics, and science. Of the 120 "back-to-basics" school systems, 85 had shown improved skills for students in the system vs only 40 with improved skills in the 120 systems which had not changed. One of the school board members took the floor to argue against the change. In her opinion, she said, there was no compelling reason to attribute the improved student skills in the "back-to-basics" systems to the specific curriculum change, for two reasons: (1) school systems that make curriculum changes probably have more energetic, adventurous administrators and faculty and thus the students would learn more in those school systems no matter what the curriculum was. (2) Any change in curriculum could be expected to produce improvement in student performance because of increased faculty interest and commitment.

Comment on the reasoning of both the superintendent and the board member. On the basis of the evidence and arguments offered, do you think it is likely that the "back-to-basics" curriculum is intrinsically superior to the district's current curriculum?
6. Bert H. has a job checking the results of an X-ray scanner of pipeline welds in a pipe factory. Overall, the X-ray scanner shows that the welding machine makes a perfect weld about 80% of the time. Of 900 welds each day, usually about 680 to 740 welds are perfect. Bert has noticed that on some days, all of the first 10 welds were perfect. However, Bert has also noticed that on such days, the overall number of perfect welds is usually not much better for the day as a whole than on days when the first 10 welds show some imperfections.

Why do you suppose the number of perfect welds is usually not much better on days where the first batch of welds was perfect than on other days?

7. Two New Yorkers were discussing restaurants. Jane said to Ellen, "You know, most people seem to be crazy about Chinese food, but I'm not. I've been to about 20 different Chinese restaurants, across the whole price range, and everything from bland Cantonese to spicy Szechwan and I'm not really very fond of any of it." "Oh," said Ellen, "don't jump to conclusions. I'll bet you've usually gone with a crowd of people, right." "yes," admitted Jane, "that's true. I usually go with half a dozen people or more from work." "Well, that may be it," said Ellen. "People usually go to Chinese restaurants with a crowd of people they hardly know. I know you, you're often tense and a little shy, and you're not likely to be able to relax and savor the food under those circumstances. Try going to a Chinese restaurant with just one good friend. I'll bet you'll like the food."

Comment on Ellen's reasoning. Do you think there is a good chance that if Jane went to a Chinese restaurant with one friend, she'd like the food? Why or why not?

8. Gerald M. had a 3-year-old son, Timmy. He told a friend: "You know, I've never been much for sports, and I think Timmy will turn out the same. A couple of weeks ago, an older neighbor boy was tossing a ball to him, and he could catch it and throw it all right, but he just didn't seem interested in it. Then the other day, some kids his age were kicking a little soccer ball around. Timmy could do it as well as the others, but he lost interest very quickly and started playing with some toy cars while the other kids went on kicking the ball around for another 20 or 30 min."

Do you agree with Gerald's reasoning that Timmy is likely not to care much for sports? Why or why not?
9. An auditor for the Internal Revenue Service wants to study the nature of arithmetic errors on income tax returns. She selects 4000 Social Security numbers by using random digits generated by an "Electronic Mastermind" calculator. And for each selected social security number she checks the 1978 Federal Income Tax return thoroughly for arithmetic errors. She finds errors on a large percentage of the tax returns, often 2 to 6 errors on a single tax return. Tabulating the effect of each error separately, she finds that there are virtually the same number of errors in favor of the taxpayer as in favor of the government. Her boss objects vigorously to her assertions, saying that it is fairly obvious that people will notice and correct errors in favor of the government, but will "overlook" errors in their own favor. Even if her figures are correct, she says, looking at a lot more returns will bear out his point.

Comment on the auditor's reasoning and her boss's contrary stand.

10. Janice is head nurse in a home for the aged. She says the following of her experiences: "There is a big turnover of the nursing staff here, and each year we hire 15-20 new nurses. Some of these people show themselves to be unusually warm and compassionate in the first few days. One might stay on past quitting time with a patient who's having a difficult night. Another might be obviously shaken by the distress of a patient who has just lost a spouse. I find though that, over the long haul, these women turn out to be not much more concerned and caring than the others. What happens to them, I think, is that they can't remain open and vulnerable without paying a heavy emotional price. They usually continue to be considerate and effective but they build up a shell."

Comment on Janice's reasoning. Do you think it is likely that she correctly identifies the nurses who are unusually warm and compassionate? Do you agree it is likely that most of the ones who are unusually warm at first later build up a shell to protect themselves emotionally?

11. At Stanbrook University, the Housing Office determines which of the 10,000 students enrolled will be allowed to live on campus the following year. At Stanbrook, the dormitory facilities are excellent, so there is always great demand for on-campus housing. Unfortunately, there are only enough on-campus spaces for 5000 students. The Housing Office determines who will get to live on campus by having a Housing Draw every year: every student picks a number out of a box over a 3-day period. These numbers range from 1 to 10,000. If the number is 5000 or under, the student gets to live on campus. If the number is over 5000, the student will not be able to live on campus.

On the first day of the draw, Joe talks to five people who have picked a number. Of these, four people got low numbers. Because of this, Joe suspects that the numbers in the box were not properly mixed, and that the early numbers are more favourable. He rushes over to the Housing Draw and picks a number. He gets a low number. He later talks to four people who drew their numbers on the second or third day of the draw. Three got high numbers. Joe says to himself, "I'm glad that I picked when I did, because it looks like I was right that the numbers were not properly mixed."

What do you think of Joe's reasoning? Explain.
APPENDIX C

Written Explanation of the Law of Large Numbers for Experiments 1 and 2

THE LAW OF LARGE NUMBERS

We are interested in studying how people go about explaining and predicting events under conditions of very limited information about the events. It seems to us to be important to study how people explain and predict under these conditions because they occur very frequently in the real world. Indeed, we often have to make important decisions based on such explanations and predictions, either because there is too little time to get additional information or because it is simply unavailable.

Experts who study human inference have found that principles of probability are helpful in explaining and predicting a great many events, especially under conditions of limited information. One such principle of probability that is particularly helpful is called the Law of Large Numbers. In this study, we will teach you the Law of Large Numbers by introducing you to the probabilistic terms and ideas associated with this principle, and then provide examples to illustrate how the Law of Large Numbers can be used to explain and predict events in the real world.

Imagine an urn that is filled with gumballs. Let's say that the urn contains a very large number of gumballs—thousands, millions, or larger. The gumballs in this urn are known collectively as the population.

Let's say that there are two types of gumballs in the urn—red gumballs and white gumballs. When we do this, we can now say that the population has two categories or groups, namely—red gumballs and white gumballs.

Now let's say that in this population of red and white gumballs there are 70% white gumballs and 30% red gumballs. If that is the case, then we know more than that the population has two categories (red and white): we now know the proportion of the gumballs in the white category (70%) and the proportion of the gumballs in the red category (30%). This is known as the population distribution (other examples of distributions are 60% red and 40% white, or 85% white and 15% red, etc., but in every distribution, the sum of the proportions must be 100%.

Let's summarize what we've covered so far:

-- A population is the entire set of objects we are interested in (all of the gumballs in the urn)
-- Categories refers to the types of objects in the population (red and white)
-- Distribution refers to the proportion of objects in each category (70% white and 30% red in this example)

One of the major goals of statistics is to find out something about a population. More specifically, we want to find out what the population distribution is. One way that we might do this would be to actually examine all of the objects in the population and count up the number of objects in each category. In our example, we would empty the entire urn and count the number of red gumballs and the number of white gumballs. Using this method, we could find out exactly what the population distribution of red and white gumballs was. But, there is a very serious problem with counting all of the objects in the population: populations, in general, are very large. If we were to count all of the objects in our gumball population, it would take more time and effort than would be practical (imagine counting a million gumballs!).
OK--so counting the entire population is impractical. What do we do instead to find out what the population distribution is?

What we do instead is to take a sample of the population. A sample is a subset of the population. We can take a sample of any size--if we pick 5 gumballs, we say that the sample size is 5; if we take 60 gumballs for our sample, we say that the sample size is 60, and so on.

When we take a sample from the population, we will get a sample distribution. The sample distribution is the proportion of objects in each category for the sample, just as the population distribution is the proportion of objects in each category for the population. For example, if we take a sample of 10 gumballs, we might get 6 reds and 4 whites. In this case, our sample distribution would be 60% red and 40% white. We also might have happened to get 9 whites and 1 red, in which case the sample distribution would be 90% white and 10% red. The important point here is that samples are estimates of populations. Since it is often impractical or sometimes impossible to examine the entire population, we instead have to draw samples to estimate what the population is like.

Some samples will have sample distributions that are closer to the population distribution than others. For instance, in our gumball example, a sample of 9 reds and 1 white would be a very poor estimate of the population, while a sample of 8 whites and 2 reds would be a pretty good estimate of the population. The critical question is: What determines how likely it is that samples will give good estimates of the population? The answer is simple: if the samples are chosen haphazardly, or randomly (by, for example, mixing the urn and reaching into the urn blindfolded and scooping out the needed number of gumballs OR by mixing the contents of a gumball machine and letting the gumballs out one by one), then there is only one factor--sample size.

This brings us to the Law of Large Numbers: as the size of a random sample increases, the sample distribution is more and more likely to get closer and closer to the population distribution. In other words, the larger the sample, the better it is as an estimate of the population.
APPENDIX D

Verbal Presentation of the Law of Large Numbers

Now that you have all read the written explanation of the Law of Large Numbers, I thought it would be nice to demonstrate, before your eyes, that the Law of Large Numbers really does work.

So I have here (pick urn up, unveil the urn, etc.) a genuine urn, filled with genuine red and black gumballs. And as in the written explanation, there happen to be 70% blue gumballs and 30% red gumballs in this urn.

A major purpose of statistics is to find out about a population from a sample of that population. Suppose it is your job to find out what proportion of the gumballs in this urn are blue and what proportion of the gumballs are red. You could dump out all of the gumballs and count all of them, but that would take quite a long time and wouldn't be worth the effort. For the sake of demonstration, this urn isn't very large. But if we had a very large urn (I like gesturing here to indicate a very large urn) filled with millions of gumballs, it's easy to see how time-consuming and impractical it would be to count the entire population of gumballs in the urn.

What you would probably do instead to find out what the composition of the urn was like would be to take a sample from the urn because the sample you chose would tell you something about the population; that is, the sample would be an estimate of the population.

According to the Law of Large Numbers, when you choose your sample randomly like this (reach into urn without looking, mix them up, and draw out a handful), the larger the sample, the better the sample is in estimating the population. To repeat -- the larger the sample is that you draw, the better that sample is in estimating the population.

Well, what I'm going to do now is to demonstrate the Law of Large Numbers (reveal blackboard with summary chart). I will pick samples of size 1, 4, and 25 (gesture to the three sections of the board as you say the numbers) to show that as the sample size increases, the sample becomes a better estimate of the population.

For each sample that I draw, I will write down on the board the number of blues in the sample, the number of reds in the sample, the percent blue in the sample, and the deviation or difference between the sample distribution and the population distribution (as you are saying the various categories, point to them on the board).

Now recall that the population distribution for this urn is 70% blue, 30% red. Therefore, for example, if the sample I draw happens to be 85% blue, the deviation of that sample will be 85 minus 70 or 15%, and I'll enter that number here (point to the deviation column).

I will draw a few samples of size 1, 4, and 25. After I'm done with drawing samples of each size I will calculate the average deviation of the samples from the population (point to all three "Average Deviation" boxes). The Law of Large Numbers states that as the sample size increases, the sample becomes a better estimate of the population. In other words the average deviation of the sample from the population will decrease as the sample size increases. So
This number (point to "Average Deviation" box) should go down as sample size (point to top of chart: "Sample Size = ") goes up.

So first I will draw samples of size 1. (take scooper out of urn). I will mix up the gumballs like this (mix gumballs while talking) and use this scooper to pick my sample (put scooper in, get some gumballs). The reason why I’m using this scooper is because the gumballs stick to your hands and turn them red and blue (put plastic lid on the scooper and shake the scooper).

OK -- the first gumball to come out of the scoop will be my sample (Shake scoop until one drops into your hand in plain view. If more than one gumball comes out, use the first one that comes out).

In this sample, I have 0 blues and 1 red (for example). (Go to the board and verbalize as you’re writing down the results of the sample). So in this sample, I had 0 blues, 1 red. That means that the percent blue in the sample was 0%. 0 minus 70 equals 70% deviation.

(got back to turn. Put the sample back, empty the scooper into the urn, and repeat the procedure. I drew 4 samples of size 1. As you go on, you don’t have to describe the process in as much detail)

(After you’ve finished, compute the average deviation). So, the average deviation for samples of size 1 is ___%.

Now I’ll pick a few samples of size 4. (Follow the same procedure as above — mix contents of urn, scooper out some gumballs, pour out the first four gumballs which fall into your open hand, summarize, put sample back in the urn, empty contents of scooper into urn, etc. Pick 4 samples of size 4 then compute average deviation.) The average deviation for samples of size 4 is ___%.

(Do the same with samples of size 25. Three samples should be enough. With samples of size 25, you can’t hold 25 gumballs in one hand. You will have to shake a few gumballs (about 7 or 8) into your hand at a time. Get two glass bowls (or clean ashtrays). When you shake a few out, separate the blues and reds and put them into different bowls. Keep track of how many you’ve already drawn so that you end up with 25.) (Example follows)

With this sample of 25 I have 18 blues and 7 reds (write results on the board). That is 72% blue and the deviation is 2%.

(If you have small deviations, you can quite at 3 samples of 25.)

(Compute average deviation for samples of size 25).

(It is possible, but not likely, that samples of size 25 will give larger deviations than samples of size 4, especially if size 4 gave you all 3-1 sample splits (average deviation = 5%). Two ways to guard against this:

1. If you’ve drawn 4 samples of size 4 and they’re all 3-1, pick more samples until you have some other sample distribution.
2. If you've drawn 3 samples of size 25 and you think your average deviation might be too close to the average deviation for samples of size 4, draw one or two more. (I only had to do this once out of the 25-30 sessions I ran.)

Summary

The Law of Large Numbers states that as the size of your sample increases, the sample becomes a better estimate of the population. This is shown here: as the samples increased in size from 1 to 4 to 25 (gesture to the top of the chart), the average deviation of the samples from the population decreased from ___% to ___% to ___% (gesture to the bottom of the chart).

I'd like to tell you something else about the Law of Large Numbers. That is, with small samples, sometimes you can't even correctly answer the simplest questions about the population.

For example, suppose you were asked to say whether there were more blue gumballs in the urn or more red gumballs. If you happened to draw this sample (point to a very bad sample of size 1: 0 blue, 1 reds. If there isn't one, then go to the samples of size 4 and point to a 1 blue, 3 red or 2 blue, 2 red) you would say "Well, from my sample, I think there are more reds than blues (if you pointed to a 2 blue, 2 red, say: "well, from my sample, I can't tell at all") and you would be wrong. But look at the larger sample of size 25: you can always correctly answer at least the most basic question -- 'Are there more blues or more reds?' With smaller samples, that is not always possible.

So I've demonstrated the Law of Large Numbers -- as the sample increases in size, the sample becomes a better estimate of the population.
<table>
<thead>
<tr>
<th>LAW OF LARGE NUMBERS DEMONSTRATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SAMPLE = 1</strong></td>
</tr>
<tr>
<td>W</td>
</tr>
<tr>
<td>------------------------------------</td>
</tr>
<tr>
<td><strong>SAMPLE = 4</strong></td>
</tr>
<tr>
<td>W</td>
</tr>
<tr>
<td>------------------------------------</td>
</tr>
<tr>
<td><strong>SAMPLE = 25</strong></td>
</tr>
<tr>
<td>W</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AVERAGE DIFF. FROM POP. =</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMPLE = 1</td>
</tr>
<tr>
<td>SAMPLE = 4</td>
</tr>
<tr>
<td>SAMPLE = 25</td>
</tr>
</tbody>
</table>
APPENDIX E

Example Problems and Analyses for the Rule Condition
and the Rule-Plus-Error Condition: Experiment 2

THREE EXAMPLE PROBLEMS

One reason that the Law of Large Numbers is important to learn is that it applies not only to urns and gumballs. The basic principles involved in the Law of Large Numbers apply whenever you make a generalization or an inference from observing a sample of objects, actions, or behaviours.

Let's look again at three of the problems you analyzed earlier. Each situation is briefly analyzed in terms of the Law of Large Numbers.

Two New Yorkers were discussing restaurants. Jane said to Ellen, "You know, most people seem to be crazy about Chinese food, but I'm not. I've been to about 20 different Chinese restaurants, across the whole price range, and everything from bland Cantonese to spicy Szechwan and I'm really not very fond of any of it." "Oh," said Ellen, "don't jump to conclusions. I'll bet you've usually gone with a crowd of people, right?" "Yes," admitted Jane, "that's true. I usually go with half a dozen people or more from work." "Well, that may be it," said Ellen. "People usually go to Chinese restaurants with a crowd of people they hardly know. I know you, you're often tense and a little shy, and you're not likely to be able to relax and savor the food under those circumstances. Try going to a Chinese restaurant with just one good friend. I'll bet you'll like the food."

Comment on Ellen's reasoning. Do you think there is a good chance that if Jane went to a Chinese restaurant with one friend, she'd like the food? Why or why not?

Analysis:

The question here is whether or not Jane is likely to enjoy Chinese food the next time she visits a Chinese restaurant. We can think of Jane's experiences in Chinese restaurants as a population. So far, Jane has drawn a sample of 20 out of this population. We can divide the members of the population into two categories, liking Chinese food and disliking Chinese food. Based on Jane's sample of size = 20 the sample distribution was 100 percent disliking the food and 0 percent liking the food. According to the Law of Large Numbers a sample size of 20 is enough to resemble the population distribution, especially since Jane has experienced quite a variety of Chinese restaurants in the sample. This latter consideration makes the sample more like a random sample. Ellen's reasoning is weak, based on the evidence.
Gerald M. had a 3-year-old son, Timmy. He told a friend: "You know, I've never been much for spurts, and I think Timmy will turn out the same. A couple of weeks ago, an older neighbour boy was tossing a ball to him, and he could catch it and throw it all right, but he just didn't seem interested in it. Then the other day, some kids his age were kicking a little soccer ball around. Timmy could do it as well as the others, but he lost interest very quickly and started playing with some toy cars while the other kids went on kicking the ball around for another 20 or 30 minutes."

Do you agree with Gerald's reasoning that Timmy is likely not to care much for sports? Why or why not?

Analysis:
The problem here is one of predicting Timmy's future attitude toward sports. Apart from other considerations, we can think of Timmy's opportunities to engage in sports activities as a population of his reactions to sports. Gerald, Timmy's father has observed a sample of two such reactions to sports. The Law of Large Numbers tells us that when sample size = 2, it is not accurately predictive of the population distribution. For example, it was certainly possible to get two red gumballs from the urn even though red gumballs were only 30 percent of the population.

Janice is head nurse in a home for the aged. She says the following of her experiences: "There is a big turnover of the nursing staff here, and each year we hire 15-20 new nurses. Some of these people show themselves to be unusually warm and compassionate in the first few days. One might stay on past quitting time with a patient who's having a difficult night. Another might be obviously shaken by the distress of a patient who has just lost a spouse. I find though that, over the long haul, these women turn out to be not much more concerned and caring than the others. What happens to them, I think, is that they can't remain open and vulnerable without paying a heavy emotional price. They usually continue to be considerate and effective, but they build up a shell."

Comment on Janice's reasoning. Do you think it is likely that she correctly identifies the nurses who are unusually warm and compassionate? Do you agree it is likely that most of the ones who are unusually warm at first later build up a shell to protect themselves emotionally?

Analysis:
Janice's theory may be correct, but we can also consider the behaviours of the nurses toward their patients as a population of behaviours over the longer period of time they spend working at the nursing home. The population distribution of these behaviours is made up of two categories, unusually compassionate behaviours and more normal behaviours. Janice has observed one or two of the newly-hired nurses exhibiting unusually warm and compassionate behaviours and then becoming more like the others. She explains this as the building of a protective shell. However, according to the Law of Large Numbers, the initial small sample of behaviours may not provide a sample distribution that accurately reflects the population distribution. Taking additional, larger samples of the nurses' behaviours will more accurately reflect their normal set of responses with patients. Therefore, Janice's arguments are weak based on the limited evidence presented in the problem.
APPENDIX F

Training Materials for Errors for the Rule-Plus-Error Condition: Experiment 2

REASONING ERRORS

The basic principles in the Law of Large Numbers may be very useful whenever you make a generalization or an inference from observing a sample. The sample might be objects, actions, or behaviours. So, it can apply to a broad range of problems in everyday life.

Sometimes people have difficulty in using the Law of Large Numbers because a population and a sample of objects, actions, and behaviours is not obvious in the situation they are thinking about or they just don't pay very much attention to that kind of information.

What else do people do when they decide or explain or predict? There are, at least, four other ways that people think when they don't have much information. These ways are not necessarily wrong for all kinds of problems. But, if a person ignores interesting and useful evidence in the problem, he or she might make an error if these four methods are applied to the problem.

"T" They appeal to a general belief. APPEAL TO TRUISM This means that they explain the problem by relying on a global belief about how things are. For example, they might say that most people are a certain way or that everybody makes mistakes or that "If it isn't broken, don't fix it".

"E" They offer a personal opinion. EGOCENTRIC BIAS This means that they explain the problem from their own personal belief, sometimes based on their own experience. For example, they start by saying "I don't agree because my brother ..." or "If that happened to me, I would ..." or "From my own experience..."

"D" They rely on a characteristic of the person. ATTRIBUTION TO DISPOSITION This means that they explain the problem by looking at the people involved in the situation and decide that those people are likely to act or think in a certain way because of the way they are. For example, they think that a person is biased, unmotivated, ambitious, experienced, talented, and so on.

"S" They make an "outside" guess. SPECULATION This means that they explain the problem by adding extra information not already in the problem and then speculating about why things are a certain way or why things might turn out a certain way based on the evidence they created.

Note that the four above methods of thinking about problems are not always wrong. But, they can be considered errors when you ignore information given in a problem, especially when that information helps you use the The Law of Large Numbers correctly.
EXAMPLES AND ERROR EXERCISES

Let's look again at three of the problems you analyzed earlier.

Two New Yorkers were discussing restaurants. Jane said to Ellen, "You know, most people seem to be crazy about Chinese food, but I'm not. I've been to about 20 different Chinese restaurants, across the whole price range, and everything from bland Cantonese to spicy Szechwan and I'm really not very fond of any of it." "Oh," said Ellen, "don't jump to conclusions. I'll bet you've usually gone with a crowd of people, right?" "Yes," admitted Jane, "that's true. I usually go with half a dozen people or more from work." "Well, that may be it," said Ellen. "People usually go to Chinese restaurants with a crowd of people they hardly know. I know you, you're often tense and a little shy, and you're not likely to be able to relax and savor the food under those circumstances. Try going to a Chinese restaurant with just one good friend. I'll bet you'll like the food."

Comment on Ellen's reasoning. Do you think there is a good chance that if Jane went to a Chinese restaurant with one friend, she'd like the food? Why or why not?

**T** Appeal to Truism  Food is food. The taste doesn't change.

**E** Egocentric Bias  I've been to Chinese restaurants with different numbers of people and I still don't like the food.

**D** Attribution to Disposition  Ellen's opinion is biased because she likes Chinese food.

**S** Speculation  No, Jane would certainly know the people from work very well.

Identify each of the following as error type T, E, D, or S.

1. Jane has now been conditioned to dislike Chinese food.  
2. Taste buds are never affected by the size of the crowd you're dining with.  
3. I disagree because when I'm feeling shy I concentrate more on the food.  
4. Jane is probably thinking about the bad rumors associated with Chinese restaurants
Gerald M. had a 3-year-old son, Timmy. He told a friend: "You know, I've never been much for sports, and I think Timmy will turn out the same. A couple of weeks ago, an older neighbour boy was tossing a ball to him, and he could catch it and throw it all right, but he just didn't seem interested in it. Then the other day, some kids his age were kicking a little soccer ball around. Timmy could do it as well as the others, but he lost interest very quickly and started playing with some toy cars while the other kids went on kicking the ball around for another 20 or 30 minutes."

Do you agree with Gerald's reasoning that Timmy is likely not to care much for sports? Why or why not?

Identify each of the following as error type T, E, D, or S.

1. No, my father and I had a similar situation. I felt threatened every time I joined a team instead of practicing the violin. _____
2. Gerald has a secret wish to compensate for his own lack of athletic ability. _____
3. From a genetic point of view, I agree with Gerald's reasoning. _____
4. Timmy has not been rewarded by his father for any interest in sports. _____

Janice is head nurse in a home for the aged. She says the following of her experiences: "There is a big turnover of the nursing staff here, and each year we hire 15-20 new nurses. Some of these people show themselves to be unusually warm and compassionate in the first few days. One might stay on past quitting time with a patient who's having a difficult night. Another might be obviously shaken by the distress of a patient who has just lost a spouse. I find though that, over the long haul, these women turn out to be not much more concerned and caring than the others. What happens to them, I think, is that they can't remain open and vulnerable without paying a heavy emotional price. They usually continue to be considerate and effective, but they build up a shell."

Comment on Janice's reasoning. Do you think it is likely that she correctly identifies the nurses who are unusually warm and compassionate? Do you agree it is likely that most of the ones who are unusually warm at first later build up a shell to protect themselves emotionally?

Identify each of the following as error type T, E, D, or S.

1. The nurses are just trying to impress their bosses during the first few days. _____
2. Janice is a very experienced nurse and therefore recognizes the emotions felt by the nurses she supervises. _____
3. My uncle is a doctor and my mother is a nurse. From my association with them, I agree with Janice's reasoning. _____
4. Most people starting a new job tend to overdo things. It's human nature. _____
APPENDIX G
Pretest Problems and Directions: Experiment 2

Problem Solving Activity (Education 220--Spring, 1991)

In British Columbia, The Year 2000 document, entitled "Enabling Learners"(1990) outlines the goals for teaching problem solving skills (e.g. "select and use information to develop solutions to problems"); for developing thinking skills; and for integrating "strands" of knowledge.

In two consecutive tutorials during March and early April, you will receive instruction and practice in solving problems. The instruction is intended as (1) a demonstration lesson of how problem solving can be taught in schools at both the Elementary and Secondary levels, and (2) a simultaneous study of what methods of instruction are most effective.

Your participation in the tutorial activity is an important part of the course. However, whether or not you wish to include the written material you generate during the activity for the study is voluntary. Any such data will be confidential, and strict anonymity will be maintained at all times.

In preparation for the first lesson, on the pages that follow, there are a number of problems that we would like you to consider. As you will see, they represent a wide range of real-life situations. We would like you to think carefully about each problem, and then write down answers that are sensible to you. You are asked to spend no more than 40 minutes working individually on the task. This means no more than an average of 4 minutes per problem. Please bring the completed booklet to the next lecture on Tuesday, March 12. It will be collected at 1:30. Use the last four numbers of your student number as identification.

Name of T.A.:_______________________________________
Day and time of tutorial:___________________________
Identification #:___________________________
Problem 1. Objective Format, Structure 4 (false alarm problem where sample data is large but biased). This problem is based on Problem 1 in Appendix B, "An economist was ...".

An education professor was arguing in favour of abolishing failing grades in secondary schools for all students. He referred to a study involving hundreds of students with high grade-point averages who attended secondary schools where failing marks had been discontinued. The research demonstrated, according to the professor, that students did not become lazy or reduce their effort to achieve in school when they knew they couldn't fail. Therefore, abolishing failing grades would not result in lower achievement.

Comment on the professor's reasoning. Is it basically correct? Are there weaknesses?

Problem 2. Objective Format, Structure 1 (conclusion from a single small sample with part 'a' as a false alarm question). This problem is based on Problem 2 in Appendix B, "A talent scout ...".

The assistant director of a major professional ice skating show visits three local ice skating competitions to observe a particular skater she is interested in recruiting. The skater displays a repertoire of outstanding skills in her skating routines. Her capabilities are equal to those of the professionals. Nevertheless, during one of the competitions and while obviously leading in points, the skater misses a crucial triple jump at the end of her routine. She ends up with a third-place finish, allowing two other skaters to exceed her total points.

The assistant director telephones the director of the show and explains that the skater would benefit from instruction at the professional level. The skater, according to the assistant director has outstanding ability, but experiences some difficulty with a specific type of jump.

Comment on the thinking apparent in the assistant director's explanation about (a) the skater has difficulty with a specific type of jump; and (b) the skater has outstanding ability. Does the thinking involved in either conclusion have any weaknesses?

Problem 3. Objective Format, Structure 3 (regression). This problem is based on Problem 3 in Appendix B, "Howard ...".

Debbie is the chief cook in a large logging camp. The men working in the camp have a reputation for rough manners and profane language. Debbie talks to her newly-hired assistants about what to expect. She tells them that a few individual loggers coming into the cookhouse will display extremely poor manners at the outset of the new season. When they do, Debbie says, she tries not to notice. She basically does not pay any attention to these individuals and finds that they don't turn out to be any worse than the others over the course of the long season in the woods. Debbie tells her assistants to use the same tactic. She tells them that the loggers will not bother them if they don't receive the attention they're seeking. Otherwise, it is likely that the loggers will persist in becoming more and more difficult to deal with.

Comment on Debbie's reasoning:
(a) Do you agree that the individual loggers with very bad behaviour are likely to persist in becoming more and more difficult to deal with?
(b) Do you agree that not paying attention to these individuals results in making their behaviour more like that of the other loggers?
Problem 4. Objective Format, Structure 2 (large sample pitted against a plausible theory not founded on data) This problem is based on Problem 5 in Appendix B, "The superintendent"

The president of a bicycle manufacturing company was talking to shareholders of the company. He argued that next year's line of bicycles should include more of the old-fashioned, three-speed street bikes and less of the company's current line of new twenty-one speed, high-tech mountain bikes. He proposed that this strategy would increase the company's profits in the coming fiscal year. He described a research study conducted with 90 manufacturing companies that had decided to sell the older style street bicycles and another 90 firms selling the latest designs in mountain bikes. He said the companies selling the street bikes had increased their net earnings by 35 percent. In fact, 65 of the 90 companies making the change to the older-style product were making more money while only 30 of the 90 companies still marketing the high-tech mountain bikes had increased their profits. One of the shareholders challenged the president's proposal. The shareholder argued that there was not enough evidence to say that selling old-fashioned bikes increased profits. The shareholder gave two reasons: (1) Manufacturing companies that embark on new products always spend more money in advertising and promotion and therefore sell more in the first year; and, (2) New product lines usually spark the imagination of sales departments creating increased profits.

What do you think about the thinking of both the company president and the shareholder? After considering the evidence, do you think that the old-fashioned street bicycles would sell better than high-tech mountain bikes?

Problem 5. Probabilistic Format, Structure 3 (regression) This problem is based on Problem 6 in Appendix B, "Bert H. has a job checking ..."

Brenda has a summer job at a well-known hamburger restaurant. The restaurant has brand new, automatic food-wrapping machines. One of Brenda's duties is to check how the new packaging equipment is working. She does this for one hour each day. Overall, the machines wrapped hamburgers and put french fries and pies into their containers perfectly about 90 percent of the time. This meant that out of 800 items done during the hour, about 700 to 740 packages are perfectly done. Brenda has noticed something she finds interesting. On some days, all of the first 15 packages done during the hour were perfect. However, Brenda also noticed that for those hours, the overall number of perfectly done packages is about the same as for any other hours when the first fifteen were not perfect.

Why do you think the overall number of perfectly done packages is the same — even for those hours when the first fifteen are perfect?
Problem 6. Subjective Format, Structure 2 (large sample pitted against a plausible theory not founded on data). This problem is identical to problem 7 in Appendix B, "Two New Yorkers were ..."

Two New Yorkers were discussing restaurants. Jane said to Ellen, "You know, most people seem to be crazy about Chinese food, but I'm not. I've been to about 20 different Chinese restaurants, across the whole price range, and everything from bland Cantonese to spicy Szechwan and I'm really not very fond of any of it." "Oh," said Ellen, "don't jump to conclusions. I'll bet you've usually gone with a crowd of people, right?" "Yes," admitted Jane, "that's true. I usually go with half a dozen people or more from work." "Well, that may be it," said Ellen. "People usually go to Chinese restaurants with a crowd of people they hardly know. I know you, you're often tense and a little shy, and you're not likely to be able to relax and savor the food under those circumstances. Try going to a Chinese restaurant with just one good friend. I'll bet you'll like the food."

Comment on Ellen's reasoning. Do you think there is a good chance that if Jane went to a Chinese restaurant with one friend, she'd like the food? Why or why not?

Problem 7. Subjective Format, Structure 1 (conclusion from a single small sample with part A as a false alarm question) This problem is identical to problem 8 in Appendix B, "Gerald M. had a 3-year-old son ..."

Gerald M. had a 3-year-old son, Timmy. He told a friend: "You know, I've never been much for sports, and I think Timmy will turn out the same. A couple of weeks ago, an older neighbour boy was tossing a ball to him, and he could catch it and throw it all right, but he just didn't seem interested in it. Then the other day, some kids his age were kicking a little soccer ball around. Timmy could do it as well as the others, but he lost interest very quickly and started playing with some toy cars while the other kids went on kicking the ball around for another 20 or 30 minutes."

Do you agree with Gerald's reasoning that Timmy is likely not to care much for sports? Why or why not?

Problem 8. Probabilistic Format, Structure 2 (large sample pitted against a plausible theory not founded on data) This problem is based on Problem 9 in Appendix B, "An auditor for the ..."

Steve works for a city newspaper in its distribution department. His job is to ensure that the papers are ready for the many different delivery routes in the city. Steve wants to know if his helpers are placing the correct advertising leaflets inside each folded newspaper. During several weeks, he checks a total of 3000 newspapers. He does this by randomly picking the papers out of the delivery stacks. Steve finds errors in a large number of the papers. Often, 2 to 6 leaflets or advertising flyers are missing from a single paper. Counting the errors again, he finds that just as many papers had extra leaflets placed in them as missing leaflets.

Steve's supervisor says that Steve must be wrong. The supervisor states that it is fairly obvious the helpers will put less leaflets in, not more. If they put in less, they'll save time on the job. He says that even if Steve's counting is correct, looking at more papers will help to show that the supervisor is right.

Write about Steve's reasoning and the supervisor's different point of view.
Problem 9. Subjective Format, Structure 3 (regression). This problem is identical to problem 10 in Appendix B, "Janice is head nurse in a home ..."

Janice is head nurse in a home for the aged. She says the following of her experiences: "There is a big turnover of the nursing staff here, and each year we hire 15-20 new nurses. Some of these people show themselves to be unusually warm and compassionate in the first few days. One might stay on past quitting time with a patient who's having a difficult night. Another might be obviously shaken by the distress of a patient who has just lost a spouse. I find though that, over the long haul, these women turn out to be not much more concerned and caring than the others. What happens to them, I think, is that they can't remain open and vulnerable without paying a heavy emotional price. They usually continue to be considerate and effective, but they build up a shell."

Comment on Janice's reasoning. Do you think it is likely that she correctly identifies the nurses who are unusually warm and compassionate? Do you agree it is likely that most of the ones who are unusually warm at first later build up a shell to protect themselves emotionally?

Problem 10. Probabilistic Format, Structure 1 (conclusion from a single small sample)
This problem is based on Problem 11 in Appendix B, "At Stanbrook University ..."

At Hollybrook Senior Secondary School, the administration has to determine which of the 1200 new students registering to attend the school each year, will have their own locker. The lockers are roomy and conveniently located, but there are only 900 of them. This means that 600 students will have their own lockers and each of the remaining 600 will have to share with a locker partner. The administration office makes the decisions by holding a "locker lottery" at the beginning of each school year. Every new student picks a number out of a drum in the administration office during the first week of school. If the number is over 600, the student obtains his/her own locker for the year. If the number is 600 or under, the student must share a locker with another student.

During the week, Stanley talks to six people who have picked a number by taking one from the top of the pile in the drum. Five out of the six people obtained numbers under 600. Stanley decides that perhaps the lower numbers are on top of the pile of numbers. He decides to reach deeply into the numbers and rushes over to the administration office to pick a number. He gets a high number. Later, he talks to five people who drew their numbers by digging deeply into the drum. Four of the five people got high numbers. Stanley says to himself, "I'm glad I reached correctly into the drum, because it looks like I was right that the numbers were not well mixed."

What do you think of Stanley's reasoning? Explain.
APPENDIX H

Posttest Problems and Directions: Experiment 2

Directions

On the pages that follow, there are a number of problems that we would like you to consider. As you will see, they represent a wide range of real-life situations. We would like you to think carefully about each problem, and then write down answers that are sensible to you. In many of the problems, you may find that the **law of large numbers** is helpful.

Since we are interested in studying instructional effects, please indicate how much time you spent with the original set of pretest problems (check one).
1. I read and wrote answers to all of the items. ___
2. I read and wrote answers to about half of the items. ___
3. I read and wrote answers to one or two of the items. ___
4. I read the problems. ___
5. I didn't read or write answers to any of the problems. ___

An economist was arguing in favour of a guaranteed minimum income for everyone. He cited a recent study of several hundred people in the United States with inherited wealth. Nearly 92% of those people, he said, worked at some job that provided earned income sufficient to provide at least a middle-class life style. The study showed, he said, that contrary to popular opinion, people will work in preference to being idle. Thus a guaranteed income policy would result in little or no increase in the number of people unwilling to work.

Comment on the economist's reasoning. Is it basically sound? Does it have weaknesses?

A talent scout for a professional basketball team attends two college games with the intention of observing carefully the talent and skill of a particular player. The player looks generally excellent. He repeatedly makes plays worthy of the best professional players. However, in one of the games, with his team behind by 2 points, the player is fouled while shooting and has the opportunity to tie the game by making both free throws. The player misses both free throws and then tries too hard for the rebound from the second one, committing a foul in the process. The other team then makes two free throws, for a 4-point lead, and goes on to win by 2 points.

Comment on the thinking embodied in the scout's opinion that the player (a) "has excellent skills" and that the player has (b) "a tendency to misplay under extreme pressure." Does the thinking behind either conclusion have any weaknesses?
Howard was a teacher in a junior high school in a community known for truancy and delinquency problems among its youth. Howard says of his experiences: "Usually, in a class of 35 or so kids, 2 or 3 will pull some pretty bad stunts in the first week -- they'll skip a day of class, get into a scuffle with another kid, or some such thing. When that kind of thing happens, I play it down and try to avoid calling the class' attention to it. Usually, these kids turn out to be no worse than the others. By the end of the term you'll find they haven't pulled any more stunts than the others have." Howard reasons as follows: "Some of these kids are headed toward a delinquent pattern of behaviour. When they find out nobody is very impressed, they tend to settle down."

Comment on Howard's reasoning:

(a) Do you agree that it is likely that the students who pull a "pretty bad stunt in the first week" are "headed toward a delinquent pattern of behaviour?"

(b) Do you agree that it is likely that the students who initially pull a "pretty bad stunt" turn out to be no worse than the others because they find no one is impressed with their behaviour?

The superintendent of schools was urging the school board to make an expensive curriculum shift to a "back-to-basics" stress on fundamental learning skills and away from the electives and intensive immersion in specialized arts and social studies topics that had recently characterized the secondary schools in the district. He cited a study of 120 school systems that had recently begun to emphasize the basics and 120 school systems that had a curriculum similar to the district's current one. The "back-to-basics" school systems, he said, were producing students who scored half a year ahead of the students in the other systems on objective tests of reading, mathematics, and science. Of the 120 "back-to-basics" school systems, 85 had shown improved skills for students in the system vs only 40 with improved skills in the 120 systems which had not changed. One of the school board members took the floor to argue against the change. In her opinion, she said, there was no compelling reason to attribute the improved student skills in the "back-to-basics" systems to the specific curriculum change, for two reasons: (1) school systems that make curriculum changes probably have more energetic, adventurous administrators and faculty and thus the students would learn more in those school systems no matter what the curriculum was. (2) Any change in curriculum could be expected to produce improvement in student performance because of increased faculty interest and commitment.

Comment on the reasoning of both the superintendent and the board member. On the basis of the evidence and arguments offered, do you think it is likely that the "back-to-basics" curriculum is intrinsically superior to the district's current curriculum?

Bert H. has a job checking the results of an X-ray scanner of pipeline welds in a pipe factory. Overall, the X-ray scanner shows that the welding machine makes a perfect weld about 80% of the time. Of 900 welds each day, usually about 680 to 740 welds are perfect. However, Bert has also noticed that on such days, the overall number of perfect welds is usually not much better for the day as a whole than on days when the first 10 welds show some imperfections.

Why do you suppose the number of perfect welds is usually not much better on days where the first batch of welds was perfect than on other days?
An auditor for the Internal Revenue Service wants to study the nature of arithmetic errors on income tax returns. She selects 4000 Social Security numbers by using random digits generated by an "Electronic Mastermind" calculator. And for each selected social security number she checks the 1978 Federal Income Tax return thoroughly for arithmetic errors. She finds errors on a large percentage of the tax returns, often 2 to 6 errors on a single tax return. Tabulating the effect of each error separately, she finds that there are virtually the same number of errors in favour of the taxpayer as in favour of the government. Her boss objects vigorously to her assertions, saying that it is fairly obvious that people will notice and correct errors in favour of the government, but will "overlook" errors in their own favour. Even if her figures are correct, he says, looking at a lot more returns will bear out his point.

Comment on the auditor's reasoning and her boss's contrary stand.

At Stanbrook University, the Housing Office determines which of the 10,000 students enrolled will be allowed to live on campus the following year. At Stanbrook, the dormitory facilities are excellent, so there is always great demand for on-campus housing. Unfortunately, there are only enough on-campus spaces for 5000 students. The Housing Office determines who will get to live on campus by having a Housing Draw every year: every student picks a number out of a box over a 3-day period. These numbers range from 1 to 10,000. If the number is 5000 or under, the student gets to live on campus. If the number is over 5000, the student will not be able to live on campus.

On the first day of the draw, Joe talks to five people who have picked a number. Of these, four people got low numbers. Because of this, Joe suspects that the numbers in the box were not properly mixed, and that the early numbers are more favourable. He rushes over to the Housing Draw and picks a number. He gets a low number. He later talks to four people who drew their numbers on the second or third day of the draw. Three got high numbers. Joe says to himself, "I'm glad that I picked when I did, because it looks like I was right that the numbers were not properly mixed."

What do you think of Joe's reasoning? Explain.
APPENDIX I

Probes of Conditional Reasoning, Domain Knowledge, and Sampling:

Experiment 2

I.D. Number __________

5-minute Pretest

1. Below are four cards. Each card has a number on one side and a letter on the other side. You are given the following rule: "If a card has an A on one side, then it has a 4 on the other side." Please indicate all and only those cards that must be turned over to determine whether or not the rule holds.

Answer: __________

A  B  4  7

2. Rate how familiar you are with the following areas. That is, how much you feel you know about these topics. Check one for each topic.

(a) Basketball
unfamiliar __ somewhat familiar __ familiar __ very familiar __ highly familiar __

(b) Student Discipline
unfamiliar __ somewhat familiar __ familiar __ very familiar __ highly familiar __

(c) Curriculum
unfamiliar __ somewhat familiar __ familiar __ very familiar __ highly familiar __

(d) Income Tax Returns
unfamiliar __ somewhat familiar __ familiar __ very familiar __ highly familiar __

(e) University Housing
unfamiliar __ somewhat familiar __ familiar __ very familiar __ highly familiar __

(f) Guaranteed Income
unfamiliar __ somewhat familiar __ familiar __ very familiar __ highly familiar __

(g) Welding
unfamiliar __ somewhat familiar __ familiar __ very familiar __ highly familiar __
3. Label the diagram below with the terms, "population" and "sample".

Circle the letter of the correct answer in the following multiple choice items.

4. Which sample gives the best estimate of the population?

5. Which sample gives the best estimate of the population?
6. Which sample gives the best estimate of the population?

(a) 

(b) 

(c) 

(d) 

7. Which sample gives the best estimate of the population?

(a) 

(b) 

(c) 

(d) 

8. Regression to the mean refers to

(a) Why an outcome with an extreme deviation is maintained in a sample.
(b) Why an outcome with an extreme deviation is not maintained in a second sample.
(c) Why an outcome with an extreme deviation is maintained in a second sample.
(d) Why an outcome with an extreme deviation is not maintained in a sample.
APPENDIX J

Law of Good Sampling:
Demonstration Script

Now that we have all read the written explanation of the Law of Good Sampling, I thought it would be exciting to demonstrate, right here before your eyes, that the Law of Good Sampling really does work.

So I have here (pick up and unveil the gumball machine) a genuine gumball machine. It's filled with genuine red and white gumballs. Just like in our written explanation, there happens to be 70% white and 30% red gumballs in this machine.

A very important purpose of statistics is to find out about a population from a sample of that population. Suppose you did not know the distribution of gumballs in the this machine. And, suppose you wanted to find out what proportion is white and what proportion is red. Well, you could take the top off and dump them all out like this (pick up the gumball machine and turn it over to simulate pouring out the gumballs) and count each and every one. But, that kind of idea would take a very long time, and it wouldn't be worth the effort.

For the sake of this demonstration, we have a small population of gumballs. But, if we had a huge container filled with millions of them, it's really easy to see how time-consuming and impractical it would be to count the entire population to find out what the distribution is.

What you would do instead is to take a sample, because the sample tells you about the population. Remember, we read earlier that the sample is an estimate of the population, or a kind of miniature imitation of it.

Now, according to the Law of Good Sampling, when you choose your sample randomly like this (shake up the machine to mix the contents and extract a few gumballs by turning the handle), the larger the sample, the better the sample is in estimating the population. In other words, the more you have in the sample, the better it is at imitating the population.

Well, what we're going to do now is to demonstrate the Law of Good Sampling. (reveal the prepared summary chart) We'll need two or three volunteers to help take out and count gumballs. (organize this making sure the volunteers, gumball machine, and chart are clearly visible to everyone) Together, we'll take out samples of size 1, 4, and 25 (gesture to the three sections of the chart) to show that as the sample size increases, the sample becomes a better estimate of the population. Since we already know what the population distribution is -- 70% whites and 30% reds -- we'll be able to tell how close we get. Remember that in the real world you don't know what the population distribution is. That's why you take a sample in the first place -- to find out. Anyway, the Law of Good Sampling says that the larger samples will be more accurate.

For each sample that we draw, I will write down on the chart the number of whites in the sample, the number of reds in the sample, the percent white in the sample, and the difference between the percent white in the sample and the percent white we already know is in the population. Remember, we call this difference the difference between the sample distribution
and the population distribution (point to the various categories on the chart as you are saying them).

Now, you'll recall that the population distribution for the gumball machine is 70% white and 30% red. Therefore, for example, if the sample we draw happens to be 85% white, the difference of that sample will be 85 minus 70 or 15%, and I'll enter that number here (point to the difference column).

We'll draw a few samples of size 1, 4, and 25. Then we'll calculate the average difference of the samples from the population (point to all three "Average Difference" boxes and provide a brief reminder about how to calculate an average).

The Law of Good Sampling states that as the sample size increases, the sample becomes a better estimate of the population. In other words, this average difference from the population will decrease as the sample size increases. The average difference becomes smaller because the sample becomes a better imitation of the population when the sample is larger. So, this number (point to "Average Difference" box) should go down as the sample size (point to top of chart: "Sample Size= ") goes up.

First, I'll ask the volunteers to draw samples of size 1 (pick up the gumball machine and shake it to mix the contents and then help the volunteers with the first sample; provide volunteers with dishes or containers for separating and counting the gumballs). OK — the first gumball to come out of the machine will be our sample. In this sample, we have 0 whites and 1 red (example). (go to the chart and verbalize as you're recording the results of the sample). So in this sample, we had 0 whites, 1 red. That means that the percent white in the sample was 0%. The difference between the 0% that we got in the sample and what we know is actually in the population is what? (wait for the correct answer). Yes, in this sample of one we have a 70% difference from the population distribution.

(Go back to the machine. Put the sample back in the open top. Repeat the procedure. Draw 6 samples of size 1 and limit the description as you proceed)

(After you're finished, compute the average difference). So, the average difference for samples of size 1 is ___%.

Now we'll take a few samples of size 4. (Follow the same procedure as above. Pick 6 samples of size 4 and then compute the average difference). The average difference for samples of size 4 is ____%.

(Do the same with samples of size 25. Three samples should be enough. With samples of size 25 the extra containers are important. An example follows.)

In this sample of 25 we have 18 whites and 7 reds (write results on the chart). That is 72% white and the difference is 2%.

(If you obtain small differences, you can quit at 3 samples of size 25. Remember to calculate the average difference for samples of size 25.)
Summary Statement:

The Law of Good Sampling states that as the size of your sample increases, the sample becomes more like the population -- that is, it becomes a better estimate of the population. This is shown dramatically on this chart which we just made from our demonstration: As the samples went up in size from 1 to 4 to 25 (point to the top of the chart), the average difference of the samples from the actual population went down from ____% to ____% to ____% (point to the bottom of the chart).

I'd like to tell you something else about the Law of Good Sampling -- something you should try very hard to remember. That is, with small samples, sometimes you can't even correctly answer the simplest questions about the population.

For example, suppose you're asked to state whether there are more white gumballs in the machine or more red gumballs. If you happen to draw this sample (point to a very poor sample of size 4: 1 white and 3 reds, or 2 whites and 2 reds. If there isn't one, go to samples of size 1 for the example). Well, from your sample, you can't tell or you get a distribution that's opposite to the actual distribution. This small sample shows that there are more reds in the gumball machine. You would not be accurate in your thinking about the population.

But now, look at any of the larger samples of 25. You can always correctly, accurately answer at least the most basic kind of question like the one we just asked. The question was, "Are there more whites or more reds?" With smaller samples you can't be sure.

So, together, we've demonstrated that the Law of Good Sampling really does work -- as the sample increases in size, the sample becomes a better estimate of the population. (remember to thank the helpers).
## Demonstration Chart (large sheet of laminated card stock)

<table>
<thead>
<tr>
<th>LAW OF LARGE NUMBERS DEMONSTRATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMPLE = 1</td>
</tr>
<tr>
<td>SAMPLE = 4</td>
</tr>
<tr>
<td>SAMPLE = 25</td>
</tr>
<tr>
<td>W</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

AVERAGE DIFF. FROM POP. =

AVERAGE DIFF. FROM POP. =

AVERAGE DIFF. FROM POP. =
APPENDIX K
Examples and Analyses:
Experiment 3

TWO EXAMPLES

One reason that the Law of Good Sampling is important to learn is that it applies not only to gumballs. The basic rules involved in the Law of Good Sampling apply whenever you think about a sample of objects, actions, or behaviours.

Let's look at two problems like the ones you wrote about earlier. Each situation is briefly analyzed using the Law of Good Sampling.

In other words, The Law of Good Sampling applies to many situations. It doesn't only apply to gumball machines. For example, the ideas about population, sample, and distribution that we demonstrated can be used when you think about general, real-life problems. You can use the Law of Good Sampling to think about what information is important in a problem.

Let's look at two examples to see how this works.

Problem 1: At a large high school, some students who were members of Student Council used a computer to pick names of other students. The computer randomly chose 450 students from the school. Then the Council members handed out a ballot to the 450 chosen students in homeroom classes. On the ballot, students were asked their opinion about increasing the cost of tickets to school dances by 2 dollars for each ticket. The extra money would help the school radio station buy new equipment and new music. After counting all the ballots, the Student Council members calculated that about 72 percent (72 out of every 100) of the 450 students said yes, they would be in favour of paying the extra money; 28 percent of the students said no.

The President of Student Council agreed that the extra 2 dollars for each ticket would help, but she said, "We have so many students in this school who really can't afford the extra money. They wouldn't go to dances if it cost more. I know that most of the students who were asked said yes, they would be willing to pay, but it's certainly far from sure that most of the students in the whole school would be willing to pay. A lot of people haven't been asked about it yet."

Do you agree with the Student Council President that it's "far from sure" that most of the whole student population would be willing to pay more for the tickets? Explain.

Please think about this problem for a few minutes. After you have considered the problem for a minute or two, turn the page for our analysis.
Analysis of Problem 1:

The members of Student Council are trying to find out about the attitude of the students at the high school toward spending more money on dance tickets. The money would be used to improve the school radio station. To use the words and ideas from The Law of Good Sampling, they are trying to find out the population distribution of high school students' attitudes toward paying more for the tickets. To do this, they randomly picked 450 students - a sample size of 450 -- and asked the people in the sample if they were in favour of the price increase. 72 percent of the 450 said yes (and 28 percent said no).

The Law of Good Sampling says that the bigger the sample is, the better it is at estimating the population. Well, here there is a lot of evidence to show that most of the students in the whole school are in favour of the idea. Because, if you remember in the gumball demonstration, randomly picked samples of 25 were very good estimates of the population. It meant that the samples were not much different from the whole population. Well, here there is a sample size of 450 randomly picked students! This is an extremely accurate estimate of the population. From this, we know that most of the students in the whole school would pay the extra 2 dollars for dance tickets.

What about the President's argument that most students probably do not agree with the price increase because they can't afford it? Although this argument sounds OK, it is not an accurate one because she ignores the information available. In fact, the president's argument is contradicted by the large sample of 450 students.

Problem 2: A very successful, professional soccer team had always picked their new players only from well-known, big-city soccer clubs. One day, one of the soccer team's coaches decided to try out two players from a small-town soccer club. The two young players, who nobody had ever heard of, were asked by the coach to join the professional team.

Although their soccer skills were not as obviously great as the skills of the big-city players, the small-town players were good team members. They played as many games as everybody else in the first season, and they scored a lot of goals. By the end of the season, both of them had proven to be excellent players.

The coach who had recruited the small-town players said to the other coaches, "This experience tells us that players from small towns are just as excellent and talented as players from the big city clubs. The main difference is in some of their skills. There isn't any difference in how many goals they score."

What do you think of the coach's conclusion? Is the argument basically strong? Does it have weaknesses?

Please think about this problem for a few minutes. After you have considered the problem for a minute or two, turn the page for our analysis.
**Analysis of Problem 2:**

The coach is trying to say something about a population. We can think of the members of this population as all soccer players from small towns. If we divide the members of the population into two categories, "excellent" and "not excellent" players from small towns, we can then think of how many there are in each category. The percent of players we find in each category is the population distribution.

The coach is arguing that the percent in the "excellent" category is very high, or just as high as it is in another population. That other population is all soccer players from big-city clubs. The coach's conclusion is based on observing a sample of size of 2. In that sample of 2, the sample distribution was 100 percent "excellent" and 0 percent "not excellent".

Even if we don't consider anything else, the sample distribution for a sample size of 2 is very likely to be different from the population distribution. A 2 against 0 sample split would not be so unusual, just as we would not be surprised to draw 2 out of 2 red gumballs from a jar with only 30 percent red gumballs. So, the coach's conclusion about all small-town players is not based on good evidence. A larger sample is needed.
APPENDIX L

Rule Group Teacher Instructions

1. Please distribute the booklet, The Law of Good Sampling Review, and direct students to carefully follow as you read aloud pages 1 - 3. Tell students, "The quiz on the Law of Good Sampling will follow our brief review of the material already introduced in the previous lesson."

2. Direct students to turn to page 9, "A New Example", and read the whole page to them as they again follow along in the print.

3. Direct students to silently read pages 4 - 8 in the booklet.

4. After everyone has completed the reading, collect the booklets.

5. Hand out the Law of Good Sampling Quiz. Administer the quiz. You may assist with any literal explanation of what is meant by the questions. ie. paraphrasing the material so that students understand what they are asked to do. Do not provide any answers or hints.

6. When everyone is done, collect the tests.

7. Take the answer key and read each question along with the correct answer. Hold up the key to point to the map questions where necessary.

8. Tell the students to remember as much as possible about the Law of Good Sampling for the next day. They will be asked to answer questions similar to the first set of six problems. They will be asked to use the Law to think about the information in the problems.
As we have learned from the previous lesson, often people have to explain things without really having very much information to help them think. After all, there are many times in life -- including at school, in jobs, with friends, at home -- when people have to solve problems or make decisions using only what they know about the situation. One reason for this is that people might not have enough time to get more information. Another reason is that there might not be more information available.

Experts who study how people think about problems have discovered special rules that people use to help their minds work more sharply. These rules (or special thinking codes) can help people solve real-life, everyday problems. In a way, the rules help us to decode more accurately what is happening in the problem. One such rule is called the Law of Good Sampling.

Imagine a huge gumball machine filled with gumballs. Let's say that the machine contains a very large number of gumballs -- thousands, millions, or larger. The gumballs in this gumball machine are called the population.

Let's say that there are two types of gumballs in the machine -- red gumballs and white gumballs. When we do this, we can now say that the population has two categories or groups, namely - red gumballs and white gumballs.

Here's a reminder before you read on: The symbol % means percent which is the same as "out of 100". For example, 25% is 25 out of 100.

Now let's say that we already know the following: In this population of red and white gumballs there are 70% white gumballs and 30% red gumballs. This means that we know a lot more about the population. We certainly know more than just the colour of the two categories, red and white. We know about something called the population distribution. The population distribution is 70% white and 30% red. (Other examples of population distributions are 60% red and 40% white; or 85% white and 15% red, and so on; but, in every distribution, the sum of the parts must be 100%).
Here's what we've said so far:

- **Population** is the complete set of things we are interested in (all the gumballs in the gumball machine).

- **Categories** means the types of things in the population (red and white).

- **Distribution** means how much there is in each category (70% white and 30% red in our example).

Statistics is an excellent way to find out something about a population when we don't know much about the population. Mainly, we want to find out what the population distribution is (how much there is in each category) when we do not already know the population distribution. One way we could do this is to count all the gumballs in that huge gumball machine we talked about earlier. Using counting, we could find out exactly what the population distribution of red and white gumballs is.

But, there is a very serious flaw with the counting idea. The difficulty is that populations, usually, are very large. Imagine if there were a million gumballs! It would take too much effort and too much time.

OK -- so what can we do instead to find the population distribution?

What we do is to take a sample of the population. A sample is a smaller version of the population. It's like an attempt to make a miniature imitation of the larger population. We can take a sample of any size. If we pick 5 gumballs, we say that the sample size is 5. If we take 60 gumballs for our sample, we say the sample size is 60, and so on.

When we take a sample from a population, we will get a sample distribution. The sample distribution means how much there is in each category in that particular miniature imitation. Remember how the population distribution told us how much there was for each category in the complete version of the population. Well, the sample distribution works in much the same way. For example, if we take a sample of 10 gumballs, we might get 6 reds and 4 whites. If this happens, our sample distribution would be 60% red and 40% blue. We also might have happened to get 9 whites and 1 red, in which case the sample distribution would be 90% white and 10% red.

The very important point is that the samples we've been talking about are estimates of populations. Since it is so often not practical or sometimes not possible to count the entire population, we have to draw samples or smaller imitations instead. These samples are estimates of what the population is like.

Some samples will have sample distributions that are close to the population distribution. Other samples will have sample distributions that are different from the population distribution. In other words, some miniature versions of the population will be very good imitations; some will be very bad imitations. For example, from our gumball machine, a sample of 9 reds and 1 white would be a very poor estimate of the population, while a sample of 8 whites and 2 reds would be a pretty good estimate of the population because it is closer to the real population distribution: 70% whites and 30% reds.
The really important question is: HOW CAN WE MAKE SURE THAT SAMPLES GIVE US GOOD ESTIMATES OF THE POPULATION?

The answer is simple: If the samples are chosen haphazardly, or randomly (by, for example, mixing all the gumballs in a jar and reaching into the jar blindfolded and scooping out the sample of gumballs; or by mixing all the gumballs in a gumball machine and then turning the handle to let a sample out), then there is only one thing that makes any difference: SAMPLE SIZE. To repeat, if we take a sample randomly, only SAMPLE SIZE will make any difference.

This brings us right back to the Law of Good Sampling: as the size of a random sample gets larger, the sample distribution is more and more likely to get closer and closer to the population distribution. In other words, the larger the sample, the better it is as an estimate of the population. If the randomly-scooped or randomly-taken sample is big enough, it is like a very good miniature imitation of the population. Therefore, we can look at it to tell us what the population is like.
OUR EXAMPLES REVIEWED

One reason that the Law of Good Sampling is important to learn is that it applies not only to gumballs. The basic rules involved in the Law of Good Sampling apply whenever you think about a sample of objects, actions, or behaviours.

Let's look at the two problems like the ones you wrote about earlier. Remember, we studied these in the previous lesson. Each situation is briefly analyzed using the Law of Good Sampling.

In other words, The Law of Good Sampling applies to many problem situations. It doesn't only apply to gumball machines. For example, the ideas about population, sample, and distribution that we demonstrated can be used when you think about general, real-life problems. You can use the Law of Good Sampling to think about what information is important in a problem. Then, you can think about how that information could help you think more sharply about the problem.

Let's look at the two examples again to refresh your memory about how this works.

Problem 1: At a large high school, some students who were members of Student Council used a computer to pick names of other students. The computer randomly chose 450 students from the school. Then the Council members handed out a ballot to the 450 chosen students in homeroom classes. On the ballot, students were asked their opinion about increasing the cost of tickets to school dances by 2 dollars for each ticket. The extra money would help the school radio station buy new equipment and new music. After counting all the ballots, the Student Council members calculated that about 72 percent (72 out of every 100) of the 450 students said yes, they would be in favor of paying the extra money; 28 percent of the students said no.

The President of Student Council agreed that the extra 2 dollars for each ticket would help, but she said, "We have so many students in this school who really can't afford the extra money. They wouldn't go to dances if it cost more. I know that most of the students who were asked said yes, they would be willing to pay, but it's certainly far from sure that most of the students in the whole school would be willing to pay. A lot of people haven't been asked about it yet."

Do you agree with the Student Council President that it's "far from sure" that most of the whole student population would be willing to pay more for the tickets? Explain.

Please think about this problem for a few minutes. After you have considered the problem for a minute or two, turn the page for our analysis.
Analysis of Problem 1:

The few members of Student Council are trying to find out about the attitude of the students at the high school toward spending more money on dance tickets. The money would be used to improve the school radio station. To use the words and ideas from The Law of Good Sampling, they are trying to find out the population distribution of high school students' attitudes toward paying more for the tickets. What are the categories in the population distribution? One category is students who agree with the idea. The other category is students who do not agree.

To find out what the population distribution is, the Council members randomly picked 450 students -- a sample size of 450 -- and asked the people in the sample if they were in favour of the price increase. 72 percent of the 450 said yes (and 28 percent said no).

The Law of Good Sampling says that the bigger the sample is, the better it is at estimating the population. Well, here there is a lot of evidence to show that most of the students in the whole school are in favour of the idea. Because, if you remember in the gumball demonstration, randomly picked samples of 25 were very good estimates of the population. It meant that the samples were not much different from the whole population. Well, here there is a sample size of 450 randomly picked students! This is an extremely accurate estimate of the population. From this, we know that most of the students in the whole school would pay the extra 2 dollars for dance tickets.

What about the President's argument that most students probably do not agree with the price increase because they can't afford it? Although this argument sounds OK, it is not an accurate one because she ignores the information available. In fact, the president's argument is contradicted by the large sample of 450 students.

Problem 2: A very successful, professional soccer team had always picked their new players only from well-known, big-city soccer clubs. One day, one of the soccer team's coaches decided to try out two players from a small-town soccer club. The two young players, who nobody had ever heard of, were asked by the coach to join the professional team.

Although their soccer skills were not as obviously great as the skills of the big-city players, the small-town players were good team members. They played as many games as everybody else in the first season, and they scored a lot of goals. By the end of the season, both of them had proven to be excellent players.

The coach who had recruited the small-town players said to the other coaches, "This experience tells us that players from small towns are just as excellent and talented as players from the big city clubs. The main difference is in some of their skills. There isn't any difference in how many goals they score."

What do you think of the coach's conclusion? Is the argument basically strong? Does it have weaknesses?

Please think about this problem for a few minutes. After you have considered the problem for a minute or two, turn the page for our analysis.
Analysis of Problem 2:

The coach is trying to say something about a population. We can think of the members of this population as all soccer players from small towns. If we divide the members of the population into two categories, "excellent" and "not excellent" players from small towns, we can then think of how many there are in each category. The percent of players we find in each category is the population distribution.

The coach is arguing that the percent in the "excellent" category in a population distribution of small-town players is very high, or just as high as it is in another population. That other population is all soccer players from big-city clubs. The coach's conclusion is based on observing a sample of size 2. In that sample of 2, the sample distribution was 100 percent "excellent" and 0 percent "not excellent".

Even if we don't consider anything else, the sample distribution for a sample size of 2 is very likely to be different from the population distribution. A 2 against 0 sample split would not be so unusual, just as we would not be surprised to draw 2 out of 2 red gumballs from a jar with only 30 percent red gumballs. So, the coach's conclusion about all small-town players is not based on good evidence. According to the Law of Good Sampling, larger sample is needed.

A New Example

Problem 3: Two New Yorkers were discussing restaurants. Jane said to Ellen, "You know, most people seem to be crazy about Chinese food, but I'm not. I've been to about 20 different Chinese restaurants, across the whole price range, and everything from bland Cantonese to spicy Szechwan and I'm really not very fond of any of it." "Oh," said Ellen, "don't jump to conclusions. I'll bet you've usually gone with a crowd of people, right?" "Yes," admitted Jane, "that's true. I usually go with half a dozen people or more from work." "Well, that may be it," said Ellen. "People usually go to Chinese restaurants with a crowd of people they hardly know. I know you, you're often tense and a little shy, and you're not likely to be able to relax and savor the food under those circumstances. Try going to a Chinese restaurant with just one good friend. I'll bet you'll like the food."

Comment on Ellen's reasoning. Do you think there is a good chance that if Jane went to a Chinese restaurant with one friend, she'd like the food? Why or why not?

Analysis:

The question here is whether or not Jane is likely to enjoy Chinese food the next time she visits a Chinese restaurant. We can think of Jane's experiences in Chinese restaurants as a population. So far, Jane has drawn a sample of 20 out of this population. We can divide the members of the population into two categories, liking Chinese food and disliking Chinese food. Based on Jane's sample of size = 20 the sample distribution was 100 percent disliking the food and 0 percent liking the food. According to the Law of Good Sampling, a sample size of 20 is enough to resemble the population distribution, especially since Jane has experienced quite a variety of Chinese restaurants in the sample. This latter consideration makes the sample more like a random sample. Therefore, Ellen's reasoning is weak, based on the evidence.
APPENDIX N

Probe of Sampling Knowledge: Experiment 3
Name: ______________________ Grade: _______

Law of Good Sampling Quiz

Read the following story.

Charles L. is a forester. This means that he knows a lot about trees, plants, and forest management. Charles is also an expert on hardwood trees. The wood from the hardwood trees, such as birch and oak, is used to make furniture.

A furniture-making company owns a large area of forest, where, among other trees, there are many birch and oak trees. The company's factory uses the wood from these trees to make furniture. The company decided to ask Charles to see how well the trees are growing in their forest. They asked Charles, "Are there more oak trees than birch trees?" Since oak is more valuable than birch, the company hopes that there are more oak trees.

The forest is large, and Charles decides to find out the answer to the above question by using the Law of Good Sampling. He doesn't want to count all the trees, so he makes a grid map of the forest area that looks like this:

[Grid Map of the Forest Area]

Charles decides to count hardwood trees in some, but not all, of the small squares on the map. He will know which square he is counting in with the help of special equipment that he carries with him. He wants to use the Law of Good Sampling correctly.
Help Charles with his problem by answering the following questions. Your answers will give him advice about how the Law of Good Sampling works.

For each question below, write your answer on the line:

1. Describe the population of this problem. __________________

2. What are the categories in the population?

3. Charles will count trees in some of the small squares. He should do his counting in about how many squares, according to the Law of Good Sampling? ______

For each question below, circle the correct answer:

4. Charles visits the black squares below to count oak and birch trees. Is he using the Law correctly? YES NO

Grid Map of the Forest Area
3. Charles visits the black squares below to count oak and birch trees. Is he using the Law correctly?  YES  NO

Grid Map of the Forest Area

6. Charles visits the black squares below to count oak and birch trees. Is he using the Law correctly?  YES  NO

Grid Map of the Forest Area
7. Charles visits the black squares below to count oak and birch trees. Is he using the Law correctly? YES NO

Grid Map of the Forest Area


10. By filling in squares on the grid map below, show the approximate location and the number of squares where Charles should count trees if he uses the Law of Good Sampling correctly.

Grid Map of the Forest Area
11. Look at the grid map in question 5. Charles counted trees in the five black squares. Four out of the five had more oak trees. What should he tell the furniture company about the accuracy of this result?

12. Look at the grid map in question 6. Twenty out of the twenty-four black squares had more birch trees. What should he tell the furniture company about the accuracy of this result?

13. Look at the grid map in question 4. All of the black squares had more oak trees. What should Charles tell the furniture company about the accuracy of this result?

14. What is the distribution of the sample in question 11 above?

15. What is the distribution of the sample in question 12 above?
APPENDIX O

Error Materials: Experiment 3

Name:_________________________ Grade:______

Wild BOARS and Thinking Mistakes
Wild BOARS and Thinking Mistakes

The ideas you just learned about in the Law of Good Sampling are very useful for thinking about certain kinds of problems. What kinds of problems are these?

The Law of Good Sampling can be used with many problems in everyday life. See if you can find something in a problem that can be counted. See if you can take a sample of something. That something might be almost anything, including things that happen over and over again, or somebody's actions, or somebody's behaviours -- anything that can make up a whole population of objects, or events, or behaviours. Sometimes people have trouble using the Law of Good Sampling because the things that can be counted are not so obvious in the problem. Sometimes people just don't pay any attention to that kind of information.

What else might they do instead if they ignore the information they could use with the Law of Good Sampling. They might explain the problem very quickly by using other, weaker types of reasons. In other words, they might jump to conclusions without thinking about the information for just a few more seconds. Picture a wild boar. It doesn't think or analyze the situation before it attacks whatever it sees. Problem solvers are sometimes like a wild boar.
Wild BOARS

B Blaming reasons  Problem solvers sometimes blame the people or the person in the problem and say that they have a prejudice to act in a certain way. For example, they might say, "That individual is narrow-minded." or "They're ambitious, and that's why they act that way...."

O Opinion reasons  Problem solvers sometimes think their own personal opinion is most important in answering the problem. For example, they might say, "I just don't agree with it." or "If that happened to me, I would...."

A Always reasons  Problem solvers sometimes believe that some things, like a situation in the problem, are always true even when there might be another explanation. For example, they might say, "It's only human nature." or "Everybody makes mistakes."

R Information that is not really in the problem. This is just reckless, because it can result in a mistake, in this case, a thinking mistake.
EXAMPLES OF MISTAKES

Let's look again at the same problems we read earlier, but this time in a different way. First, read Problem 1, and then study the four examples of thinking mistakes.

Problem 1: At a large high school, some students who were members of Student Council used a computer to pick names of other students. The computer randomly chose 450 students from the school. Then the Council members handed out a ballot to the 450 chosen students in homeroom classes. On the ballot, students were asked their opinion about increasing the cost of tickets to school dances by 2 dollars for each ticket. The extra money would help the school radio station buy new equipment and new music. After counting all the ballots, the Student Council members calculated that about 72 percent (72 out of every 100) of the 450 students said yes, they would be in favour of paying the extra money; 28 percent of the students said no.

The President of Student Council agreed that the extra 2 dollars for each ticket would help, but she said, "We have so many students in this school who really can't afford the extra money. They wouldn't go to dances if it cost more. I know that most of the students who were asked said yes, they would be willing to pay, but it's certainly far from sure that most of the students in the whole school would be willing to pay. A lot of people haven't been asked about it yet."

Do you agree with the Student Council President that it's "far from sure" that most of the whole student population would be willing to pay more for the tickets? Explain.

B Blaming I don't agree with the President, because she's obviously biased and doesn't want the price of dance tickets to go up.

O Opinion I agree with the President, because in my opinion, it's not worth the extra two dollars.

A Always The President is right, because all people in any high school would not care about a radio station.

R Reckless The President probably likes to listen to classical and country music. The people running the school radio station do not play the songs she likes.
Next, for Problem 1, identify each of the following as mistake type B, O, A, or R. Circle the letter of the correct answer below, and then mark it on the attached sheet. You may look back to the previous pages for reminders.

1. Asking people for their opinion never proves anything.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

2. The Student Council Members are just crazy about music.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

3. The President of the Council is obviously a penny-pincher.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

4. School dances should never ever be used to raise money.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

5. Personally, I'm not interested in better radio stations.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

6. I'm in student government, and I wouldn't agree to raise prices.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

7. I don't believe in voting.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

8. People who didn't get asked are probably very angry.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

9. School radio stations are always wanting more money.
   (a) Blaming  (b) Opinion  (c) Always  (d) Reckless

10. The President probably hates dances.
    (a) Blaming  (b) Opinion  (c) Always  (d) Reckless
Problem 2: A very successful, professional soccer team had always picked their new players only from well-known, big-city soccer clubs. One day, one of the soccer team's coaches decided to try out two players from a small-town soccer club. The two young players, who nobody had ever heard of, were asked by the coach to join the professional team.

Although their soccer skills were not as obviously great as the skills of the big-city players, the small-town players were good team members. They played as many games as everybody else in the first season, and they scored a lot of goals. By the end of the season, both of them had proven to be excellent players.

The coach who had recruited the small-town players said to the other coaches, "This experience tells us that players from small towns are just as excellent and talented as players from the big city clubs. The main difference is in some of their skills. There isn't any difference in how many goals they score."

What do you think of the coach's conclusion? Is the argument basically strong? Does it have weaknesses?
For Problem 2, identify each of the following as mistake type B, O, A, or R. You may look back to the previous pages

11. The coach is wrong. Big cities always have better players.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

12. The coach is right, because I'm from a small town and am great in sports.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

13. The coach is the type of person who is looking for a promotion.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

14. The other coaches have bad attitudes toward players from small towns.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

15. From my own experience in soccer, I know he's wrong.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

16. The argument is weak because town size does not matter.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

17. Soccer is certainly not my favourite game.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

18. The two players probably had sponsors and got better equipment.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

19. The coach's thinking is limited, because he's from a small town.
   (a) Blaming (b) Opinion (c) Always (d) Reckless

20. It's likely that the players were the coach's nephews.
   (a) Blaming (b) Opinion (c) Always (d) Reckless
APPENDIX P

Pretest Problems, Posttest Problems, and Directions: Experiment 3

Name: ___________________________
Grade: _________

DIRECTIONS

On the pages that follow, there are six problems to think about. As you will see, the problems come from a wide range of real-life situations.

Please read each problem. Then think carefully about it and write down an answer that makes sense to you. You are asked to write several sentences to make each complete answer. You are also asked to do two more things:

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
A talent scout for a college basketball team goes to two senior high school basketball games. His purpose is to observe a player. The player looks generally excellent in both games, making great plays over and over again. However, in one of the two games, when his team is losing by two points, the player is fouled by a player from the opposing team. Because of this foul, the player gets a chance to tie the game score by completing two free throws. But, he misses both free throws. Then, he tries too hard and lets the other team get a chance for free throws. The other team makes both shots for a 4-point lead and later ends up winning the basketball game by 2 points.

The talent scout writes a report where he says, "The player has excellent skills, but is not good under pressure. He can probably learn to get over this with better coaching."

What do you think about the scout's opinions on (a) that the player "has excellent skills" and (b) that the player "is not good under pressure"? Does the scout's thinking have any weaknesses? Write your answers below.

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
It was the first week of school, and again Debbie had volunteered to help monitor a primary class. She watched over the grade 3's on rainy "inside" days while their teacher went for lunch in the staff room. Debbie talks about her previous years' experiences monitoring other grade 3's: "Usually, in a class of 25 or so kids, one or two will 'goof off' in the first week. They might throw chalk, stand on desks, things like that. When it happens, I ignore it and try not to call attention to their behaviour. Usually, these kids don't turn out to be any worse than the others." Debbie also says, "Some of these kids are headed toward a pattern of bad behaviour. When they find out nobody is very impressed, they tend to settle down."

Write about Debbie's thinking: (a) Do you agree that the kids in this problem who "goof off" are headed toward a pattern of bad behaviour? (b) Do you agree that the kids who "goof off" don't turn out to be any worse than the others because nobody is very impressed with their behaviour?

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
The manager of a toy-making company was talking to owners of the company. She said the plan for next year's toys should include more "basic" toys like dolls, teddy bears, and dump trucks and less of the "high tech" electronic toys like video games. She said this would help to make more money for the company. The manager described a new research study comparing 120 toy companies that had decided to sell such "basic" toys with another 120 companies that had decided to sell "high tech" electronic toys. She explained that the basic toy companies were selling more than the others. In fact, 85 of the 120 basic toy companies were making more money while only 40 of the 120 high tech companies had increased their profits.

One of the owners argued against the manager's plan. The owner said there was no good reason to say that selling basic toys would make more money. The owner gave two reasons: (1) Toy companies that make big changes probably have better sales people and would make more money no matter what they did; and, (2) Any change in the type of toys would make more money because everybody who worked for a company would get interested in something different and work harder.

What do you think about the reasons given by both the toy company manager and the owner? After considering the information given in the problem, do you think that basic toys would sell better than high tech toys?

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
Brenda has a summer job at a well-known hamburger restaurant. The restaurant has brand-new, automatic food-wrapping machines. One of Brenda’s duties is to check how the new packaging equipment is working. Overall, the machines wrapped hamburgers and put french fries and pies into their containers perfectly about 80 percent of the time. This meant that out of 900 items done every day, about 680 to 740 packages are perfectly done. Brenda has noticed something. On some days, all of the first 10 packages were perfect. However, Brenda also noticed that for those days, the overall number of perfectly done packages is about the same as on any other days, about 80 percent. This overall number was the same as when the first ten were not perfect.

Why do you think the overall number of perfectly done packages is the same -- even for those days when the first ten are perfect?

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
Kelly works for a city newspaper in its distribution department. Her job is to make sure the papers are ready for the many different boys and girls with delivery routes in the city. Kelly wants to know if the workers in her department are placing the right advertising leaflets inside each folded newspaper. During several weeks, she checks a total of 4000 newspapers. She does this by randomly picking the papers out of the delivery stacks. Kelly finds errors in a large number of the papers. Often, 2 to 6 leaflets or advertising flyers are missing from a single paper. Counting the errors again, she finds that many papers had extra leaflets placed in them. The number of papers with extra leaflets is about the same as the number of papers with missing leaflets.

Kelly's boss says that Kelly must be wrong. The boss states that it is fairly obvious the helpers will put less leaflets in, not more. If they put in less, they'll save time on the job. The boss argues that even if Kelly's counting is correct, looking at more papers will help to show that the boss is right.

Write about Kelly's reasoning and the boss's different point of view.

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
At Hollybrook Senior Secondary School, the administration office has to make an important decision. They have to decide which of the 1,000 new students registering to attend the school each year, will get their own locker. The lockers are great, but there are only 750 of them. This means that 500 students will have their own lockers and each of the remaining 500 will have to share with a locker partner. The administration office decides by holding a "locker lottery" at the beginning of each school year. Every new student picks a number out of a box during a 3-day period. If the number is 500 or under, the student gets his/her own locker for the year. If the number is over 500, the student must share a locker.

On the first day of the draw, Joe talks to five people who have picked a number. Four out of the five people got a low number. Because of this, Joe suspects that the numbers in the box were not properly mixed. He decides the early numbers are lower and rushes over to the administration office to pick a number. He gets a low number. Later, he talks to four people who drew their numbers on the second or third day of the lottery. Three of the four people got high numbers. Joe says to himself, "I'm glad I picked early, because it looks like I was right that the numbers were not properly mixed."

What do you think of Joe's thinking? Explain.

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
Name: ________________________________

Grade: ________________

DIRECTIONS

On the pages that follow, there are six problems to think about. As you will see, the problems come from a wide range of real-life situations.

Please read each problem. Then think carefully about it and write down an answer that makes sense to you. You are asked to write several sentences to make each complete answer. You may find that using the Law of Good Sampling is helpful.

You are also asked to do two more things:

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
DIRECTIONS

On the pages that follow, there are six problems to think about. As you will see, the problems come from a wide range of real-life situations.

Please read each problem. Then think carefully about it and write down an answer that makes sense to you. You are asked to write several sentences to make each complete answer. You may find that using the Law of Good Sampling is helpful. You may also find that it is helpful to know about thinking mistakes like the four BOARS.

You are also asked to do two more things:

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
The director of a large travelling ice show goes to watch two skating competitions in the town where the show is performing. She wants to observe a local young figure skater who is in both of the competitions. The young skater shows many difficult moves in her skating and does the difficult moves like a professional. But, in one of the two competitions, she misses an important triple jump. Then the skater tries really hard and almost stumbles. She loses points and ends up in third place in that competition.

The director telephones her office and explains that the skater would do well in the ice show, especially with the show's professional coaching. The director also says the skater has excellent skills, but she has a tendency to try too hard.

What do you think about the director's opinions on (a) that the skater has excellent skills; and (b) that the skater tries too hard? Does the Director's thinking have any weaknesses? Write your answers below.

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
Tom is an experienced leader of a group of cub scouts. The scouts come from an area of the city known for its problems with juvenile delinquency. For the annual summer camp, Tom is put in charge of a new group of 25 scouts. He also gets a new assistant leader. He tells the new assistant that some of these boys will try to pull some pretty bad stunts in the first few days. When they do, Tom says, he tries not to notice. He basically does not call any attention to the ones who misbehave. Tom says these one or two scouts don't turn out to be any worse than all the others during the rest of their time at the camp. He tells his assistant that "some of these kids are headed toward delinquency. When they don't get attention for poor behaviour, they tend to settle down".

Write about what you think of Tom's reasoning:

(a) Do you agree that the scouts who "pull some pretty bad stunts in the first few days" are headed toward delinquency?
(b) Do you agree that the scouts who "pull stunts" turn out to be no worse than the others because they don't get attention for their poor behaviour?

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
One of the coaches of a national bicycle racing team was making a speech at a special meeting. She said that the team's racers should have their bikes equipped with the new ZX gearshift system. According to the coach, the new gearshift would give the racers the extra advantage needed to win in the next Olympic Games. She described a recent research study of 90 bicycle racers from other countries who used the new ZX system. The study also looked at another 90 racers who did not use the new system. The study showed that 65 of the 90 ZX racers won medals in their last races. Only 30 of the 90 racers using older equipment won medals.

Another coach stood up to argue with the first coach. In his opinion, there was no good reason to switch to the new ZX gearshift system. He gave two reasons why he thought it would make any difference: (1) Bike racers who try out new equipment are probably better racers anyway; and (2) Any new equipment would help because a bike racing team would really believe in it and therefore team members might try harder.

Write about what you think of the reasons given by the two coaches. After considering the evidence and the arguments offered by the coaches, do you think the ZX gearshift system would help the national bicycle racing team?

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
Bert H. has a job in a factory. He checks the work of a machine that makes running shoes. Generally, over one whole day, the machine makes perfect running shoes about 85 percent of the time. Of a total of 1800 shoes made each day, usually about 1400 to 1560 shoes are perfect. Bert noticed that on some days the first twelve shoes were perfectly made. However, Bert also noticed that it didn't seem to make any difference for the rest of the day, because the overall number of perfect shoes for the day was still about the same as on other days when the first batch showed mistakes. The mistakes included crooked stitching or labels in the wrong places.

Why do you suppose the number of perfect shoes is the same as usual (about 1400 to 1560 perfect shoes out of a total of 1800) for those days when the first twelve shoes are perfect?

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
Barbara works in the office of an old-fashioned department store. The cashiers in the store still add up their bills by hand. Barbara wants to study the arithmetic errors that cashiers sometimes make when they add up customers' purchases. She randomly selects 3000 bills from the previous year. During the next month, she carefully checks each bill for arithmetic errors. She finds mistakes in a very large percentage of the bills, often 2 to 3 errors in one bill.

Barbara wants to find out if the errors are in favour of the store or in favour of the customers. She finds that there are about the same number of each kind of adding mistake. Some mistakes made money for the store and some for the customers. Her boss says that she must be wrong. The boss says the cashiers would notice mistakes where the store lost money to the customers. Even if Barbara's figures are correct, the boss says, Barbara should look at many more bills and these would probably prove she was wrong.

Discuss Barbara's thinking and her boss's opposite opinion.

| 1. Underline the part of the problem that has the most important information for helping you answer the question. |

| 2. Underline the most important part of your written answer. |
A community centre is holding a fundraising raffle as part of an evening fun fair. At a table, people pay their money and then reach into a drum to pick their ticket. The grand prize is a trip to Hawaii, and there are many other prizes. To be eligible for the grand prize, a person must get a ticket with a special scratch code.

Early in the evening, George talks to six of his friends who have already bought raffle tickets. Five out of the six people did not get tickets with the special scratch code. They did not become eligible for the grand prize. As a result, George guesses that the tickets in the drum are not properly mixed up. He waits until later in the evening to buy a ticket. He reaches into the drum and gets a ticket with the special code. Then, he talks to three other people who also bought their tickets later. All three people got tickets with the scratch code. George decides he did the right thing. The drum was certainly not properly mixed.

What do you think of George's thinking? Explain.

1. Underline the part of the problem that has the most important information for helping you answer the question.

2. Underline the most important part of your written answer.
APPENDIX Q

Letter of Permission and Consent Form

April 24, 1991

Dear Parent or Guardian:

During this week I am visiting Mr./Miss __________ Social Studies classes to teach problem solving strategies. Part of the instruction includes a research project involving several schools in the District.

I am writing to request your permission to have your son or daughter participate in the research project. All findings from the project will be analyzed anonymously.

The research investigates what teaching methods are most effective in helping students use problem solving strategies in analyzing information. Students will be taught specific strategies as a whole class and will later be asked to answer questions regarding what they learned from the strategies.

Please return the attached consent form to Mr./Miss __________ by April 30th. If you have any questions about the project, please telephone me at the S.F.U. Faculty of Education, 291-3395.

Thank you for your assistance.

Sincerely,

Pete Kosonen, M.A.
Faculty of Education, Simon Fraser University
Please have your son or daughter return this form to _________ by May 3rd. Thank you.

I give permission for ___________________________ to participate in the research project described in the letter from Mr. Kosonen. I understand that all information about individual participants will be kept in strictest confidence.

__________________________
Parent/Guardian

__________________________
Date

A report on the findings of the study will be available in October, 1991. If you would like to receive a report about the project, please include your name and address below.

__________________________
Name

__________________________
Apt/Street

__________________________
City  Postal Code