

STATISTICS STUDENTS REASONING WHEN
COMPARING DISTRIBUTIONS OF DATA

by

MATTHEW ALAN CIANCETTA

A dissertation submitted in partial fulfillment of the
requirements for the degree of

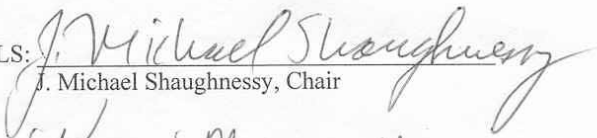
DOCTOR OF PHILOSOPHY
in
MATHEMATICS EDUCATION

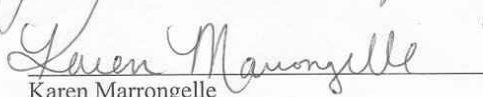
Portland State University
©2007

DISSERTATION APPROVAL

The abstract and dissertation of Matthew Alan Ciancetta for the Doctor of Philosophy in Mathematics Education were presented January 10, 2007, and accepted by the dissertation committee and the doctoral program.

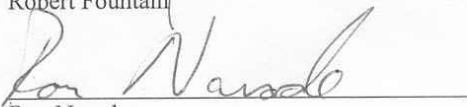
COMMITTEE APPROVALS:


J. Michael Shaughnessy, Chair


Karen Marrongelle


Luis Saldanha


Robert Fountain


Ron Narode
Representative of the Office of Graduate Studies

DOCTORAL PROGRAM APPROVAL:



Karen Marrongelle, Director
Mathematics Education Ph.D. Program

TABLE OF CONTENTS:

List of Tables.....	iv
List of Figures.....	x
Chapter 1: Introduction.....	1
<i>Statistics Education</i>	2
<i>Enculturation</i>	6
<i>Statistical Perspectives</i>	7
<i>Informal Inferences</i>	9
<i>Local vs. Global Views of Data</i>	11
<i>Data set as a Distribution</i>	12
<i>Comparing Data Sets</i>	15
<i>Research Questions</i>	17
Chapter 2: Literature Review and Framework	21
Literature Review.....	21
<i>Intuitive strategies when predicting and making informal inferences when</i> <i>comparing data sets</i>	22
<i>Acknowledgment, understanding and reasoning about variation when</i> <i>comparing data sets</i>	41
<i>Reasoning About Distributions</i>	58
<i>Literature Review Discussion</i>	73
Framework.....	75
<i>Initial Framework by Shaughnessy and Colleagues</i>	77
<i>Lattice Structure Framework By Shaughnessy and Colleagues</i>	79
<i>Framework by Watson and Moritz</i>	81
<i>Framework by Bakker and Gravemeijer</i>	82
<i>Expanded Lattice Structure Framework</i>	83
<i>Conclusion</i>	91
Chapter 3: Methodology	93
Introduction.....	93
Subjects and Data Collection.....	93
Research Design.....	99
<i>Data Collection</i>	99
<i>Task Development: Data Set Comparison Survey</i>	101
<i>Pilot Study</i>	110
<i>Task Development: Interview Protocol</i>	113
<i>Data Analysis</i>	115
Conclusion	119

Chapter 4: Framework Refinement, Results and Analysis	121
Refinement of the Expanded Lattice Structure Framework	121
<i>Level 0 (Idiosyncratic)</i>	125
<i>Level 1 (Local)</i>	127
<i>Level 2 (Transitional)</i>	131
<i>Level 3 (Initial-Distributional)</i>	137
<i>Level 4 (Distributional)</i>	142
<i>Reliability Assessment</i>	144
<i>Cross Task Numeric Codes</i>	147
<i>Framework Refinement Summary</i>	154
Survey Results by Group.....	155
<i>Survey Results: Task 1, the Yellow/Brown task</i>	157
<i>Survey Results: Task 2, the Movie Wait-Time task</i>	165
<i>Survey Results: Task 3, Pink/Black survey task</i>	172
<i>Survey Results: The Pink/Black task – Without descriptive statistics (Task 3)</i> <i>vs. With descriptive statistics (Task 4)</i>	179
<i>Survey response summary: Pink/Black tasks without and with descriptive</i> <i>statistics</i>	204
<i>Survey Results: Task 5, Ambulance task</i>	205
<i>Survey Results: The Ambulance task – Without descriptive statistics (Task 5)</i> <i>vs. With descriptive statistics (Task 6)</i>	212
<i>Survey response summary: Ambulance tasks without and with descriptive</i> <i>statistics</i>	231
<i>Survey Responses: Cross Task Numeric Codes</i>	233
Interview Results.....	237
<i>Analysis of Interviews</i>	237
<i>Background of the six interviewees</i>	237
<i>Cross case analysis of interviewees</i>	240
<i>The interviewees’ understandings of statistical terms</i>	241
<i>Responses to task 1: the Yellow/Brown task</i>	251
<i>Responses to Survey task 2: the Movie-Wait-Time task</i>	258
<i>Responses to task 3 and task 4: the Pink/Black task – Without</i> <i>descriptive statistics and With descriptive statistics</i>	266
<i>Responses to task 5 and task 6: the Ambulance task – Without</i> <i>descriptive statistics and With descriptive statistics</i>	296
<i>Summary of Cross Task Numeric Code assignment</i>	323
Chapter 5: Discussion and Conclusion	328
Research Goal: Expand and refine the interpretive framework.....	328
Research Questions	333
<i>Research Question 1</i>	334

<i>Research Question 2</i>	337
Limitations of the research.....	342
Implication for future research and teaching.....	344
Conclusion	347
References	349
Appendix A: Informed Consent forms	358
Appendix B: Text version of survey tasks	361
Appendix C: Detailed survey results	368

LIST OF TABLES:

Table	Page
1. Response strategies for two ‘comparison of data sets’ interviews, by Watson and Moritz.....	31
2. Course enrollment of participants	95
3. Participants’ Major Fields of Study	96
4. Statistics backgrounds of the participants	97
5. Educational Level breakdown: Counts of each group.....	98
6. Pilot Study results	113
7. Examples of Idiosyncratic type responses.....	125
8. Distribution of Idiosyncratic responses across survey tasks.....	127
9. Examples of Local type responses	128
10. The distribution of Local responses across the survey tasks	130
11. Distribution of Transitional responses across the survey tasks	133
12. Examples of Transitional-shape type responses	134
13. Examples of Transitional-center type responses.....	135
14. Examples of Transitional-variation type responses	137
15. Distribution of Initial-Distributional responses across the survey tasks	139
16. Examples of Initial-Distributional: proportional type responses	140
17. Examples of Initial-Distributional: initial-global type responses	141
18. Examples of Distributional type responses.....	143
19. Distribution of Distributional responses across the survey tasks.....	144
20. Inter-rater reliabilities for coding the survey tasks	146
21. Distribution of Cross Task Numeric Lattice Codes	154

22. Statistics backgrounds of the participants	156
23. The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for all groups.....	162
24. The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for all groups.....	169
25. The distribution of framework level codes, for responses from task 3 (the Pink/Black task), for all groups.....	176
26. Decisions shifts by group for both Pink/Black tasks.....	181
27. Distribution of responses from the 1-GS group for the Pink/Black task: Without statistics vs. With statistics.....	186
28. Distribution of Level 2 and Level 3 responses, from the 1-GS group, for the Pink/Black task: Without statistics and With statistics	186
29. Distribution of responses from the 1-SE group for the Pink/Black task: Without statistics vs. With statistics.....	189
30. Distribution of Level 2 and Level 3 responses, from the 1-SE group, for the Pink/Black task: Without statistics and With statistics	189
31. Distribution of responses from the 2-GS group for the Pink/Black task: Without statistics vs. With statistics.....	192
32. Distribution of Level 2 and Level 3 responses, from the 2-GS group, for the Pink/Black task: Without statistics and With statistics	192
33. Distribution of responses from the 2-SE group for the Pink/Black task: Without statistics vs. With statistics.....	195
34. Distribution of Level 2 and Level 3 responses, from the 2-SE group, for the Pink/Black task: Without statistics and With statistics	196
35. Distribution of responses from the GRAD group for the Pink/Black task: Without statistics vs. With statistics.....	198
36. Distribution of Level 2 and Level 3 responses, from the GRAD group, for the Pink/Black task: Without statistics and With statistics	199
37. The distribution of framework level codes, for responses from task 5 (the Ambulance task), for all groups.....	209

38. Decisions by group for the Ambulance tasks: Counts for Without statistics vs. With statistics	214
39. Distribution of responses from the 1-GS group for the Ambulance task: Without statistics vs. With statistics.....	217
40. Distribution of Level 2 and Level 3 responses, from the 1-GS group, for the Ambulance task: Without statistics and With statistics	218
41. Distribution of responses from the 1-SE group for the Ambulance task: Without statistics vs. With statistics.....	221
42. Distribution of Level 2 and Level 3 responses, from the 1-SE group, for the Ambulance task: Without statistics and With statistics	222
43. Distribution of responses from the 2-GS group for the Ambulance task: Without statistics vs. With statistics.....	224
44. Distribution of Level 2 and Level 3 responses, from the 2-GS group, for the Ambulance task: Without statistics and With statistics	225
45. Distribution of responses from the 2-SE group for the Ambulance task: Without statistics vs. With statistics.....	227
46. Distribution of Level 2 and Level 3 responses, from the 2-SE group, for the Ambulance task: Without statistics and With statistics	228
47. Distribution of responses from the GRAD group for the Ambulance task: Without statistics vs. With statistics.....	230
48. Distribution of Level 2 and Level 3 responses, from the GRAD group, for the Ambulance task: Without statistics and With statistics	230
49. Overall reasoning levels across groups. Cross Task Numeric Codes	234
50. Background information of interviewees	238
51. Interviewees' decisions and response levels for task 1: the Yellow/Brown task ..	256
52. Interviewees' decisions and response levels for task 2: the Movie Wait-Time task	264
53. Interviewees' decisions, estimates and response levels for tasks 3 and 4: the Pink/Black task (without statistics) and the Pink/Black task with statistics	294

54. Interviewees' decisions and response levels for tasks 5 and 6: the Ambulance task (without statistics), and the Ambulance task with statistics	322
55. Interviewees' cross task numeric framework levels	324
56. The distribution of responses from task 1 (the Yellow/Brown task), coded across framework levels, for group 1-GS	371
57. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 1, for group 1-GS.....	371
58. The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for group 1-SE	373
59. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 1, for group 1-SE.....	373
60. The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for group 2-GS.....	375
61. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 1, for group 2-GS.....	376
62. The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for group 2-SE	378
63. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 1, for group 2-SE	379
64. The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for group GRAD	380
65. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 1, for group GRAD.....	380
66. The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group 1-GS	381
67. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 2, for group 1-GS.....	382
68. The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group 1-SE	384

69. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 2, for group 1-SE	385
70. The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group 2-GS	386
71. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 2, for group 2-GS.....	387
72. The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group 2-SE	389
73. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 2, for group 2-SE	389
74. The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group GRAD	390
75. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 2, for group GRAD.....	391
76. The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group 1-GS.....	393
77. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 3, for group 1-GS.....	395
78. The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group 1-SE	398
79. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 3, for group 1-SE	398
80. The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group 2-GS.....	401
81. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 3, for group 2-GS.....	402
82. The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group 2-SE	404
83. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 3, for group 2-SE	404

84. The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group GRAD	406
85. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 3, for group GRAD	407
86. The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group 1-GS	409
87. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 5, for group 1-GS	410
88. The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group 1-SE	412
89. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 5, for group 1-SE	414
90. The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group 2-GS	416
91. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 5, for group 2-GS	418
92. The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group 2-SE	420
93. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 5, for group 2-SE	422
94. The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group GRAD	424
95. The distribution of responses coded at level 2 (transitional) and level 3 (initial distributional) from survey task 5, for group GRAD	424

LIST OF FIGURES:

Figure	Page
1. Two of Gal, Rothschild, and Wagner's data set comparison tasks.....	24
2. Stem and Leaf Plot of Heights of Students and Basketball Players.....	26
3. Graphs used in two of the four tasks from Watson and Mortiz's interview protocol.....	27
4. Item 1: Comparing two samples of measurement data	36
5. Item 2: Comparing two samples of measurement data	36
6. Hypothetical data generated by students	38
7. Movie Wait-time Task.....	46
8. Meletiou and Lee's example histograms.....	49
9. Lann and Falk's data set comparison task	50
10. Points Per Game Scored by Two Basketball Players	60
11. Heights of Trees Grown in Light and Dark Soil After 3 Months	61
12. Cobb's first computer mini-tool.....	63
13. Cobb's second computer mini-tool	65
14. Initial Framework developed by Shaughnessy, Ciancetta and Canada.....	78
15. Lattice Structure Framework	80
16. Bakker's and Gravemeijer's framework.....	83
17. Expanded Lattice Structure, before refinement	84
18. Data Set Comparison Survey: Task 1, the Yellow/Brown task.....	102
19. Data Set Comparison Survey: Task 2, the Movie Wait-Time task.....	103
20. Data Set Comparison Survey: Task 3, the Pink/Black task.....	104
21. Data Set Comparison Survey: Task 4, the Pink/Black task with statistics.....	105

22. Data Set Comparison Survey: Task 5, the Ambulance task	108
23. Data Set Comparison Survey: Task 6, the Ambulance task with statistics	109
24. Lattice Structure Framework	111
25. Expanded Lattice Structure, before refinement	112
26. The Expanded Lattice Structure.....	122
27. Expanded Lattice Structure, before and after refinement.....	124
28. Flow chart #1 for assigning Cross Task Numeric Codes	149
29. Flow chart #2 for assigning Cross Task Numeric Codes	150
30. Flow chart #3 for assigning Cross Task Numeric Codes	151
31. The distribution of response levels across all groups for the Yellow/Brown task	163
32. The distribution of response levels across all groups for the Movie Wait-Time task	170
33. The distribution of response levels across all groups for the Pink/Black task	176
34. Response levels of the 1-GS students to both Pink/Black tasks: Without statistics and With statistics.....	185
35. Response levels of the 1-SE students to both Pink/Black tasks: Without statistics and With statistics.....	188
36. Response levels of the 2-GS students to both Pink/Black tasks: Without statistics and With statistics.....	191
37. Response levels of the 2-SE students to both Pink/Black tasks: Without statistics and With statistics.....	194
38. Response levels of the GRAD students to both Pink/Black tasks: Without statistics and With statistics.....	198
39. The distribution of response levels to the Ambulance task, separated by group.....	210
40. Response levels of the 1-GS students to both Ambulance tasks: Without statistics and With statistics.....	217

41. Response levels of the 1-SE students to both Ambulance tasks: Without statistics and With statistics.....	220
42. Response levels of the 2-GS students to both Ambulance tasks: Without statistics and With statistics.....	223
43. Response levels of the 2-SE students to both Ambulance tasks: Without statistics and With statistics.....	226
44. Response levels of the GRAD students to both Ambulance tasks: Without statistics and With statistics.....	229
45. The distribution of Cross Task Numeric codes across groups.....	234
46. Eduardo's sketch of standard deviation.....	245
47. Ann's sketch of standard deviation	247
48. Amber's comparison of shapes	274
49. Amber's estimation of "mid-range" times.....	302
50. Partial evolution of the Lattice Structure Framework used to describe students' statistical reasoning	329
51. The distribution of response levels of the 1-GS students to the Ambulance task, separated by recommendation.....	409
52. The distribution of response levels of the 1-SE students to the Ambulance task, separated by recommendation.....	413
53. The distribution of response levels of the 2-GS students to the Ambulance task, separated by recommendation.....	417
54. The distribution of response levels of the 2-SE students to the Ambulance task, separated by recommendation.....	420
55. The distribution of response levels of the GRAD students to the Ambulance task, separated by recommendation.....	423

Chapter 1

Introduction

The role of probability and statistics is gaining greater importance in today's society. This can be attributed to the sheer volume of numerical and graphical information people are bombarded with in daily life. It is now commonplace for people to make decisions based on information provided by political polls, educational achievement, economic forecasts, interest rates, drug effectiveness, sale prices, crime rates and taxes. "The numbers that surround these issues arise from processes that are understandable, at least in a general sense, by someone with a little knowledge of statistics. These same processes are used in business and industry to measure productivity, improve quality, and manage systems, so that a knowledge of statistics is essential for both good citizenship and productive employment" (Scheaffer, 2000, p. 158).

Citizens need to be aware of and understand valid ways that predictions, comparisons and, ultimately, decisions are made from data, whether it is in numerical form or displayed graphically (Fendel & Doyle, 1999; NCTM, 2000). Probabilistic and statistical knowledge gives people the ability to "think statistically" (Pfannkuch, 1997). By acquiring the ability to think statistically, people will be well-equipped to make and evaluate the complex decisions, interpretations, and inferences, all based on data, that are so necessary in today's world. A working knowledge of probability and statistics contributes to making informed decisions when evaluating claims made by

others, such as politicians and drug companies (Gal, 2004). Other situations in which thinking statistically can be beneficial are when deciding whether or not to buy lottery tickets or when deciding to purchase insurance or when attempting to comprehend medical advice. Although there is general consensus that the ability to analyze situations from a statistical perspective is not one that is acquired naturally (Fischbein & Schnarch, 1997; Kahneman & Tversky, 1972; Konold, 1989; Tversky & Kahneman, 1974, 1983) it can potentially be gained through statistics education.

Statistics Education

Probability and statistics is gaining wide recognition in the general education of the population and is thus being incorporated into the mainstream of mathematics curriculums within the United States as well as other countries (Batanero, Godino, Vallecillos, Green, & Holmes, 1994; Moore, 2004). In the U.S., this is evidenced principally by the National Council Teachers of Mathematics which has included standards for Probability, Statistics and Data Analysis for the K-12 curriculum in their 1989 and 2000 documents (NCTM, 1989, 2000). This trend has been noted by David S. Moore, former program director for statistics and probability at the National Science Foundation and former president of both the American Statistical Association and the International Association for Statistical Education, who claimed that “in the United States, working with data is now an accepted strand in school mathematics curricula” (Moore, 2004). Further evidence comes from high school students’ enrollment in Advanced Placement (AP) Statistics courses and their participation in AP statistics exams. The College Entrance Examination Board’s (CEEB) Advanced

Placement Program® web site describes the AP Program as “a cooperative educational endeavor between secondary schools and colleges and universities” (CEEB, 2005). Students who participate in the AP Program take college-level courses in a high school setting. These students potentially gain college-level skills and in many cases earn college credit. The first AP Statistics exam was offered in 1997 and was completed by 7,667 students. Since then the number of students taking the AP Statistics exam has increased faster than any other subject exam in the AP Program’s history. The AP Statistics exam has seen increases of between 6,100 and 9,800 every year except 2005 and 2006. In those 2 years, the number of high school students who took the AP Statistics exam increased by approximately 11,000 students each year and thus in 2006 more than 88,000 students had taken the exam. In 1997 the AP Statistics exam had the 18th highest total enrollment (out of the 32 exams offered) and by 2006 that ranking had increased to 10th out of 35 (CEEB, 2007).

Students who participate in an AP Program are generally considered college bound. Thus the increases in participation in the AP Statistics program could be an indicator that enrollment in undergraduate statistics courses is also on the rise. Additionally the U.S. department of education’s National Center for Education Statistics reports that from 1972 to 1992 postsecondary courses in Statistics were among the top 30 courses completed by bachelor’s degree recipients, rising from the 25th in 1972 to the 18th in 1992 (Wirt et al., 2004). Data from the 2000 CBMS survey (Lutzer, Maxwell, & Rodi, 2002) indicates that enrollment in undergraduate Statistics courses taught in Mathematics Departments and Statistics departments of four-year

colleges and universities and in Mathematics programs of two-year colleges has increased from 175,000 in the fall of 1980, to 319,000 in the fall of 2000. That is an increase in enrollment of approximately 82.3%. These increases in enrollment could be indicators that knowledge of probability and statistics is becoming more valued for both high school graduates and college students.

Past research has continued to verify what Kahneman and Tversky noticed in the early 1970s, that is, although during the course of normal life many people may be exposed to data along with numerous examples of variability in data, very few people discover the fundamental statistical rules governing it (Fischbein & Schnarch, 1997; Kahneman & Tversky, 1972; Konold, 1989; Tversky & Kahneman, 1974, 1983). In addressing how statistics instruction can aid in remedying this problem, Scheaffer notes the following:

Conventionally, statistics has been taught as a series of techniques rather than a process of thinking about the world. Teachers and students tend to emphasize particulars rather than principles, narrow mechanics rather than broad methodologies, and specific formulas rather than general formulations. Techniques are useful, and perhaps that is where instruction in a discipline must begin, but now the instruction in and practice of statistics must move beyond the magical use of textbook or technological procedures to clear understanding of analyses and communication of results-beyond rote to reflection. At the introductory college level and, indeed, at the grades K-12 level, the guidelines set out by the ASA-MAA Focus Group (G. Cobb, 1992) in the early 1990s provide a means to effect change in statistics education of the twenty-first century. These guidelines are built around the three-point foundation shown below.

- *Emphasize statistical thinking*
- *Use more data and concepts, less theory and fewer recipes*
- *Foster active learning (Scheaffer, 2000, pp. 158-159)*

Thus, Scheaffer appears to be making a claim about the learning of statistics that is analogous to a claim that Schoenfeld (1988, p. 86) made about learning mathematics: “Mastering formal procedures of mathematics is a far cry from learning mathematics.”

Garfield and Gal (1999, pp. 210 - 211) describe statistical reasoning as a broad goal in statistics education with several specific types of reasoning that they advocate students need to develop as they learn statistics. Those types of reasoning are:

- ***Reasoning about data:*** Recognizing or categorizing data as quantitative or qualitative, discrete or continuous, and knowing how the type of data leads to a particular type of graph, or statistical measure
- ***Reasoning about representations of data:*** Understanding the way in which a plot is meant to represent a sample, understand how to read and interpret a graph, knowing how to modify a graph to better represent a data set, and being able to see beyond random artifacts in a distribution to recognize general characteristics such as shape, center and spread
- ***Reasoning about statistical measures:*** Understanding what measures of center, spread, and position tell about a data set; knowing which are best to use under different conditions and how they do or do not represent a data set; knowing that using summaries for predictions will be more accurate for large samples than for small samples; knowing that a good summary of data includes a measure of center as well as a measure of spread; and knowing that summaries of center and spread can be useful for comparing data sets
- ***Reasoning about uncertainty:*** Understanding and using ideas of randomness, chance, and likelihood to make judgments about uncertain events; knowing that not all outcomes are equally likely; knowing how to determine the likelihood of different events using an appropriate method
- ***Reasoning about samples:*** Knowing how samples are related to a population and what may be inferred from a sample; knowing that a larger, well chosen sample will more accurately represent a population and that there are ways of choosing a sample that can make

it unrepresentative of the population; and being cautious when making inferences made on small or biased samples

- ***Reasoning about association:*** *Knowing how to judge and interpret a relationship between two variables, knowing how to examine and interpret a two-way table or scatter plot when considering a bivariate relationship, and knowing that a strong correlation between two variables does not mean that one causes the other*

So, the education goals of Garfield and Gal show considerable alignment with Scheaffer's educational goals of emphasizing statistical thinking, using more data and concepts, less theory and fewer recipes, and fostering active learning with the goal for statistical reasoning by Garfield and Gal as stated above.

Enculturation

Schoenfeld (1992) cited an expanding base of literature to claim that mathematics learning can be conceived of as “an inherently social (as well as cognitive) activity, and an essentially constructive activity instead of an absorptive one.” Central to this perspective is the notion of *enculturation*. Enculturation refers to a process where upon entering a community or culture, one can acquire the values and “point of view” of that community. Resnick (1988, p. 58) articulates enculturation as part of conceptualization of thinking and learning that proposes that “becoming a good problem solver – becoming a good thinker in any domain – may be as much a matter of acquiring the habits and dispositions of interpretation and sense-making as of acquiring any particular set of skills, strategies or knowledge.” Ben-Zvi (2004) claims that enculturation is particularly important with regard to statistical thinking as the domain of statistics has its own values, belief systems, and habits of questioning,

representing, concluding, and communicating. “Thus for *statistical enculturation* to occur, specific thinking tools are to be developed along side collaborative and communicative processes taking place in the classroom” (Ben-Zvi, 2004, p. 43).

This study is, in part, intended to contribute to the understanding of these processes, specifically reasoning about data, reasoning about representations of data, and reasoning about statistical measures. In the sections that follow the normative perspectives of the statistics community, applicable to this study, will be described.

Statistical Perspectives

Making judgments, decisions, and predictions from data requires an understanding of variation in data (Shaughnessy & Pfannkuch, 2002). Those judgments, decisions and predictions are formed informally and formally through a process of statistical inquiry; that process involves thinking and reasoning in a statistical way. Specifically, Moore posits five core elements of statistical thinking:

1. *The omnipresence of variation in processes.*
2. *The need for data about processes.*
3. *The design of data production with variation in mind.*
4. *The quantification of variation.*
5. *The explanation of variation* (Moore, 1990, p. 135).

Similarly, in their analysis of interviews with practicing applied statisticians, Pfannkuch (1997) and Pfannkuch and Wild (2000) found that those statisticians viewed accounting for variation as a key element in statistical thinking. Based on their findings, Pfannkuch and Wild (2004, p. 19) submitted five types of thinking, fundamental to statistical thinking, that involve: i) the *recognition of the need for data*; ii) *transnumeration*, meaning a transformation of contextual data into or across

numerical or graphical representations that reveal previously hidden features of the data and consequently impact how the data is interpreted (Shaughnessy, 2006); iii) *consideration of variation*, including noticing, acknowledging, measuring, modeling, explaining, and dealing with variation; iv) *reasoning with statistical models*, particularly using aggregate-based reasoning; and v) *integrating the statistical and contextual*.

The Moore list and Pfannkuch and Wild's list have considerable overlap. In particular both lists prominently use the terms *data* and *variation*. When referring to *data* I will use Moore's (1990, p. 96) definition that "data are not merely numbers, but numbers with context." When referring to *variation*, I will follow the distinction that Reading and Shaughnessy (2004) make between the terms *variability* and *variation*. *Variability* will mean the tendency for something to be apt to vary or change, while *variation* will mean the description or measurement of that change. Some of the referenced research may treat these terms interchangeably, but I will endeavor to keep the terms distinct. For example a discussion of students' *reasoning about variation* would specifically "deal with the cognitive processes involved in describing the observed phenomena in situations that exhibit variability, or the propensity for change." (Reading & Shaughnessy, 2004, p. 202).

Statistical thinking, including reasoning about variation, is integral to making decisions and judgments in the context of statistical inquiry. To make valid decisions and judgments based on data is to make statistical inferences. These inferences are *formal inferences* when they are based on a statistical test, such as a t-test. Statistical

thinking is essential when interpreting statistical tests to make decisions about data.

Informal inferences are made based on a person's statistical knowledge and intuition, but not necessarily on the results of a statistical test. Watson and Moritz (1999), among others, advocate that thinking statistically, in particular reasoning about variation, is integral to making good informal inferences which then can better inform which statistical test to apply. Good informal inferences can also aid in the interpretation of statistical tests and the formal inferences made about the data.

Informal Inferences

Making inferences and predictions about data that has been gathered from an uncertain situation is a major topic in university level introductory statistics courses. When asked to make an inference or prediction about data gathered from an uncertain situation, people (both trained and untrained in probability and statistics) use a combination of intuitive and natural assessments and perhaps some formal conceptual knowledge (Konold, 1989). Researchers Lovie and Lovie (1976) make a distinction between making inferences and predictions on an intuitive level, called intuitive statistics, and making inferences and predictions based on formal knowledge of statistics, called inferential statistics. The fundamentals of inferential statistics include “making decisions and drawing conclusions from data, especially where such decisions and conclusions are uncertain because of the variability of chance” (Sanders, 1981, p. 195). The fundamentals of intuitive statistics include understanding and conjecturing about the data from a statistical perspective before formal, inferential statistical methods are employed. Watson and Moritz (1999) argue for the importance

of students learning and using intuitive statistical methods prior to applying formal inferential statistics to make data-based decisions. They claim that intuitive methods such as making visual comparisons of data displays or estimating the center of a collection of data provide a valuable back drop for “confirming inferences made using theoretical tests, hopefully avoiding the tendency to apply a formula without first getting a feel for the data sets involved” (Watson & Moritz, 1999, pp. 166-167). Whether making decisions at the level of intuitive statistics or at the formal level of inferential statistics or somewhere in between, utilization of Statistical Thinking as outlined by Moore (1990) and Pfannkuch and Wild (2004) is important in making valid, data-based decisions. A key to making inferences or decisions from this statistical perspective is understanding and handling the variability in the data from a global perspective.

Statistical inference and sampling distributions are generally considered as two of the main topics in university level statistics courses. Students’ difficulties with these topics are becoming well documented (Batanero, Tauber, & Sánchez, 2004; Chance, delMas, & Garfield, 2004; Meletiou & Lee, 2002). These difficulties may stem from a lack of global perspective of data sets, that is, considering sets of data as whole entities, with their own characteristic trends and patterns. Makar and Confrey (2004) and Watson and Moritz (1999), among others, advocate that students need to build strong intuitive foundations of prior statistical concepts, in particular the concept of distribution, in order to avoid a “recipe-like” application of equations to solve

inference problems and to discourage “black-and-white” deterministic reasoning as opposed to encouraging probabilistic reasoning.

Local vs. Global Views of Data

“Until a data set can be thought of as a unit, not simply as a series of values, it cannot be described and summarized as something that is more than the sum of its parts.” That quote by Mokros and Russell (1995, p. 35) exemplifies one of the first steps in moving to a statistical way of viewing data. When considered as a whole entity, the data can be described with trends and patterns such as shape, center and spread. Those trends and patterns are also used to make statistical comparisons between groups of data (see Bakker & Gravemeijer, 2004; Ben-Zvi, 2002; Konold & Higgins, 2003). Using trends and patterns of whole groups of data to communicate descriptions, to make comparisons and to ultimately make decisions is one of the most basic and powerful elements of handling data in a statistical way.

There is a growing body of evidence that in student’s initial experiences with data they “tend to focus on describing individual data points, or clusters of similar individuals.” (Konold & Higgins, 2003, p. 202). For example, consider a collection of measurement data of heights of a group of people. If the height measurement of “five feet” is considered as only a personal characteristic of the person or people who are that tall, then that is a local view of the data. When a person focuses on individual data points to answer a question or views that data set as an amalgam of individual points, each with their own characteristics, I will call that a *local view (or individual view)* of the data (Bakker & Gravemeijer, 2004; Ben-Zvi, 2002; Konold & Higgins, 2003).

An important step in considering data from a statistical perspective is to make a shift from thinking about the data in local ways to considering the data as a whole entity. The foundation to being able to view data globally is considering a data set as a whole entity. Beyond that foundation, a hallmark of a *global view (or aggregate view)* is the recognition that the data are distributed in the space of all possible outcomes (Bakker & Gravemeijer, 2003). Consider the previous example concerning a collection of measurement data of heights of a group of people. A global view of ‘five feet’ is exemplified when it is thought of as a measurement, associated with one or several individuals, set on a scale of all possible height measurements, say, from less than two feet (a baby) to more than seven feet (a professional basketball player). Another hallmark of the global view is a recognition that the group itself has its own properties, patterns, and relationships that are not evident in any one individual (Ben-Zvi, 2002; Konold & Higgins, 2003; Pfannkuch & Wild, 2004). These properties, patterns and relationships such as shape, center, and spread, are expressed via the concept of a distribution. Although this *global view* of data may have many levels of finer gradations, they will not be addressed here.

Data set as a Distribution

I have previously used Moore’s (1990) definition of “numbers with context” when using the term *data*. Data are associated with a characteristic of an object or with the results of repeated measurements of a process. For example, data can be collected on the heights of people, the weights of objects or the amount of time it takes events to occur. Such characteristics are called *variables*. In most collections of data, the value

of a variable will differ from case to case. When data are displayed, either graphically or in tabular form, the displays often show how the variable varies. When used in this research, the term *frequency distribution* or *data distribution* refers to a tabular or graphical display of the frequencies of the variable's values. A frequency distribution displays a pattern or patterns of variability of the values of the variable (Albert & Rossman, 2001). The term *distribution*, on its own, refers to a data set as a global entity or unit that can be illustrated tabularly or graphically as a frequency distribution on a scale of possible values. Statisticians also use the term *probability distribution* when specifying all possible values of the variable along with the probability of occurrence associated with each value (McClave & Sincich, 2003). Two common probability distributions used to model data are the normal distribution and the binomial distribution (Bakker & Gravemeijer, 2004).

As a global entity, a distribution has its own characteristics apart from the values that a variable takes on. Three of the most commonly referred to characteristics are center, spread (or variation) and shape. For example, mean, median and mode calculations describe the characteristic of center. Range and standard deviation calculations can be used to describe the characteristic of variation, while a skewness calculation is often used to describe shape. Other, more informal, ways to describe a distribution utilize words such as “bumpy,” “spread out,” and “bunched up.”

According to many researchers distribution is an “organizing conceptual structure” that can be used to perceive data in a global way as opposed to locally (Bakker & Gravemeijer, 2004; P. A. Cobb, 1999; Konold, Pollatsek, Well, & Gagnon,

1997; Petrosino, Lehrer, & Schable, 2003; Utts, 1999). The importance of being able to perceive a collection or set of data as a distribution and to possess the skill to switch between a local perspective and a global, distributional perspective comes to light when formulating and evaluating statistical arguments. Garfield and Ben-Zvi (2004, p. 399) define *Distribution* as “a representation of quantitative data that can be examined and described in terms of shape, center, and spread, as well as unique features such as gaps, clusters, outliers, and so on” and they identify it as one of the core ideas of statistics that the educational research community is giving increased attention to in terms of research, instruction, and assessment. As an organizing conceptual structure the idea of a distribution intrinsically promotes aggregate-based reasoning about data. The concept of distribution provides a tool to notice, acknowledge, measure, model, explain, and deal with variation. Considering data from a distributional perspective is also a tool for integrating the statistical and the contextual. Integrating the statistical and contextual, as well as dealing with variation and reasoning in a global way, are all major components of statistical thinking. Thus the concept of distribution comes to light as an organizing conceptual structure because it naturally incorporates those previously mentioned major components of statistical thinking. When the concept of distribution is understood it can aid in providing a foundation to base examinations, analysis, decision-making and inferences about situations involving data. Conclusions based on the organizing conceptual structure of distribution are considered to be statistically enculturated.

The importance of distributions in understanding statistics has been well articulated by researchers such as Bakker and Gravemeijer (2004), Ben-Zvi and Garfield (2004), and Konold and Higgins (2003). Tasks and investigations involving comparisons of two distributions have been used by researchers such as Bakker and Gravemeijer (2003), Ben-Zvi (2004), Konold and colleagues (see Konold & Pollatsek, 2002; Konold, Pollatsek, Well, & Gagnon, 1997; Konold et al., 2002), Makar and Confrey (2002; , 2004), Watson and Moritz (1999), and Gal Rothschild, and Wagner (1989). Those researchers have provided further insight into students' and teachers' reasoning about distributions, in particular reasoning about variation, as well other aspects of statistical reasoning. Konold and Pollatsek (2002) and Makar and Confrey (2002; , 2004) advocate that the richness of tasks and investigations involving comparisons of distributions make them accessible to beginning learners of statistics as well as advanced learners. Questions that involve comparisons of distributions have great potential for being set in interesting and authentic contexts that promote a focus that is not only on central tendency but also on other distributional aspects such as variation and shape.

Comparing Data Sets

The ability to compare data sets by thinking and reasoning statistically about them is a critical skill from a statistics education standpoint. This ability leads directly to making statistically valid decisions and inferences about situations involving data. Konold and colleagues (e.g., Konold & Higgins, 2003; Konold & Pollatsek, 2002) argue that making group comparisons is at the heart of statistics. Specifically, one of

the most basic questions in statistics is to examine differences in two sets of data in order to ascertain if some factor has produced a difference or differences between the data sets. They also advocate that the ability to address questions that involve comparing distributions, from a statistical perspective, should be seen as a major step in statistics instruction as it is the foundation from which the ability to answer further statistical questions arise. Similarly, other researchers such as Watson and Moritz (1999, p. 146) argue that “if encouraged to explore situations involving two or more data sets, those who eventually reach more sophisticated statistics courses will be familiar with the idea of comparison and elementary ways of carrying it out.” Additionally, they encourage that if students first build an intuitive foundation for comparing data sets that is based on a distributional perspective, they may be able to avoid recipe-like inferential reasoning where formulas are blindly applied without consideration of their meaning or appropriate usage.

There is a growing body of research that utilizes the context of data set comparisons to understand grade K-12 students’ views of data, their conceptions of distribution and how their views and conceptions are related to how they make decisions concerning data sets (For example, see Bakker & Gravemeijer, 2004; Ben-Zvi, 2002, 2004; Ben-Zvi & Arcavi, 2001; Gal, Rothschild, & Wagner, 1989, 1990; McClain, Cobb, & Gravemeijer, 2000; Watson, 2001; Watson & Moritz, 1999). Similarly, there is a growing body of research on university students’ difficulties in understanding and learning about statistical inference and sampling distributions (For example, see Batanero, Tauber, & Sánchez, 2004; Chance, delMas, & Garfield, 2004;

Meletiou & Lee, 2002). These studies highlight the difficulties that students have with the transition from data analysis to inference and the potentially important role that an intuitive understanding of distribution plays in that transition.

Makar and Confrey (2002, 2004) have also shown that pre-service teachers have similar difficulties with the transition from data analysis to inferences, specifically in gaining the ability to view data distributionally. Lee, Meletiou and colleagues (see Lee, 1998, 1999; Lee, Zeleke, & Wachtel, 2002; Meletiou, 2000; Meletiou & Lee, 2002) have conducted several teaching experiments in undergraduate introductory statistics courses. Their class designs are generally built around actively considering, explaining, quantifying and dealing with variation. Several of their investigations and assessment tasks required students to reason about data set comparisons. While some of their students clearly demonstrated improved and sophisticated reasoning from a statistical perspective, their results also indicated that numerous beginning statistics students at the university level had many of the same difficulties with reasoning from a global perspective as students in the K-12 levels. My review of the literature related to understanding and reasoning about distributions revealed no studies on university students who are enrolled in classes beyond the introductory level, specifically at the graduate level, or who are enrolled in statistics courses specifically designed for engineering majors.

Research Questions

The above discussion serves to motivate the broad research question for this study: What are university-level statistics students' informal conceptions of

distribution? More specifically, how do they reason when comparing data sets?

Konold et al. (1997) has cited the need to better understand the kinds and nature of problems that emerge during instruction and persist throughout instruction as students encounter new concepts, methodologies, representational systems, and forms of argument. Yet there is a striking absence of research into graduate-level statistics students' understandings of statistical concepts as well as the statistical conceptions of undergraduate students with diverse educational backgrounds. Graduate level statistics students have had, potentially, many years of statistics instructions and may have considerably different conceptions and use different reasoning strategies than novice statistics students. Similarly, other groups of students with different backgrounds, such as science and engineering students who commonly study statistics for their majors, may also have different conceptions and use different reasoning strategies than the often studied K-12 students or college freshman statistics students. Thus, this research involves both undergraduate and graduate-level statistics students, who have a diversity of declared majors, including engineering majors. These students are in various stages of their education and have diverse educational backgrounds, in particular some have had very little statistics education and some have had considerable statistics education. As reasoning about data set comparisons is such a vital step in statistics education, the specific context of this study is set in the realm of students comparing, making inferences and making decisions about pairs of data sets. In investigating the overarching question that drives this research project, the following specific questions arose and were investigated:

1. What types of reasoning strategies do students use when making comparisons of data sets? Specifically, are the strategies global or local or in transition from local to global?
2. What aspects of distribution (i.e. center, shape, spread) do students attend to when comparing data sets?

Beyond answering the specific questions above, an important goal of this research was to further expand and refine the conceptual framework, originally developed by Shaughnessy, Ciancetta, Best, and Noll (2005), for describing middle and high school students' statistical reasoning. To meet this goal and to address the research questions, this research was designed to be a descriptive study with a major component focused on extending and then refining an interpretive framework. In order to observe a wide spectrum of responses and reasoning on the tasks, this research involved a large number of participants, from a diverse group of college students: Undergraduate, post-baccalaureate, and graduate students. The results of this study also contribute to deepening the baseline information on undergraduate statistics students' types of informal conceptions of distributions and to collect initial information on graduate students' reasoning about distributions.

Chapter two includes a description of some of the previous research that informed this study and contributed to the building of the interpretive framework. The chapter focuses mainly on studies investigating students' learning and understanding of distributions and specific features of distributions as they compare data sets. Chapter two also contains a discussion of the evolving conceptual framework, initially

developed by Shaughnessy, Ciancetta, Best, and Canada (2004). The initial framework was expanded as a result of the literature review and analysis of data from the pilot study for this thesis. The framework was then again further refined as a result of the current study. Chapter three details the methodology used in this investigation. Chapter four contains an in-depth description of the evolving conceptual framework and how it was expanded, refined, and used to interpret the survey and interview data. Chapter four also summarizes the survey and interview results and implications as interpreted through the final version of the framework. Chapter five includes a summary of the results and how they give insight into the overarching question about students' informal conceptions of distribution. Implications for teaching and recommendations for future research are also addressed.

Chapter 2

Literature Review and Framework

Literature Review

The purpose of the literature review is to present some of the existing literature that has provided a framework for my study. Three main themes emerge from my examination of research related to people's strategies, reasoning and conceptions when comparing distributions: (1) Research focused on intuitive strategies when predicting and making informal inferences about data sets; (2) Acknowledgment, understanding and reasoning about variation; (3) Reasoning about distribution. The first theme concerns investigations into students' descriptions of data sets, strategies for comparing data sets and their uses and understanding of some of the features of distributions when describing and comparing data sets, such as the use of average, range or standard deviation. The second theme concerns studies on how people understand, interpret, and estimate measures of variation, such as range, standard deviation or variance, when they examine and compare data sets. The third theme relates to studies that have investigated how students and their teachers perceive data sets and reason about distributions of data.

The studies in these themes have considerable overlap as a person's understanding of data and statistics influences the strategies employed by that person to compare data sets. In particular, understanding data sets as distributions is a key concept linking data and statistics. The features of a distribution, such as center, variation and shape, are all interconnected. Thus, understanding a data set as a

distribution that is comprised of several features can influence how descriptions, comparisons, and predictions are made from that data. The themes of understanding distribution and variation, center, shape, and density as features of distribution underlie all the reviewed investigations.

*Intuitive strategies when predicting and making
informal inferences when comparing data sets*

The research cited in this section focuses on people's reasoning strategies when they describe and make predictions and informal inferences from their comparisons of data sets. All studies cited in this section deal with students in grades K-12. Many of the participants in these studies compare data sets in strikingly different ways. Thus indications about how students reason about distributions and understand the various characteristics of distributions can be gleaned from the results of these studies. Similar results across these studies imply that while comparing data sets is a challenging task for all students. Younger students tend to have the most difficulties and commonly make inappropriate comparisons based purely on context, on isolated data points, or on a specific feature, such as mode. As students' ages increase, so does the frequency of use of more sophisticated strategies, such as using proportional reasoning and incorporating and relating several features of the distributions to make comparisons.

Elementary and Middle School Students

Gal, Rothschild, and Wagner (1989) studied third graders' and sixth graders' natural intuitions and naive statistical reasoning strategies as they made comparisons of data sets. The students completed task-based surveys and corresponding task-based interviews. The data sets and the questions used in the study were given to the students in one of two different contexts, that is distances that frogs jump in jumping contests and student scores on a school test. In each task the students were asked to decide if either the groups did equally well or if one group did better than the other. The students were also asked to explain the reasons for their decisions. The characteristics such as size, shape, center and variation of each of the distributions were manipulated, thus each of the pairs of data sets had some similar characteristics and some different characteristics. For example, the most basic comparison was of two data sets with the same range, the same size, and similar shape, yet one was shifted so far to the right that the two data sets had no overlapping values. Another pair had the same range, the same size and the same end points (so all the values overlapped), yet one was skewed left and the other was skewed right so there was a clear difference in the locations of the centers. Two particularly interesting and challenging comparisons are shown in Figure 1.

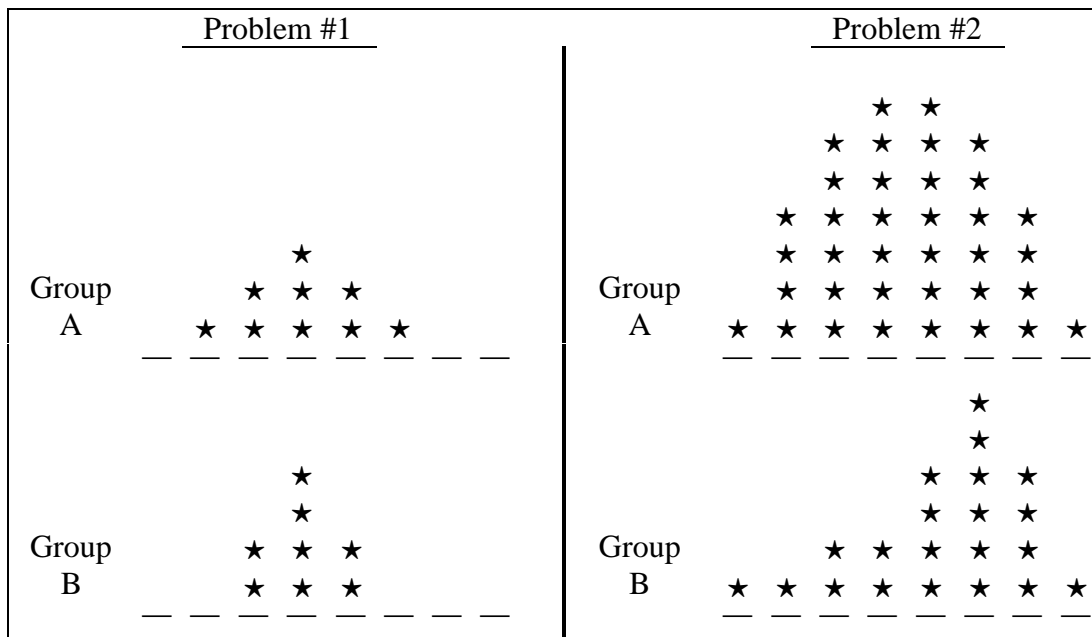


Figure 1. Two of Gal, Rothschild, and Wagner's data set comparison tasks. The contexts of 'distances jumped by frogs' and 'class test scores' were used for both problems 1 and 2. For each problem, students compared group A and B then decided if the groups did equally well or if one group did better.

The first comparison contains data sets of equal size, with equal center locations and shapes but slightly different ranges. The second comparison contains data sets of different sizes with the same ranges, but the smaller data sets is skewed and the larger is bell shaped, thus the shapes and centers are different.

The students' decisions and reasons for if one group did better if they both did equally well were divided into the categories of *statistical*, *protostatistical*, and *other/task-specific* methods. When the students compared summaries of the data such as calculating means, their responses were classified as statistical. Statistical responses went beyond a focus on individual data points and generally involved students integrating several kinds of group features including range, dispersion, shape, and central tendency. Students responding in a protostatistical way either ignored some

features of the data or had trouble synthesizing information. For example, 3rd graders often compared only the modes, ignoring all other aspects of the distributions. When students added or totaled the results, or gave qualitative information such as inferring that the team with the smaller number of frogs is better because they try harder, those responses were classified as other/task-specific methods.

The difficulties those children had in drawing correct conclusions rose as the number of features that needed to be attended to was increased. Also comparisons that required the use of proportional reasoning were difficult for most of these children. For example, problem #2 in Figure 1 was considered by the researchers to be a comparison that required proportional reasoning because of the difference in group size, but $\frac{2}{3}$ of the 3rd graders and $\frac{1}{3}$ of the 6th graders did not give any indication that they attended to the difference in group sizes. A common type of response that these students provided was to choose group A as better, “because they have more students with higher grades.”

Results indicated that a majority of students used more than one solution strategy over the course of answering all the questions. Additionally, those students who stuck with only one solution strategy tended to be less successful in terms of making an appropriate choice and supporting it with an appropriate strategy. The 6th graders were more successful than the 3rd graders, but it was unclear how maturational changes, school and cultural effects interacted to create that phenomenon.

Bright and Friel (1998) reported on their study of ways students in grades six, seven, and eight make sense of graphs and make connections between pairs of graphs.

In the portion of their study where they investigated students' understanding of stem and leaf plots the students were first asked questions about describing two individual sets of data. The sets included heights of students and heights of basketball players, each with different centers and different variation. Then the students were shown one plot (see Figure 2) that included both sets of data and asked, "Just how much taller are the basketball players than the students in this class?" Student responses indicated that they could make sense of the individual data sets. For example, they could separately describe 'typical' heights of basketball players and students. But these middle school students could not seem to make inferences about the typical difference in height between the two groups.

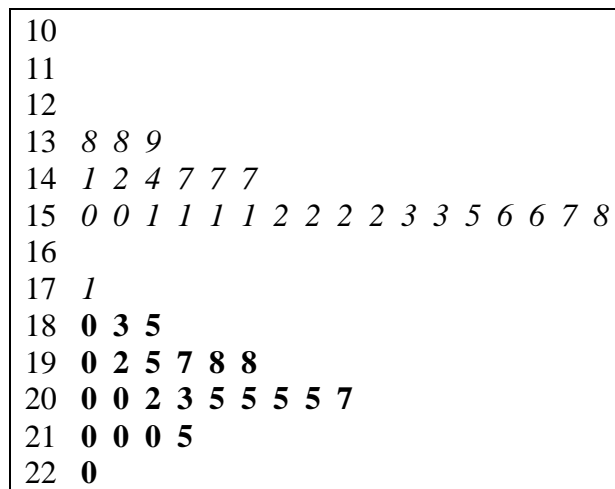


Figure 2. Stem and Leaf Plot of Heights of Students and Basketball Players. Note: Students' heights are in *italics* and basketball players' heights are in **bold**.

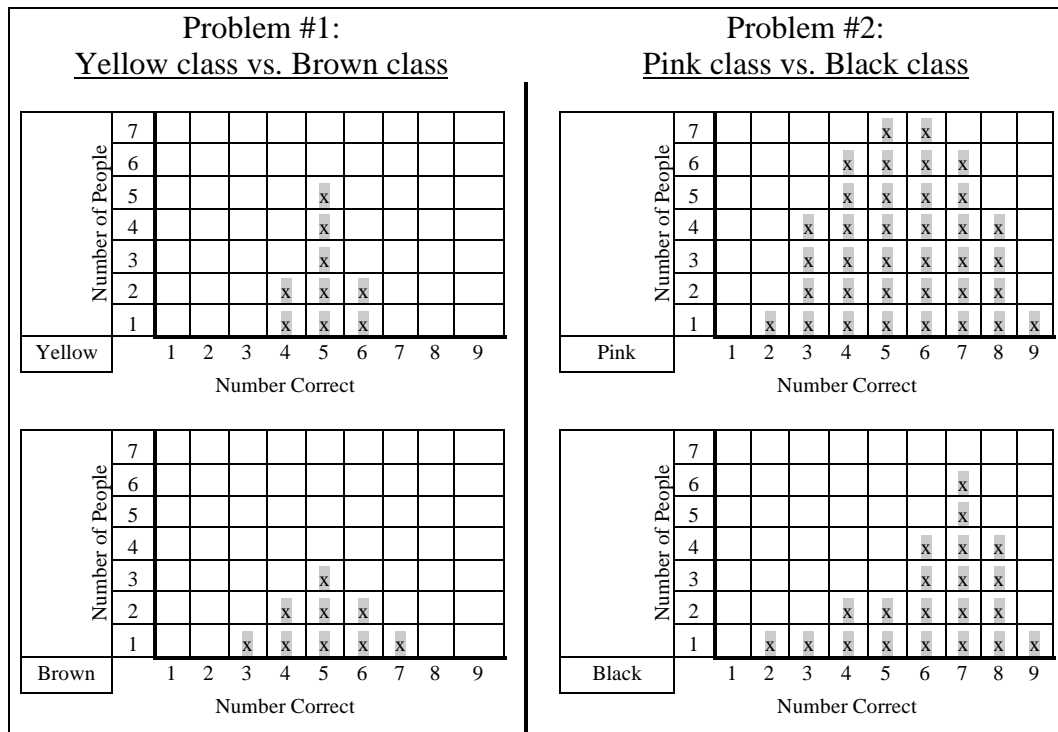


Figure 3. Graphs used in two of the four tasks from Watson and Mortiz's interview protocol. Two schools are comparing some classes to see which is better at quick recall of 9 math facts. Compare the different classes' scores and decide if they scored equally well or if one of the classes scored better. Explain how you decided.

Watson and Moritz (1999) interviewed 88 students from Tasmania and South Australia in grades three through nine, to investigate the structure of student thinking when comparing two data sets in graphical form. Four tasks were adapted from a set of nine similar tasks used by Gal (1989). Figure 3 displays the two tasks that were adopted from the tasks displayed in Figure 1. Watson and Moritz exclusively used the context of test scores for all four of their tasks.

Watson and Moritz's analysis of responses differed from Gal's in that their analysis was not only concerned with students' numeric comparisons of the data, such

as comparisons of centers, but also visual comparisons of the graphical data. Although the authors did not set out to investigate variability issues, variation was inherent in the tasks, and consequently addressed by some of the students' strategies. Students used both visual and numeric strategies for comparing the data sets.

The detailed analysis of students' responses that Watson and Moritz employed was based on the structure of observed learning outcomes (SOLO) model, a neo-Piagetian model of cognitive functioning (Biggs, 1992; Biggs & Collis, 1982). In the original SOLO model respondents' responses were described according to three levels of observed outcomes.

1. *Unistructural responses (U) represent the use of only one relevant aspect of the domain of the task presented.*
2. *Multistructural responses (M) involve the processing of several disjoint relevant aspects, usually in sequence, but not all aspects are integrated.*
3. *Relational responses (R) demonstrate an integrated understanding of the relationships between the different aspects of the domain, so that the whole has a coherent structure and meaning. (Watson & Moritz, 1999, p. 149)*

The levels of understanding displayed by the students' responses in Watson and Moritz's study reflected a similar cognitive model. The levels of Watson and Moritz's model were derived from the cognitive model of Biggs and Collis (1982) where the numerical and visual strategies that the students used formed the substance of the aforementioned levels, and the observed responses encompassed two cycles of these levels. These first and second cycles can be summarized as:

U₁: A single feature of the graph was used in simple group comparisons.

M₁: Multiple step visual comparisons or numerical calculations were performed in sequence on absolute values for simple group comparisons.

R₁: All available information was integrated for a complete response for simple group comparisons; appropriate conclusions were restricted to comparisons with groups of equal size.

U₂: A single visual comparison was used appropriately in comparing groups of unequal sample size.

M₂: Multiple step visual comparisons or numerical calculations were performed in sequence on a proportional basis to compare groups.

R₂: All available information, from both visual comparison and calculation of means, was integrated to support a response in comparing groups of unequal sample size. (Watson & Moritz, 1999, p. 158)

Typical U₁ responses included the word “more” without further explanation, such as indicating that one of the data sets “got more points.” Numerical M₁ responses frequently calculated total scores and then compared those totals. Visual M₁ responses frequently exhibited a consideration of particular individual scores, such as noting the mode for one set of scores versus noting the highest score for the other set and no definitive conclusion about which set was better. Common R₁ responses were based on the shapes of the distributions. The U₂ responses frequently used visual insights about shapes along with a naive or informal form of proportional reasoning, such as when comparing groups of unequal size (see Figure 3, problem # 2), noting that the smaller group, the Black class, did better because “for the amount of people in their class they have got a higher number” (p. 156). The M₂ responses made specific use of

calculation of the mean of each group, and R_2 responses not only used the mean calculations but also noted other aspects, such as the difference in the sizes of the groups or using the means to predict what test scores one may expect from each group. In general, responses categorized as U_1 , U_2 , M_1 , or M_2 displayed one predominant strategy in arriving at a solution, whereas responses categorized as either R_1 or R_2 displayed mixed strategies.

The students' responses showed evidence of higher levels of reasoning, in a U-M-R cycle, with increased grade levels. Grade three students did not engage in proportional reasoning (i.e., multiplicative reasoning as described in the review of Cobb, 1999) when comparing data sets. A majority of students from grades five, six, and seven responded in the first U-M-R cycle with an almost even split between numerical and visual strategies. Half of the ninth graders reasoned in the first cycle and half in the second. Only students responding in the second U-M-R cycle may have made use of strategies that incorporated proportional reasoning, for example comparing means. Watson and Moritz suggest that these results, in part, indicate that students may need considerably more data handling experiences with a variety of data sets to gain an understanding how and when it is appropriate to use the mean.

Watson (2001) followed up the study by Watson and Moritz (1999), described above, by investigating school students' abilities to draw inferences when comparing two data sets presented in a graphical form. Forty-two of the original students from Watson and Moritz's 1999 study were interviewed three to four years later using the

same protocol. Figure 3 displays two of the questions posed to the students. These questions were adapted from those seen in Figure 1.

In addition to re-addressing Watson and Moritz's original research questions with longitudinal data concerning the students' strategies for comparisons and the fit of the data to their developmental framework, they also investigated the following research question: "What evidence is shown that variation displayed in the data sets is explicitly considered in making decisions about which group did better?" (p. 343).

Watson categorized strategies into visual, numerical or mixed (the same as Watson and Moritz, 1999). Responses coded at the SOLO performance levels U_1 , M_1 , U_2 , and M_2 displayed predominantly either visual or numerical strategies, whereas R_1 and R_2 responses displayed a mixture of visual and numerical strategies. Results for comparisons of strategies used across both interviews are displayed in Table 1.

Table 1

Response strategies for two 'comparison of data sets' interviews, by Watson and Moritz

		Response Strategy (first interview)			Total
		Visual	Numerical	Mixed	
Response Strategy (second interview)	Visual	9	1	3	13
	Numerical	3	4	3	10
	Mixed	7	7	5	19
	Total	19	12	11	42

Almost 43% of the students used the same type of strategy for both interviews. Six of the students switched from using mixed strategies to just using either visual or

numerical, while 14 students initially used only one type of strategy then combined the other strategy in their longitudinal interview.

Watson also examined the responses for evidence indicating that variation displayed in the data sets was explicitly considered in making decisions about which group did better. The students' strategies that specifically had indications that they considered the variation in the data sets clustered into six categories. Those categories are: *No acknowledgement of variation*; *Individual features – single columns* [of data]; *Individual features – multiple columns* [of data]; *Global features – 'more'* [assumed to be based on visual comparisons]; *Global features – multiple features*; and *Global features – integrated, compared and contrasted*.

Responses categorized under *individual features – single columns* consisted of comparisons of one or two columns, without taking into account any relationship between them. *Individual features – multiple columns* responses took into account more than two columns but no other features of the graphs. Responses were coded *Global features – 'more'* when they referred to one global feature of the graph, expressed in the term 'more,' without further explanation, such as when responding to problem #2 (Figure 3) a student explained, "Yes, that one [Pink] did [better] because more people got higher scores than people in that class [Black]" (p. 352). *Global features – multiple features* responses combined several features of the graphs, such as, including multiple columns with global features. Finally, the hallmarks of the responses categorized as *Global features – integrated, compared and contrasted* are that they were the most sophisticated in that they integrated, compared and/or

contrasted multiple features of the graphs. An example of this type of response to problem 2 in Figure 3 is:

Okay, by averaging it Black scored better. They got 6.2 and the Pink class got 5.5. So even though they [Pink] had more people, they have more people who scored lower, like it kind of goes in an archish kind of shape [points to Pink], like they had more people score around the middle kind of range. Whereas bearing on the numbers in the class, they had more people score around the middle kind of area [Pink]. The Black class had more people score around the top kind of area. So, averaging it Black still scored better (Watson, 2001, p. 363).

Even though students from all grades used strategies classified in the no acknowledgement of variation category, there was a tendency for students from the higher grades to use more complex strategies. There was also a trend for more students to acknowledge more variation in the longitudinal interview and to use more mixed strategies in the longitudinal interview.

Watson (2002) built on her previous research, described above (see Watson, 2001; Watson & Moritz, 1999), by introducing a new methodology to study students' development of inferential reasoning. Interviews based on both problems seen in Figure 2 were conducted with sixty students from grades three, six and nine. The investigation, in part, focused on whether or not improved responses can be induced with some educational intervention. The intervention used was a presentation of cognitive conflict, presented in the form of video clips from earlier students whose ideas conflicted with the interviewee's, in an attempt to change conceptions. For the Yellow vs. Brown task (see Problem 1, Figure 1), students who decided that either the Yellow class or the Brown class scored better were shown prompts from students who

chose 'Equal,' and those students who chose 'Equal,' were shown prompts from students who chose either 'Yellow' or 'Brown.' For the Pink vs. Black task (see Problem 2, Figure 1) only students who decided that the Pink class scored better or that the classes scored equally well were shown prompts from earlier students who chose 'Black.'

The distribution of the initial levels of response (prior to cognitive conflict) was in similar proportions to the findings of Watson and Moritz (1999) at the five SOLO levels of U_1 , M_1 , R_1 , U_2 , M_2 where the R_2 level was combined with the M_2 level. Results concerning the presentation of cognitive conflict for Problem 1: Yellow vs. Brown as seen in Figure 2 indicated that 57 of the students who could improve their responses actually did. Results concerning the presentation of cognitive conflict for Problem 2: Pink vs. Black only 30 percent of students who chose either 'Pink' or 'Equal' switched to 'Black.'

Finally, after an analysis of the students' responses with respect to the variation displayed in the data sets, Watson found that the responses fit within the developmental framework related to individual features of variation and global features of variation as described by Watson (2001). Further, analysis indicated that a student's acknowledgment of variation, particularly in the Pink vs. Black task (problem 2, Figure 2) is not a good predictor of ultimate success on the task. Those students who demonstrated success on the Pink vs. Black task were more likely to utilize the shape of the graphs in decision making processes. An important implication that Watson made as a result of this research is that the results indicate that students

should be encouraged to build “understanding of basic features of graphical representations, such as basic shape, columns, clumps and humps, rather than expecting that means will be understood and employed” (Watson, 2002, p. 251).

Secondary School Students

Konold, Pollatsek, Well, and Gagnon (1997) interviewed two pairs of high school students who had recently completed a year-long course in probability and statistics. The interviews mainly consisted of questions that required the students perform basic statistical analyses using the computer program *DataScope*.

DataScope's capabilities include finding descriptive statistics, outputting frequency tables, bar graphs, box plots, and scatter plots of whole data sets or groups within data sets. Results indicated that when the students were asked to compare two groups using representations of their own in *Data Scope*, they mostly chose to examine the data in frequency tables and rarely used statistically appropriate methods such as using means, percents or medians in their comparisons. The researchers conjectured that these students were making comparisons based on the sizes of the sets or on attributes of individual data points, while neglecting the variability in the data and also neglecting to consider group propensities of the data. Failure to use group propensities implies that students were not making comparisons based on intensities or rates of occurrences within each data set.

Estepa, Batanero, and Sánchez (1999) studied secondary school students' strategies and judgments of association when comparing two samples of measurement data. The students were given a questionnaire with the two items shown in Figures 4

and 5. In measurement data such as this, there is not only special cause variation caused by measuring devices, techniques, and procedures, but there also is common cause variation due to the natural fluctuations in people's blood pressure and sugar, not to mention possible variation caused by the treatments.

	Mrs. A	Mrs. B	Mrs. C	Mrs. D	Mrs. E	Mrs. F	Mrs. G	Mrs. H	Mrs. I	Mrs. J
Blood Pressure before treatment	115	112	107	119	115	138	126	105	104	115
Blood Pressure after treatment	128	115	106	128	122	145	132	109	103	117

Figure 4. **Item 1:** Comparing two samples of measurement data. The following data were obtained when measuring the blood pressure of a group of 10 women, before and after applying medical treatment: Using the information contained in this table, do you think that the blood pressure in this sample depends on the time of measure (before or after the treatment)? Explain your answer.

Pupil	A	B	C	D	E	F	G	I	J	K	L	M	N	O	P	Q	R	S	T	U
Gender	M	M	M	M	M	M	M	M	M	M	F	F	F	F	F	F	F	F	F	F
Sugar level	9	0	9	8	6	7	4	9	8	9	6	0	7	0	8	3	6	7	7	3

Figure 5. **Item 2:** Comparing two samples of measurement data. The following data were obtained when measuring the sugar level in the blood of male and female school children (M = Male, F = female): Using this information, do you think that the sugar level in the blood in this sample depends on the sex? Explain your answer.

As part of their analysis, the researchers classified the students' intuitive strategies as correct strategies or partially correct strategies or incorrect strategies. Correct intuitive strategies involved *comparing means*, *comparing totals*, *comparing percentages* and *comparing distributions*. Partially correct intuitive strategies involved *comparing the values of the response variable for each case in related samples*, taking

into account exceptional cases, finding out differences (between the data values for each person in the data set), and *global comparison* (qualitative comparisons without mention of any related statistics). Finally, the incorrect intuitive strategies included *expecting similar values* (and basing judgments of association on the fact that those similar values did not occur), *comparing highest and lowest values* (of both distributions), *comparing ranges, assessing coincidences, basing on previous theories,* and *others* (such as using incorrect calculations and mis-interpretations of the context).

Approximately 84% of the students made a correct associative judgment for Item 1 and approximately 74% made a correct associative judgment for Item 2. Although this is not to say that all the correct judgments were accompanied by correct procedures. In responding to Item 1 about 21% of the students used correct strategies, while about 66.7% used partially correct strategies, and about 13% used incorrect strategies. In responding to Item 2 approximately 45% used correct strategies, approximately 31% used partially correct strategies, and 23% used incorrect strategies. Analysis of responses also indicated that 11.7% of the students expected the variation in the data to always have the same sign, a *determinist conception* of association, while a little more than 30% of the students demonstrated a *local conception* of association by assessing exceptional cases or comparing highest and lowest values.

Konold et al., (2002) interviewed students in 7th and 9th grade science classes to explore how they reasoned when comparing data sets. The students had been participating in the ‘Road kill’ collaborative science project where they observe the

number and type of animals killed on local roads, then share that information, via the Internet, with partner schools.

The participating students examined the ‘Road Kill’ data, developed questions and hypothesis, and investigated their questions by analyzing subsets of the data. In the interview setting, when students were asked to predict the number of animals they tend to see every day, responses frequently incorporated ranges with qualifiers such as ‘around’ and ‘probably,’ thus incorporating ideas of variability along with centers. Next, the students, in teams, were asked to make up data they would reasonably expect to observe over a 15-day period. Figure 6 shows examples of some of the ‘made-up’ data. That data the teams generated was reasonably consistent with the previous ‘real’ data collected.

Team														
D4		x	x	x	x	x	x	x	x	x	x	x	x	x
D3		x	x	x	x	x	x							
D2		x			x	x	x							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	# animals per day													

Figure 6. Hypothetical data generated by students. Each data set (D2, D3, D4) represents the number of animals the teams of students might observe on town roads over a span of about 15 days. Each x represents one day. Highlighted in gray are the ranges the teams gave as summaries (i.e. modal clumps) of their made-up data.

The students were asked to summarize the fabricated data for someone who could not see the graphs (see Figure 6). (The data had been organized into stacked dot plots.) All the teams included the range of values in their descriptions. The teams also incorporated the idea of what Konold et al., (2002) deemed ‘modal clumps’ into their summaries. The modal clumps were generally ranges in the middle of the data sets (midrange) that included the mode and had a higher percentage of data than either of the other two partitions. Although when students were asked to compare groups of data, they did not make use of modal clumps.

Konold and his colleagues suggest that the idea of modal clumps may have value as a starting point for students who are learning to summarize and compare data. In particular, aiding the development and use of the modal clump idea could be an integral part of helping students to view shape, center, and spread as salient features of a distribution.

These seven studies only included work with elementary, middle and high school students with a noticeable absence of reasoning strategies employed by undergraduate and graduate students. When the participants of the reviewed studies described distributions, they frequently referred to centers. Students generally encountered difficulties when asked to compare distributions and then make decisions based on their comparisons. Generally, students had more difficulties making appropriate comparisons as the number of features of the distributions that needed to be attended to increased. Students were particularly challenged when making comparisons of data sets that had equal centers but different variation and/or shape as

well as when making comparisons of unequally sized data sets. In the instances, when students referred to variation, it was generally equated to the range of the data.

Although, some students, such as the seventh and ninth graders, who were interviewed by Konold et al. (2002), may have intuitively used interquartile range in their descriptions of ‘modal clumps.’ In the instances where students compared unequal sized data sets, those who employed proportional reasoning strategies consistently made appropriate comparisons. None of the students from grade 3 used proportional reasoning, but as students’ ages increased so did their use of proportional reasoning strategies.

On the whole, the results from the seven studies reviewed in this sub-section indicate that tasks involving comparisons of data sets are challenging yet engaging for students in grades K-12. Each of the studies utilized task-based interviews with the students, while some also incorporated data from classroom activities. Overall there were indications that the sophistication of the responses increased with age, although it is not clear if this was due to natural maturation or education or both. There were students at each grade level that provided responses at the lowest level. For example, these responses were called ‘task-specific,’ ‘non-appropriate,’ ‘incorrect’ or ‘unistructural’ and often utilized qualitative assessments of the data. A commonality among these classifications was the use of very basic, local, strategies, such as comparing individual data points or differences between extremes, comparing totals (with sets of unequal size), and in general neglecting the variation in the data sets. Responses classified at the highest level were called ‘statistical,’ ‘statistically

appropriate,’ ‘relational’ or ‘correct.’ Commonalities among these responses were uses of more sophisticated, global, type strategies. For example, some students’ responses, particularly older students, had indications that they reasoned proportionally by using strategies that, say, included percentages or rates. Other students responses were categorized somewhere between the lowest and highest levels. These students frequently attended to and made use of individual group features in isolation of each other.

*Acknowledgment, understanding and reasoning about variation
when comparing data sets*

The research reviewed in this section was performed with a variety of students in K-12 grades as well as with students in university level introductory statistics classes. All of the studies, in whole or in part, attempt to describe students’ understandings of some of the various measures and concepts of variation, such as range, variance and standard deviation. Similar to the studies from the previous section, most of the studies cited in this section place the participants in an environment where they are comparing data sets in contextualized situations. This section was included because of the inseparability of the ideas of variation and distribution, that is, “without variation there is no distribution” (Bakker & Gravemeijer, 2004, p. 149). Thus, what we learn about students’ understanding of variation is directly related to their understanding of distribution.

Middle and High School Students

Ben-Zvi (2004) investigated how junior high school students begin to reason about variability as part of an open-ended group-comparison task given in a rich supportive classroom context. The study followed the behavior and discourse of two novice seventh grade students engaged with an Exploratory Data Analysis task involving the comparison of the surname lengths of a group of Americans versus the surname lengths of an equally sized group of Israelis. The researcher's goal was, "to trace the emergence of beginners' reasoning about variation in a comparing distributions situation, including the development of cognitive structures and the socio-cultural processes of understanding and learning" (Ben-Zvi, 2004, p. 45).

As part of their regular classroom activities, the two students took part in lessons that introduced three methods to compare distributions: (a) absolute and relative frequency distributions presented in tables; (b) basic measures of variation and center, such as range, mode, mean, and median; and (c) graphical representations, such as a double bar chart. The data collection included videotaping the two students during their regular classroom periods, interviewing them after class and analyzing their notebooks. Analysis of this data resulted in a description of the students' novice perspective of the data. From their novice perspective, the students were either unable to describe variability or focused on local information to make informal descriptions of variability. The analysis also indicated that the students' perspective grew towards an expert perspective, that is a global view of the data as a distribution that can be

used for noticing, acknowledging, describing and explaining the variability within and between the groups.

Ben-Zvi identified seven developmental stages that captured these students' reasoning about variability. In the initial developmental stages, that is stage 1: *On what to focus: Beginning from irrelevant and local information*, and stage 2: *How to describe variability informally in raw data*, the students did not notice any global features of the raw data, nor did they notice the variability within those features. At these stages the students either did not focus on variability or only on "local deviations" from one data point to the next. At the next stage, stage 3: *How to formulate a statistical hypothesis that accounts for variability*, the students were able to conjecture about informal rules that described the variability between the groups. A characteristic of stage 3 was the use of the language "usually" or "not always." When the students added observations about the ends of the distributions, for example including comments about the "least" or "most" they were classified at stage 4: *How to account for variability when comparing groups using frequency tables*. Stage 5: *How to use center and spread measures to compare groups*, was characterized by "the students' insignificant and monotonous use of statistical measures" (p. 57) to make group comparisons. In this stage the students constructed a statistical measures table for each group that included the counts, mode, maximum, minimum, range, mean, median, and "outlying values." Then the students did not appear to make comparisons of those statistics in meaningful ways, as Ben-Zvi noted, "their actions seem to be merely procedural, missing both the meaning of measures as representative numbers

(Mokros & Russell, 1995), and the distinction between center and spread measures” (p. 53). When the students were classified at stage 6: *How to model variability informally through handling outlying values*, they began to compare several outliers that led them to a simple view of the distributions. Their comparison began by separating each group into two parts. The students then noted that the majority of the name lengths in the 1st group were concentrated on a lower interval with outliers positioned higher while the name lengths for the other group the majority of the name lengths were concentrated higher with the outliers positioned lower. The students based their comparisons on the opposite patterns, i.e. oppositely skewed distributions. At stage 7: *How to notice and distinguish the variability within and between the distributions in a graph*, the students were guided to construct a double bar chart; then using that bar chart they were to describe an emerging trend. The students struggled with interpreting the graph and also with making unclear and contradictory statements, but they did eventually settle on a statement. Their statement, “*The emerging trend is that frequency of relatively short names (up to 5 letters) is higher in Israel than in the USA, but the frequency of relatively long names is higher in the USA than in Israel*” (p. 56) was a description of the variability between the groups, yet it was based on local methods, i.e., frequency assessments. Overall, Ben-Zvi (2004, p. 57) noted that

The students’ development of reasoning about variability in comparing groups was accompanied by somewhat of a global perception of distribution as an entity that has typical characteristics such as shape, center and spread. This perception seems to be a precondition to being able to describe the two distributions as generally similar in shape and variability, but horizontally shifted.

Shaughnessy, Ciancetta, Best, and Canada (2004) reported on the results of two questions from a task-based interview that focused on variability when comparing distributions. Of the 24 students interviewed, eight were middle school students and 16 were high school students. All were members of classes that, together with their teachers, were participating in the Development of Conceptions of Variability Project (Shaughnessy, 2003). Prior to the interview the classes had completed a task-based survey designed to gauge the students' thinking about variability in three principle contexts: in sampling situations, in probability experiments, and in data sets. After the survey was administered, 24 randomly selected students were interviewed. The interview protocol was based on a subset of the survey tasks. The classes then participated in weeklong teaching episodes on sampling distributions. Twenty-three of the original 24 students who were selected for the first interview also completed the second interview after the conclusion of the teaching episodes. The second interview included the task shown in Figure 7 on the next page.

Responses to the *Movie Wait-Time* task were coded into six categories: *Specific Data Points*; *Variation*; *Centers*; *Distribution*; *Informal Inferences*; and *Context*. It was possible for responses to fall into several categories. Except when students referred to distribution, their reasons did not get re-coded under center and variation.

When students compared or examined specific data points in the distributions, such as comparing the data points at the low ends of the distributions, their responses were coded *Specific Data Points*. The code of *Context* was used when students would

Movie Waiting Time. A recent trend in movie theaters is to show commercials along with previews before the movie begins. The *wait-time* for a movie is the difference between the advertised start time (like in the paper) and the ACTUAL start time for the movie.

A class of 21 students investigates the wait-times at two popular movie theater chains: Maximum Movie Theaters and Royal Movie Theaters. Each student attended two movies, a different movie in each theater, and recorded the wait-times in minutes below.

<p>Maximum Movie Theaters:</p> <p>5.0 12.0 13.0 5.5 9.5 13.0 5.5 11.5 8.0 8.5 14.0 13.0 8.5 7.0 8.5 12.5 13.5 11.5 9.0 10.0 11.0</p> <p>Mean=10 minutes; Median=10 minutes</p>	<p>Royal Movie Theaters:</p> <p>11.5 11.0 9.0 10.5 8.5 11.0 9.0 10.5 9.5 8.5 10.0 11.5 10.5 8.5 9.0 11.0 11.0 9.5 10.0 9.0 11.0</p> <p>Mean = 10 minutes; Median = 10 minutes</p>
--	---

Wait-Time for Movies

Maximum Movie Theaters

				X								X
	X			X				X		X		
X	X		X	X	X	X	X	X	X	X	X	X
5	6	7	8	9	10	11	12	13	14			

Minutes (rounded to the nearest half-minute)

Royal Movie Theaters

				X		X						
				X		X						
			X	X		X	X					
			X	X		X	X	X	X			
			X	X	X	X	X	X				
5	6	7	8	9	10	11	12	13	14			

Minutes (rounded to the nearest half-minute)

a) What can you conclude about the wait-times for the two theaters?

b) One student in the class argues that there is really no difference in wait-times for movies in both theaters, since the averages are the same. Do you agree or disagree? Why?

c) Which of these theater chains would you choose to see a movie in? Why?

46

The *Variation* code was used for responses that included comparisons of relative amounts of variation, such as “their [Royal] wait-times are kind of bunched up together and Maximum movie theaters is more spread out.” The *Variation* code was also used when students calculated or approximated the ranges of the data sets or informally referred to the spread, i.e., “Well, it’s [Royal] within a 8.5 and 11.5 range. Up here [points to Maximum graph] you’ve got some all the way down to 5 and all the way up to 14.” Responses that referred to the means and/or medians were coded *Centers*. When students reasoned using both centers and variation simultaneously, their responses were coded *Distribution*. For example, “you also have to, like, look at how much it’s weighted because there’s [pause] because this one [point to the Maximum graph] is so much higher and that would account for the smaller ones, like, that would bring down the average versus these ones [points to the Royal graph]. There is just more of them, like, these ones [pints to the Maximum graph] were more spread out and it kind of evened out.” Some students speculated about probabilities of experiencing certain wait-times at each theater or used language such as ‘predictable,’ ‘consistent,’ ‘reliable,’ ‘chances’ or ‘luck.’ Those responses were coded under the *Informal Inferences* category.

Results from part (a) of the Movie Wait-Time task indicated that most of the interviewed students attended to both centers and variation. In responding to part (b), about two-thirds of the students said the data sets were different despite having the same mean and median. Most provided reasons that could be categorized in multiple ways with a majority linked to variation. Just over 70% of the students chose to go to

the Royal Theater when responding to part (c). They generally indicated a higher confidence in experiencing a 10-minute wait time, an informal inference about the distributions. Also about one-third of the students included personal contexts and past experiences in their responses.

College Students

Loosen, Lioen, and Lacante (1985) conducted an experiment on 154 psychology freshman, none of whom received any instruction on variability. The students were shown two different sequences of blocks, A and B. The blocks were organized in increasing lengths. Sequence A contained block lengths of 10, 20, 30, 40, 50, and 60 cm while the blocks in sequence B had only two different lengths, that is, three blocks of 10 cm and three of 60 cm. Half of the students said A was more variable, 36% chose B as more variable, and 14% said there were equally variable. These results implied that these students' intuitive concept of variability was concerned mostly with how much the values in each data set differed from each other as opposed to differing from a mean, i.e. how much they were unlike.

Meletiou and Lee (2002) conducted a teaching experiment with an introductory college statistics class that was based on variation and on modeling realistic statistical investigations. Their goal was to "increase students' awareness of variation by helping them realize that the need for statistical investigations is created due to the existence of variation." Special emphasis was placed on interpreting and understanding histograms. One of the tasks on their pre-assessment asked which of the histograms shown in Figure 8 has more variability.

Eight of the 24 students incorrectly responded that A had had more variability because it was ‘bumpier.’ At the end of the course the majority of students continued to have trouble with the concept of sampling distribution, but most students did show improvement in moving away from uni-parameter thinking by incorporating both center and variation into their analyses and predictions.

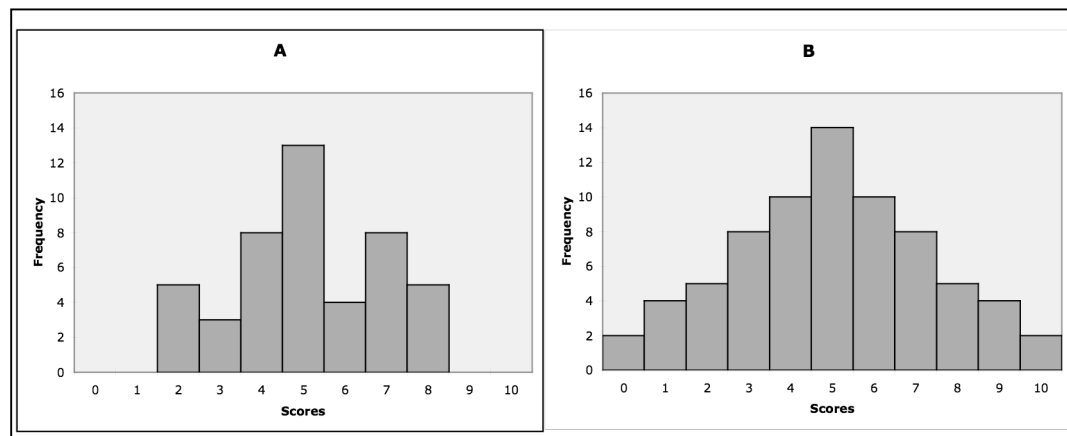


Figure 8. Meletiou and Lee’s example histograms.

Lann and Falk (2003) studied the conceptions and informal magnitude assessments of variability held by first-year university students. The students made intuitive comparisons of pairs of small data sets, each with the same number of data points and equal means and medians. The authors anticipated that comparing data sets with equal centers would push the students to consider variation. Each student compared eight pairs of data sets, according to their intuitions (i.e. no calculations), and decided which set had a higher heterogeneity. The data was presented in two different contexts: salaries and test scores. One ‘test score’ comparison is shown in Figure 9.

How confident are you of your choice?	In which subject is the heterogeneity higher? (Circle your choice)	Students' Scores	Subject
5 4 3 2 1 Very Very unsure sure	X Y	12, 22, 35, 36, 40, 42, 43, 44, 88 19, 20, 33, 34, 40, 41, 44, 47, 82	X Y

Figure 9. Lann and Falk's data set comparison task. Explanations/remarks: What did you consider when answering? How did you choose your answer? What influenced your confidence level? etc.

In their analysis of responses, Lann and Falk reduced the set of possible measures of variation that the students used to four: Variance (**V**), Range (**R**), Mean

Absolute Deviation (**MAD**) calculated as $MAD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$, and Interquartile Range

(**IQR**). Students' responses most frequently coincided with the **R** classification, followed by **V** with **MAD** and **IQR** the least frequent. Lann and Falk found that the students on the whole exhibited complex intuitive ideas of variability with no one clear and dominant consensus.

Researchers delMas and Liu (2003) studied a variety of ways that introductory college statistics students understand standard deviation. In their study, they gave the participating students activities based on a computer program that allowed them to explore factors in discrete data sets that affect the size of the set's standard deviation. These activities were set in the context of comparing data sets and observing which had the larger standard deviation. After completion of the activities the students were presented with a test that required the students to examine ten pairs of histograms that had various features manipulated, such as means, shapes, skewness, ordering of the

heights of the bars, contiguous vs. spacing between bars, and distributions that were mirror images of each other. The students then indicated which histogram in each pair had either a higher or lower standard deviation.

During the activities most of the students responded using simple, rule oriented approaches. They indicated that the histograms that were mirror images of each other or having the same arrangement of bars on different locations on the horizontal axis were generally thought of as having the same standard deviation. Histograms thought to have larger standard deviations had bars equally spread out along the number line, or had bars placed as far away from each other as possible, or were observed to have bars as far away from the mean as possible. Histograms thought to have smaller standard deviations generally had bars placed next to each other with the students indicating that the distributions should be either “bell shaped” and symmetrical or having ascending, descending or no apparent order of heights. Some students also indicated that histograms with smaller standard deviations had the tallest bars in the middle or close to the mean.

The test results indicated that students correlated range with standard deviation, i.e. distributions with larger ranges had larger standard deviations. Those students would decide a U-shaped histogram would have a smaller standard deviation as compared to a bell shaped histogram with a larger range. Another strategy or rule that the students used included correlating the distribution that simply had more values with having a higher standard deviation.

Lee, Zeleke and Wachtel (2002) reported findings on how college students in introductory statistics classes learned the concept of variation. They specifically focused on the students' understanding of variation and how it is related to other important statistical concepts. Two groups of nine students were interviewed. The students were selected from two larger groups, one that was taught using a traditional lecture format and the other from a group that was taught using the PACE model (Projects, hands on Activities, Cooperative learning and Exercises). The PACE class was conducted in a computerized classroom where teams of students worked cooperatively on 'hands-on' statistical projects and activities. The PACE students were actively engaged in solving real world problems with statistics and presenting their solutions both in written form and by oral presentation (Lee, 1998).

Part of the interview entailed having the students answer questions posed in the context of investigating the hypothesis that students entering their university had weaker quantitative backgrounds than students from 20 years ago. The students were asked questions such as "How do you compare two groups?", "What do you do first?", "What is the target population?", "How do you describe the center and/or variation numerically, graphically?".

When responding to "How do we differentiate between data sets that are widely scattered or are cuddled together?", 85% of the students from the traditional class did not remember any numerical quantities used to describe variation, but 63% of the PACE class mentioned standard deviation and 10% mentioned range. When asked, "How can you describe variation or spread graphically?", 25% of the students

from the traditional class mentioned using histograms while 90% of the PACE students mentioned using histograms. The students were also asked to make comparisons of two institutions based on SAT statistics from each institution. The students were given the information that the distribution of SAT scores from the first institution had a mean of 540, a median of 535, and a standard deviation of 40. They were also given the information that the distribution of SAT scores from the second institution had a mean of 540, a median of 535, and a standard deviation of 1. Thus the two distributions of scores had the same centers but different variation. The students made implications about the symmetry of each group and the relative closeness of the mean and median. Representative responses from the ‘traditional’ classes were:

- *I think they are both symmetric, the difference of five units in hundreds of scores is not much.*
- *The one with standard deviation 1 is symmetric. Isn't there something like small standard deviation means less variation?*
- *I think it is very hard to get standard deviation 1. There must be some kind of mistake. We can't compare the two groups.*
- *Both are roughly symmetric because to be perfectly symmetric the mean and median have to be equal.*

Representative responses from the ‘PACE’ students were:

- *With standard deviation 40, the distance between the mean and median is very small.*
- *If the standard deviation is 1, then the mean is 5 standard deviation units away from the median. That is far and it is not symmetric.*
- *In the first case the mean is within one standard deviation away from the median. In the second case it is away about 5 standard deviation*

units. That indicates to me an outlier pulled the mean and hence skewed the distribution.

- *Even though the difference between the mean and the median is 5 units I think the standard deviation makes a difference here. I guess the second seems skewed.*

In general the students that were taught using the PACE model indicated that they relied on more complex thinking and were more articulate. Both groups of students had difficulties connecting different concepts of variation; many students felt that there was an association between skewness and higher frequencies of values on one side of a graph as opposed to an association with outliers.

Pre-Service Teachers

Makar and Confrey (2005) interviewed seventeen prospective secondary mathematics and science teachers as part of a study designed to document the different types of language used by pre-service teachers when they are engaged in the statistical task of comparing distributions. Eight of the pre-service teachers had not previously studied statistics, five had taken a traditional university-based statistics course, and the remaining had not taken a formal statistics course but indicated that they had previous experiences with statistics from other mathematics or science courses that they had taken.

The interview task was set in the context of a middle school trying to assess the effectiveness of a semester-long mathematics ‘Enrichment’ program that provided extra help for eighth-grade students who were preparing for the state exam. The teachers were shown a pair of dot plots of authentic data taken from students in an

enrichment class and a regular eighth grade class. The data consisted of the change (difference) in the students' scores from their eighth grade state exam score in mathematics to their scores on a practice test given near the end of eighth grade.

The authors found that the language used by the teachers naturally fell into two categories: Standard statistical language and non-standard statistical language. The standard statistical language included the categories of *Proportion or number improved, Mean, Maximum/Minimum, Sample size, Outliers and extreme values, Range, Shape* (e.g., skewed, bell-shaped), and *Standard deviation*. The language used by the teachers to articulate statistical concepts that were categorized as non-standard statistical terminology included the categories of *Spread* and *Distribution chunks*. The terms used that were classified as *Spread* were *spread out, scattered, evenly distributed, dispersed*, along with the antonyms *grouped, bunched up, and clustered*. When a respondent partitioned a distribution into a triad of 'improving,' 'not improving,' and 'about the same,' that language was categorized as *Distribution chunk*. Other language categorized in this way occurred when the respondents focused specifically on a middle chunk, similar to a modal clump as described by Konold et al. (2002). Finally some of the respondents indicated that they saw a subset of the distribution, say a group of high values, as more than just individual points. The respondents saw these values as a contiguous subset with dynamic borders. The researchers conjectured that the teachers who used the notion of distribution chunks viewed the data as more than just individual values yet did not have a view of the data as a single entity or aggregate. They concluded that

The three perspectives of seeing partial distributions – triads, modal clumps, and distribution chunks – indicate that there are more than just two perspectives of distribution that are usually discussed in the literature: single points and aggregate. This third perspective – partial distributions or “mini-aggregates” – deserves further research to investigate the strength of its link to statistical thinking about distributions. (Makar & Confrey, 2005, p. 48).

The eight studies reviewed in this sub-section underscore the wide range understandings of variation that students from the elementary grades through college and prospective teachers possess. These studies also highlight the spread of abilities that these students have in estimating and using the various measures of variation when examining and comparing data sets. From these studies it is not clear that understanding and sophistication of uses of variation naturally increases with maturation, yet there is some evidence that focused instruction about variation, such as standard deviation, or as part of a wider focus on distribution does help to increase the level of sophisticated use and understanding about the measures of variation.

These studies also revealed that students have a wide variety of misconceptions about variation. In particular, some students utilize simplistic rules that are not always true, when dealing with issues of variation. For example, some students equated range and the generic term ‘variation.’ Those students also may have used range as an indicator of other measures of variation, such as variance and standard deviation. Some students also relied on erroneous shape assessments to give them insight about the variation of graphically displayed data, such as relating the degree of ‘bumpiness’ of a histogram with the variation of the data or believing that regular patterns in the

data provide indications about the variation of the data. Other students relate the size of a data set with its variation, i.e. sets with more data points have more variation.

Students who understand variation in more sophisticated ways tended to consider, and even integrate, variation along with other features of the distribution when comparing or describing data sets. For example, students who had experiences collecting, summarizing and comparing data sets not only compared means but also included shape and spread when making decisions and inferences about those sets. In the studies where students had hands on experiences working with data and comparing data sets, those students tended to have more success when comparing estimates of measures of variation, such as standard deviation. Also, the results from Makar's and Confrey's (2005) study with prospective teachers indicated that statistical language referring to variation is also interconnected with language referring to other distribution characteristics, particularly shape. Of particular interest is Makar's and Confrey's speculation that there is a perspective of distribution between local and global.

As with the research reviewed in the previous section, the research reviewed in this section is devoid of studies dealing with upper level undergraduate students and graduate students. Investigating how this population understands variation and distribution may indicate which misconceptions cited above persist and if there are any more sophisticated conceptions that develop.

Reasoning About Distributions

The studies reviewed in this section all involve either elementary or middle school students, or middle school teachers. An integral part of each study was the participants' engagement in lessons, each of which incorporated data comparison tasks. The common overarching theme to the lessons was to promote thinking and reasoning about data sets as distributions, i.e. from a statistical perspective. As with the previous studies cited in this review, there were indications that the participants of these studies found that working on data comparison tasks was challenging yet engaging.

Elementary and Middle School Students

Petrosino, Lehrer, and Schable (2003) reported on an eight week teaching and learning experiment on fourth grade students' thinking about distribution. Understanding both centers, and variation about centers are essential components of distributional thinking. The students participating in this experiment worked on class activities that were set in the general context of data modeling. The specific context of measurement was used because centers of distributions can indicate values of attributes, and variation of data about those centers can indicate variations in the measurement process. Additionally, by working with measurement data, students had the opportunity to gain experiences trying to distinguish between random and systematic variation. I have previously described random variation as common cause variation that cannot be explained, sometimes referred to as 'noise' in a system. My

previous description of systematic variation was called special cause variation or causal variation and is variation that results from an identifiable source.

In this experiment, the students performed the various tasks that included observing, measuring, collecting data, as well as describing data and making decisions based on data comparisons. Distribution was introduced as a means of displaying and structuring the data. Surveys and interviews were conducted at the completion of instruction. Several interview questions were designed to assess the students' reasoning when comparing distributions of unequal sizes, reasoning about measurement variability, and comparing two distributions with different shapes, but equal ranges.

One of the lessons was specifically focused to give students the opportunity to learn about distinguishing between random and systematic variation. In this lesson, the students experimented with two design types of model rockets (rounded vs. pointed nose cones). The students were assigned the task to determine whether or not the difference in the distributions of maximum heights attained by each type of rocket was due to random variation or systematic variation (i.e. cone design). Causes of variation identified and discussed by the students included procedures for measuring, precision of the measurement tools, and trial-by-trial variation. Students initially noticed and commented on outliers, but eventually the class agreed to use the median of the distribution of height measurements as an indicator of 'true height.' The class also used a calculation they called the 'spread numbers' (i.e. the averages of differences from centers) as an indication of accuracy.

In an assessment task that Petrosino, Lehrer, and Schable (2003) gave to the students involved comparing distributions of unequal size, the students were given the number of points two basketball players scored over several games then asked to decide which of the two players should be selected to an All-star game (see Figure 10).

Player	Scores
Bob	21, 16, 23, 21, 20, 17, 16, 22
Deon	24, 18, 21, 25, 22, 28

Figure 10. Points Per Game Scored by Two Basketball Players.

Eleven of the 15 students initially selected Bob because he scored more points (an additive way of reasoning). Then after the interviewer probed with a follow-up question about Bob also having played in more games, more than half of those who chose Bob switched and only three students stayed with the “more points” reasoning while the others reasoned using the mean or median. Two students chose Deon because his range of scores was higher and two other students chose Bob and reasoned that because he had more scores (Bob had eight scores recorded for him while Deon only had six), Bob is more experienced so he should be chosen.

Another assessment task that Petrosino, Lehrer, and Schable (2003) administered involved comparing two distributions with different shapes, but equal ranges. It was designed to investigate if the students’ comparisons were influenced by the evident difference in variability. The scenario presented to the students involved tree heights at a certain nursery. The students were presented with tree height data

from trees grown in light soil and trees grown in dark soil (see Figure 11). Then each student had to decide which soil they grew better in or if they grew the same in each.

Of the 14 students who were asked this question, six decided that soil did not matter. Four of the 14 students only calculated and compared the averages, which were equal, while the remaining ten students made use of the similar centers while also being influenced by the differing variation in the data sets. Other strategies the students employed include comparing the extremes (4 students), comparing medians (2 students), comparing the percentage of trees that are average height or taller (2 students), and comparing average spreads (2 students).

<u>Light Soil</u>									
10	20	30	40	50	60	70	80	90	100
19	23		43	51	60	71			109
				52	61	75			
				53	61				
				54	62				
				56	65				
				57	65				
				58					
<u>Dark Soil</u>									
19	21	32	41		61	75	81	91	109
	22	34	41		61		82	92	
	23							92	
	25							93	

Figure 11. Heights of Trees Grown in Light and Dark Soil After 3 Months

Throughout the tasks that these fourth grade students performed during this experiment the students generally interpreted centers of distributions as representative of the true values of the attributes they were investigating, such as maximum height attained by model rockets. The overarching context of measurement served to aid the students in considering center and variation jointly in their descriptions and

comparisons of distributions. Although some students continued to rely on individual descriptors of distributions such as extremes, range, median, mean, etc., the result that some students came to find it difficult to interpret center without spread and vice versa is encouraging given that that type of thinking is critical to the development of reasoning from a distributional perspective.

Cobb (1999) analyzed 29 seventh grade students' mathematical reasoning during a ten-week teaching experiment designed around the overarching idea of distribution. Lessons were specifically designed to promote the emergence of using features of a distribution such as mean, mode, median, skewness, spread-out-ness, and relative frequency as ways of describing how specific data sets are distributed within a space of values. Over the course of 34 lessons the students made decisions or judgments based on their descriptions and comparisons of data sets. The initial lessons required the students to describe single data sets and later the students compared two or more sets of data. Both types of activities required the use of two computer-based mini-tools. The tools allowed the students to organize and partition the data.

The first mini-tool could handle a maximum of 40 data points and allowed the students to sort the data by its various characteristics and partition the data into two groups, green and pink. This computer mini-tool represented each data point as a horizontal bar. The lengths of the bars directly correspond to the numerical values of the data points they represent. Figure 12 illustrates an example of data collected on the life spans of two different brands of batteries. The length of each bar represents the life span of each battery that was tested.

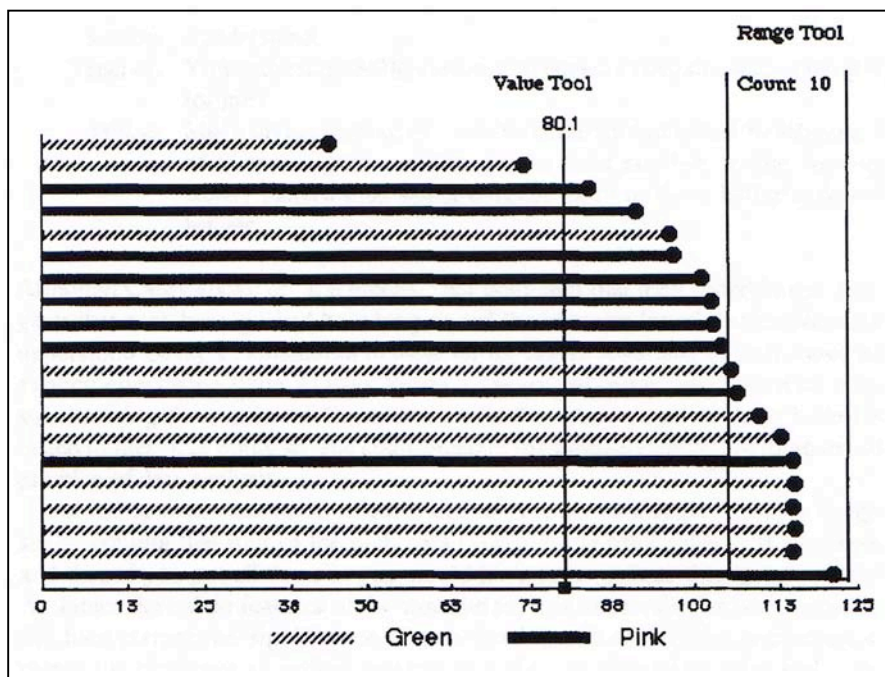


Figure 12. Cobb's first computer mini-tool

After the initial set of lessons that utilized this first mini-tool, the students identified significant characteristics of data sets such as range, maximum, minimum, number of points above and below a certain value, defined by the researchers as 'cut points,' and the median along with its relation to the mean. These can be indications of additive reasoning. A strong indicator of additive reasoning is when a student partitions the data set into two or more pieces that could appropriately aid in the description or decision about the data, then the student reasons about the number of data points in the various parts of the data set in part-whole terms. A common qualitative characteristic of the data that the students identified and shared was that of being "bunched up." This intuitive observation could be indicative of noticing

variation in the data, which contributes to the shape of the distribution. When a data set was identified as being bunched up about a certain value, the students considered this an indication of consistency.

The second mini-tool included all the features of the first and could additionally handle one or two data sets of up to 400 points. The second mini-tool also allowed the students to partition the data into groups of a specified size, partition the data into groups with a specified interval width, partition the data into two equal groups, and partition the data into four equal groups (see Figure 13). The remaining set of lessons utilizing the second mini-tool incorporated discussions that focused on reasoning about the data multiplicatively as opposed to reasoning about the data additively. Multiplicative reasoning occurs when a student partitions the data set into two or more pieces that could appropriately aid in the description or decision about the data, then the student reasons about the proportion (or percentage) of data points in the various parts of the data set in part-whole terms, i.e. reasoning proportionally.

For example, a student who is examining a data set that contains 40 data points would be reasoning additively if he observed that 20 values are above the median, but a student who is reasoning multiplicatively would observe that half (or 50%) of the data is above the median. Proportional reasoning becomes essential when a comparison of unequal sized data sets is required. The authors describe the ‘mathematical practice’ that emerged from the second set of lessons as that of “exploring qualitative characteristics of distributions.” As part of this mathematical practice, the students reasoned multiplicatively about the data, used the computer

mini-tool to identify global patterns in the data, and the students described those patterns in qualitative terms, such as using the phrase ‘hill shaped’ to describe a data set. In particular, the ‘hill shaped’ description was also seen as a shift from focusing on individual data points to a more global view of the data.

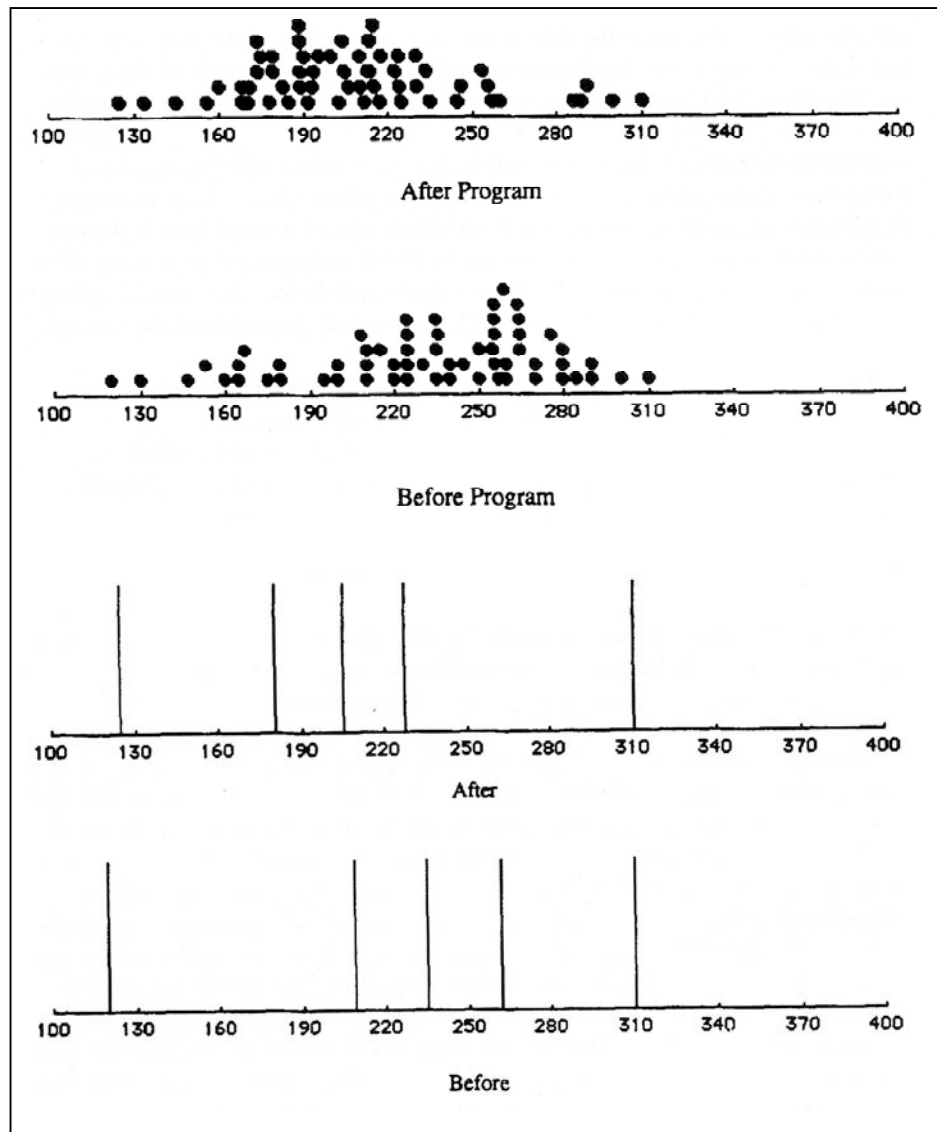


Figure 13. Cobb's second computer mini-tool.

Results indicated that the lessons using the first mini-tool were accompanied by predominantly additive reasoning, while the transitions to using the second mini-tool accompanied by the incorporation of multiplicative reasoning. The nature of the class discussion about the data sets also shifted from the first to the second set of lessons. Discussion during the lessons that utilized the first mini-tool focused on the practical decision or judgment that needed to be made while discussions during the lessons that utilized the second mini-tool shifted to ways to organize the data that supported particular decisions or judgments.

In an investigation of ways to support middle school students' development of statistical reasoning McClain, Cobb, & Gravemeijer (2000) developed an instructional sequence to support students' gradual development of the single, multifaceted notion of distribution. They conjectured that if students began to think about data in terms of distribution, then investigations into structuring data could help to identify trends and patterns in the data. The instructional tasks incorporated the use of two computer mini tools designed to support students' ability to either described a data set or analyze two or more data sets in order to make a decision or a judgment.

By the end of the experiment, students did begin to reason about global trends in the data sets, whereas previously they focused on number of points in particular regions of the sets. This was evidenced by over half of the students routinely describing a part of the distribution as a proportion or percentage of the whole. That type of reasoning is what Konold and colleagues (Konold, Pollatsek, Well, & Gagnon, 1997) have called *group propensities* (i.e., the rate of occurrence of some data value

within a group that varies across a range of data values). Further, some students also included justifications for the statistics that they used in their arguments. For example, a student justified partitioning a data set based on the location of a ‘hill’ as opposed to arbitrarily partitioning the data at the midpoint of the range.

Bakker & Gravemeijer (2004) explored how middle school students’ informal reasoning about distribution can be developed in a technological learning environment. The teaching experiments were conducted during 12 to 15 lessons in four seventh-grade classes in the Netherlands. The researchers envisioned “that reasoning with shaped forms the basis for reasoning about distributions” (p. 149), and thus created lessons that utilized specially designed software tools to encourage students’ to create graphs and make predictions about data sets based on informal assessment and reasoning about shapes.

The lessons were audio recorded, student work and final tests were collected in all classes, field notes were taken, and a set of mini-interviews were held during the lessons as well as videotapes and pretests in the last two experiments. The analysis of this data resulted in the identification of patterns of student answers that were similar in all teaching experiments. These patterns were categorized as a learning trajectory that evolved in three stages correlating to the representations of the data used in the lessons.

The activities correlating to stage 1 – Data Represented by Bars – promoted students to reason about different aspects of distributions in informal ways. The students used terminology such as majority, center, extreme values, spread-out-ness,

consistency, chance, and reliability as they compared two data sets on battery life spans for two brands of batteries (10 batteries for each brand). The students investigated the data and prepared a report on the life spans using a computer mini-tool, the same “first computer mini-tool” as previously described in the research by Cobb (1999), where each battery life span was represented by the length of a horizontal bar. The mini-tool also allowed for the students to make visual estimations of the mean of each data set. Some of these students explained that they cut off the longer bars then gave the bit to the shorter bars. The researchers interpreted this activity as using the mean as a representative value for a data set and reasoning about the brand instead of the individual data values.

The activities correlating to stage 2 – Dots Replace Bars – encouraged the students to reason about shapes of distributions and to quantify informal notions such as frequency and the majority. In this second stage, the students engaged in activities where they represented their data in stacked dot plot form using a second computer mini-tool, the same “second computer mini-tool” as previously described in the research by Cobb (1999). The dot plot representations created by the students who used mini-tool 2 were closer to conventional representations of distributions than the horizontal bars from mini-tool 1, and students could organize data in ways that came closer to histogram and box plot representation. An example of a statistical problem that students solved with mini-tool 2 was on jeans sizes. Students had to report to a factory their recommendation for the percentage of each size that should be made, based on a data set of the waist measurements (in inches) of 200 men. The intent of

this activity was to refocus students' attention on the whole distribution as opposed to the mean as well as providing an opportunity for students' to reason about absolute and relative frequencies. In the jeans sizes activity the researcher saw that students tend to divide unimodal distributions into three groups of 'low,' 'average,' and 'high values.' The 'average' group was characterized by the majority in the middle (similar to what Konold and colleagues (2002) called *modal clumps*) and seemed to be more meaningful to students than the single value of the mean.

During the activities correlating to stage 3 – Symbolizing Data as a Bump – the students created their own graphs, without data, to analyze and describe what they expect in various scenarios. Because up until this point the students had not explicitly reasoned about and with informal shapes of the data, the teacher introduced the term “bump” to draw their attention to the shape. As the students used this new terminology, some of them began to make predictions about density and shape in terms of the whole distribution, however most still focused on subsets of the distribution.

Middle School Teachers

McClain (2003) reported on a one-year collaboration between researchers and middle school teachers where the teachers participated in activities on reasoning about data in preparation for giving the same activities to students. The collaboration occurred over one academic year during monthly work sessions and a week-long summer work session. The work sessions were centered around an instructional sequence that was designed to support middle school students' beginning to reason

about data in terms of distribution by focusing on multiplicative ways of structuring data. The goal of the research was to provide an analysis of the development of one group of middle school teachers' understandings of statistical data analysis.

The instructional sequence contained activities that involved using the same computer mini-tools as previously described in the research by Cobb (1999) to describe, compare and analyze data sets in order to make a decision or judgment. Each session was video recorded, field notes were taken by a research assistant, copies of teachers' and students' written work were collected, and interviews with each teacher were audio taped. The analysis of the work sessions revealed a progression of normative ways that the teachers reasoned over the course of the year. Many of the teachers structured the data by placing a vertical line in the data to create a 'cut-point,' and when they reasoned about the relative number of data points in each part of the distribution, the researchers claimed that the teachers were seeing the data as an aggregate. In tasks involving a comparison of unequal sized data sets, some teachers began to see the data from a "density perspective," that is, from the perspective of simultaneously coordinating the reading of frequencies over each axis as rates where the total accumulates as the data is read from left to right.

Makar and Confrey (2002) investigated middle school teachers' statistical reasoning when comparing two groups as the teachers participated in an immersion model of professional development that involved doing statistics like statisticians. The authors posit four constructs of the concepts necessary for meaningful comparisons of two groups: measurable conjectures, tolerance for variability, understanding of the

context, and an ability to draw conclusions and/or inferences based on data. The measurable conjecture construct occurs when the teachers moved from working on problems to conjecturing about the data. Makar and Confrey claimed that a tolerance for variability, which includes variability within a group, variability between groups, and variability from one sample to the next, is essential for subjects to make descriptive comparisons. They also claimed that when the data being examined is set in a context that is understood, the subjects are more likely to make in-depth analyses of the data. Finally, Makar and Confrey conjectured that the ability to compare two groups is a powerful tool that can be used to draw conclusions and leads into the arena of inferential reasoning, an important skill at all levels of learning statistics.

The teachers demonstrated the greatest competence with descriptive statistics and graphical representations of data, particularly when they worked with some statistical software. Using the software the teachers generally demonstrated adequate skills in choosing appropriate graphs and use of summary statistics to describe differences between groups. The teachers had considerable difficulty in providing non-deterministic, data-based evidence for their conjectures about group comparisons beyond descriptive statistics. For example, when assessing whether or not group differences were significant, the teachers relied on intuitive judgments of “big enough” rather than using statistical tests available to them through the computer software. Although these teachers appeared to have a tolerance for variability, it was only in an informal and naive sense.

Each of the studies reviewed in this sub-section gathered data from students' work on classroom activities, as well as task-based surveys and interviews. Overall there were indications that the lessons, activities, and possibly the interviews themselves helped to promote a move towards a distributional perspective of data sets. There was also ample evidence that students and teachers who had minimal experiences with data tended to hold local views of distributions. When students had experiences and focused instruction dealing with variation in data in conjunction with other features of a distribution such as center or shape or size (see Petrosino, Lehrer, & Schable, 2003), some of those students compared data sets in a distributional way by incorporating both centers and variation into their comparisons.

The participants' local views of distributions came to light through descriptions and comparisons of data sets that were based on individual points such as maximums or minimums or on the frequency of occurrence of a specific data value or on a sum or a count of values above or below a "cut-off point." These methods were seen as particularly problematic when a comparison of unequal sized data sets was required. Slightly more than a strictly local perspective was characterized by use of individual descriptors or individual characteristics such as only range, or only median, or only mean.

Researchers generally agreed that the beginning of a shift away from a local view was characterized by qualitative shape assessments. Other moves toward global views were characterized by multiplicative reasoning, i.e., quantifying proportions with respect to the whole data set above or below a specific value. Global views were

generally characterized when there were indications of using centers as a representative characteristic of a distribution. Participants, who perceived data sets globally, found it difficult to reason with a center without a corresponding spread. They also may have justified the mean as representative by shifting data in a leveling process similar to that described by Mokros and Russell (1995). In the Makar and Confrey (2002) study where teachers had access to descriptive statistics to aid in their descriptions and decisions, it was unclear how those teachers perceived the data sets that they were working with.

Literature Review Discussion

This review has focused on studies related to investigations of people's strategies, reasoning and conceptions when comparing distributions: Research focused on intuitive strategies when predicting and making informal inferences; Acknowledgment, understanding and reasoning about variation; and Reasoning about distribution. Some common results that are seen in the previous sections are that many students in grades K-12 as well as their teachers hold views of data that ranges from local in nature to global in nature. Each of the studies collected research data through utilizing task-based interviews with the participants, while some also incorporated data from classroom activities during experimental lessons. Overall there were inconsistent indications that older participants tended to have views of data that were closer aligned with a global perspective. The participants whose strategies for making comparisons and informal inferences were more sophisticated, tended to compare the data sets in distributional ways by basing their comparisons on several features of the

distributions, such as center, spread and shape. The less sophisticated strategies were based on simplistic aspects of the data sets, such as individual data points, frequencies, and sums.

Students' understanding of variation has mostly been investigated with undergraduate university students. Some of these students clearly have sophisticated concepts of variation, particularly that of standard deviation. It appears from these studies that a local/global perspective of variation is closely tied to a corresponding local/global perspective of distributions. The naïve, local conceptions of variation included equating range and the general term of 'variation,' or equating the 'bumpiness' of a bar graph to its variation also tended to use less sophisticated strategies to make data set comparisons. Alternately, some studies suggested that students who may have a more sophisticated understanding of variation, particularly as a feature of a distribution, tend to employ more sophisticated strategies for comparing data sets. Overall this literature review supports the assertion by Bakker (2004, p. 64) that,

Despite differences between the curricula in different countries, the underlying problem remains the same: students generally lack the necessary conceptual understanding for analyzing data with the statistical techniques they have learned. The problem many statistics educators encounter is that students tend to perceive data just as a series of individual cases (a case-oriented view), and not a whole that has characteristics that are not visible in any of the individual cases (an aggregate view).

Given the importance of understanding data sets from a global perspective, as distributions (Bakker & Gravemeijer, 2004; Ben-Zvi & Garfield, 2004; Konold,

Pollatsek, Well, & Gagnon, 1997) and that the ability to compare two groups is at the heart of statistics (Konold & Higgins, 2003), an understanding of how students and teachers reason about distributions and their characteristics is of vital importance. Research in this area with students at the K-12 level, their teachers and introductory statistics students at the university level is growing, yet conclusions are still tenuous at best. There has been no research in this area done with upper level undergraduate or graduate statistics students. Research with undergraduates has the potential to confirm and extend initial results, while investigations with graduate students has the potential to inform us about continuing difficulties as well as indicate possible useful conceptions and strategies that could be fostered in students taking introductory classes.

Framework

As described in the introduction and expanded upon in the literature review, there is general agreement among many educational researchers that a global understanding of data is critical in the development of the concept of distribution. Also, when engaging in tasks that require a comparison of data sets, applying that conceptual understanding about distribution to make valid comparisons, descriptions and decisions about data sets is vital. Results from the studies cited in the literature review as well as the development of different frameworks used by Shaughnessy and colleagues (see Shaughnessy, Ciancetta, Best, & Noll, 2005; Shaughnessy, Ciancetta, & Canada, 2004) and Watson and Moritz (see Watson, 2001, 2002; Watson & Moritz, 1999) and Bakker and Gravemeijer (2003, 2004), respectively, indicate that students

commonly perceive and compare distributions in ways that can be categorized into hierarchical levels, from reasoning additively to distributionally, or from unistructural strategies to relational strategies, or from viewing data sets as individual points to viewing data sets as distributions. At the lowest level students do not understand or appropriately engage in the task, while at the higher levels students transition from making comparisons about individual data points to global comparisons. Although each of the following frameworks in this section do not necessarily reference “local” and “global” perspectives, the commonalities are striking.

The framework that has been used in the analysis of survey and interview data from this study is an expansion and refinement of the frameworks developed by Shaughnessy, Ciancetta, and Canada (2004) and Shaughnessy, Ciancetta, Best, and (2005). The initial framework (see Shaughnessy, Ciancetta, & Canada, 2004) and a refinement of the initial framework, the lattice structure framework (see Shaughnessy, Ciancetta, Best, & Noll, 2005) were developed to categorize grades 6-12 students’ responses to tasks that involve comparing distributions of repeated random samples. An expansion of the lattice structure framework was hypothesized based on the frameworks of Watson and Moritz (1999) and Bakker and Gravemeijer (2003, 2004) and the results from the research of Konold and colleagues (see Konold & Pollatsek, 2002; Konold et al., 2002) and much of the other research cited in the literature review. The results from those studies indicated that the lattice structure framework could be expanded to describe students’ reasoning when they compare data sets resulting from non-sampling situations and it also could describe the reasoning of a

broader range of students, particularly undergraduate and post-baccalaureate college students.

Initial Framework by Shaughnessy and Colleagues

The initial framework developed by Shaughnessy and colleagues (Shaughnessy, Ciancetta, & Canada, 2004) for describing students' reasoning when they were engaged in tasks involving repeated sampling was a result of one of the phases of a three-year research project on students' understanding of variability in statistics. Part of the study consisted of giving task-based surveys to 272 students from grades 6 – 12 at the beginning of their school year. The surveys were comprised of a series of tasks involving repeated sampling. Students' reasoning on the tasks predominantly fell into three types: *additive* when their explanations were driven by frequencies, *proportional* when their explanations were driven by relative frequencies, or *distributional* when their explanations were driven by both expected proportions and spreads. These three types of reasoning were considered as related hierarchally (see Figure 14) with additive the lowest reasoning level and distributional the highest reasoning level.

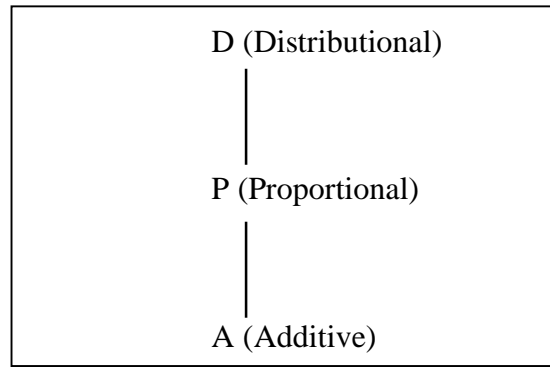


Figure 14. Initial Framework developed by Shaughnessy, Ciancetta and Canada (2004) used to develop the framework in Figure 15.

The tasks were specifically related to dichotomous sampling experiments involving red and yellow candies. When explanations relied on absolute numbers or frequencies of reds in an original mixture such as, “because there are more reds,” they were classified as additive. Proportional reasons fell into two subgroups, implicit proportional reasons and explicit proportional reasons. When students had difficulty putting their reasoning into words, yet their responses implicitly suggested that their thinking involved sample proportions or population proportions, or probabilities, or percents such as, “Most of them will be around 6, but I just can’t explain why,” their reasons were classified as *implicit proportional*. Students’ reasons classified as *explicit proportional* explicitly mentioned ‘ratio of reds,’ ‘percent of reds,’ ‘probability of reds’ in their explanations, and connected it back to the original mixture. Distributional reasons integrated both centers, and variation around those centers, into their reasoning on these tasks.

Lattice Structure Framework By Shaughnessy and Colleagues

The *Lattice Structure* framework (see Figure 15) was developed by Shaughnessy, Ciancetta, Best and Noll (2005) as a refinement of Shaughnessy, Ciancetta and Canada's (2004) initial framework. The expansion of the initial framework to the lattice structure framework was, in part, a result of one of the phases of a three-year research project on students' understanding of variability in statistics. Part of the study consisted of giving task-based surveys to 232 students from grades 6-12 at the end of their school year. The lattice structure framework resulted during the process of categorizing students' reactions to a series of four survey tasks that involved repeated sampling. Two of the tasks involved examining sampling distributions from a known mixture of 750 red candies and 250 yellow candies. One task involved deciding which sampling distributions, from the same mixture, out of a group of four, could be made-up and which could be real, and another task involved predicting what the population could be after examining four real sampling distributions from an unknown mixture. Student responses were described with reference to the lattice structure framework, a five-tiered framework for statistical reasoning: *other* (0), *additive* (1), *transitional* (2), *proportional* (3), and *distributional* (4).

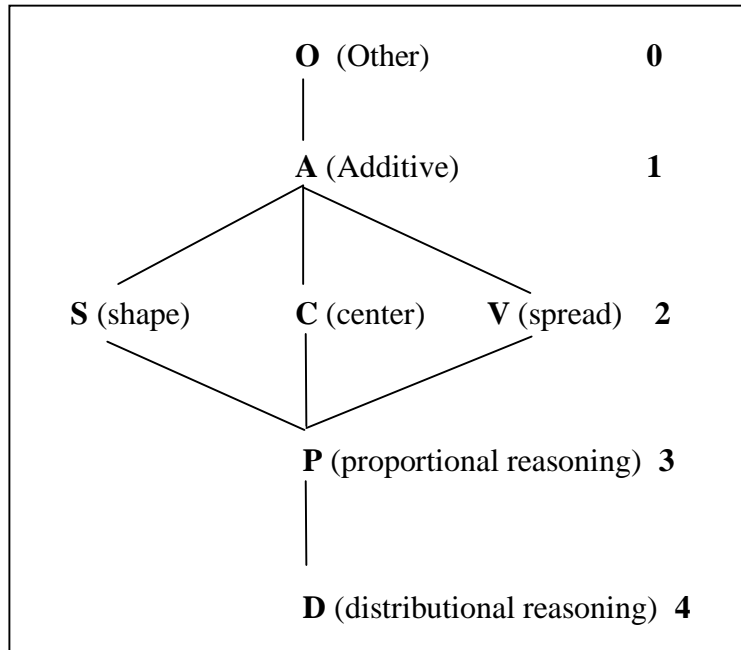


Figure 15. Lattice Structure Framework

A rubric score of 0 was assigned generally when a reason was not given, unclear, not pertinent, or contained no real information. These types of responses were classified as *other* and also included responses that were nonsense responses, off task, left blank, or indicated that the experiment was “random” so outcomes could not be predicted or evaluated. When the dominant type of reasoning was *additive* (A), a rubric score of one was assigned. *Additive* type responses were generally characterized by “more red” arguments, that is, primarily additive or frequency only type reasoning. The collection of *transitional* response types appeared to be in a transitional zone from additive to proportional reasoning. The *transitional* responses tended to focus on one particular aspect of the distribution in a fairly explicit way; shape (S) reasons involved citing the overall shape, such as skewed, normally distributed, center (C) reasons

involved statements or strongly implications of the use of centers (e.g., modes, %, probability) but not necessarily explicit proportional reasoning, and spread (**V**) type reasons involved range, spread, or a general attention to variation. When the dominant type of reasoning was one of the transitional types, a rubric score of two was assigned. *Proportional* (**P**) type responses were primarily focused on the population proportion and *explicitly* connected the population proportion with sample proportion in some way, e.g., “7 of 10 because 750 of 1000” or “the mixture is $\frac{3}{4}$ red, so 7 or 8 out of ten is expected.” When the dominant type of reasoning was proportional a rubric score of three was assigned. The *distributional* (**D**) type responses used multiple aspects of the distribution, *explicitly* integrating at least two of the attributes of a distribution (S, C, V). A rubric score of 4 was assigned to explicit distributional reasoning.

Framework by Watson and Moritz

Watson and Moritz (1999) employed a detailed analysis of students’ responses from tasks that required a series of comparisons of test scores, represented graphically, from two different groups of students. Some comparisons involved groups of equal size and some involved groups of differing size. The levels of understanding displayed by the students’ responses in Watson and Moritz’s study reflected a more general cognitive model. These levels were derived from the cognitive model of Biggs and Collis (Biggs, 1992; Biggs & Collis, 1982) where the numerical and visual strategies that the students used formed the substance of three levels, *Unistructural* (U), *Multistructural* (M), and *Relational* (R) and the observed responses encompassed two

cycles of these levels. The first cycle specifically applied to comparisons of equal sized groups. Watson and Moritz (1999, p. 158) summarized this cycle as:

U₁: A single feature of the graph was used in simple group comparisons.

M₁: Multiple step visual comparisons or numerical calculations were performed in sequence on absolute values for simple group comparisons.

R₁: All available information was integrated for a complete response for simple group comparisons; appropriate conclusions were restricted to comparisons with groups of equal size.

The second cycle specifically applied to comparisons of unequal sized groups.

Watson and Moritz (1999, p. 158) summarized this cycle as:

U₂: A single visual comparison was used appropriately in comparing groups of unequal sample size.

M₂: Multiple step visual comparisons or numerical calculations were performed in sequence on a proportional basis to compare groups.

R₂: All available information, from both visual comparison and calculation of means, was integrated to support a response in comparing groups of unequal sample size.

Framework by Bakker and Gravemeijer

Bakker and Gravemeijer (2003, 2004) have developed and used the structure, shown in Figure 16 below, to distinguish different layers and aspects of distributions. They claim that the structure can be read both upward and downward. When reading the structure upward, the data is seen as individual values that are used to calculate the numerical quantities of mean, median, range, etc. From the upward perspective the calculated value of, for example, the mean is merely the result of an operation on the individual values of the data set. When reading the structure downward, data is

examined through the lens of distribution. That is, distribution is drawn on as on organizing conceptual structure to conceive of center, spread, and skewness as characteristics of the distribution. Bakker and Gravemeijer argue that a statistical expert could combine the perspectives, where the upward perspective leads to a frequency distribution and the downward perspective results from using probability distributions.

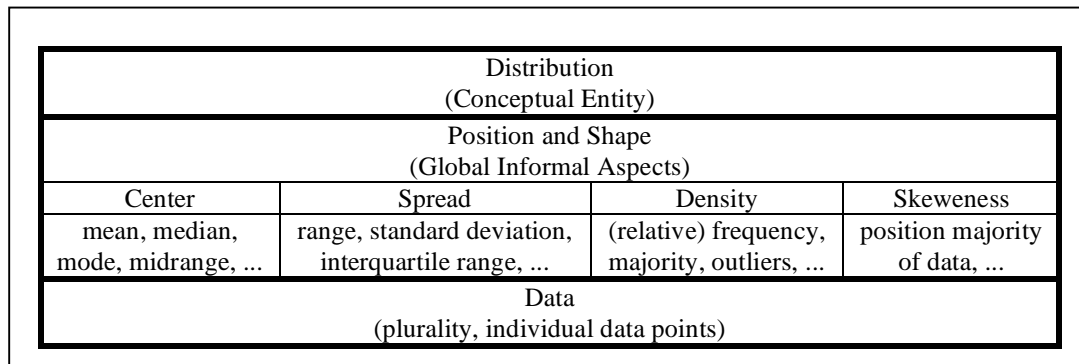


Figure 16: Bakker's and Gravemeijer's framework for describing the relationship between data and distribution.

Expanded Lattice Structure Framework

Based on the frameworks described above, it was anticipated that the analysis of responses to the tasks associated with this research could be interpreted with reference to a five-tiered framework for statistical reasoning that was primarily based on the framework of Shaughnessy, Ciancetta, and Canada (2004), and Shaughnessy, Ciancetta, Best, and Noll (2005) and modified with respect to the frameworks of Watson and Moritz (1999) and Bakker and Gravemeijer (2003, 2004). This framework is organized in a five-tiered lattice structure (see Figure 17). It is comprised of the following levels: *idiosyncratic* (Level 0), *additive* (Level 1), *transitional* (Level 2),

proportional and *global types (not completely distributional)* (Level 3), and *distributional* (Level 4).

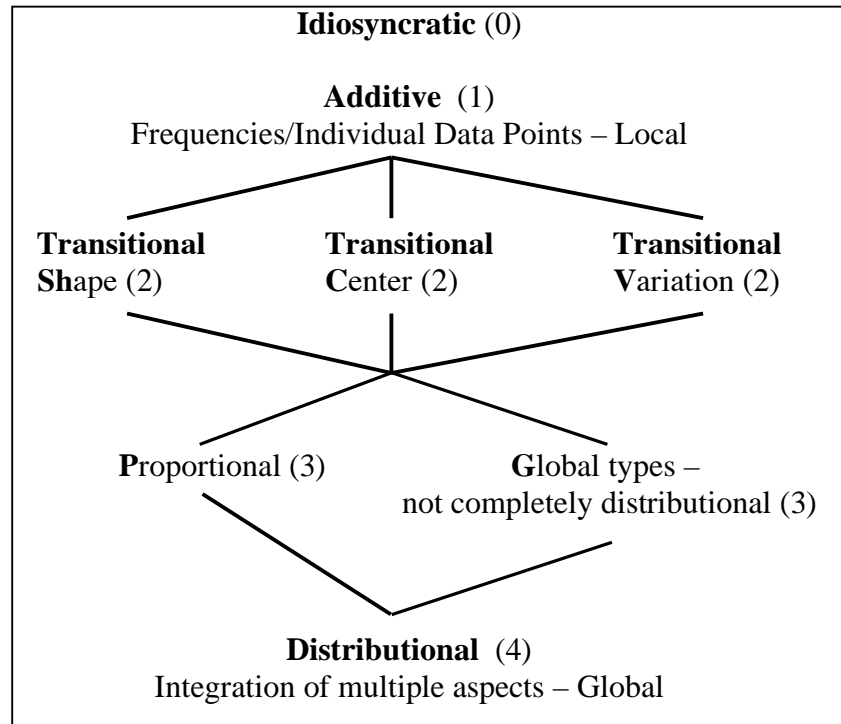


Figure 17. Expanded Lattice Structure, before refinement

Level 0 – Idiosyncratic Responses

The characterization of the *idiosyncratic* level began with level 0, *other* from the framework of Shaughnessy and colleagues with contributions from three of the other studies in the literature review section. These three studies are: Gal, Rothschild and Wagner's (1989) work with investigating young students' natural intuitions and naive statistical reasoning strategies as they made comparisons of data sets; Estepa, Batanero and Sánchez's (1999) study of secondary school students' strategies and judgments of association when comparing two samples of measurement data; and

Ben-Zvi's (2004) investigation of how junior high school students begin to reason about variability as part of an open-ended group-comparison task given in a rich supportive classroom context.

Some of the students in Gal, Rothschild and Wagner's (1989) study inferred qualitative information to make comparisons of data sets. These types of comparisons involved inferring that one group in the comparison is 'trying harder' so they are better. In Estepa, Batanero and Sánchez's (1999) study where students judged the association between two groups, some students relied on their contextual preconceptions of what they expected. These student conjured theories about the data based on their previous knowledge yet devoid of any data-based assessments of association. Finally, Ben-Zvi's (2004) investigation of how students begin to reason about variability as part of an open-ended group-comparison task revealed that students without enough understanding of the context of the task frequently focused on irrelevant information. These students were unable to meaningfully engage in the task.

The *idiosyncratic* level of the framework for this research was initially characterized by responses that were off task, un-codeable, and/or very contradictory or inconsistent. Students might have indicated they were not sure or guessing. Student work would not be helpful in categorizing reasoning, such as a reasons that were completely based on the context of the task or did not refer to the data.

Level 1 – Additive Responses

The main characterizations of the *additive* level are indications in a response that a student holds a *local* view of the data sets. The additive level from the framework of Shaughnessy and colleagues is characterized by responses that focus on frequencies of specific values or individual data points. This additive level correlates to Bakker and Gravemeijer's (2003, 2004) lowest level on their framework, i.e. the perception of a data set as a plurality or individual data points, and also closely matches with Watson and Moritz's (1999) first unistructural level, U_1 , and first multistructural level, M_1 . Typical U_1 responses included the word "more" without further explanation, such as indicating that one of the data sets "got more points." Numerical M_1 responses frequently calculated total scores and then compared those totals. Visual M_1 responses frequently exhibited a consideration of particular individual scores such as noting the mode for one set of scores versus noting the highest score for the other set and no definitive conclusion about which set was better. Similar descriptions of additive type responses were included in the majority of the research cited in the literature review.

In the comparison tasks used by Gal, Rothschild and Wagner (1989), they noted that many of the participating students added up the values in each group to make their comparisons. Using comparisons of specific values in each data set, such as comparing extreme values, was noted in the studies by Konold, Pollatsek, Well, and Gagnon (1997), Estepa, Batanero, and Sánchez (1999), Shaughnessy, Ciancetta, Best, and Canada (2004), and Bakker (2004). Also Konold, Pollatsek, Well, and Gagnon

(1997), noted that some students would base their comparisons solely on the sizes of the data sets they are comparing and categorized this as a “non-statistical” comparison. Finally, Cobb (1999) students who reasoned additively in his study partitioned the data sets, and then, as they made their comparisons, they reasoned about the number of data points in the partitions as opposed to reasoning about the percentage of data points in the partitions.

At the *additive* level, there are strong indications of a local view of the data. Additive responses are only focused on frequencies, individual data points, sums or a comparison of sizes of each data set. In general, there are indications in these responses that the student does not have a global view of the data. It is important to note that when making comparisons of equal sized data sets, some additive type reasons can be appropriate, such as comparing sums, or comparing the number of points in corresponding partitions.

Level 2 –Transitional Responses

At the *transitional* level, there are indications in the responses that the student’s view of the data is transitioning from local to global. These responses tend to focus on one particular aspect of the distribution such as only shape or only center or only variation. Overall there may be fluctuation between global and local views of the data, with the local view being inappropriate. There could also be a monotonous use of formulaic procedures to calculate measures of characteristics, such as either center or variation, with no justification or indication of why it is appropriate to do so or what the measures imply. The separation of the transitional level into three divisions

corresponds with the distributional characteristics of *shape*, *center* and *variation*. Similar descriptions of transitional type responses were also included in the majority of the research cited in the literature review.

Gal, Rothschild, and Wagner (1989) categorized students comparison strategies as *protostatistical* when they ignored multiple features and appeared to consider only part of the data sets. Bright and Friel (1998) found that while many of the young students they studied could describe data sets in terms of what is typical, those students were unable to quantify typical differences between data sets. Common R_1 students' comparisons of data sets that were based on the shapes of the distributions were classified at the first relational level, R_1 , in Watson and Moritz's (1999) framework. Cobb (1999) found that students' qualitative shape assessments, such as 'bunched up' were precursors to identifying global trends and patterns in the data. Bakker and Gravemeijer (2003), and Shaughnessy, Ciancetta, Best, and Canada (2004) students used the informal terminology of 'spreadout-ness,' 'consistency,' and 'reliability' to sometimes assess shape and sometimes assess variation. Petrosino, Lehrer, and Schable (2003) found that some of the students they worked with used only the averages or only the medians or only the spreads or ranges to make comparisons of data sets. Finally Ben-Zvi (2004) found that some students would use measures of both center and spread to compare data sets but their comparisons relied on "insignificant or monotonous use of statistical measures."

Transitional responses are an indication that the student may be transitioning from a local view to a global view of the data. These types of responses tend to focus

on one particular aspect of the distribution such as only attending to shape, center or variation in a fairly explicit way. Reasoning could be predominantly based on informal shape descriptions. Reasons could also refer to average or middles and strongly imply the use of centers for comparisons but not necessarily reflect explicit proportional reasoning, that is, the explanation may focus on algorithmic calculations of aspects such as the mean or median. Transitional reasoning could also be based on range, spread, or informal variability assessments without incorporating other aspects of the distribution.

Level 3 – Proportional and Global Type Responses

The level three responses have indications that students view the data globally but possibly not at the *distributional* level. This level emerged by considering the *proportional* level in the Shaughnessy and colleagues framework and the global informal aspect of *density* in the Bakker and Gravemeijer (2003, 2004) framework along with Konold and colleagues' (1997) assertion that group propensities, i.e., the intensity or rate of occurrence of data values. The other possible level three type responses may provide an informal global “picture” of the data sets yet did not clearly integrate multiple aspects of the distribution and did not make a proportional type argument; these were tentatively called *Global type responses*. This level emerged mainly from some of Watson's (2001) categorizations of uses of variation when comparing data sets, that is using global features but not necessarily integrating, comparing and contrasting them, and from Konold and colleagues' (2002) description of students' use of “modal clumps” to describe data sets, as well as from Makar's and

Confrey's (2005) speculation that there is a perspective of distribution between local and global, that is, a perspective of partial distributions or "mini-aggregates."

The proportional category is supported by the results from Gal et al (1989) who classified strategies that employed proportional reasoning to make comparisons as *statistical*. Also, students' comparisons of data sets by assessing the percentage or proportion of data above a specific value was documented by Petrosino, Lehrer, & Schable (2003), Cobb (1999), McClain, Cobb, & Gravemeijer (2000), and McClain (2003).

Global type responses – not completely distributional was essentially a 'place holder' name and description that was anticipated to be refined during the data analysis process of this research. Results from the reviewed research studies indicated students' explanations could show indications that they are at least in the initial stages of seeing data sets from a global perspective. Such responses may be similar to the use of "modal clumps" as noted by Konold, et al (2002) or "mini-aggregates" as noted by Makar's and Confrey (2005). For example, Bakker and Gravemeijer (2004) reported students using the strategy of partitioning data sets into three groups: a low group, an average group, and a high group. Although this strategy is not explicitly proportional, it does show indications of a global nature.

Level 4 – Distributional Responses

Responses categorized at the *distributional* level show strong indications that the students made global comparisons of the data sets and hence have a global view of the data. This level corresponds to the highest level of Shaughnessy and colleagues'

framework, i.e. *distributional*, where reasoning *explicitly* involves integration of at least two of the attributes of a distribution. This level further corresponds to viewing data as a distribution (or entity) in the framework of Bakker and Gravemeijer (2003, 2004) and also corresponds to the highest level of Watson and Moritz's (1999) framework, that is the second Relational level, R_2 , where both visual and numerical strategies are integrated. The R_2 responses not only used the mean calculations to make comparisons but also integrated other aspects such as the difference in the sizes of the groups to support the comparison of the means. The *distribution* categorization was also included in results from the studies of Estepa, Batanero and Sanchez (1999) and Shaughnessy, Ciancetta, Best and Canada (2004).

Conclusion

The *Expanded Lattice Structure Framework* described in this section was constructed from normative perspectives of the statistics community and from research results of studies investigating students' strategies and reasoning when they describe and compare data sets. It was anticipated that this framework would be refined during the data analysis phase of this research and, in fact, was refined. In this framework, prior to refinement, students' responses to data set comparison tasks are described with reference to a five-tiered framework for statistical reasoning:

idiosyncratic (0), *additive* (1), *transitional* (2), *proportional* and *global types (not completely distributional)* (3), and *distributional* (4). Responses are categorized as *idiosyncratic*, in general, when the student's work was not helpful in categorizing reasoning, such as a reason that was completely based on the context of the task and

does not refer to the data. *Additive* type responses have strong indications that the student has a local view of the data. In *transitional* type responses, there are indications that the student's view of the data is transitioning from local to global. These responses tend to focus on one particular aspect of the distribution such as only shape or only center or only variation. When responses have some indications that the student has of a global view of the data, but the reasoning for the decision does not fully integrate the various characteristics of the distributions, then the response is categorized at level three. These types of responses can be primarily focused on the population density such as proportion above or below a specific value. Other level three type responses can provide an informal global "picture" of the data. The highest level of response is *distributional*. These responses have strong indications that the student holds a global view of the data. Distributional responses integrate two or three aspects of the distribution, such as center and spread or center and shape.

Chapter 3

Methodology

Introduction

Chapter three describes the procedures for gathering and analyzing the data. The first section describes the students who participated in this study and provides a more general description of the research design. The second section offers more detail about the data gathering, including what data was collected and how. The third section illustrates the specific way that data was analyzed.

Participating undergraduate and graduate students who were enrolled in at least one of nine different statistics courses offered at a university in a large metropolitan area located in the Pacific Northwest completed a web-based survey and a subset of those students participated in follow-up in-depth interviews. In the survey, called the *Data Set Comparison Survey* (see Appendix B), the students compared pairs of data sets and decided whether they differ or not and whether one was “better” than the other within the contexts provided. The students provided reasons for their decisions. Those students who participated in the follow-up in-depth interviews had the opportunity to further detail and explain some of their survey responses and respond to follow-up questions.

Subjects and Data Collection

Volunteers from the university’s student population were solicited. The process of soliciting volunteers to take the survey began with the researcher contacting each of the instructors of the courses listed in table 2. Each of the instructors agreed to

provide 5-10 minutes at the beginning of a class for the researcher to come in and describe his research project and the data comparison survey and provide instructions for accessing the survey and soliciting volunteers to take the survey. The survey was available for completion and submission through a web site provided by the Instruction and Research Services department at the university. For those classes that used Web CT (Web Course Tools), a link was provided after the student logged on. Web CT is a web based course management system. For other classes, a web address was given to the students or a link on the instructor's home page was provided. A total of two hundred seventy five students completed the Data Comparison Survey. See Table 3 for the courses that these students were enrolled in.

Students from courses (1), (2), (3), and (4) received extra credit for participating in the study. An alternate extra credit opportunity was available for students who did not wish to participate. Students enrolled in either STAT 243 or STAT 244 were in a variety of stages of degree requirements, from students who were beginning their degree to students who were completing their degree. They had a wide range of ages, from late teenage years to early forties. The mathematical backgrounds of these students varied widely. The prerequisite for enrollment in STAT 244 is STAT 243. For many of the STAT 244 students, STAT 243 was the first math class they had taken in many years and it might have been the first math class they had taken since high school. Also, many of the students enrolled in STAT 244 because it is a requirement of their major field of study. At the other extreme, some of the STAT 244 students may have been mathematics majors who were interested in statistics.

Table 2

Course enrollment of participants.

<u>Statistics Course</u>	<u>Level</u>	<u>Number of participants</u>
(1) Introduction To Probability And Statistics I (STAT 243 for Business Majors)	Undergraduate	26
(2) Introduction To Probability And Statistics I (STAT 243 for Non-Business Majors)	Undergraduate	113
(3) Introduction To Probability And Statistics II (STAT 244 for Business Majors)	Undergraduate	34
(4) Introduction To Probability And Statistics II (STAT 244 for Non-Business Majors)	Undergraduate	37
(5) Applied Statistics for Engineers and Scientists I (STAT 451/551 – 2 sections)	Undergraduate/ Graduate	50
(6) Applied Statistics for Engineers and Scientists II (STATS 452/552)	Undergraduate/ Graduate	3
(7) Introduction to Mathematical Statistics III (STAT 463/563); Applied Regression Analysis (STAT 466/566); Theory of Linear Models (STAT 666)*	Undergraduate/ Graduate	12
Total Participants		275

*Note that Theory of Linear Models was strictly a graduate level course.

The STAT 451/551 courses are required for the undergraduate engineering majors and it is possible that those students have never taken a previous statistics class. All the STAT 560 level and higher classes are required for the graduate statistics majors. All the other upper level statistics classes are comprised of mostly graduate students majoring in statistics or mathematics. A breakdown of the major fields of study of the participants, as reported on by themselves on their surveys, is shown in table 3:

Table 3

Participants' Major Fields of Study

<u>Major Field of Study</u>	<u>Number of Participants</u>
Business Administration Related Majors	76
Science Related Majors	40
Mathematics & Statistics	17
Fine & Performing Arts Related Majors	3
Engineering & Computer Science	51
Economics	5
Social Sciences Related Majors (Not Psychology)	24
Psychology	39
Pre-Medicine/Nursing/Pharmacy	21
Speech and Hearing Sciences	3
Health Science Related Majors	6
Undeclared	10

The participants were divided into five distinct groups: 1-GS, 1-SE, 2-GS, 2-SE, and GRAD. Group 1-GS students [Beginning 1st general statistics course] were enrolled in their first statistics class. Group 1-GS students generally were enrolled in statistics courses (1) or (2). Group 1- students [Beginning 1st statistics for engineers and scientists course] were also enrolled in their first statistics class, but this class was specifically designed for engineers and scientists, i.e. statistics course (5). Group 2-GS students [Beginning 2nd general statistics course] were enrolled in their second

statistics class. Group 2-GS students generally were enrolled in statistics courses (3) or (4). Group 2-SE students [Beginning 2nd statistics for engineers and scientists course or more] were enrolled in course (5) or (6) with students who were enrolled in course (5) indicating that they had completed other statistics courses that were not necessarily designed for engineers and scientists. Group GRAD students [Many senior or graduate level statistics courses] were enrolled in one, two or all three of the courses listed in (7). Tables 4 and 5 display descriptions of each group.

Table 4

Statistics backgrounds of the participants

<u>Group</u>	<u>Description of students' statistics background</u>	<u>Number of participants</u>
1-GS	Beginning 1 st general statistics course	137
1-SE	Beginning 1 st statistics for engineers course	37
2-GS	Beginning 2 nd general statistics course	74
2-SE	Beginning 2 nd statistics for engineers course or more	15
GRAD*	Many senior or graduate level statistics courses	12

*Of the 12 students in the GRAD group, 9 reported their major as Statistics, 2 reported their major as Mathematics, and 1 reported his/her major as Economics.

Table 5

Educational Level breakdown: Counts of each group

<u>Educational Level*</u>	<u>Group 1-GS</u>	<u>Group 1-SE</u>	<u>Group 2-GS</u>	<u>Group 2-SE</u>	<u>Group GRAD</u>	<u>Total</u>
Freshman	35	0	16	0	0	51
Sophomore	37	6	20	0	0	63
Junior	42	17	18	5	0	82
Senior	12	11	17	6	2	48
Postbac	11	1	3	1	1	17
Graduate	0	2	0	3	9	14

*Educational level was self-reported. Freshman = 1-44 quarter credits; Sophomore = 45-89 quarter credits; Junior = 90-134 quarter credits; Senior = 135 or more quarter credits.

The next phase of the study involved video taping follow-up interviews with some of the students who took the survey. This phase was described to the students at the conclusion of the survey. The interviews were expected to last about 30 minutes and those students who participated in the interview phase would be reimbursed \$10 for their time. Students checked on their survey either “Yes, I am willing to be considered for a follow-up interview” or “No, please do not consider me for a follow-up interview.” After the students completed the survey, two students each from courses (1), (2), (3), (4) and three from course (5) (see table 3) were selected for interviews. This group of interviewees was chosen to reflect as many different decisions as possible that could be made on the survey tasks. Only one student agreed to be interviewed from course (6). Ten students from the three courses in (7) (see table 3) agreed to be interviewed. Twenty-two interviews were completed. Surveys were

completed during the first two weeks of the spring term of 2005 and the interviews were completed during the 3rd and 4th weeks of the same term.

Research Design

A mixed-method approach was used in this study where a large amount of qualitative written task-based survey data was both collected and used to inform the collection of a small amount of in-depth interview data. Then both qualitative methods and quantitative methods were utilized in the data analysis stage. Data analysis will follow Onwuegbuzie and Teddlie's seven-stage conceptualization of the mixed methods data analysis process, followed by interpretation and legitimization of the data and then conclusions. The seven stages are as follows: (1) data reduction, (2) data display, (3) data transformation, (4) data correlation, (5) data consolidation, (6) data comparison, and (7) data integration.

Data Collection

As previously described, the task-based survey was administered to a large convenience sample of 275 statistics students consisting of 244 undergraduates (51 freshmen, 63 sophomores, 82 juniors, and 48 seniors), 17 post-baccalaureate students and 14 graduate students. The data set comparison survey was designed to address the main research questions: What are university-level statistics students' informal conceptions of distribution, and how do they reason when comparing data sets? For the survey, students were asked to examine pairs of data sets immersed in three contexts: Test scores, movie wait-times and ambulance response times. The tasks required the students to decide if one set of data was better than the other in a

particular situation and context. Then the students responded to the open-ended question about what their reasons were for their decision, generating qualitative data. See the *Task Development: Data Set Comparison Survey* section for complete details concerning the data set comparison survey.

The task-based interviews followed up on the students' responses to the tasks from the survey. These interviews focused on the survey tasks that required comparisons of data sets of unequal sizes. Those tasks were chosen for the interview because they require more complex types of reasoning, as determined from the results of the literature review. The interviews also included additional questions regarding the students' interpretations of the questions from the survey and explanations of their understanding of the meanings of various statistical terms used in the survey. Data recorded in the video (and audio) format include verbal explanations and other visual data such as hand gestures. Students also had the opportunity in the interview to produce written data. Only a few students did this by sketching graphs to aid in their explanations. These interviews were considered task-based because the interviewees interacted with both the tasks' environments and the interviewer throughout the interview (Goldin, 2000). One of my main purposes for conducting the interviews is aligned with Patton's (2002, p. 341) described purpose for conducting interviews, that is, "to allow us to enter into the other persons perspective. Qualitative interviewing begins with the assumption that the perspective of others is meaningful, knowable, and able to be made explicit." My other main purpose for conducting the interviews is to gain a source of triangulation of the collected survey data.

Task Development: Data Set Comparison Survey

The tasks in this survey (see appendix B) are set in three different contexts: Test scores in Tasks 1, 3 and 4 , wait times at movie theaters in Task 2 , and response times for ambulance services in Tasks 5 and 6 (developed by the researcher for this study). These tasks are designed to address the two specific research questions: “What aspects of distribution (i.e. center, shape, spread) do students attend to when comparing data sets?” and “What types of strategies do students use when making comparisons of data sets?”

Each of the six tasks contain two data sets that are comprised of quasi-real data derived from the particular context of the task. The data sets are presented in graphical form and students are asked to make a decision based on that data and explain their decision. Previous researchers have used versions of Task 1 which involves comparing test scores from the Yellow and Brown classes (see Gal, Rothschild, & Wagner, 1989; Watson & Moritz, 1999). The particular version most similar to the one used in this survey can be found in Watson and Moritz (1999).

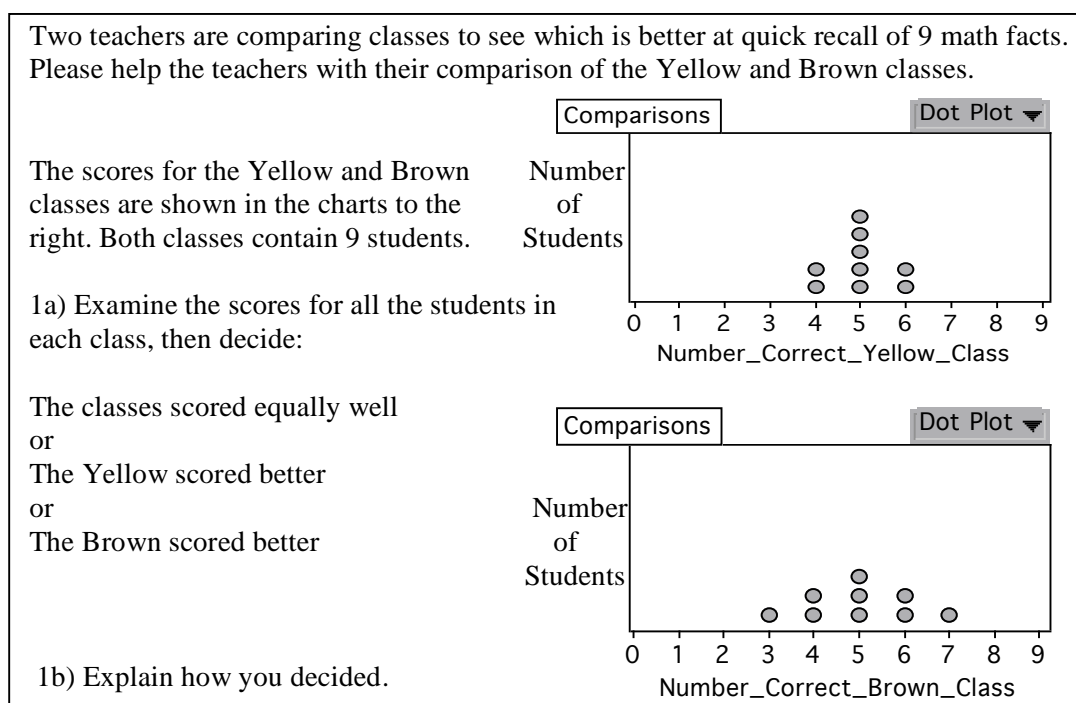


Figure 18. Data Set Comparison Survey: Task 1, the Yellow/Brown task.

Task 1 (see figure 18) will be referred to as the *Yellow/Brown task*. The data sets in the Yellow/Brown task have equal centers (mean, median, mode) and the equality of their centers is assumed to be visually evident so there is no mention of that fact. The data sets also have similar unimodal shapes but differ in their range. Tasks from previous studies that required students to make comparisons concerning data sets with equal centers and similar shape have promoted a variety of reasoning, particularly about shape and variation.

The version of task 2 used in this study (see Figure 19), comparing movie wait-times from the Maximum and Royal Theaters, was previously used with students in grades 6-12 (see Shaughnessy, Ciancetta, Best & Canada, 2004). Task 2 will be

referred to as the *Movie Wait-Time task*. In the Movie Wait-Time task students were given the information that the means and medians are equal. These facts along with the bi-modal shapes and different ranges of the graphs were intended to promote reasoning about the variation in the comparison. The follow up question asking the student to choose which theater to attend allows for contextual aspects of the comparison to be brought to light as well as to provide an opportunity for the student to indicate other features of the data sets that may influence his or her decision.

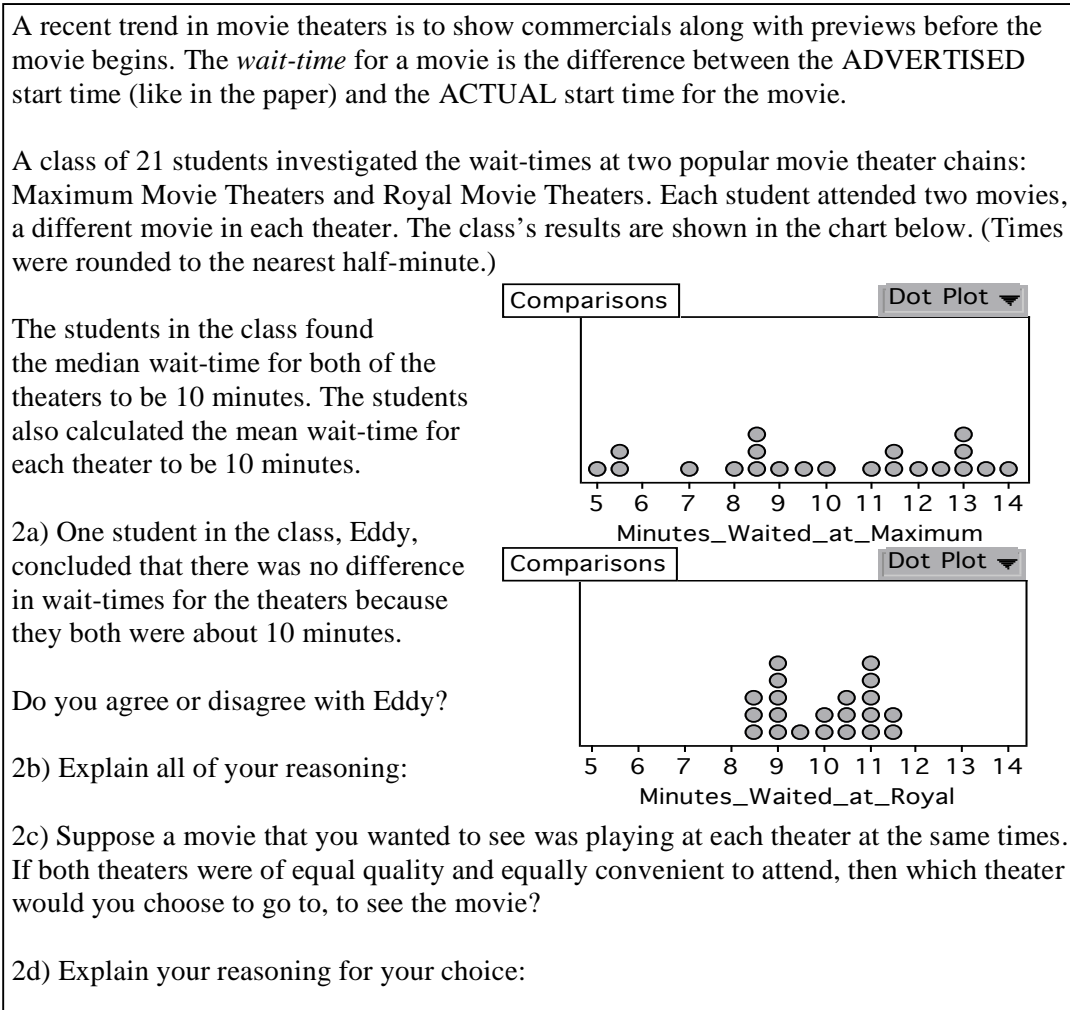


Figure 19. Data Set Comparison Survey: Task 2, the Movie Wait-Time task.

Several versions of Task 3, comparing test scores from the Pink and Black classes, have previously been used with students in grades 3-12 (see Gal et al., 1989; Watson & Moritz, 1999). The particular version most similar to the one used in this survey can be found in Watson and Moritz (1999). Task 3 will be referred to as the *Pink/Black task* (see figure 20).

Two teachers are comparing classes to see which is better at quick recall of 9 math facts. Please help these teachers with their comparison of the scores for the Pink and Black classes.

The scores for the Pink and Black classes are shown in the charts to the right. The Pink class contains 36 students and the Black class contains 21 students.

3a) Examine the scores for all the students in each class, and then decide:

The classes scored equally well.
or
The Pink class scored better.
or
The Black class scored better.

3b) Explain how you decided.

3c) If you decided that one of the classes scored better, then estimate how much better. (If you decided the classes scored equally well please enter 0.)

3d) Explain how you determined your estimation. (If your estimation in part 3c was 0, then enter the word 'equal' for your estimation.)

Comparisons

Dot Plot ▼

Number of Students

Number_Correct_Pink_Class

Comparisons

Dot Plot ▼

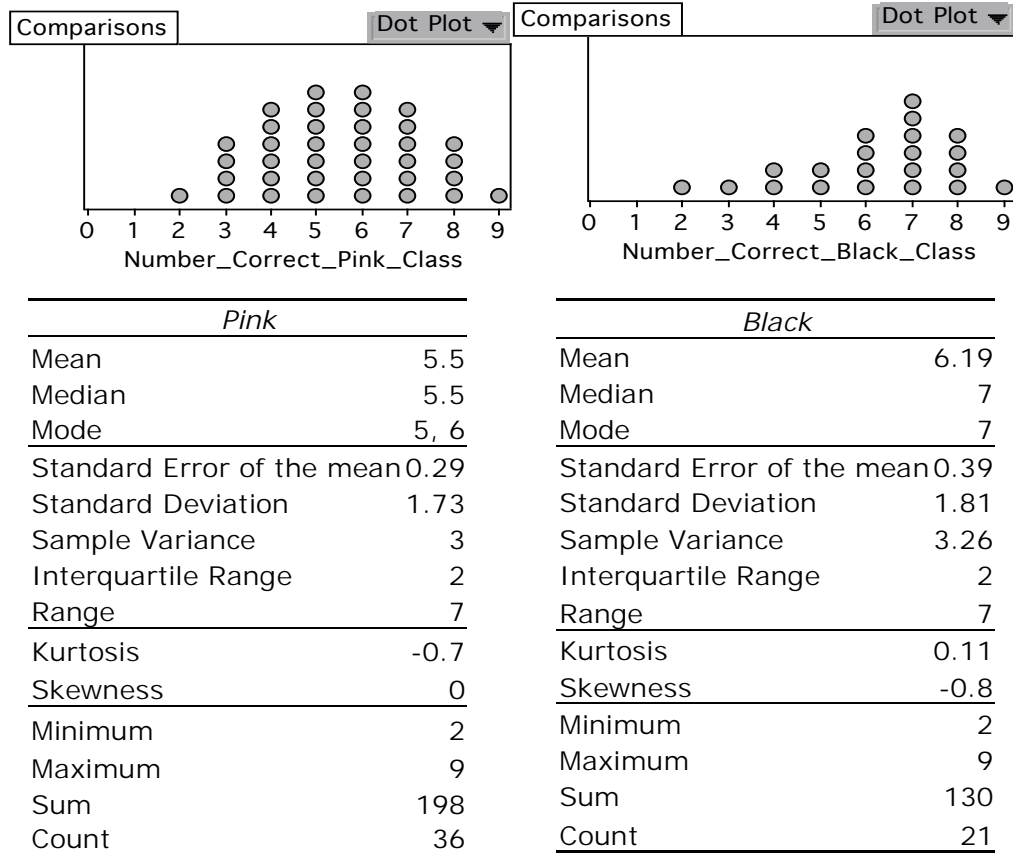
Number of Students

Number_Correct_Black_Class

Figure 20. Data Set Comparison Survey: Task 3, the Pink/Black task.

104

Now you have the opportunity to re-examine the Pink and Black class's scores. The same charts containing the data are shown below along with some of their corresponding descriptive statistics.



4a) Re-examine the data sets along with their corresponding descriptive statistics. You now have the opportunity to change or amend your responses to the previous question. In light of these descriptive statistics, decide:

The two classes scored equally well.

or

The Pink class scored better.

or

The Black class scored better.

4b) Explain how you decided:

4c) If you decided that one of the classes scored better, then estimate how much better. (If you decided the classes scored equally well please enter 0.)

4d) Explain how you determined your estimation. (If your estimation in part 3c was 0, then enter the word 'equal' for your estimation.)

Figure 21. Data Set Comparison Survey: Task 4, the Pink/Black task with statistics.

In the Pink/Black task there is a clear difference in shapes and centers between these sets. Each of the mean, median, and mode are higher in the Black class. The ranges are the same but the Pink class's data is bell shaped while the Black class's data is skewed. The Pink class is larger in size. Students need to address the issue of the difference in the size of each data set, either explicitly or implicitly, to effectively answer this question. In previous studies this question has illustrated the proportional reasoning skills, or lack there of, that students may have. The second part of the task applies to students who decided that one of the classes did score better; they are asked to estimate how much better. This task follows a line of questioning similar to one used by Bright and Friel (1998) who asked middle school students to estimate the difference in the typical height of two groups of people, based on comparing two distributions of height data.

Task 4 was newly designed for this study (see Figure 21), and is intended to give students the opportunity to re-think their responses to task 3 in light of the provided descriptive statistics for each data set. It was expected that the students might use the statistics to confirm and support their initial judgment, or possibly change their minds and/or change their reasons for their decision. Task 4 will be referred to as the *Pink/Black task with statistics*.

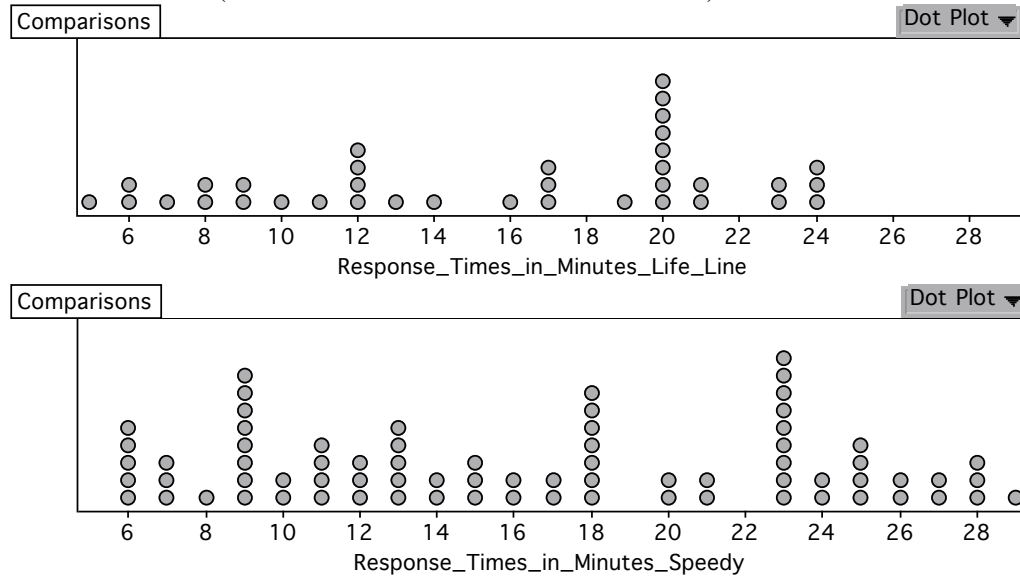
The data sets in Task 5 were newly designed, for this study, and were intended to be challenging to compare (see Figure 22). The data set for the Speedy Ambulance service response times is more than twice the size of the data set for the Life Line Ambulance service response times. Both data sets are unimodal with Life Line's mode

located at a lower time than Speedy's mode. Speedy's data set has several other 'peaks' located at lower times than Life Line's mode, and students may also consider those as modes. Life Line has a smaller mean while Speedy has a smaller median. Life Line has a smaller range than Speedy. Life Line has a slightly smaller minimum and a smaller maximum than Speedy. The data sets have different shapes. The question about which service to recommend is more open ended than the others in an effort to elicit reasoning about context and the data, although it is assumed that quicker response times are more desirable. No descriptive statistics are provided for Task 5 in an attempt to promote reasoning about aspects other than centers, such as shape, variation and clustering. Task 5 will be referred to as the *Ambulance task*.

Task 6 was newly designed for this study, and is intended to give students the opportunity to re-think their responses to Task 5 in light of the provided descriptive statistics for each set (see Figure 23). It was expected that students might use the statistics to confirm and support their initial judgment or possibly change their minds and/or change their reasons. Task 6 will be referred to as the *Ambulance task with statistics*.

The school board for BIG School had to make a decision about which one of two ambulance service companies to call when emergencies arise at their school. The two ambulance companies in the area of the school are Life Line Ambulance Service and Speedy Ambulance Service.

The school board members obtained the most recent 36 response times for Life Line and the most recent 74 response times for Speedy. These response times are shown in the charts below. (Times are rounded to the nearest minute.)



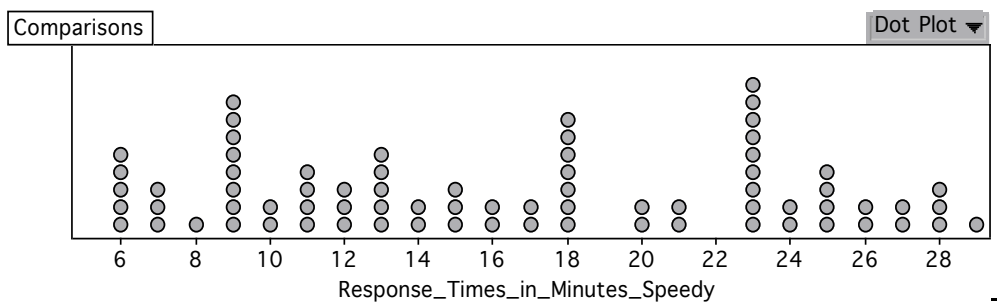
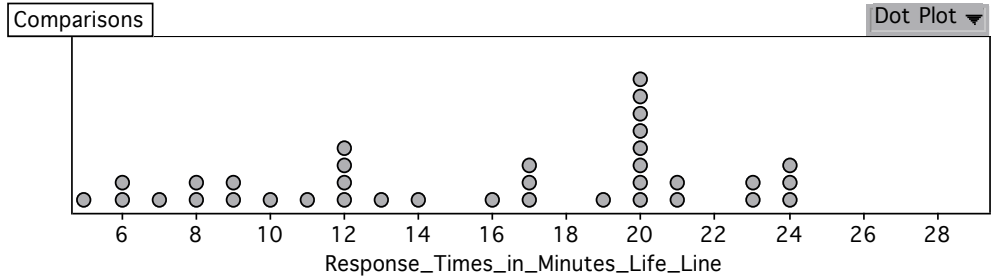
5a) Before the school board members began their debate as to which ambulance service to use, they requested that you look at the data and give your intuitive opinion as to which ambulance service they should choose. Examine the data, then decide:

Recommend Life Line Ambulance Service.
or
Recommend Speedy Ambulance Service.

5b) Explain your reasoning on your choice:

Figure 22. Data Set Comparison Survey: Task 5, the Ambulance task.

After you gave the BIG School board members your intuitive opinion, one member calculated some descriptive statistics and asked you to re-examine the response times for both ambulance services. The same charts displaying the response time data for both Life Line and Speedy ambulance services are shown below along with the corresponding descriptive statistics.



<i>Life Line</i>		<i>Speedy</i>	
Mean	15.56	Mean	16.45
Median	17	Median	16
Mode	20	Mode	23
Standard Error of the mean	0.992	Standard Error of the mean	0.806
Standard Deviation	5.95	Standard Deviation	6.93
Sample Variance	35.4	Sample Variance	48.1
Interquartile Range	9.25	Interquartile Range	12.75
Range	19	Range	23
Kurtosis	-1.295	Kurtosis	-1.276
Skewness	-0.249	Skewness	0.1304
Minimum	5	Minimum	6
Maximum	24	Maximum	29
Sum	560	Sum	1217
Count	36	Count	74

6a) Re-examine the data and the corresponding descriptive statistics, then determine which ambulance service you would recommend.

Recommend Life Line Ambulance Service OR Recommend Speedy Ambulance Service.

6b) Explain your reasoning on your choice:

Figure 23. Data Set Comparison Survey: Task 6, the Ambulance task with statistics.

Pilot Study

The Data Set Comparison survey that was used in this research, was first given to a group of 41 AP Statistics high school students in December 2004 as part of research supported by NSF role grant REC-0207842. The students were from a school located in the suburbs of a metropolitan area in the Pacific Northwest. Of the 41 students, 23 were female and 18 were male. The majority of the students, 26 of them, were 17 years old, while 12 were 18 years old and one student each was 14, 15, and 16 years old. The students were enrolled in two sections (classes) of the same AP statistics course that were taught by the same instructor. It was possible for all the students in both classes to take the survey, although participation was optional. As an incentive for participation, extra credit was offered to those students who completed the survey. The students completed the written form of the survey only. The web-based version of the survey was not available at the time of the pilot study, thus the students took a paper-and-pencil version that had the same wording as the web-based version. The students' decisions and explanations were transcribed, then coded using the lattice structure framework described in the *Framework* section of Chapter 2 (see Figure 24).

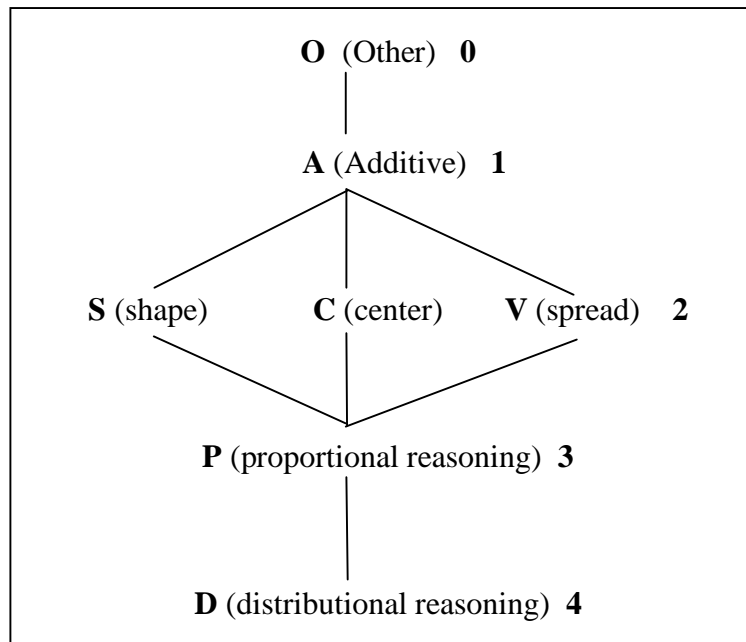


Figure 24. Lattice Structure Framework

The details of the framework were still evolving at the time of the pilot study. The responses from the AP students were categorized into four levels for each of the tasks. Not all of the level 3 responses could be classified as proportional. Based on the literature review and assessment of the pilot study responses, other types of level 3 responses were hypothesized. There were possibly level 3 type responses on the pilot that were not necessarily proportional, but seemed to provide an informal global “picture” of the data sets. Those responses were not level 4 because they did not clearly integrate multiple aspects of the distribution. These types of responses were tentatively called *Level 3 – Global type responses*. Next the Lattice Structure Framework was modified, as seen in Figure 25, and the responses from the pilot study were re-coded using this expanded framework.

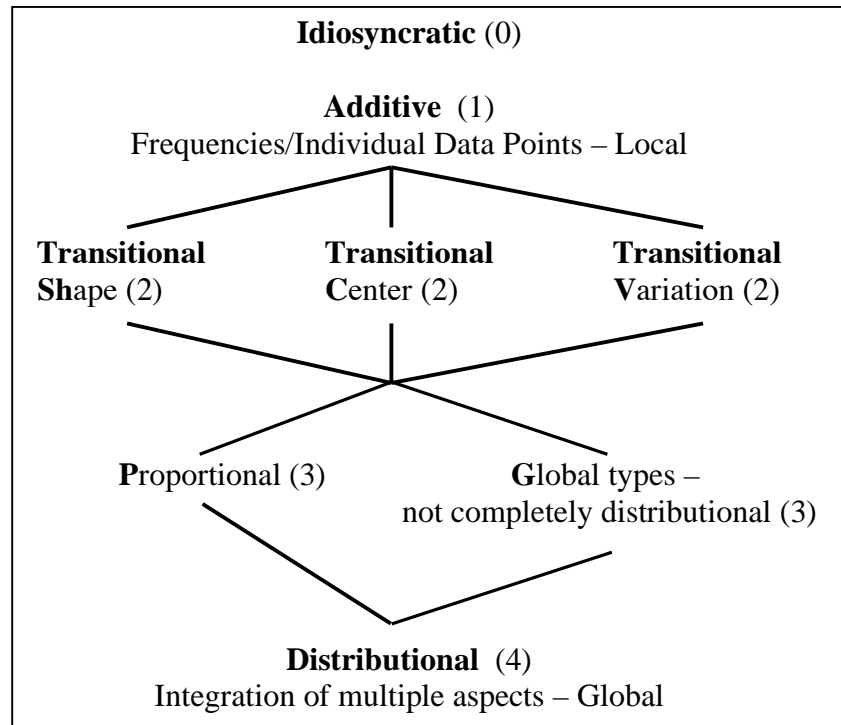


Figure 25. Expanded Lattice Structure, before refinement

The overall results of the pilot are displayed in Table 6. From the results of the pilot study, it was deemed not necessary to make any changes to the Data Set Comparison Survey, because there were responses for each task classified at each of the four upper levels and, for each task, only zero, one or two students provided idiosyncratic, level 0, responses.

Table 6

Pilot Study results

<u>Task</u>	<u>Response Level</u>				
	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
Task 1 Yellow/Brown	1(2.4)	9(22.0)	7(17.1)	9(22.0)	16(39.0)
Task 2 Movie Wait-Time	2(4.9)	8(19.5)	21(51.2)	10(24.4)	12(29.3)
Task 3 Pink/Black	0(0)	3(7.3)	9(22.0)	12(41.5)	17(41.5)
Task 4 Pink/Black w/stats	2(4.9)	0(0)	2(4.9)	19(46.3)	18(43.9)
Task 5 Ambulance	0(0)	9(22.0)	16(39.0)	12(29.3)	4(9.8)
Task 6 Ambulance w/stats	1(2.4)	1(2.4)	4(9.8)	8(19.5)	27(65.9)

*Counts occurring in each cell are given with % out of 41 total students in parentheses: count (% of row).

Task Development: Interview Protocol

After completion of the pilot study, an interview based on the last four survey tasks along with follow up questions about the meaning of some statistical terminology was constructed. After the university students who participated in this study completed the survey, interviews were conducted with 22 of those students. Due to time constraints only the last four of the six task questions were addressed in the interview. These were Tasks 3 and 4, i.e., both of the Pink/Black tasks (before and after being supplied with descriptive statistics), and Tasks 5 and 6, both of the Ambulance tasks (before and after being supplied with descriptive statistics).

Interviews were expected to run approximately 30 minutes, but most ran between 45

minutes and one hour. Participants were compensated \$10 for their participation in the interview. Each interviewee's completed on-line survey was brought to the interview and made available for reference. A blank, hard copy of each survey was also brought to each interview. (See appendix C for the complete interview protocol). In general, the script and protocol were organized according to four stages described by Goldin (2000) as: 1) Posing the question (free problem solving), with sufficient response time allotted followed by only nondirective probes; 2) Minimal heuristic suggestions, if the response is off-task, vague, or the student does not understand the initial question; 3) Guided use of heuristic suggestions, again if the response is off-task, vague, or the student does not understand the initial question or if the anticipated response or description does not occur; 4) Exploratory, metacognitive questions, such as "Do you think you could explain how you thought about this problem?" or "Another student answered this problem by only comparing the means, what do you think their reasons were for doing that?"

The interview proceeded in two phases: The first phase applies to Tasks 3, 4, 5, and 6. Tasks 3 and 5, without the descriptive statistics, were addressed first, then Tasks 4 and 6, with the descriptive statistics, were addressed. The second phase involved an exploration of the meanings that the students associate with several descriptive statistics terms. Students were instructed not to be concerned about providing the same responses as they did when they took the survey, but to respond in whatever way they currently felt was appropriate. After each task, the researcher attempted to introduce cognitive conflict by asking each student to evaluate some

survey responses given other students. These responses were paraphrased from some actual survey responses. A similar technique was used by Watson (2002, p. 252), who claimed that although the responses from other students may not be ideal they, “reflect exactly the type of argumentation one might expect from students interacting with each other in a classroom.” The introduction of potential cognitive conflict was intended to promote the re-thinking of incorrect arguments, to challenge correct arguments and to document the stability of the students’ reasons across their arguments. In the second phase, each interviewee was shown a list of the names of the descriptive statistics used in tasks four and six: Mean, Median, Mode, Standard Error, Standard Deviation, Sample Variance, Interquartile Range, Range, Minimum, Maximum, Sum, and Count. The students were asked which of the terms they felt they understood the meaning of. Then each student was asked to briefly explain the meaning of all those terms that he or she claimed to be familiar with. For those terms that the student was not familiar with, the interviewer offered to provide a brief definition at the conclusion of the interview.

Data Analysis

The data collected in this study was analyzed using Onwuegbuzie and Teddlie’s (2003) seven data analysis stages in the mixed method data analysis process. The seven stages are as follows: (1) data reduction, (2) data display, (3) data transformation, (4) data correlation, (5) data consolidation, (6) data comparison, and (7) data integration. Onwuegbuzie and Teddlie are careful to note that although these stages are presented in a sequential fashion, they are by no means linear as it is

possible that only some of the steps may be used for any specific analysis, such as the data reduction and data display steps and no others. The aim in the analysis is to describe and categorize the data, resulting in a confirming and/or expanding theory of how the participants reasoned when making decisions based on comparisons of data sets.

In stage 1, Onwuegbuzie and Teddlie (2003) describe quantitative data reduction as including computation of descriptive statistics or other exploratory data analysis, while qualitative data reduction may include writing summaries, coding, writing memos, making clusters, and making partitions. Stage 2 involves organizing the reduced Stage 1 data into displays, such as tables, graphs, rubrics or Venn diagrams. The goal behind the choices of displays is to make it easily understandable. The process of qualitizing the quantitative data and/or quantizing the qualitative data is accomplished in stage 3, the data transformation stage. If both types of data are collected, then stage 4, the data correlation stage, is needed. In this stage quantitative data is correlated with any quantitized qualitative data. Stage 6, the data comparison stage, is used particularly when correlation or consolidation of the two types of data is not possible, so that the researcher at least compares data obtained from different sources. In stage 7, the data is integrated into a coherent whole. After integration, initial data interpretation can be made, as well as possible conclusions and inferences.

Data Reduction and Display

For this study, the data were organized into tables and then data reduction began with an examination of the responses to each separate task. To complete the

various phases of this analysis process, I relied on techniques similar to some of the those used in grounded theory as described by Strauss and Corbin (1990) and Dey (1999): open coding (categorizing the data), axial coding (connecting the categories to their subcategories), and constant comparison (category refinement). The Expanded Lattice Structure Framework as described in the Literature Review and Framework chapter guided this process and was itself then refined as a result of this process.

In describing open coding Dey (1999) cites Strauss and Corbin's (1990, p. 62) definition as "the process of breaking down, examining, comparing, conceptualizing and categorizing data." In the process of open coding I proceeded by examining all the responses, task by task and then formed broad initial groupings of the responses. This process was largely influenced by descriptions of the five levels of the Expanded Lattice Structure framework and by the results of the other research studies presented in the Literature Review. Although the initial grouping of responses was influenced by the results of previous research, the groupings were not restricted by it.

Axial coding followed open coding. Strauss and Corbin (1990, p. 62) define axial coding as "a set of procedures whereby data are put back together in new ways after open coding, by making connections between categories." I used axial coding to focus on the responses grouped at each level and developed fine-grained groupings. As with the open coding, the fine grained groupings that emerged in axial coding were influenced by the description of the Expanded Lattice Structure framework and by the other research results reported in the Literature review, but new fine grained groups also emerged that had not been previously reported. The fine grained groups

contributed to the evolving definition of the Levels of the framework and to the sub-categories that comprise the levels in the lattice framework.

The last phase in the process was constant comparison, where each core category (level) was selected and systematically related to all the other categories. Using the constant comparison process, I re-examined the responses and their initial codes and repeatedly went back and forth between the data and the categorizations looking for confirming and disconfirming evidence and consistency from all the data, refined the categories, and repeated the process. This process helps to validate the relationships between categories and fills in categories that need further refinement and development .

Data Transformation

Quantifying the qualitative ‘reasons data’ occurred after the students’ reasons from the data comparison survey and corresponding interview segments were fully categorized. Codes across all the responses for each student were examined and an overall response level was assigned to each student.

Data comparison and integration

In this stage the interviewees were grouped into lattice reasoning levels according to their cross-task survey codes. Of the 22 interviewees, two were classified overall at *Level 1*, five were classified overall at *Level 2*, 11 were classified overall at *Level 3*, and four were classified overall at *Level 4*. Then six case study interviewees were chosen as follows: Both interviewees were chosen from the level 1 overall survey codes, two interviewees were randomly chosen from the level 2 overall survey

codes, one interviewee was randomly chosen from the level 3 overall survey codes, and one interviewee was randomly chosen from the level 4 overall survey codes. These six interviews were transcribed. Then the responses from the second phases of the interview, where students explained their understanding of the meanings of the descriptive statistics used, were examined and compared to the normative meanings of the terms. Then the last four survey questions and follow-up questions were coded according to the framework developed during the analysis of the survey responses. Next the reduced interview data from each of the randomly selected interviewees was compared to the reduced data from their survey for triangulation purposes. This data comparison phase was used to check for internal validity. Differences between results from the interview data and results from the surveys were resolved, explained or noted as potential threats to internal validity.

Data Interpretation

After the data analysis stages were completed, the data interpretation began whereby inferences were made. Both intra-task interpretations and inter-task interpretations were completed. Interpretations included descriptions of apparent trends in strategies, reasoning, and conceptions among the various sub-groups of participants.

Conclusion

Volunteers from the university's student population formed a convenience sample from four undergraduate statistics classes, four upper level undergraduate and graduate statistics classes and one advanced level graduate class. By virtue of their

enrollment and participation in college statistics courses, it was assumed that all the students would have had at least some exposure to the terminology used in the instruments and would have no trouble understanding the questions. Also, as the students were dispersed over courses that range from introductory to advanced, a wide range of responses was expected.

The instruments used for data collection were a task-based survey and an in-depth interview, based on the survey tasks. All the participants completed the survey, then a small group of 22 were purposefully chosen for interviews. Participants were chosen for interviews to reflect a wide range of responses, based on initial analysis of the survey responses. A sample of six interviews was selected (randomly when possible) from the 22 to develop in-depth case studies to support the five course groupings.

The use of Onwuegbuzie and Teddlie's seven stage mixed methods data analysis process, together with employing the coding phases of open, axial, and selective coding from grounded theory, allowed the participants reasoning when comparing data sets to be described through distinct yet linked categories in the refined Expanded Lattice Structure framework.

Chapter 4

Framework Refinement, Results and Analysis

This chapter contains three sections. The first section articulates the refinement of the Expanded Lattice Structure Framework that resulted from the survey analysis, as described in the Methodology chapter. The second section presents the survey results, as interpreted through the refined Expanded Lattice Structure Framework. In the last section, six of the follow-up interviews are analyzed in detail, as interpreted through the refined Expanded Lattice Structure Framework. The interview analyses are intended to provide triangulation evidence for the interpretation of the survey results as well as to explore some of the limitations of using the Lattice Structure Framework to interpret responses to the data set comparison tasks.

Refinement of the Expanded Lattice Structure Framework

The framework used to interpret the survey responses for this research grew out of the framework proposed by Shaughnessy, Ciancetta, and Canada (2004) to interpret student responses to tasks related to comparing data sets in a sampling environment. As a result of my literature review of related research I expanded this framework and hypothesized that the expanded framework could be refined and extended beyond the context of a sampling environment for use in interpreting student responses to tasks related to comparing data sets. This subsection discusses the expansion of Shaughnessy, Ciancetta, and Canada's (2004) framework, including a detailed description and examples of the refinement of that framework, based on the analysis process previously described in the Methodology chapter.

The *Expanded Lattice Structure Framework* described in the Literature Review chapter is a five-tiered framework for statistical reasoning that was constructed from normative perspectives of the statistics community and from research results of studies investigating students' strategies and reasoning when they describe and compare data sets. I hypothesized that this framework, shown in Figure 26, could be used to interpret students' responses to data set comparison tasks with reference to the following tiers: *idiosyncratic* (0), *additive* (1), *transitional* (2), *proportional* and *global types (not completely distributional)* (3), and *distributional* (4).

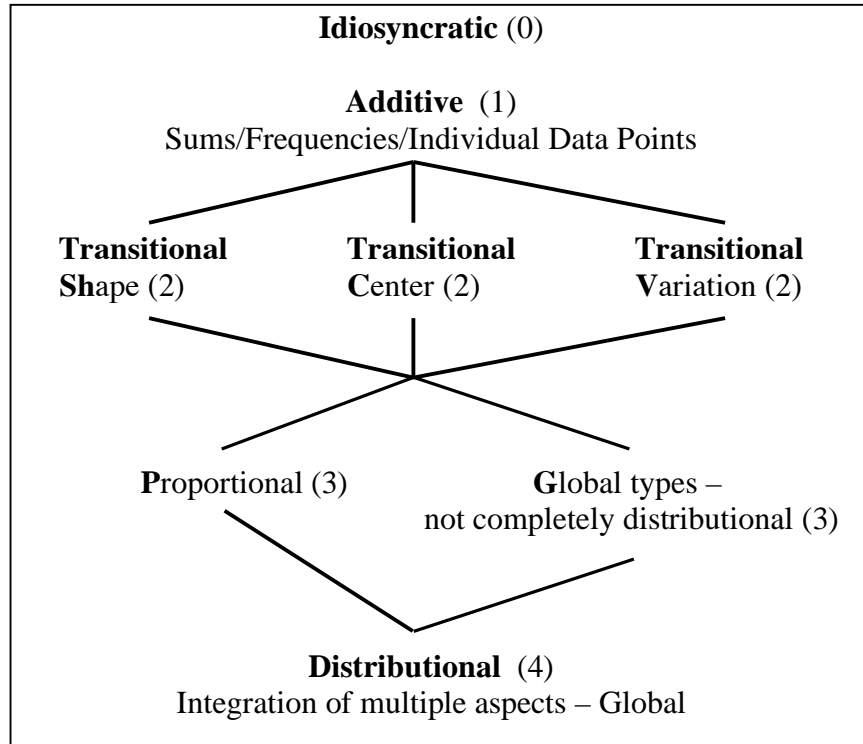


Figure 26. The Expanded Lattice Structure.

Idiosyncratic, level 0 type responses, in general, are not helpful in categorizing reasoning. For example, if a student were to provide a reason that was completely

based on the context of the task and did not refer to the data, that would be categorized as *idiosyncratic*. *Additive*, level 1-type responses tend to rely on local type reasoning, such as reasons based on comparisons of sums. Other level 1-type responses can be based on comparisons of absolute frequencies or individual data points. These responses indicate that, for the particular task that the student has reasoned about, the student has a local view of the data as opposed to considering unifying aspects of the entire distribution. In *transitional*, level 2 type responses, there are indications that the student's view of the data is transitioning from local to global. These responses tend to focus on one particular aspect of the distribution such as only shape or only center or only variation. When responses have some indications that the student has of a global view of the data, but the reasoning for the decision does not fully integrate the various characteristics of the distributions, then the response is categorized at level three. Level three type responses can be primarily focused on the population density such as proportion above or below a specific value. Other level three type responses are part of a hypothesized category of responses that provide an informal global "picture" of the data. For example, responses that focus on more than one measure of center or a focus on the range and mode, but not the mean or median, may contain implications about the overall shape, spread or location of a distribution but do not explicitly attend at least two of those three aspects. The highest tier of response is *distributional*. These responses have strong indications that the student holds a global view of the data. Distributional, level 4 type responses integrate two or three aspects of the distribution, such as center and spread or center and shape.

Through the process of constant comparison (Dey, 1999; Strauss & Corbin, 1990) of students' responses to tasks, as part of the survey analysis, the Expanded Lattice Structure Framework was refined as shown in Figure 27. Levels 0, 1, and 2 were further detailed with their basic descriptions remaining unchanged. The overall name for level 1 was changed to *Local* to more accurately reflect the commonalities among the various types of reasoning at level 1. Level 3 was named *Initial Distributional* with two subcategories: *Proportional* and *Initial Global* (the former hypothesized category). Level 4 was slightly expanded to include proportion as an aspect of a distribution that might be attended to and integrated in a response. These levels were assigned to responses from the tasks on a very conservative basis and thus may be an underestimation of a student's potential maximal reasoning level on a given task. A detailed description of the refinement of each of the levels follows.

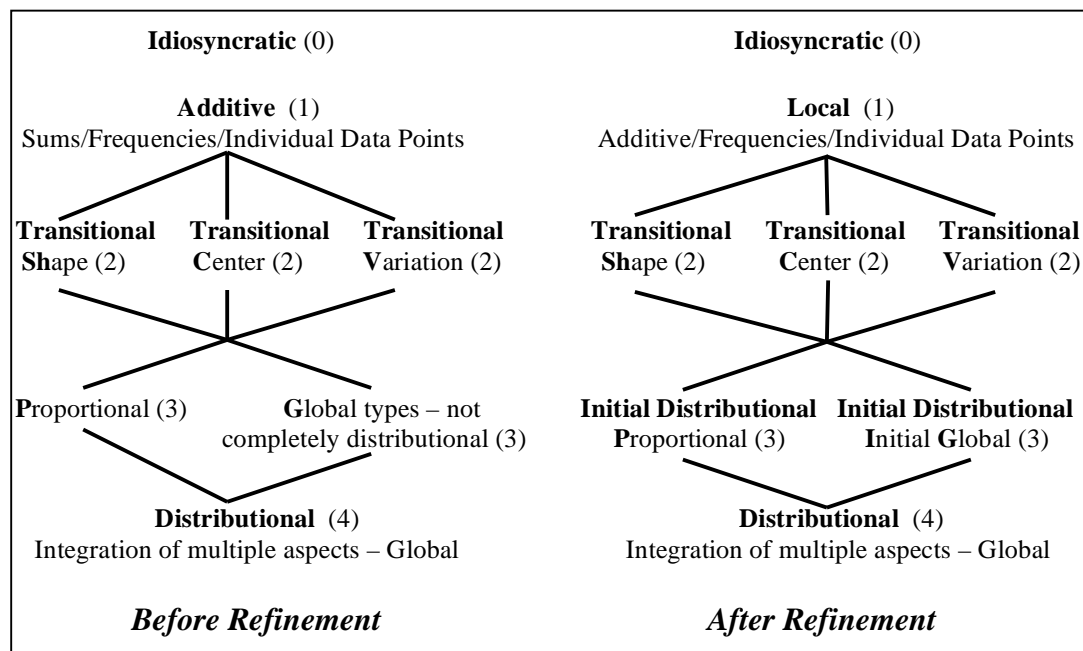


Figure 27. Expanded Lattice Structure, *before* and *after* refinement

Level 0 (Idiosyncratic)

Responses that were generally un-codeable are interpreted as *idiosyncratic* under this framework. Refer to table 7 for examples of this type of response.

Idiosyncratic responses can be off task, contradictory or inconsistent (example 3).

Reasons for decisions might indicate that the student was guessing (example 4). Also, students' work may not be helpful in categorizing reasoning (examples 2 and 5).

Idiosyncratic responses could also be completely based on the context of the task and not refer to the data (example 6). Responses in which there are indications that the student misread the graph or the question were also coded Idiosyncratic (example 1).

Table 7.

Examples of idiosyncratic type responses.

<u>Survey Task</u>	<u>Decision</u>	<u>Reason</u>
Task 1: Yellow/Brown Example 1	Yellow	<i>The students must have learned how to do questions 4, 5 and 6 in the yellow class and were more consistent than the brown class.</i>
Task 2: Movie Wait-Time Example 2	Disagree	<i>The difference between the maximum time and the royal time.</i>
Task 3: Pink/Black Example 3	Black (7 points better)	<i>In quality control. (The mode = 7 is maximum)</i>
Task 4: Pink/Black w/Stats Example 4	Black (2.5 points better)	<i>It tells you. (Guess)</i>
Task 5: Ambulance Example 5	Speedy	<i>Estimating</i>
Task 6: Ambulance w/Stats Example 6	Speedy	<i>Because when I need ambulance, I need speedy ambulance which focus on take patients to hospital quickly. [sic]</i>

Except for task 5, the Ambulance task, fewer than 10% of responses were interpreted as idiosyncratic for each task. This is evidence that the participating students generally understood the survey tasks. Task 5, the Ambulance response time task, was designed to be the most challenging decision of all the tasks as its data sets were designed to not have obvious centers or standard shapes. Also, included in idiosyncratic response were those reasons that were solely based in the context of the problem and did not attend to the data in any way. That type of response was most common for task 5. Table 8 shows the distribution of idiosyncratic responses across all the survey tasks. The Yellow/Brown task and Ambulance task have the highest rate of idiosyncratic responses. This could be due to several factors: The data sets in the Yellow/Brown task have similar shapes and equal centers thus possibly leading students to look for differences not related to the data; The data sets in the Ambulance task are difficult to compare as the many differences between the data sets are not consistent, thus possibly leading students to make their decisions based on contextual assumptions, such as the name “Speedy” implying that ambulance service drives faster.

Table 8.

Distribution of idiosyncratic responses across survey tasks.

<u>Survey Task</u>	Count (percent) of <u>Idiosyncratic responses</u>
Task 1: Yellow/Brown	25 (9.1)
Task 2: Movie Wait-Time	18 (6.5)
Task 3: Pink/Black	11 (4.0)
Task 4: Pink/Black w/statistics	12 (4.4)
Task 5: Ambulance	31 (11.3)
Task 6: Ambulance w/statistics	23 (8.4)

Percentage of total count of participants (n = 275) in parentheses.

Level 1 (Local)

Responses that contained indications that the student may have a local view of the data were classified at level 1. By using the constant comparative method during the survey analysis, level 1 of the Expanded Lattice Structure Framework was refined to more accurately reflect the local perspective of data sets that is commonly associated with these responses. The level 1 descriptor was renamed *Local*, and examples of various Local responses are in Table 9. The original descriptions of level 1 responses were verified. This includes additive type responses, i.e., specifically referring to sums and can be related to comparing data sets of equal size (example 2) or of unequal size (example 5). Additive type reasoning can be appropriately applied to comparisons of equal sized data sets because calculations involving proportions and sums will yield equivalent results in that case (example 2), however, when a

comparison based on sums is applied to data sets of unequal size, erroneous conclusions can be formed (example 5).

Table 9.

Examples of Local type responses.

<u>Survey Task</u>	<u>Decision</u>	<u>Reason</u>
Task 1: Yellow/Brown Example 1	Yellow	<i>Yellow had more that scored 5's</i>
Task 1: Yellow/Brown Example 2	Equal	<i>They both got a total of 45 correct answers.</i>
Task 2: Movie Wait-Time Example 3	Disagree	<i>Certainly there is a difference. The data is not identical. I'm confident that they're different because using the data above I would choose to go to a Maximum movie.</i>
Task 3: Pink/Black Example 4	Equal	<i>it is hard to say which class scored better since the number of students in each class is not the same. 15 students is a large difference.</i>
Task 3: Pink/Black Example 5	Pink (37% better)	<i>More students got correct answers in the pink class.(I looked at how many more correct answers pink had, and considered the # value.)</i>
Task 5: Ambulance Example 6	Life Line	<i>it [Life Line] has the best time (5.5 min) and Speedy has the worst time (29min).</i>
Task 5: Ambulance Example 7	Speedy	<i>Speedy because there were more trials done on Speedy so it is more accurate, you cannot rely on Life Line's 36 responses.</i>

Other types of responses that did not necessarily refer to sums when supporting a particular decision, such as reasoning about individual data points (example 6), an amalgam of individual data points (example 3), and frequencies of occurrence of specific outcomes (example 1) are also categorized at level 1. Finally, the Local

categorization was expanded to include responses that tended to focus exclusively on the size difference of the data sets (examples 4 and 7) because in the absence of reference to any other aspect of the distributions, this type of reasoning appeared to be related to viewing the distributions as amalgams of individual data points.

Table 10 displays the distribution of Local response types across all the survey tasks. Tasks 5 and 1 had the highest percentage of students providing Local level responses. Task 5, the Ambulance task required a comparison of unequal sized data sets, thus most Local type reasoning strategies that would be applied to the data sets would not be considered statistical. Also, the data sets in task 5 had few clear differences between their characteristics and thus the comparison was potentially the most difficult to make. This may be a reason that the highest percentage of students resorted to using Local type reasoning strategies in task 5. Task 1, the Yellow/Brown task required a comparison of equal sized data sets, thus some Local type strategies, in particular reasoning with the sums are appropriate and may be a reason why a higher percentage of students employed Local type reasoning strategies to that comparison.

Table 10.

The distribution of *Local* responses across the survey tasks.

<u>Survey Task</u>	Count (percent) of <u><i>Local</i> responses</u>
Task 1: Yellow/Brown	81 (29.5)
Task 2: Movie Wait-Time	24 (8.7)
Task 3: Pink/Black	42 (15.3)
Task 4: Pink/Black w/stats	10 (3.6)
Task 5: Ambulance	102 (37.1)
Task 6: Ambulance w/stats	24 (8.7)
<u>Percentage of total count of participants (n = 275) in parentheses.</u>	

Another interesting trend seen in the table is the rather dramatic decrease of local type responses from the tasks without descriptive statistics to the tasks with descriptive statistics. When descriptive statistics were included with both the Pink/Black and Ambulance tasks only about one-fourth as many students provided local type responses as compared to when the descriptive statistics were not included. The descriptive statistics seem to have influenced students to move away from local type responses, yet it is not clear that the descriptive statistics influenced these students to reason in more sophisticated ways. It is possible that students felt compelled to refer to the descriptive statistics in their responses merely because the statistics were included with the tasks. By referring to the descriptive statistics in their explanations, many students' responses would no longer be categorized as local.

Level 2 (Transitional)

Responses interpreted as *transitional* are an indication that the student's perception of the data sets may be transitioning from a local view to a global view. The subcategory assigned reflects the primary focus of the student's explanation. Transitional responses tend to focus on one particular aspect of the distribution such as attending only to shape, center or variation in a fairly explicit way. The explanation may focus on algorithmic calculations of aspects such as the mean or median. It is possible for transitional type responses to address several aspects of a distribution; in this case they are made separately, such as listing the values of the characteristics and not integrating or relating them to support the decision, similar to responses classified as *Multistructural* under the SOLO model (see Biggs, 1992; Biggs & Collis, 1982; Watson & Moritz, 1999).

Some of the responses to the two tasks that display some descriptive statistics associated with each data set, that is task 4, Pink/Black w/stats and task 6, Ambulance w/stats, were not entirely captured by this framework. Responses such as those referred to "because of the stats" or contained explanations that were essentially a list of statistical terms, some of which could provide evidence contrary to the decision made. While students who provided these types of responses potentially had a sophisticated global view of the data and may have integrated multiple aspects of the distribution to support their decision, the responses themselves do not provide enough evidence to categorize them higher than at the transitional level. These responses were categorized at level 2 of the framework but were not assigned one of the three

descriptive subcategories. In table 6 they are identified under the heading *N/A* because they are not explicitly associated specifically with the *Center*, *Shape*, or *Variation* subcategories. Unfortunately these types of responses provide little to no insight into students' reasoning other than the possibility that those students may have little or no understanding of the various descriptive statistics, yet they referred to the descriptive statistics in their explanation because the statistics were included with the task.

Table 11 displays the distribution of transitional response types across all the survey tasks. For tasks 1 and 2 the highest percentage of transitional type responses are Shape and Variation, respectively. This could be because the most distinctive features of the distributions in each of those tasks are the similar “bell shapes” of the distributions in the Yellow/Brown task and the strikingly different spreads of the distributions in the Movie Wait-Time task. Thus, students reasoning at the transitional level may be drawn to reason about those comparisons. For the Pink/Black and Ambulance tasks, the highest percentage of transitional type responses were categorized under Shape, yet when students re-examined those tasks with descriptive statistics there was a clear shift away from Shape and to Center. It is possible that because measures of center, such as the mean, median and mode, are commonly introduced to students before any other statistics, students are more familiar with these measure and hence are drawn to refer (or even defer) to them in the explanations for their decisions. A detailed description of the Center, Shape, and Variation subcategories follows with examples.

Table 11.

Distribution of *transitional* responses across the survey tasks.

<u>Survey Task</u>	<u>Count (percent) of <i>Transitional</i> responses</u>				
	<u>Center</u>	<u>Shape</u>	<u>Variation</u>	<u>N/A*</u>	<u>total</u>
Task 1: Yellow/Brown	26 (9.5)	69 (25.1)	17 (6.2)	--	112 (40.7)
Task 2: Movie Wait-Time	23 (8.4)	47 (17.1)	67 (24.4)	--	137 (49.8)
Task 3: Pink/Black	68 (24.7)	77 (28.0)	4 (1.5)	--	149 (54.2)
Task 4: Pink/Black w/stats	86 (31.3)	15 (5.5)	8 (2.9)	30 (10.9)	139 (50.5)
Task 5: Ambulance	36 (13.1)	62 (22.5)	13 (4.7)	--	111 (40.4)
Task 6: Ambulance w/stats	90 (32.7)	9 (3.3)	14 (5.1)	33 (12.0)	146 (53.1)

Percentage of total count of participants (n = 275) in parentheses.

*N/A = Not explicitly associated with the subcategories, Center, Shape, or Variation, due to unclear references to "stats."

Transitional: Shape

Transitional-Shape reasoning is predominantly based on informal shape descriptions. Examples of these types of responses follow in Table 12. Such descriptions commonly use language such as normal, not normal, skew, bell, perfect, or symmetry without incorporating other aspects of the distributions (example 2). There may be indications of an attempt at proportional reasoning but it is not explicit (example 5). A combination of referring to shape and an outlier was categorized under *Shape*. Reasons about distribution(s) after 'canceling out' the ends, without indication of 'averaging' the ends were also categorized under *Shape* (example 1). Although descriptions of distribution(s) using language such as 'scattered,' 'spread out' or 'compact,' 'close together,' 'tighter' show some indications of attention to range, they were considered primarily as attending to *Shape* (examples 3 and 4). Finally,

whenever mode was referred to in combination with the location of one of the ends of a distribution, that type of response was considered as an informal Shape type of reasoning (example 6).

Table 12.

Examples of *transitional-shape* type responses.

<u>Survey Task</u>	<u>Decision</u>	<u>Reason</u>
Task 1: Yellow/Brown Example 1	Equal	<i>If you cancel out the 7 and 3, the result for Brown is the same as the result for Yellow omitting two samples at the mean. This leads me to believe, that they are the same.</i>
Task 1: Yellow/Brown Example 2	Brown	<i>The bell curve is more appropiate [sic] for representing a classes scoring.</i>
Task 2: Movie Wait-Time Example 3	Disagree	<i>Because the time waited for Royal is very compacted compared to Maximum. Maximum waiting time is spread out somewhat evenly.</i>
Task 2: Movie Wait-Time Example 4	Agree	<i>Both theatres seem to have waiting times that are all over the place, however, chances are, after they are all grouped up, the waiting time seems to be closer to 10 minutes more often than not.</i>
Task 3: Pink/Black Example 5	Black (10% better)	<i>The plot seems to shift more to the right than that of the other class. (I think it is a slight percentage better because they have fewer students, and the graph shifts over only about 2 places.)</i>
Task 5: Ambulance Example 6	Life Line	<i>All Life Line response times are 24 or less, where Speedy has response times over 28 minutes. The most common response time for Speedy is 23 minutes, compared to Life Line's 20 minutes.</i>

Transitional: Center

Transitional-center reasoning is predominantly based on centers. Examples of these types of responses follow in Table 13. Transitional-center type responses most often refer to location of modes (example 3), means (example 2), or medians (example 5) without reference to other characteristics and not explicitly proportional type reasoning. Center type reasons could also refer to average or middles (example 4), particularly when reasons solely focused on explanations of the algorithms for mean or median (example 1).

Table 13.

Examples of *transitional-center* type responses.

<u>Survey Task</u>	<u>Decision</u>	<u>Reason</u>
Task 1: Yellow/Brown Example 1	Equal	<i>just add the scores up and divide by the # of students.</i>
Task 2: Movie Wait-Time Example 2	Agree	<i>There is no difference in the wait times if you figure the mean wait time of both theaters.</i>
Task 3: Pink/Black Example 3	Black	<i>I decided this because the greatest frequency in Black class is 7 while the Pink class seems to be between 5 and 6. So this let me know that the Black class scored better.</i>
Task 4: Pink/Black w/stats Example 4	Black (0.69 points better)	<i>It appears that the average score for the black class is higher. (The difference in the mean.)</i>
Task 5: Ambulance Example 5	Speedy	<i>The median is shown at a lesser time.</i>

Transitional: Variation

Transitional-variation type reasoning is predominantly based on variation without incorporating other aspects of the distribution. Examples of these types of responses follow in Table 14. These types of reasons could specifically refer to ‘range,’ ‘spread,’ specific calculations of the distance between end-points or refer to the values of the end-points, then a qualitative assessment of the distance between (examples 2, 5 and 6). Other variation type responses specifically use the language ‘standard deviation,’ ‘standard error,’ ‘sample variance’ or ‘interquartile range’ and include a relative comparison of those quantities (examples 3 and 4). Explanations that refer to the consistency of the data might be partially related to shape but without more evidence were categorized under variation (example 1). Reasoning that attends to combinations of variation and outliers were categorized as variation.

Table 14.

Examples of *transitional-variation* type responses.

<u>Survey Task</u>	<u>Decision</u>	<u>Reason</u>
Task 1: Yellow/Brown Example 1	Yellow	<i>Yellow had more consistent scoring than brown.</i>
Task 2: Movie Wait-Time Example 2	Disagree	<i>There is a wide spread of data at the maximum theater. There can be a wait time between 5-14 mins. But in the royal theater is between 8-12 mins.</i>
Task 2: Movie Wait-Time Example 3	Disagree	<i>I disagree because the variance at Maximum Theaters is greater.</i>
Task 3: Pink/Black w/stats Example 4	Pink (0.08 better)	<i>The standard deviation is smaller in the pink class than the black class. (The difference between the standard deviation.)</i>
Task 5: Ambulance Example 5	Life Line	<i>There numbers didn't spread out that far as the other service. [sic]</i>
Task 6: Ambulance w/stats Example 6	Life Line	<i>Lowest response time for life line was under six and nothing above 24</i>

Level 3 (Initial-Distributional)

The Initial-distributional category is separated into proportional and initial-global. Responses are based on proportions or population density observations or have global indications but do not include a sophisticated incorporation of multiple aspects of the distributions. Initial-global reasoning is strongly based on centers with a weak or only implied incorporation of shape (or variation) or is strongly based on shape (or variation) with a weak or only implied incorporation of centers. Also, initial-distributional responses could be considered distributional except that they include an

erroneous qualification that it may be inappropriate to make a particular comparison because of the unequal size of the data sets.

Table 15 displays the distribution of Initial-Distributional responses across all the survey tasks. Tasks 1 and 2, Yellow/Brown and Movie Wait-Time, required comparisons of equal sized data sets and those tasks had considerably fewer students that provided Proportional type responses as opposed to Initial Global type responses. Some students may not have explicitly referred to proportions or densities on those tasks because the data sets to be compared have equal size and thus reasoning about sums would yield the same conclusions as reasoning about proportions. That is not the case for tasks 3 and 5, Pink/Black and Ambulance, as they required comparisons of unequal sized data sets. Thus, utilizing proportional arguments to reason about tasks 3 and 5 is more appropriate and may be why students who reasoned about those tasks at the Initial-Distributional level provided a greater percentage of Proportional type responses. However, there was also a clear shift from Proportional type responses to Initial Global type responses after students re-examined the Pink/Black and Ambulance tasks with the inclusion of descriptive statistics. A contributing factor to this trend is the possibility that with the inclusion of descriptive statistics, students may feel compelled to provided explanations for their decisions that reference some of the statistics provided. As none of the statistics included specific proportional measures, this could lead students away from their initial proportional reasoning. A detailed description of the Proportional and Initial Global subcategories follows with examples.

Table 15.

Distribution of *initial-distributional* responses across the survey tasks.

<u>Survey Task</u>	<u>Count (percent) of <i>Initial-Distributional</i> responses</u>		
	<u>Proportional</u>	<u>Initial Global</u>	<u>Total</u>
Task 1: Yellow/Brown	3 (1.1)	32 (11.6)	35 (12.7)
Task 2: Movie Wait-Time	2 (0.7)	28 (10.2)	30 (10.9)
Task 3: Pink/Black	49 (17.8)	8 (2.9)	57 (20.7)
Task 4: Pink/Black w/stats	9 (3.3)	79 (28.7)	88 (32.0)
Task 5: Ambulance	20 (7.2)	3 (1.1)	23 (8.4)
Task 6: Ambulance w/stats	1 (0.4)	40 (14.5)	41 (14.9)

Percentage of total count of participants (n = 275) in parentheses.

Initial-Distributional: Proportional

Proportional type reasoning is primarily based on density observations, such as comparing proportions of data above or below a “cut-point” similar to reasoning described by McClain (2003). Examples follow in Table 16: In examples 1, 3, and 4 the “cut-points” are at 60%, 6 and 20, respectively, while in example 2 there are 2 “cut-points” at 8 and 12. This reasoning is essentially global yet has a singular focus because it focuses on a subset of the entire data (though in relation to the whole).

Table 16.

Examples of initial-distributional: proportional type responses.

<u>Survey Task</u>	<u>Decision</u>	<u>Reason</u>
Task 1: Yellow/Brown Example 1	Brown	<i>A higher percentage of students in the brown class "passed" the recall with at least a 60%.</i>
Task 2: Movie Wait-Time Example 2	Disagree	<i>There is a 50% chance that you will wait for less than 10 minutes, about 25% chance that you will wait for more than 12 minutes or less than 8 at the Maximum Theater. At the Royal Theater there is no chance of waiting for less than 8 minutes or more than 12.</i>
Task 3: Pink/Black Example 3	Black	<i>Two-thirds of the black class scored at least 6 compared to only half of the pink class.</i>
Task 5: Ambulance Example 4	Life Line	<i>Visually taking about half the responses off of the Speedy Ambulance chart, Speedy Ambulance had more response times greater than 20 minutes.</i>

Initial-Distributional: Initial-Global

Table 17 displays some examples of initial-distributional responses with an initial global focus. Initial global type reasoning has some indications of a global perspective but does not include a sophisticated incorporation of multiple aspects of the distributions. Without this new category, these types of responses would have to be coded at the Transitional level or higher, but not at the Local or Idiosyncratic levels. Possible initial-global type reasons can reference two or three measures of center (mean, median, mode) but not refer to any other characteristics (example 2) or describe averaging ends or moving ends to the center or “balancing-out” the ends

(example 1). Other initial-global responses focus on the mean or median with incorporation of the endpoints as related to location, such as comparing medians and indicating that both of the endpoints of one distributions are lower than the endpoints of the other distribution. Similarly these types of reasons could still focus on the mean or median but with an incorporation of endpoints as related to shape or variation (example 4). Finally, some initial-global type responses would be categorized as distributional except for incorrect implications or incorrect use of terminology (example 3).

Table 17.

Examples of initial-distributional: initial-global type responses.

<u>Survey Task</u>	<u>Decision</u>	<u>Reason</u>
Task 1: Yellow/Brown Example 1	Equal	<i>5 is obviously the average for both classes, because I can see that they are both symmetrical around 5. The brown class had a student that did worse and a student that did better, so it balances out.</i>
Task 2: Movie Wait-Time Example 2	Agree	<i>The mean, and/or median gives the true summary of the overall 'picture' especially when they coincide most of the time.</i>
Task 3: Pink/Black w/stats Example 3	Black	<i>I would still say the black class did better based on skew and mode and median but realistically speaking the pink class really did learn more, or rather retained more knowledge.</i>
Task 6: Ambulance w/stats Example 4	Life Line	<i>Life Line had a lower mean response time than Speedy. They [Life Line] had a smaller min/max range in response times. Speedy had over twice as many minutes (sum) as Life Line.</i>

Level 4 (Distributional)

Responses interpreted as *Distributional* are an indication that the student's perception of the data sets, for a particular task, may be global. Table 18 displays some examples of Distributional type responses across all the survey tasks. Responses classified as Distributional are based on a combination and integration of at least two of the aspects of shape, center, variation or proportion, such as an integrated use of average and variation (example 5), center and shape (example 2 and 3) or center and proportion (examples 1 and 4). The possibility of incorporating a proportional aspect with any of the other three aspects was added as a result of the constant comparison process. Merely mentioning more than one aspect is not sufficient, as a Distributional response is similar to *Relational* responses under the SOLO model (see Biggs, 1992; Biggs & Collis, 1982; Watson & Moritz, 1999) in that it should “demonstrate an integrated understanding of the relationships between the different aspects of the domain, so that the whole has a coherent structure and meaning” (Watson & Moritz, 1999, p. 149).

Table 18.

Examples of *Distributional* type responses.

<u>Survey Task</u>	<u>Decision</u>	<u>Reason</u>
Task 1: Yellow/Brown Example 1	Yellow	<i>Assuming I calculated correctly (in my head) they have the same mean, median and mode, but the variance of the yellow class is smaller. In the yellow class we have a greater percentage of the class knowing at least 50% of the material than in the brown class.</i>
Task 2: Movie Wait-Time Example 2	Agree	<i>The reason I agree because the data sets are evenly distributed with 10 data points on each side of the value 10. So this kind of let me know that the average of the data should be 10.</i>
Task 3: Pink/Black Example 3	Black	<i>Black Class seems to have higher mean score and skew to the right</i>
Task 5: Ambulance Example 4	Life Line	<i>75% of calls are answered in 20 minutes or less versus 23 minutes or less for Speedy, and the average wait time is shorter for Life Line.</i>
Task 6: Ambulance w/stats Example 5	Life Line	<i>In the case of Life Line, the data is more compressed about a lower mean, the overall data range is lower, the minimum is lower the max is lower, and based off their standard deviation they are more consistent.</i>

Table 19 displays the distribution of Distributional responses across all the survey tasks. The Movie Wait-Time task had the highest percentage of Distributional responses possibly because both data sets were relatively small, equal in size, with similarities and differences that are easily articulated. There were modest increases in Distributional responses from the tasks without descriptive statistics, tasks 3 and 5, to

the tasks with descriptive statistics, tasks 4 and 6. It is possible that for those students who understood the terminology, having the various measures to reference helped to formulate more sophisticated responses.

Table 19.

Distribution of *Distributional* responses across the survey tasks.

<u>Survey Task</u>	Count (percent) of <u><i>Distributional</i> responses</u>
Task 1: Yellow/Brown	22 (8.0)
Task 2: Movie Wait-Time	57 (20.7)
Task 3: Pink/Black	16 (5.8)
Task 4: Pink/Black w/stats	26 (9.5)
Task 5: Ambulance	8 (2.9)
Task 6: Ambulance w/stats	41 (14.9)
Percentage of total count of participants (n = 275) in parentheses.	

Reliability Assessment

After the survey responses were coded and the framework refined through constant comparison, four volunteers were briefly trained in coding the survey responses by using the expanded and refined Lattice Structure Framework. Each coder was provided with a description of the framework along with examples similar to the previous description and examples given in this section. Then each coder was given the survey responses from three participants (the same three participants were given to each coder). The responses from the three participants were used for trial coding, that is, first, each coder practiced assigning codes to the responses from those participants, second, the researcher and coder reviewed the coder's decisions and third, together

they addressed interpretation issues and questions that the coder may have had. The responses used for the trial coding were purposefully selected because, all together, they reflected a diverse variety of levels in the framework, some of which were anticipated to be easily coded and others were anticipated to be difficult to code.

Coders #1 and #2 had previous experience using the Lattice Structure framework (see Figure 15) that was developed by Shaughnessy, Ciancetta, Best and Noll (2005). Coder #3 had prior experience using the initial framework (see figure 14) developed by Shaughnessy, Ciancetta, and Canada (2004) and coder #4 had neither prior experience with coding responses of this nature nor experience working with a similar type of framework. The trial coding responses were distributed to each of the coders approximately one day after each coder was provided with a description of the framework along with examples similar to the previous description and examples given in this section. When coders #1, #2, and #4 indicated that they had completed the trial coding, each, separately, conferred with the researcher for about an hour to address any interpretation issues that arose while each assigned codes. Coder #3 was only available through email communication, so any interpretation issues that coder #3 had were addressed via email.

After the trial coding was completed, approximately 20% of the surveys were chosen at random and distributed among the four volunteers, so each volunteer coded only 14 surveys (about 5%). The volunteers' codes and the researcher's codes were then checked for agreement and the inter-rater reliabilities for coding each task was found to generally be between 60% and 80% as shown in table 20.

Table 20.

Inter-rater reliabilities for coding the survey tasks.

Volunteer coders	<u>Task 1</u>	<u>Task 2</u>	<u>Task 3</u>	<u>Task 4</u>	<u>Task 5</u>	<u>Task 6</u>
Coder #1	79%	79%	71%	79%	79%	86%
Coder #2	93%	79%	71%	79%	93%	79%
Coder #3	79%	64%	79%	36%	50%	43%
Coder #4 (novice)	64%	57%	71%	86%	85%	36%
Total	79%	70%	73%	70%	77%	61%

Percentages rounded to the nearest whole number.

As displayed in table 20, coders #1 and #2 have consistency rates of 0.7 or greater with the researcher's code assignments for each task. Generally, 0.7 is considered an acceptable inter-rater reliability (Stemler, 2004). However, coders #3 and #4 did not meet this standard for each task. Coder #3 did have some prior experience in coding procedures using a similar framework, but had the least amount of training in using the framework associated with this research and coder #4 had a similar amount of training as coders #1 and #2, but had no prior experience in assessing responses to tasks such as the ones used in this research. Even though each coder examined only about 5% of the surveys, the success of coders #1 and #2 indicate the potential for reproducibility of code assignments to similar tasks using the Expanded Lattice Structure Framework as refined from this research.

Cross Task Numeric Codes

After all the coding was completed for each individual task, a *Cross Task Numeric Code* (Lattice Framework Level 0 – 4) was assigned to each student for the dominant type of reasoning that they exhibited across all the tasks. Several studies, such as Gal et al (1989) and Watson (2001, 2002), noted that frequently students did not consistently use the same type of strategy for comparing data sets across a series of tasks. While that phenomenon was true for this study as well, it was also true that some students' responses were categorized at consistent levels of the framework across the tasks. The researcher of this study surmised that if a cross task code could be assigned, this code would provide insight into the students' perspectives of distributions.

A *Cross Task Numeric Code* for each student was determined by examining the coded survey responses. As with the codes for the individual responses, the cross task code is assigned conservatively and is thus a possible lower bound of each student's reasoning level across the survey tasks. The process for assigning cross task numeric codes is displayed in the flow charts in figures 28, 29 and 30. A description of the flow charts follows.

For each student, if the code for each task was at the same, consistent level in the framework, then that framework level was assigned as the student's Cross Task Numeric Code. In the event that a student provided responses that were not coded at a consistent framework level across the tasks, then each task was not given the same weight for determining the Cross Task Numeric Code.

The procedure for determining the Cross Task Numeric Codes for the survey responses that were coded at inconsistent framework levels across the tasks was separated into two stages. Stage 1 began with assigning an initial code aligned with the codes from the responses from the Pink/Black and Ambulance tasks (tasks 3 and 5, without statistics). In stage 2 the initial code is adjusted based on the responses to the remaining tasks. Each of the Pink/Black and Ambulance tasks required comparisons of distributions of different sizes, thus local types of reasoning strategies were generally not reasonable. So those tasks were given the most “weight” as they were used to assign initial codes. The initial codes were adjusted at most one level, up or down, depending on responses to the remaining tasks. In rare cases, the initial code was adjusted two levels. The Yellow/Brown Task and the Movie Wait-Time task both required comparisons of distributions of equal size, thus some Local level types of reasoning strategies and Transitional strategies were reasonable and sufficient to make the comparison. Responses for those tasks at levels 1 and 2, respectively, generally did not warrant adjusting the initial code down.

Key: Task 1 = Yellow/Brown; Task 2 = Movie Wait-Time; Task 3 = Pink/Black; Task 4 = Pink/Black with statistics; Task 5 = Ambulance; Task 6 = Ambulance with statistics

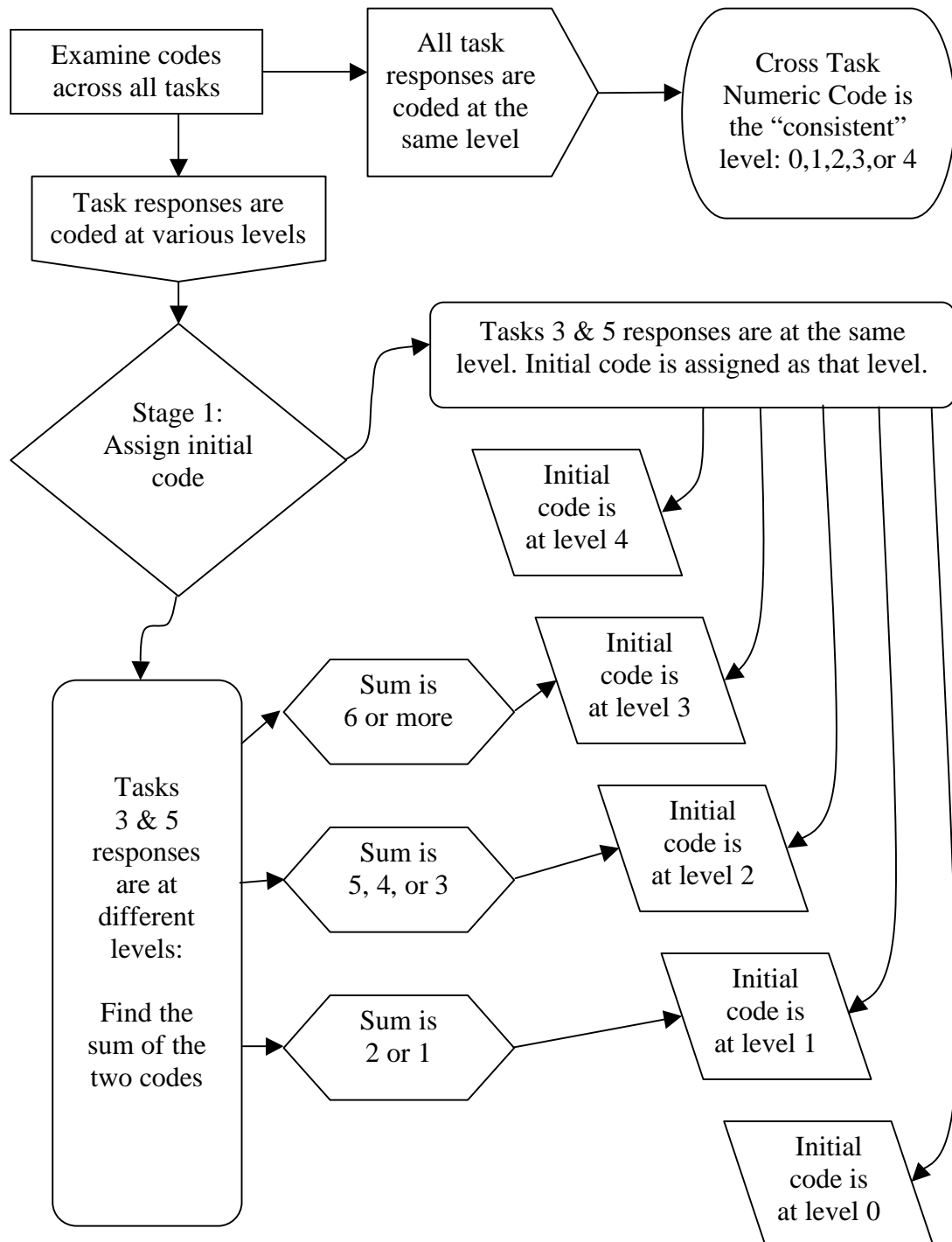


Figure 28. Flow chart #1 for assigning Cross Task Numeric Codes

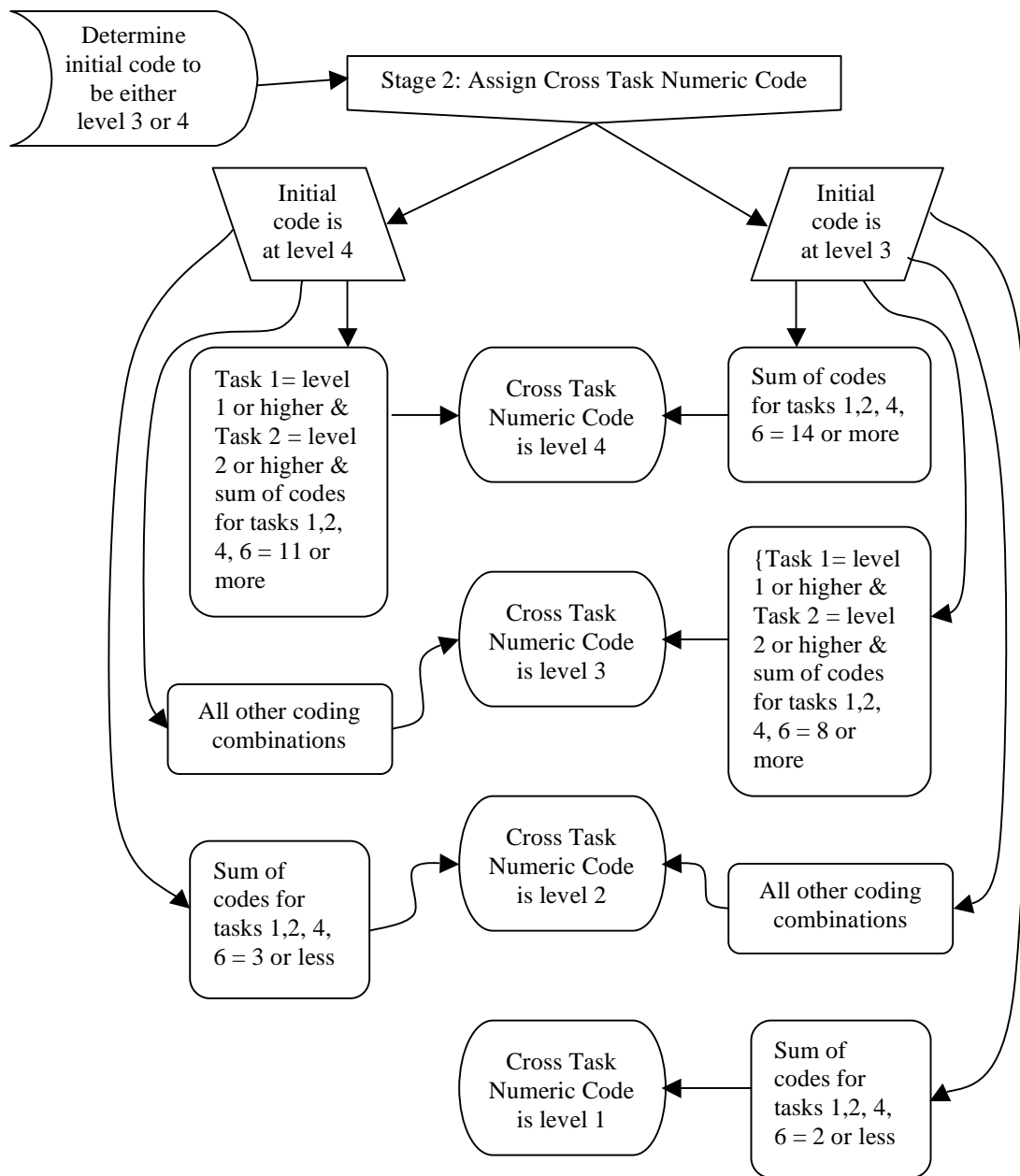


Figure 29. Flow chart #2 for assigning Cross Task Numeric Codes



In stage 1 (see Figure 28), if both codes were at the same level, then that level was assigned as the student's initial code. If both codes were at different levels, then: If the sum of the codes for tasks 3 and 5 equaled six or more, then the initial code was level 3 (an average of 3 each); if the sum of the codes for tasks 3 and 5 equaled five, four or three, then the initial code was level 2 (an average of 2 ± 0.5); if the sum of the codes for tasks 3 and 5 equaled two or one, then the initial code was level 1.

In stage 2 (see Figure 29), if the initial code was level 4, then: If the code for task 1 was level 1 or higher, the code for task 2 was level 2 or higher and the sum of codes for tasks 1, 2, 4 and 6 equaled 11 or more, then a cross task numeric code of 4 was assigned; if the sum of codes for tasks 1, 2, 4 and 6 equaled 3 or less, then a cross task numeric code of 2 was assigned; for any other combinations of codes, a cross task numeric code of level 3 was assigned.

In stage 2 (see Figure 29), if the initial code was level 3 then: If the sum of codes for tasks 1, 2, 4 and 6 equaled 14 or more, then a cross task numeric code of 4 was assigned; if the code for task 1 was level 1 or higher, the code for task 2 was level 2 or higher and the sum of codes for tasks 1, 2, 4 and 6 equaled 8 or more, then a cross task numeric code of 3 was assigned; If the sum of codes for tasks 1, 2, 4 and 6 equaled 6 or less, then a cross task numeric code of 1 was assigned; for any other combinations of codes a cross task numeric code of level 2 was assigned.

In stage 2 (see Figure 30), if the initial code was level 2, then: If the sum of codes for tasks 1, 2, 4 and 6 equaled 15 or more, then a cross task numeric code of 4 was assigned; if the codes for tasks 1 and 2 were each level 3 or higher with tasks 4

and 6 coded at any level, then a cross task code of level 3 was assigned; if tasks 1, 2, 4, and 6 were all coded at level 1 or lower, then a cross task numeric code of 1 was assigned; for any other combinations of codes a cross task numeric code of level 2 was assigned.

In stage 2 (see Figure 30), if the initial code was level 1, then: If the codes for tasks 1, 2, 4, and 6 were all at level 2 or higher, then a cross task numeric code of level 2 was assigned; if the codes for the tasks 1 and 2 were also at level 1 then, irrespective of the code assignments for tasks 4 and 6, a cross task numeric code of level 1 was assigned; for any other combinations of codes it was left to the coder to make a judgment to decide which cross task numeric code best aligned with the responses.

Finally, in stage 2 (see Figure 30), if the initial code was level 0, then: If any two codes for tasks 1, 2, 4, and 6 were idiosyncratic, then a cross task numeric code of level 0 was assigned; if less than four responses were idiosyncratic then, it was left to the coder to make a judgment to decide which cross task numeric code best aligned with the responses.

Table 21 displays the distribution of cross task numeric codes for the survey responses. The lowest percentages of students were coded at cross task levels of 4 and 0. This is not surprising as students needed to consistently provide responses at those levels for those cross task numeric codes to be assigned. The cross task code of level 1 was also rather limited at about 9.5% of participants. This is also not surprising, as the relative consistency of level 1 codes was required for the assignment of the cross task code of level 1. The cross task code of level 2 was assigned to the majority of

participants, followed by level 3. This may be due to the possibility of assigning a cross task code of level 2 or 3 to responses consistent at either level as well as responses that vary above and below each level.

Table 21.

Distribution of Cross Task Numeric Lattice Codes

<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>
6(2.2)	26(9.5)	182(66.2)	56(20.4)	5(1.8)	275(100)

Quantities in parentheses represent percent of total participants (n = 275).

Framework Refinement Summary

As part of the analysis of survey responses, the Expanded Lattice Structure Framework was refined. The levels of this framework are intended for use in describing the reasoning expected of university students as they engage in tasks that require comparisons of distributions of data. The levels were primarily adapted and refined from the framework by Shaughnessy and colleagues (2005) with additional influence from the frameworks developed by Bakker and Gravemeijer (2004) and Watson and Moritz (1999), among others. The previous sections describe and illustrate responses to the data sets comparison tasks that are classified at each level. After the refinement was completed, all the surveys were recoded using the final version of the Expanded Lattice Structure Framework. The initial results show that, for each of the survey tasks, students can potentially reason at any of the framework levels: Level 0 (Idiosyncratic); Level 1 (Local); Level 2 (Transitional); Level 3 (Initial Distributional); and Level 4 (Distributional). Initial steps towards verification of reliability were taken with encouraging results. In the following sections survey

responses will be summarized and analyzed in relation to groups of students with similar statistical backgrounds. Then interviews from selected students will be analyzed and related to their survey responses, along with a discussion of implications about the validity Expanded Lattice Structure Framework when it is used for describing students' reasoning about comparisons of distributions of data.

Survey Results by Group

Two hundred seventy five undergraduate and graduate students who were enrolled in statistics courses completed the on-line Data Comparison Survey. The 275 participants' survey responses were divided into five discrete groups for analysis: 1-GS, 1-SE, 2-GS, 2-SE, and GRAD. Group 1-GS contained responses from undergraduate or post-baccalaureate students who were beginning their first general statistics course. Group 1-SE contained the responses from undergraduate or post-baccalaureate students who were beginning their first statistics course, but the course was specifically designed for engineering and science majors. Group 2-GS contained the responses from undergraduate or post-baccalaureate students who were beginning their second general statistics course. Group 2-SE contained the responses from undergraduate, post-baccalaureate or graduate students who were science or engineering majors, had completed at least one college level statistics course, and were at least enrolled in their second statistics course. Group GRAD contained the responses from senior undergraduate, post-baccalaureate or graduate students who had completed and were also enrolled in advanced level undergraduate statistics courses

and/or graduate level statistics courses. Table 22 displays the number of students in each group.

Table 22

Statistics backgrounds of the participants

<u>Group</u>	<u>Description of students' statistics background</u>	<u>Number of participants</u>
1-GS	Beginning 1 st general statistics course	137
1-SE	Beginning 1 st statistics for engineers course	37
2-GS	Beginning 2 nd general statistics course	74
2-SE	Beginning 2 nd statistics for engineers course or more	15
GRAD*	Many senior and/or graduate level statistics courses	12

*Of the 12 students in the GRAD group, 9 reported their major as Statistics, 2 reported their major as Mathematics, and 1 reported his/her major as Economics.

All of the students who completed the survey were volunteers. All of the students from groups 1-GS, 1-SE, and 2-GS received extra credit in their statistics class for completing the survey. In an attempt to assess how representative of their statistics classes the volunteers are, one sample t-tests were performed to compare the mean grade of the volunteers in each group to the fixed mean grade of all the students enrolled in those statistics classes. These tests revealed that the mean grade of the 1-GS group was significantly higher ($t = 3.87, p < 0.01$) than the mean grade of all the students from the statistics classes the 1-GS students were enrolled in. But the mean grade of the 1-SE group was not significantly different ($t = 0.79, p > 0.05$) than the mean grade of all the students from the statistics classes the 1-SE students were enrolled in, and similarly the mean grade of the 2-GS group was not significantly

different ($t = 0.69, p > 0.05$) than the mean grade of all the students from the statistics classes the 2-GS students were enrolled in. Thus, while the 1-SE and 2-GS groups can be considered to be representative of the students in their statistics classes, the 1-GS students were not necessarily representative, as they tended to earn higher grades in their statistics classes than their classmates.

Survey Results: Task 1, the Yellow/Brown task

The data sets in the Yellow/Brown task have equal centers (mean, median, mode) and the equality of their centers is assumed to be visually evident, so there is no mention of that fact. These data sets have similar uni-modal shapes but differ in their variation (see Appendix B). Participants' responses on this task were categorized at each level of the framework. Across all the groups the most common decision was that the classes scored equally well, followed by the decision that the yellow class scored better. For each of the groups, except GRAD, the most common reasons for the 'equal' decision were coded at level 2 and most of those reasons focused on a comparison of center, i.e., the means (or medians or modes) are equal. This was not a surprising result as determining that the centers are equal is an easy visual assessment. For the groups that were in general statistics classes, 1-GS and 2-GS, the most common reasons for supporting either a 'yellow' or 'brown' decision were coded at level 1. Many of these reasons compared either the heights of the columns at the center or the 'number of dots' (or sums of scores) at the score of 5 and above or at the score of 6 and above. For the groups that were comprised of students in statistics for engineers courses, 1-SE and 2-SE, the most common reasons for supporting either a

‘yellow’ or ‘brown’ decision were coded at level 2 with either a focus on variation or shape. Most of the reasons that the GRAD group cited were coded at Level 4, irrespective of their decisions. Students from the GRAD group provided reasons that tended to incorporate either center and spread or center and shape. A detailed breakdown and discussion of the results for each group can be found in appendix C.

For each of the groups the most frequent decision was that the classes scored equally well, and when combined, 53.45% of all participants chose ‘equal.’ The decision that the Yellow class scored better was consistently the second most frequent decision for each group, and when combined, 36.73% of all participants chose ‘Yellow.’ Hence, the decision that the Brown class scored better was consistently the least frequent decision for each group, and when combined, only 9.82% of all participants chose ‘Brown.’ On the whole, students understood the Yellow/Brown task as responses categorized as idiosyncratic were minimal at about 9%, with most coming from the large 1-GS group. Generally, the idiosyncratic responses for the Yellow/Brown survey question provided no information on student thinking and had few instances of mis-interpretation of the task or associated graph.

Students from groups 1-GS, 1-SE, 2-GS and 2-SE supported the ‘equal’ decision, by using a clear majority of Transitional responses, focused exclusively on comparing a measure of center. This was not a surprising result as determining that the centers are equal is an easy visual assessment because the data sets’ obvious ‘bell shape,’ equal centers, small size and equal size. Some of these Transitional responses specifically described the process of calculating the mean or median and thus appear

to be considering the data as an amalgam of individual points. Other responses merely stated that they compared either the mean or median and thus could have taken into account a global “picture” of each data set and then used the respective centers as group representatives to be compared.

The majority of Local responses, in support of ‘equal,’ came from the 1-GS students, with a few from students in groups 1-SE and 2-GS and none from the 2-SE and GRAD groups. The Local type reasons often cited a comparison of sums, a potentially reasonable type strategy because of the equal size of the data sets. However, none of the responses that compared sums included any extra information, such as citing that the data sets have equal size, and thus it is likely that the students providing these responses were not thinking proportionally, but were thinking additively.

Students from the GRAD group were the only ones who supported the ‘equal’ decision with more Distributional reasons than Initial Distributional reasons. Also, when those level 3 and 4 responses, in favor of ‘equal’ are combined, the GRAD group is the only one to have those responses out-number the Transitional responses. None of the level 3 responses, from any group that supported the “equal” decision, focused on comparing proportions or densities. All of the Initial Distributional responses in favor of ‘equal’ were the Initial Global type. Many of those responses alluded to all three measures of center (mean, median, mode) coinciding or described averaging ends or “balancing-out” the ends along with citing the equal means. While these responses encompassed utilizing more than one feature of the distributions, it is

not clear that the students who provided these responses were viewing the data sets as whole entities. Most of the level 4 responses integrated a comparison of means (and/or medians) with either variation or shape. Many cited a comparison of means and an assessment that both data sets had an equal distribution of scores on each side of its mean and thus it is likely that the students providing these responses were considering their comparison from a global perspective where the data sets are whole units.

The two groups of students who were enrolled in general statistics courses, 1-GS and 2-GS, provided the majority of level 1 responses in support of either ‘Yellow’ or ‘Brown’ with only a few from the 1-SE and 2-SE groups and none from the GRAD group. Many of the level 1 reasons in favor of ‘Yellow’ tended to focus either on the higher number of scores at 5 and above, that is, comparing the sums of scores at 5 and higher or on the feature that the column height for 5 was higher for the Yellow class than for the Brown class, that is, deciding that the Yellow class scored better because it had a “taller” mode. As with those students who compared total sums to decide ‘equal,’ the students who compared partial sums to decide ‘Yellow’ made no mention of the appropriateness of this strategy because of the equal size of the data sets. Thus, it is likely that the students providing these responses were not thinking proportionally but were thinking additively, similar to the additive reasoning described by Cobb (1999).

Of the students who decided in favor of ‘Yellow’ and provided responses at or above level 2, most included a ‘consistency’ assessment of scores as part of their determination of which class “scored better.” This type of assessment was used

exclusively in level 2 responses and was often a part of the level 3 and 4 responses. Only three students out of all the participants provided Initial Distributional responses focused on proportion, two supported the ‘Yellow’ decision and one supported the ‘Brown’ decision. The one student who provided a proportional response supporting the Brown class wrote, “A higher percentage of students in the brown class "passed" the recall with at least a 60%.” This response is considered by the researcher as a prototypical level 3, proportional, type response. Just a few students from groups 1-GS, 1-SE, or 2-GS provided level 4, distributional type responses. No students from group 2-SE gave distributional type responses while most of the group GRAD students, seven of 12, provided distributional responses for the Yellow/Brown task.

Table 23 and Figure 31 display the results for the Yellow/Brown survey task across all groups for each level of the Expanded Lattice Structure Framework. In contrast to the 1-GS students whose responses were concentrated as levels 1 and 2, the responses provided by the 1-SE students are concentrated at level 2 with about the same amount of responses at higher and lower levels. When the mean response level of the 1-GS group at 1.47 is compared to the mean response level of the 1-SE group at 2.13, a 2-sample t-test shows that the mean for the 1-GS group is significantly lower than the mean for the 1-SE group ($t = -3.71, p < 0.01$). While both 1-GS and 1-SE students were enrolled in their first statistics class, the 1-SE students were engineering majors and were either enrolled in or had completed a first course in calculus; while 1-GS students may have only had the equivalent of high school algebra as their most

recent mathematics course. Thus the difference in mathematics backgrounds could have contributed to the difference in mean response levels.

Table 23.

The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for all groups.

<u>Group</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>	<u>Mean</u>
1-GS	17 (12.4)	55 (40.1)	52 (38.0)	9 (6.6)	4 (2.9)	137 (100)	1.47
1-SE	1 (2.7)	7 (18.9)	20 (54.1)	4 (10.8)	5 (13.5)	37 (100)	2.13
2-GS	5 (6.8)	18 (24.3)	29 (39.2)	16 (21.6)	6 (8.1)	74 (100)	2.00
2-SE	2 (13.3)	1 (6.7)	9 (60.0)	3 (20.0)	0 (0.0)	15 (100)	1.87
GRAD	0 (0.0)	0 (0.0)	2 (16.7)	3 (25.0)	7 (58.3)	12 (100)	3.42
Total	25 (9.1)	81 (29.5)	112 (40.7)	35 (12.7)	22 (8.0)	275 (100)	1.81

Percent of group responses in parentheses.

From Figure 31, it also appears that the 2-GS students' mean response level could be significantly higher than that of the 1-GS students'. A 2-sample t-test reveals that the mean of 1.47, from the 1-GS group, is significantly lower than the mean of 2.00, from the 2-GS group ($t = -3.68$, $p < 0.01$). Thus, it appears that the one semester of general statistics that the 2-GS students completed did contribute to increasing their response level on the Yellow/Brown task, away from Local type reasons and toward Transitional type reasons.

Although a statistical test using the 2-SE group is inappropriate because it is too small, it does appear that the 1-SE students and the 2-SE students responded similarly, yet the mean response level of the 1-SE students was 2.13 while the mean response level for the 2-SE students was lower, at 1.87. Both groups did have a majority of responses at level 2, but the 2-SE students provided more responses at

level 3 than level 1, while the 1-SE students gave more responses at level 1 than level 3. As the 2-SE students have completed a statistics course and the 1-SE students have not, it is surprising to the researcher that the 2-SE responses are not noticeably at higher levels than the 1-SE students.

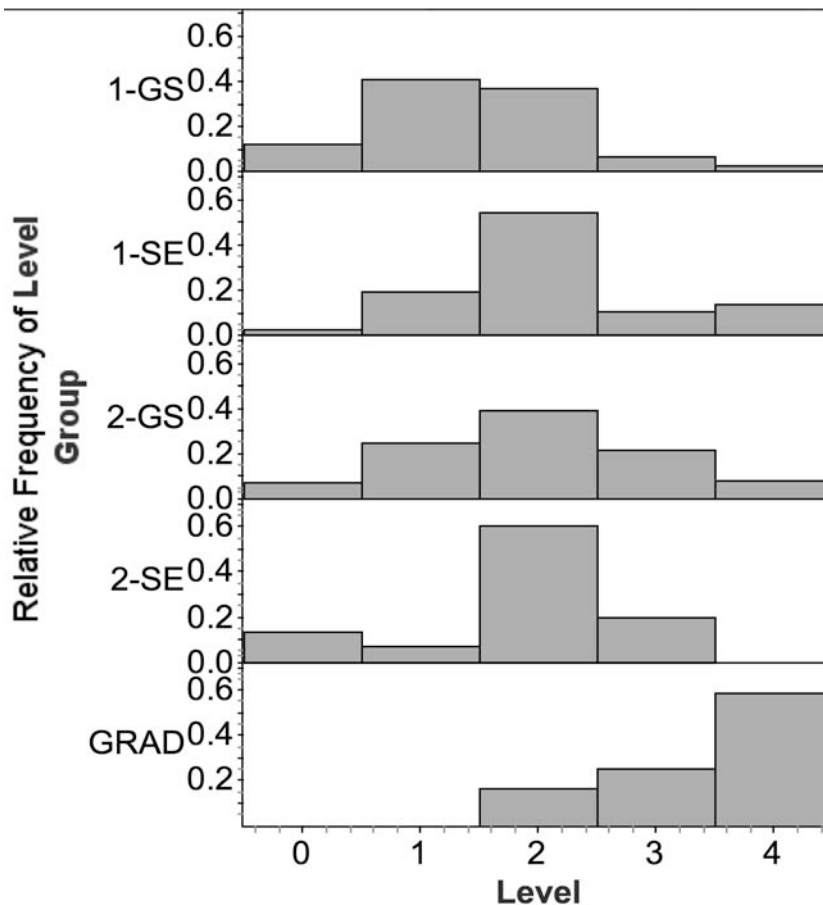


Figure 31. The distribution of response levels across all groups for the Yellow/Brown task.

The students in the GRAD group provided responses at noticeably higher levels, on the whole, than any other group. Although a statistical test to confirm this is inappropriate, the GRAD students did give the highest (group) percentages of level 4

and level 3 responses with the lowest (group) percentage of level 2 responses and no level 1 or 0 responses. This trend is not surprising as the GRAD students had the most sophisticated mathematics and statistics backgrounds.

Very few students reasoned with proportions on the Yellow/Brown task. This was particularly evident with the 1-GS group, as they provided the lowest percentage of responses at levels 3 and 4 with the highest percentage of responses at level 1, across all groups. Several factors may have combined to produce this result, such as the relatively small size of the data sets, the equal size of the data sets, the equal centers, and the similar shapes. The statistically naïve background of the 1-GS students could also contribute to those students finding non-proportional type arguments more accessible and thus more convincing. The 1-GS students also gave a high percentage of idiosyncratic responses for task 1, many of which simply provided no reasonable information to base a code on, such as explanations similar to “just by looking at the numbers” were common for 1-GS students and coded at level 0.

Finally, while all groups consistently had the highest frequency of decisions in favor of ‘equal’ and the second highest frequency of decisions in favor of ‘Yellow,’ the 1-GS students tended to approach the task from perspectives aligned with lower framework levels, Local and Transitional, while the GRAD students tended to approach the task from perspectives aligned with higher framework levels, Initial Distributional and Distributional, and the students from the remaining groups tended to approach the task from a perspective aligned with the Transitional framework level.

Survey Results: Task 2, the Movie Wait-Time task

For the Movie Wait-Time task students were given the information that the means and medians are equal. These facts, along with the bi-modal shapes and different ranges of the graphs, were intended to promote reasoning about the variation in the comparison (see appendix B). Although it was not explicitly stated, the data sets were of equal size as the same group of kids collected wait-times for each theater. Participants' responses on this task were categorized at each level of the framework. Across all the groups the most common decision was to 'Disagree' thus implying that the data sets are different, as opposed to 'Agree' implying that the data sets are the same.

Irrespective of their decision, across all groups, students chose to go to the Royal Theater over the Maximum Theater at about a 2 to 1 rate. Most of the students who chose the Royal Theater made mention that the higher consistency or predictability of the wait-times at Royal was more desirable. Many of those who chose the Maximum theater indicated that they were willing to "take a chance" on waiting a very short time.

For each of the groups, except 1-SE, the most common reasons for the 'Disagree' decision were coded at level 2 and most of those reasons focused on a comparison of either variation or shape. This was not a surprising result as determining that the range of the data sets are not similar is an easy visual assessment. For each of the groups, except GRAD, the most common reasons for the 'Agree' decision were coded at level 2 and most of those reasons focused on a comparison of

centers, i.e., noting that that data sets had either the same mean or median. This was not a surprising result as the information about the equal centers was given in the problem statement. There were relatively few Idiosyncratic responses across all the groups, implying that the task was understood by most students. As with the Yellow/Brown task, most Idiosyncratic responses simply contained no information about how the student made his or her decision. However, a few students gave completely contextual responses to the Movie Wait-Time task that were also coded at level 0. In these responses, students based their decision on their own experiences in attending movies and disregarded the data. A detailed breakdown and discussion of the results for each group can be found in appendix C.

Across the groups, consistently, about 70-80% of respondents disagreed with Eddy, that is, they decided that there was a difference in wait-times between the two theaters, while about 20-30% agreed with Eddy, that is, they decided that there was no difference in wait-times between the two theaters. Irrespective of their decision, students chose to go to the Royal Theater over the Maximum Theater at about a 2 to 1 rate across the groups. Overwhelmingly, the students who chose to go to the Royal Theater did so because of its “consistency” or “predictability” in wait-times. Most of those who chose to go to the Maximum Theater did so to “take a chance” on getting a short wait-time, while a few chose Maximum because they enjoyed the commercials and wanted the potential for a long wait-time so that they might see many commercials.

More than half of the responses to the Movie Wait-Time task, from all participants, were categorized at level 2. The trend for these responses was that most supported 'Disagree' based on the difference in variation while those that supported 'Agree' generally repeated Eddy's claim, that the means were equal. That those two strategies were often used is not surprising. To argue that the wait-times are different, it seems reasonable to reference the most obvious difference between features of the data sets, that is, the difference in spread. If one agrees with Eddy, that the wait-times are the same because the means are about the same, one may not feel a need to provide an even more detailed argument.

In particular, the 2-GS students gave a rather large number of level 2 – 'Disagree' responses that focused on center. After a closer inspection of these responses, it was found that there was no common reasoning trend in support of the decision. For example, some students recalculated the means and/or medians and made errors doing so, some misinterpreted the meaning of the median, and some correctly recalculated the means and found that there was indeed a two second difference and decided that the wait-times were different because of that small difference.

Similar to all the other groups, the students from the GRAD group who provided Transitional responses also largely favored 'Disagree.' All of the reasons that the GRAD students provided for the 'Disagree' decision cited the difference in variation (see table 95 in appendix C) and most of those responses specifically referred to "standard deviation" or "variance." There was one student who did write that the

difference in variance is “sufficient” to determine that the data sets are different. It is quite possible that this student was considering the data sets from a global perspective, yet his or her given response was coded at level 2. This situation highlights the nature of the coding process for this research, that is, codes were assigned conservatively and thus may represent a lower bound for the students’ reasoning.

As with the high percentage of Transitional responses, the relatively high percentage of Distributional responses could, in part, be due to the task description. Whether students ‘Agree’ or ‘Disagree,’ at level 4 they tended to begin their explanation by addressing Eddy’s claim about the equal means and then expanded their analysis to include assessments of dispersion or shape.

Overall, there were about half as many Level 3 responses as Level 4 responses. Only two of the level 3 responses focused on comparing proportions while the rest had an Initial Global focus. Whether the students who provided the Initial Global responses agreed or disagreed, they generally cited the equal means, then those who agreed added that the medians coincided with the means; those who disagreed often added a comparison of the ends, that is each set had different shortest wait-times and/or different longest wait-times. These students seem to be attempting to incorporate a comparison of more than one feature of the distributions, but it is also not clear that they were comparing the distributions as whole entities, i.e., from a global perspective.

The trend of considerably more Initial Global than Proportional responses was also seen in the Yellow/Brown Task. Similarities between the two tasks are that the

data sets to be compared are both fairly small in size, both have means and medians that coincide, and both are of equal size. Further investigation using tasks that require comparisons when each of those conditions is changed separately may reveal which has a greater impact on promoting the use proportional strategies.

The fewest responses were at level 1 (except for the Idiosyncratic responses) and they largely were in favor of ‘Disagree.’ This does not seem surprising as most only compared a single wait-time, such as the shortest or longest or the frequency of times at 10 minutes, then concluded that the times (or frequencies) were different thus the wait-times were different for the theaters. The comparison of single data points or of frequencies of a specific wait-time indicate that these students appeared to consider these data sets as collections of individual times not as whole units.

Table 24.

The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for all groups.

<u>Group</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>	<u>Mean</u>
1-GS	12 (8.8)	13 (9.5)	77 (56.2)	18 (13.1)	17 (12.4)	137 (100)	2.11
1-SE	2 (5.4)	4 (10.8)	12 (32.4)	5 (13.5)	14 (37.8)	37 (100)	2.67
2-GS	2 (2.7)	5 (6.8)	42 (56.8)	6 (8.1)	19 (25.7)	74 (100)	2.47
2-SE	2 (13.3)	2 (13.3)	7 (46.7)	1 (6.7)	3 (20.0)	15 (100)	2.07
GRAD	0 (0.0)	0 (0.0)	7 (58.3)	1 (8.3)	4 (33.3)	12 (100)	2.75
Total	18 (6.6)	24 (8.7)	145 (52.7)	31 (11.3)	57 (20.7)	275 (100)	2.31

Percent of group in parentheses

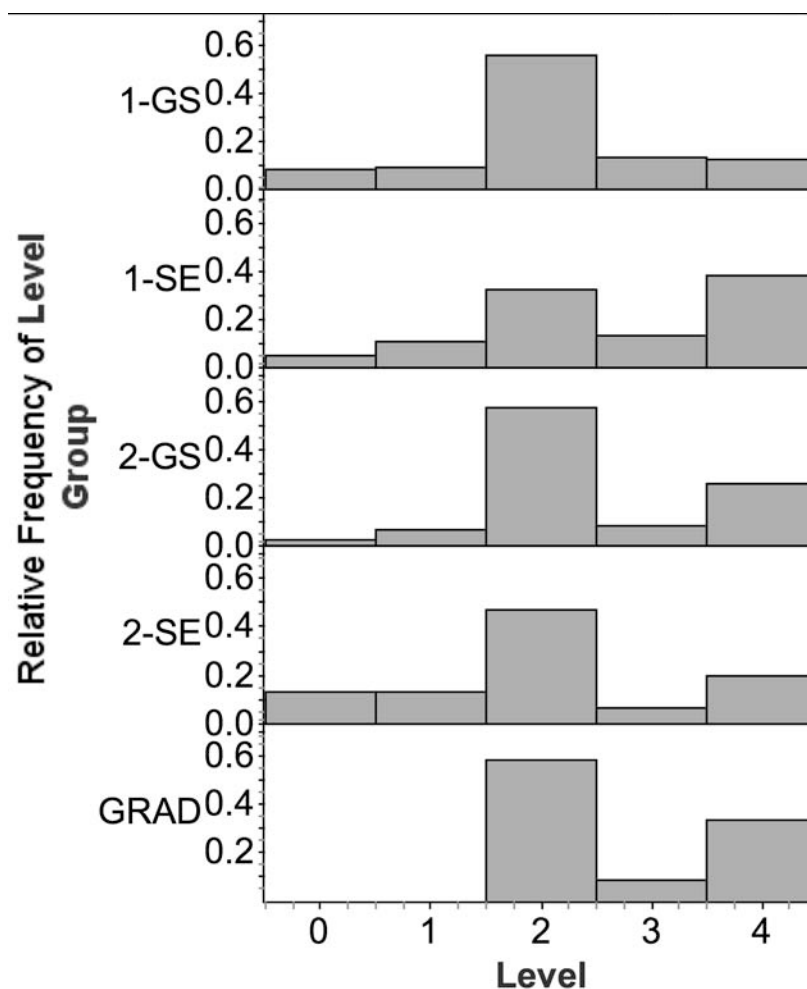


Figure 32. The distribution of response levels across all groups for the Movie Wait-Time task.

Table 24 and figure 32 display the results for the Yellow/Brown survey task across all groups, for each level of the Expanded Lattice Structure Framework. Comparing the results of the response categorization across groups reveals that the responses provided by the 1-GS students were spread across all levels with a large concentration at level 2; groups 1-SE, 2-GS, and 2-SE also provided responses categorized across all levels, their responses where primarily at levels 2 and 4; and the

students from the GRAD group appeared to provide the highest level of responses, overall, as none were at levels 0 and 1 with most at levels 2 and 4.

When the mean response level of the 1-GS group at 2.11 is compared to the mean response level of the 1-SE group at 2.68, a 2-sample t-test shows that the mean for the 1-GS group is significantly lower than the mean for the 1-SE group ($t = -2.53$, $p < 0.01$). At the time when the survey was taken, the 1-SE students generally had taken more mathematics than the 1-GS students, which could contribute to their higher mean response level.

From Figure 32, it also appears that the 2-GS students' mean response level could be significantly higher than that of the 1-GS students'. A 2-sample t-test reveals that the mean of 2.11, from the 1-GS group, is significantly lower than the mean of 2.47, from the 2-GS group ($t = -2.43$, $p < 0.01$). Thus, it appears that the one semester of general statistics that the 2-GS students completed did contribute to increasing their mean response level, on the Movie Wait-Time task.

Although a statistical test using the 2-SE group is inappropriate, it does appear that the 1-SE students and the 2-SE students responded similarly. Both have a majority of responses at levels 2 and 4 but the 2-SE students provided higher percentage of Transitional responses. The mean response level for the 1-SE students was 2.68 while the mean response level for the 2-SE students was 2.07. As the 2-SE students have completed at least one statistics course prior to their current statistics course for engineers, it is surprising to the researcher that the 2-SE responses are not prominently at higher levels than the 1-SE students.

Although the students in the GRAD group provided a majority of responses at the Transitional level, on the whole their responses appeared higher than any other group as they had no level 0 or level 1 responses and the second highest percentage of level 4 responses. This trend is somewhat surprising as the GRAD students had the most sophisticated mathematics and statistics backgrounds, yet it could be the case that many of these students considered a level 2 response as sufficient evidence to support their decision, particularly for those who decided that the wait-times were indeed different.

Finally, while all groups consistently had the highest frequency of decisions in favor of ‘disagree,’ implying that they considered the wait-times different at each theater, the 1-GS students tended to approach the task from perspectives aligned with Transitional framework level, while all the other groups tended to approach the task from perspective aligned with either the Transitional or Distributional framework levels.

Survey Results: Task 3, Pink/Black survey task

The context of the Pink/Black task is the same as the Yellow/Brown task, that is, comparing test scores from two classes (see appendix B). The number of scores in each data set is not the same; the Pink class is larger than the Black class. There is a clear difference in shapes and centers between these sets. Each of the mean, median, and mode are higher in the Black class. The ranges are the same but the Pink class’s data is bell shaped while the Black class’s data is skewed. The second part of the task

applies to students who decided that one of the classes did score better; they are asked to estimate how much better.

Participants' responses on the first part of this task were categorized at each level of the framework. Across all the groups the most common decision was that the Black class scored better, followed by the decision that the classes scored equally well, and the least common decision was the Pink class scored better. For each of the groups, except 1-SE and GRAD, the most common reasons for the 'Black' decision were coded at level 2. The 1-SE group provided slightly more responses at level 3 than level 2, and the GRAD group provided slightly more responses at level 4 than at either level 2 or 3. The majority of the level 2 responses focused on a comparison of either center or shape and at level 3 there was a trend towards a Proportion focus as opposed to an Initial Global focus. These were not surprising results as the centers of the scores for the Black class are all visibly higher than the centers of the scores for the Pink class and the skewed shape of the scores for the Black class provides for a visually obtainable assessment that the Black class has proportionally higher scores.

The researcher expected most of the Transitional responses for the Pink/Black task to focus on center, but the 1-GS students had more students who focused on shape (see table 77 in appendix C). These shape responses generally described a "shift," a "slide" or a "curve," to the right, for the Black class's scores. Others explained that the Black class had fewer total scores but more scores to the right. Although the previous example is a preliminary proportional type argument, it was classified as level 2 – shape because of its inarticulateness. For statistically naïve students, such as the 1-GS

students, the prominent difference between the shapes may have been easier to describe than reasoning about the centers. The Transitional – Center responses mostly cited the higher “average” or “mean” of the Black class, however a few cited the median and some cited the mode, such as, “...the greatest frequency in Black class is 7 while the Pink class seems to be between 5 and 6.”

The somewhat high frequency of level 2 responses from the 1-GS students in support of ‘equal’ was also a bit surprising (see table 77 in appendix C). Most cited the difference in shapes or means but then referenced the difference in class size and claimed that because the Black class had fewer students, the classes’ shapes (or the averages) were equal (or about equal). For example, “The graphs were both following the same kind of pattern, one with just fewer students or sample size. If they had been equal it looks like they would appear the same.”

All of the 2-GS students who supported ‘equal’ or ‘Pink,’ with reasons focused on shape or center, attempted to compare the data sets not as individual data points but as groups, yet their understanding of how to do this contained serious flaws. For example, those who chose ‘Pink’ based on their miscalculations of the means appear to have relied solely on their calculations without regard to the shapes and those who decided in favor of ‘equal’ or ‘Pink,’ in an attempt to account for the difference in class size, appear to have difficulty reasoning proportionally about the required comparison.

The second part of the task where students were asked to estimate the difference between the classes’ scores, proved to be difficult for many students, except

those from the GRAD group and, surprisingly, the 1-SE group. There were relatively few Idiosyncratic responses across all the groups for the first part of the task, implying that part was understood by most students. As with the Yellow/Brown task, most Idiosyncratic responses simply contained no information about how the student made his or her decision. The second part of the task elicited considerably more Idiosyncratic responses highlighting their potential non-global understanding of the data sets. A detailed breakdown and discussion of the results for each group follows.

Reasons for the decisions on the Pink/Black task were coded at each of the five lattice structure levels (see table 25 and Figure 33), but unlike the Yellow/Brown task, reasons at the Local level were not appropriate due to the difference in sizes of the data sets. This feature of the Pink/Black task is highlighted by the trend that for each of the groups, all of the students who provided reasons at levels 3 and 4 argued that the Black class scored better, while almost all of the students who responded at level 1 supported either the 'equal' or 'Pink' decisions. Although proportional reasoning was not explicitly required to successfully determine that the Black class scored better, it was explicitly referenced by many of those who responded at level 4 and an overwhelming majority, 49 of 57, who responded at level 3. Across all the groups, except for the GRAD group, the level 2 type responses were the most common and about 75% of those responses supported the 'Black' decision. It was assumed that participating students generally understood that task as only 4% provided Idiosyncratic responses.

Table 25.

The distribution of framework level codes, for responses from task 3 (the Pink/Black task), for all groups.

<u>Group</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>	<u>Mean</u>
1-GS	6 (4.3)	27 (19.7)	81 (59.1)	21 (15.3)	2 (1.5)	137 (100)	1.90
1-SE	2 (5.4)	3 (8.1)	16 (43.2)	15 (40.5)	1 (2.7)	37 (100)	2.27
2-GS	2 (2.7)	10 (13.5)	40 (54.1)	18 (24.3)	4 (5.4)	74 (100)	2.16
2-SE	1 (6.7)	1 (6.7)	9 (60.0)	0 (0.0)	4 (26.7)	15 (100)	2.33
GRAD	0 (0.0)	1 (8.3)	3 (25.0)	3 (25.0)	5 (41.7)	12 (100)	3.00
Total	11 (4.0)	42 (15.3)	149 (54.2)	57 (20.3)	16 (5.8)	275 (100)	2.09

Percent of each group in parentheses.

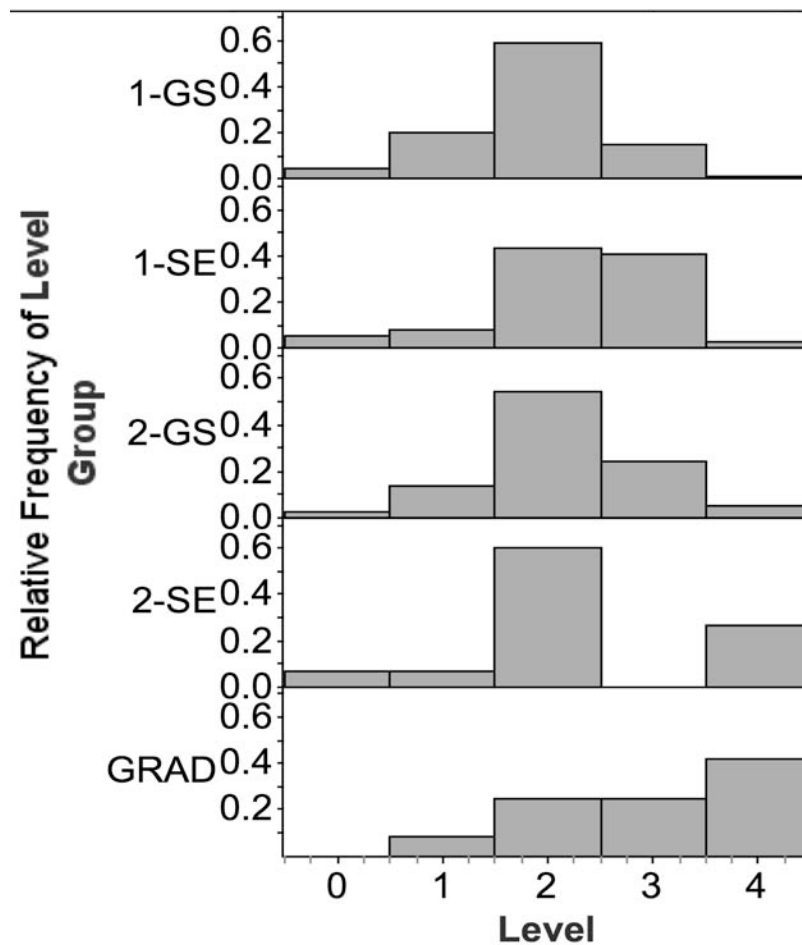


Figure 33. The distribution of response levels across all groups for the Pink/Black task.

The students from the GRAD group appeared to consistently provide responses at higher levels of the Lattice Structure Framework than all other groups. Except for one student, all the GRAD students responded at level 2 or higher, with the highest percentage at level 4. The 1-GS students and the 2-GS students responded at quite similar levels of the framework. Both had more than 50% of their students respond at level 2 and only a few level 4 and a few Idiosyncratic responses. The 2-GS group did have slightly more students respond at level 3 than at level 2 whereas slightly more 1-GS students responded at level 1 than at level 3. Those two opposing trends may account for the mean level, of 2.16, for the 2-GS student responses to be significantly higher than the mean level, of 1.90, for the 1-GS students ($t = -2.28, p = 0.012$).

Although the difference in mean levels is significant, it does seem a bit surprising that the difference is not larger considering that the 2-GS students have had quite a bit more statistics instruction than the 1-GS students. The 1-SE students responded at a mean framework level of 2.27, also significantly higher than the mean response level of the 1-GS students ($t = -2.37, p = .011$). As students from both the 1-GS and 1-SE groups had only completed about 2 or 3 statistics classes at the time they took the survey, the 1-SE students had a stronger mathematics background, which may be a contributing factor to the 1-SE students' higher mean response level. When the 1-SE response levels are compared to the 2-SE response levels using figure 25, the distribution of responses across framework levels appear quite different for each group as the 1-SE responses are clustered at levels 3 and 4 while the 2-SE responses are primarily at level 2 and 4. Yet the mean response level for the 2-SE students is 2.33,

only 0.06 higher than the 1-SE students. As the 2-SE students generally have more statistics and mathematics in their background, those trends are a surprising result.

Only about 15% of students who gave Initial Distributional or Distributional type reasons for determining that the Black class scored better were unable to then make a reasonable estimation for how much better the Black class scored. The success rate of about 85%, for students who responded at either level 3 or 4, at making reasonable estimations for how much better the Black class scored is at least partial evidence that these students viewed the Pink and Black data sets from a global perspective. The prevalence of level 1 responses that focused on comparing frequencies or sums or scores even though the classes were of unequal size, highlights the potential that those students viewed the Pink and Black data sets from a Local perspective. Although the success rate for deciding in favor of 'Black' was high, for those who responded at level 2, the picture of how well those students estimated how much better the Black class scored varied considerably between groups. All of the students from the GRAD group and about 85% of the 1-SE students, who responded at level 2 in favor of 'Black,' also provided reasonable estimates for how much better the Black class scored. However, of those students who responded at level 2 in favor of 'Black,' in groups 1-GS, 2-GS and 2-SE, each group had only about 45% of their students also provide reasonable estimates for how much better the Black class scored. This result was a rather surprising, as students from each of the 2-GS and 2-SE groups had completed at least one statistics course and the 2-SE students had considerably more mathematics courses in their background than either of the 'GS' groups. The

students who focused on comparing a single group feature but were unable to also make a reasonable estimate for how much better the Black class scored, potentially did not consider the feature they used for comparison as a group representative and thus, likely did not consider the Pink and Black data set from either a purely local or global perspective.

Finally, while all groups consistently had the highest frequency of decisions in favor of 'Black,' the 1-GS students tended to approach the task from perspectives aligned with lower framework levels, Local and Transitional, while the Grad students and the 2-SE students tended to approach the task from perspectives aligned with higher framework levels, at the Transitional level and above, and the students from the 1-SE and 2-GS groups tended to approach the task from perspectives aligned with the Transitional framework level and Initial distributional level.

Survey Results: The Pink/Black task –

Without descriptive statistics (Task 3) vs. With descriptive statistics (Task 4)

The results from survey task 4 and how they compare to the results from task 3 are separated into three subsections for convenience purposes. First, the decisions about which class scored better or if they scored equally well are compared and discussed for both tasks. Second, the levels of responses that the students provided for both tasks are compared and discussed. Third, the estimation strategies that the students used in both tasks are compared and discussed. Overall, the “with statistics” responses tended to be briefer than the “without statistics” responses as many of the

“with statistics” responses merely referenced certain descriptive statistics with no further explanation.

Pink/Black survey task decisions: Without statistics vs. With statistics

Of the students who initially decided that ‘the Black class scored better’ an overwhelming majority (95.8%) stayed with their ‘Black’ decision after viewing some descriptive statistics associated with each data set. Just over two-thirds (67.3%) of the students who initially decided that ‘the classes scored equally well, switched their decision to ‘Black’ after viewing the descriptive statistics and more than half (54.8%) of the students who initially decided that ‘the Pink class scored better’ switched their decision to ‘Black’ after viewing the descriptive statistics. Overall, 85.8% of students decided on ‘Black’ after viewing the descriptive statistics. Table 26 displays the specific decision shifts, by group, for before and after viewing the descriptive statistics. For each group, a two-proportion z test was used to compare the proportion of successful responses (‘Black’) from before and after the students had access to descriptive statistics.

Table 26.

Decisions shifts by group for both Pink/Black tasks.

<u>Pink/Black Decision without Stats</u>	<u>Pink/Black Decision with Stats</u>			<u>Group</u>
	<u>Black</u>	<u>Equal or Pink</u>	<u>Total</u>	
	Black	78	6	84
	Equal or Pink	31	22	53
	Total	109	28	
				1-GS (n=137)
	Black	32	0	32
	Equal or Pink	2	3	5
	Total	34	3	
				1-SE (n=37)
	Black	50	1	51
	Equal or Pink	17	6	23
	Total	67	7	
				2-GS (n=74)
	Black	13	1	14
	Equal or Pink	1	0	1
	Total	14	1	
				2-SE (n=15)
	Black	11	0	11
	Equal or Pink	1	0	1
	Total	12	0	
				GRAD (n=12)

The two ‘GS’ groups were the only ones to have a significant increase of the proportion of students who switched their initial decision from either ‘equal’ or ‘Pink’ to ‘Black.’ For the 1-GS students, $z = -3.31$ and $p < 0.01$, while for the 2-GS students, $z = -3.27$ and $p < 0.01$. It was not surprising that the students from the 1-SE, 2-SE and GRAD groups did not switch to ‘Black’ in significant proportions as a large majority of those students initially had decided in favor of ‘Black.’

Of the group 1-GS students who initially decided that ‘the Black class scored better’ an overwhelming majority (92.9%) stayed with their ‘Black’ decision after viewing some descriptive statistics associated with each data set and almost 60% of

the students who initially decided in favor of either ‘equal’ or ‘Pink’ switched their decision to ‘Black’ after viewing the descriptive statistics. From group 2-GS, 98% of those students initially decided in favor of ‘Black’ and then stayed with their ‘Black’ decision after viewing some descriptive statistics associated with each data set and almost 74% of the students who initially decided in favor of either ‘equal’ or ‘Pink’ switched their decision to ‘Black’ after viewing the descriptive statistics. The percentage of students who either did not switch from their ‘equal’/‘Pink’ decision or actually switch from ‘Black’ to ‘equal’/‘Pink’ was the highest among all groups at about 20% for 1-GS and second highest at about 9% for 2-GS. For the seven 1-GS and 2-GS students who switched away from their ‘Black’ decision, the reasons they provided for their switches were generally either Idiosyncratic or a misinterpretation or possibly a misreading of the available descriptive statistics. For example, two of the students who switched to ‘Pink’ wrote, “because there are more people [in the Pink class] and the mean score” and “standard error of mean was lower than the Black class.” One of the students who switched from ‘Black’ to ‘equal’ wrote,

*dang, well it seems that many of their statictics [sic] were equal and
dispite [sic] the slightly lower score mean of the pink class I feel that the
larger population was a contributor and thus conclude they are equal*

Similar to those students who switched their decision away from ‘Black,’ those students from both ‘GS’ groups who did not switch to ‘Black’ from their initial ‘equal’ or ‘Pink’ decision generally provided responses that indicated that those students had

difficulty reasoning proportionally and/or interpreting the available descriptive statistics.

The vast majority of students from groups 1-SE, 2-SE and GRAD either initially decided in favor of 'Black' and did not switch or switched to 'Black' after having access to some descriptive statistics associated with each data set. From all these three groups only three students (from group 1-SE) initially chose 'equal' or 'Pink' and did not switch to 'Black' and one 2-SE student actually switched from 'Black' to 'Pink.' The one 1-SE student who initially chose 'Pink' and did not switch and the 2-SE student who switched to 'Pink' from 'Black' both referenced the smaller standard deviation of the Pink class as "better" with no further explanation. The other two 1-SE students both initially chose 'equal' and did not switch. Those two students referenced the size difference between the classes as the reason why the classes scored equally well. These four students seemed to have similar difficulties as the 'GS' students who did not switch to 'Black.' Irrespective of whether students, from all groups, switched their decision or not, most students reasons for their decisions, after having viewed the descriptive statistics, seemed to have shifted classification in the Expanded Lattice Structure Framework. Trends in the levels of responses are explored in the next section.

Pink/Black task response levels: Without statistics vs. With statistics

For this follow-up task, the Pink/Black task with descriptive statistics, the code 'N/A' was added at level 2 to account for responses that appeared to merely recite the statistical terms or refer to "the statistics" in a meaningless and uninformative way.

For example, “the scores in the Black class still had better statistics” was coded N/A. Also, responses that list some terms yet the values for some are clearly not “better,” such as, “I look at the mean, median, mode and the standard deviation. Black's class got the higher values in these criteria.” In that example, it is not clear to the researcher why a higher standard deviation supports the decision that ‘the Black class scored better.’ In general, there were relatively few of these types of responses and thus were not addressed on the following discussion and analysis.

The distributions, across framework levels, of responses given by students from the 1-GS group to both tasks 3 and 4 are shown in Figure 34. From those graphs, it appears that the 1-GS students tended to provide responses at higher levels after having access to some descriptive statistics. The shift appears to be away from level 1 type responses and toward level 3 type responses. Overall, a paired two-sample t-test reveals that the mean response level of 2.22 with statistics is significantly higher than the mean response level of 1.90 without statistics ($t = -4.11, p < 0.01$). Thus, the 1-GS students not only decided in favor of ‘Black’ in greater proportions but also correspondingly responded at higher levels of the Expanded Lattice Structure Framework. The details of this shift in responses are shown in table 27.

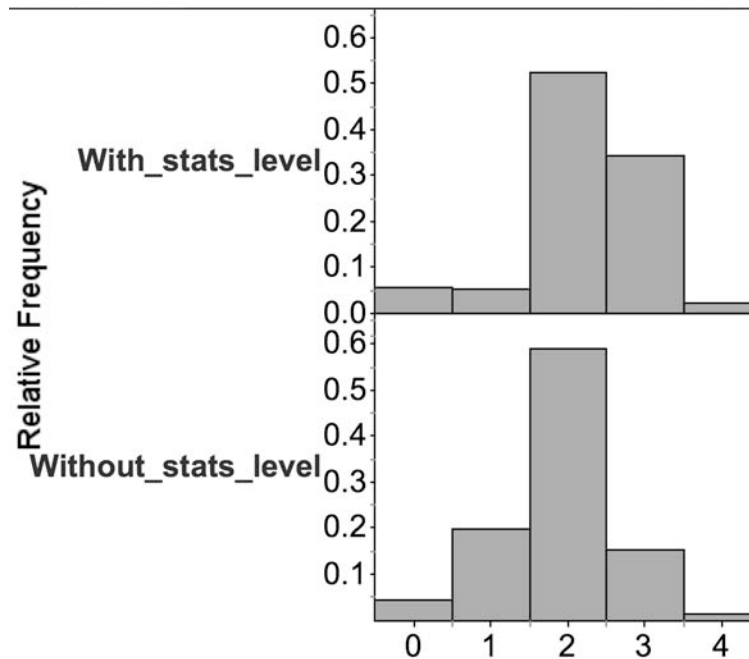


Figure 34. Response levels of the 1-GS students to both Pink/Black tasks: Without statistics and With statistics.

The outlined cells in table 27 contain the counts of student responses that did not shift levels. About half of the 1-GS students supported their decision with reasons at a level equal to the reasons given before they viewed the descriptive statistics. Almost 38% of the group 1-GS students, after viewing the descriptive statistics, provided reasons for their decision at a higher level than their initial reason. Also, after viewing the descriptive statistics, only slightly more than 10% of group 1-GS students provided reasons for their decisions that were at a lower level than their initial reasons. Two of the more noticeable trends were that about 17% of 1-GS students' responses shifted from other levels to level 2 and about 25% shifted from other levels to level 3. Those shifts also accompanied high success rates for deciding in favor of 'Black' as all of the 1-GS students who responded at either level 3 or 4 also decided in favor of

‘Black’ and 75% of 1-GS students who responded at level 2 also decided in favor of ‘Black,’ a 15% increase compared with task 3 responses at level 2.

Table 27.

Distribution of responses from the 1-GS group for the Pink/Black task: Without statistics vs. With statistics.

		<u>Response Level with stats</u>					
		<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>Total</u>
<u>Response Level without stats</u>	0:	-	-	5	1	-	6
	1:	4	6	11	6	-	27
	2:	4	1	49	26	1	81
	3:	-	-	6	13	2	21
	4:	-	-	1	1	-	2
Total		8	7	72	47	3	137

Table 28.

Distribution of Level 2 and Level 3 responses, from the 1-GS group, for the Pink/Black task: Without statistics and With statistics.

Response	Level 2					Level 3		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
without Stats	29	50	2	-	81	19	2	21
with Stats	41	9	5	17	72	7	40	47

Table 28 displays the pattern of what features of the data the 1-GS students, who responded at the Transitional and Initial Distributional levels, focused on. At level 2, after having access to the descriptive statistics, considerably fewer students focused on Shape and Proportion and considerably more students focused on Center or had an Initial Global focus. Of those who focused on Center, all but four exclusively

cited either the mean or “average” of the Black class as higher (as opposed to either the median or mode) with no other information. All of the Initial Global responses cited at least two or all three measures of center, mean, median and mode, for the Black class as either being “higher” or “better” and no other information. These two groups of students, combined, represent a majority of 1-GS students, who exclusively focused on measures of center after they had access to descriptive statistics. As the 1-GS students had limited statistical instruction, it is likely that measures of center were the statistics that they were most comfortable with and thus relied on when they had access to descriptive statistics.

The distributions across framework levels of responses given by students from the 1-SE group to both tasks 3 and 4 are shown in Figure 35. From those graphs, it appears that the 1-SE students tended to provide responses at higher levels after having access to some descriptive statistics. The shift appears to be away from levels 0 and 1 types of responses. Overall, a paired two-sample t-test reveals that the mean response level, of 2.59, with statistics is significantly higher than the mean response level, of 2.27, without statistics ($t = -1.97, p = 0.028$). Thus, although the 1-SE students decided in favor of ‘Black’ in about the same proportions with or without descriptive statistics, they responded at higher levels of the Expanded Lattice Structure Framework with descriptive statistics. The details of this shift in responses are shown in table 29.

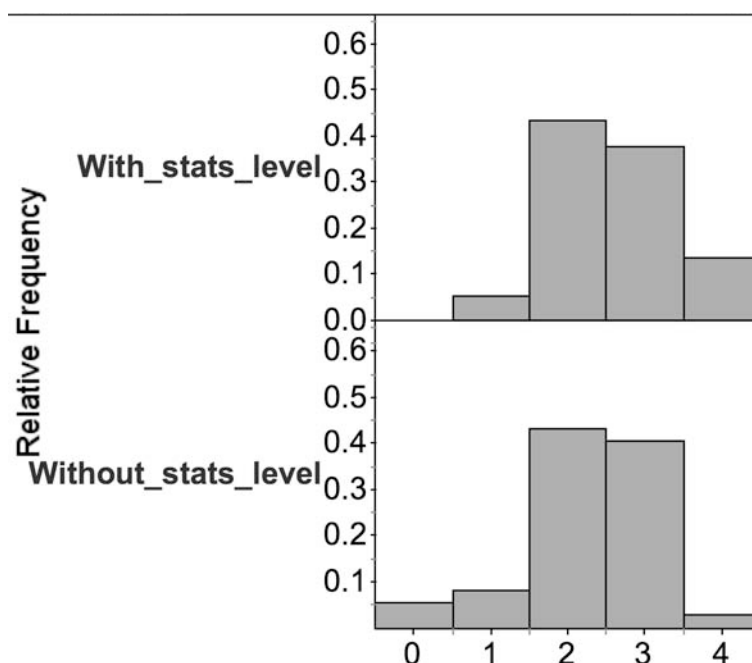


Figure 35. Response levels of the 1-SE students to both Pink/Black tasks: Without statistics and With statistics.

After viewing the descriptive statistics, almost 60% of all students from group 1-SE supported their decision with reasons at a level equal to the reasons given before they viewed the descriptive statistics, as shown in the outlined cells of table 29. About 27% of group 1-SE students, after viewing the descriptive statistics, provided reasons for their decision at a higher level than their initial reason. Also, after viewing the descriptive statistics, only slightly more than 13% of group 1-SE students provided reasons for their decisions that were at a lower level than their initial reasons. About equal amounts of students' "with statistics" responses shifted to level 2 from other levels, shifted to level 3 from other levels, and shifted to level 4 from other levels. However, most of the shift to level 2 was from level 3 while the shifts to levels 3 and 4

were from lower levels. All of the level 3 and 4 “with statistics” responses favored “Black” as well as all but one of the level 2 responses.

Table 29.

Distribution of responses from the 1-SE group for the Pink/Black task: Without statistics vs. With statistics.

		<u>Response Level with stats</u>					
		<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>Total</u>
<u>Response Level without stats</u>	0:	-	-	-	2	-	2
	1:	-	1	1	1	-	3
	2:	-	1	11	2	2	16
	3:	-	-	4	9	2	15
	4:	-	-	-	-	1	1
Total		0	2	16	14	5	37

Table 30.

Distribution of Level 2 and Level 3 responses, from the 1-SE group, for the Pink/Black task: Without statistics and With statistics.

Response	Level 2					Level 3		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
without Stats	11	3	2	-	16	11	4	15
with Stats	12	0	1	3	16	0	14	14

Table 30 displays the pattern of what features of the data the 1-SE students, who responded at the Transitional and Initial Distributional levels, focused on, a somewhat similar pattern as the 1-GS students. At level 2, after having access to the descriptive statistics, fewer students focused on Shape and considerably fewer focused on Proportion. While only one more student focused on Center, considerably more had

an Initial Global focus. Of those who focused on Center, all but one exclusively cited either the mean or “average” of the Black class as higher (as opposed to either the median or mode) with no other information. All of the Initial Global responses cited at least two or all three measures of center, mean, median and mode, for the Black class as either being “higher” or “better” and no other information, except for one student. That one student wrote, “The pink class had a smaller Std. deviation and variance. But the black had higher mean, median and mode.” Although this student attended to two features, they seemed to be attended to separately, not in an integrated way. Again, similar to the 1-GS group, the two groups of students, combined, who responded at levels 2 and 3, and focused exclusively on measures of center, represent a majority of 1-SE students. As the 1-SE students also had limited statistical instruction, it is likely that measures of center were the statistics that they were most comfortable with and thus relied on when they had access to descriptive statistics.

The distributions, across framework levels, of responses given by students from the 2-GS group to both tasks 3 and 4 are shown in Figure 36. From those graphs, it appears that the 2-GS students tended to provide responses at higher levels after having access to some descriptive statistics. The shifts appear quite similar to that of the 1-GS students, that is away from level 1 type responses and towards levels 3 and 4 types of responses. Overall, a paired two-sample t-test reveals that the mean response level, of 2.47, with statistics is significantly higher than the mean response level, of 2.16, without statistics ($t = -2.95, p < 0.01$). Thus, the 2-GS students not only decided in favor of ‘Black’ in greater proportions but also correspondingly responded at higher

levels of the Expanded Lattice Structure Framework, just as the 1-GS students did. The details of this shift in responses are shown in table 31.

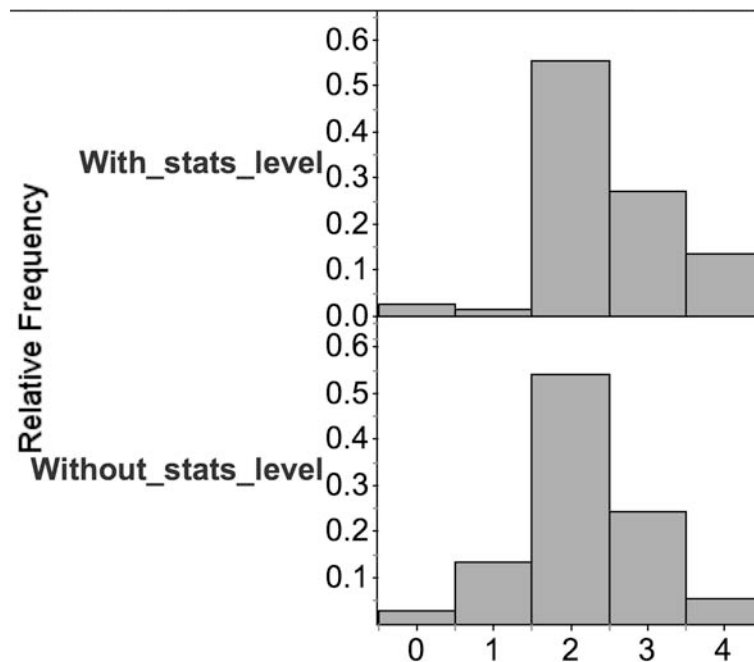


Figure 36. Response levels of the 2-GS students to both Pink/Black tasks: Without statistics and With statistics.

The outlined cells in table 31 contain the counts of student responses that did not shift levels. After viewing the descriptive statistics, almost 50% of all students from group 2-GS supported their decision with reasons at an equal level, almost 40% provided reasons for their decision at a higher level, and slightly more than 10% provided reasons for their decisions at a lower level, to their initial reasons given before they viewed the descriptive statistics. Those percentages are almost identical to the trend from group 1-GS. Some of the more noticeable trends were that about 20% of 2-GS students' responses shifted from other levels to level 2, about 19% shifted from other levels to level 3 and almost 11% shifted from lower levels to level 4. Those

shifts also accompanied high success rates for deciding in favor of ‘Black’ as all of the 2-GS students who responded at either level 3 or 4 also decided in favor of ‘Black’ and about 85% of 2-GS students who responded at level 2 also decided in favor of ‘Black,’ a 15% increase compared with task 3 responses at level 2, the same percentage increase as the 1-GS students who responded at level 2.

Table 31.

Distribution of responses from the 2-GS group for the Pink/Black task: Without statistics vs. With statistics.

	<u>Response Level with stats</u>					<u>Total</u>
	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
<u>Response Level without stats</u> 0:	1	-	1	-	-	2
1:	-	1	7	2	-	10
2:	-	-	26	11	3	40
3:	1	-	6	6	5	18
4:	-	-	1	1	2	4
Total	2	1	41	20	10	74

Table 32.

Distribution of Level 2 and Level 3 responses, from the 2-GS group, for the Pink/Black task: Without statistics and With statistics.

<u>Response</u>	<u>Level 2</u>					<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
without Stats	19	21	0	-	40	16	2	18
with Stats	29	6	1	5	41	1	19	20

Table 32 displays the pattern of what features of the data the 2-GS students, who responded at the Transitional and Initial Distributional levels, focused on. At

level 2, after having access to the descriptive statistics, similar to groups 1-GS and 1-SE, considerably fewer students focused on Shape and Proportion and considerably more students focused on Center or had an Initial Global focus. All of those who focused on Center exclusively cited either the mean or “average” of the Black class as higher (as opposed to either the median or mode) with no other information. All, except for one, of the Initial Global responses cited at least two or all three measures of center, mean, median and mode, for the Black class as either being “higher” or “better” and no other information. The one student who did not exclusively cite measures of center wrote, “I would still say the black class did better based on skew and mode and median but realistically speaking the pink class really did learn more, or rather retained more knowledge.” While this response had potential to be classified as Distributional, the student’s qualification that “the pink class really did learn more” was a cue that the student may not fully understand the comparison he or she made, from a global perspective. These two groups of 2-GS students who responded at either levels 2 - center or level 3, combined represent a majority of 2-GS students, who exclusively focused on measures of center after they had access to descriptive statistics. Although the 2-GS students had completed one statistics course and were enrolled in their second, most apparently tended to rely exclusively on measures of center when comparing the Pink/Black distributions with access to descriptive statistics.

The distributions, across framework levels, of responses given by students from the 2-SE group to both tasks 3 and 4 are shown in figure 29. From those graphs,

it appears that the 2-SE students tended to provide responses at higher levels after having access to some descriptive statistics, shifting away from giving responses at levels 1 and 2 and toward level 3. Although the 2-SE students provided a higher mean response level, of 2.6, for task 4 compared to a mean response level of 2.33 for task 3, a paired two-sample t-test revealed no significant difference between those means ($t = -0.72, p > 0.05$). Thus, although the 2-SE students decided in favor of ‘Black’ in about the same proportions with or without descriptive statistics and responded at approximately the same mean level of the Expanded Lattice Structure Framework, with descriptive statistics, Figure 37 reveals that their responses were distributed across the levels, for each task, in slightly different patterns. The details of those shifts in responses are shown in table 33.

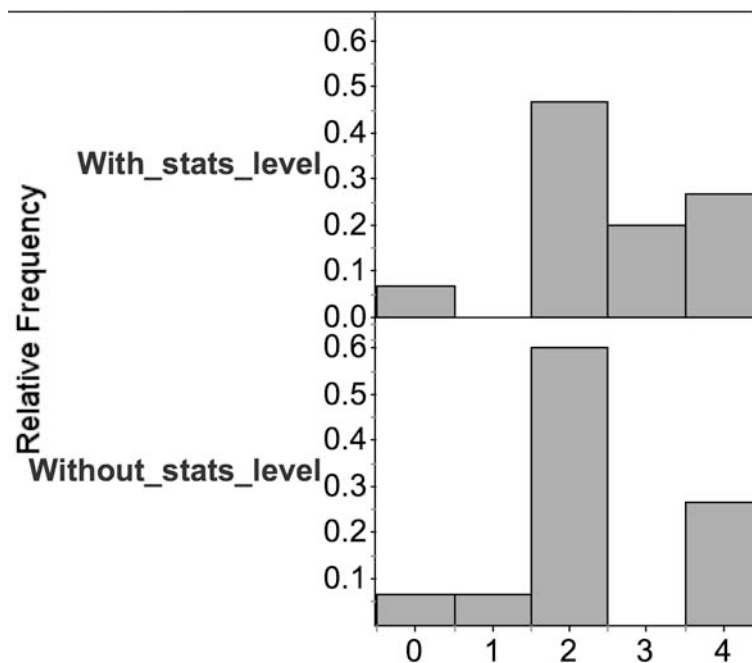


Figure 37. Response levels of the 2-SE students to both Pink/Black tasks: Without statistics and With statistics

After viewing the descriptive statistics, six of the 15 students from group 2-SE (40%) supported their decision with reasons at a level equal to the reasons given before they viewed the descriptive statistics, as shown in the outlined cells of table 33. Another 40% of group 2-SE students, after viewing the descriptive statistics, provided reasons for their decision at a higher level than their initial reason. Those five students provided level 0 or 1 responses to task 3 but then after considering the descriptive statistics provided responses at levels 3 or 4. Also, after viewing the descriptive statistics, only three students from group 2-SE (10%) provided reasons for their decisions that were at a lower level than their initial reasons. All three of those students provided level 4 responses without the descriptive statistics. Despite the shifts to providing responses at different levels, all but one of the 2-SE students continued to decide in favor of 'Black.' Thus, almost half of all group 2-SE students decided 'Black' and also supported their decision with a reason at level 3 or 4.

Table 33.

Distribution of responses from the 2-SE group for the Pink/Black task: Without statistics vs. With statistics.

	<u>Response Level with stats</u>					<u>Total</u>
	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
<u>Response Level without stats</u> 0:	1	-	-	-	-	1
1:	-	-	-	1	-	1
2:	-	-	4	2	3	9
3:	-	-	-	-	-	0
4:	-	-	3	-	1	4
Total	1	0	7	3	4	15

Table 34.

Distribution of Level 2 and Level 3 responses, from the 2-SE group, for the Pink/Black task: Without statistics and With statistics.

<u>Group</u>	<u>Response</u>	<u>Level 2</u>					<u>Level 3</u>		
		<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
2-SE (n=15)	without Stats	7	2	0	-	9	0	0	0
	with Stats	3	0	1	3	7	0	3	3

Table 34 displays the pattern of what features of the data that the 2-SE students, who responded at the Transitional and Initial Distributional levels, focused on. There were only a few similarities to the reasons provided by the 1-SE and ‘GS’ students. At level 2, after having access to the descriptive statistics, fewer students did focus on Shape but also considerably fewer focused on Center. At level 3, no 2-SE students focused on Proportion, without or with descriptive statistics, but three students did shift their responses to have an Initial Global focus with statistics, whereas without statistics their responses were lower than level 3. Of the three who focused on Center, all exclusively cited either the mean or “average” of the Black class as higher (as opposed to either the median or mode) with no other information. All of the Initial Global responses cited at least two or all three measures of center, mean, median and mode, for the Black class as either being “higher” or “better.” Similar to the 1-SE and ‘GS’ groups, the majority of 2-SE students who responded at levels 2 or 3 focused exclusively on measures of center, and although these students do not represent a majority of 2-SE students, they are 40% of the 2-SE group, a fairly high percentage. So, although the 2-SE students had completed at least one statistics course and generally had strong mathematics backgrounds, the group did show a slight

trend, similar to the 1-SE and ‘GS’ groups, of tending to rely exclusively on measures of center when comparing the Pink/Black distributions with access to descriptive statistics.

The distributions, across framework levels, of responses given by students from the GRAD group to both tasks 3 and 4 are shown in Figure 38. From those graphs, it appears that the GRAD students tended to provide responses at approximately similar levels after having access to some descriptive statistics. Although the GRAD students provided a lower mean response level, of 2.83, for task 4 compared to a mean response level of 3.0 for task 3, a paired two-sample t-test revealed no significant difference between those means ($t = 0.52$, $p = 0.307$). Thus, the GRAD students decided in favor of ‘Black’ in about the same proportions with or without descriptive statistics and responded at approximately the same mean level of the Expanded Lattice Structure Framework, for both tasks without and with descriptive statistics. The details of the shifts in responses that did occur are shown in table 35.

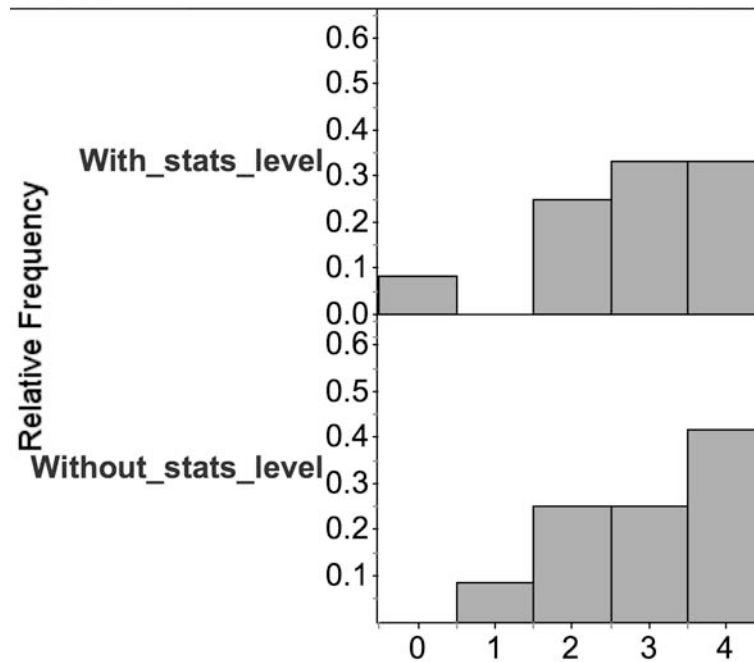


Figure 38. Response levels of the GRAD students to both Pink/Black tasks: Without statistics and With statistics

Table 35.

Distribution of responses from the GRAD group for the Pink/Black task: Without statistics vs. With statistics.

	Response Level with stats					Total
	0	1	2	3	4	
Response Level without stats	0:	-	-	-	-	0
	1:	-	-	1	-	1
	2:	-	-	1	2	3
	3:	1	-	1	1	3
	4:	-	-	-	1	4
Total	1	0	3	4	4	12

After viewing the descriptive statistics, half of the 12 students from the GRAD group supported their decision with reasons at a level equal to the reasons given before they viewed the descriptive statistics, as shown in the outlined cells in table 35. Three of the group GRAD students (25%) provided reasons for their decision at a higher level and three provided reasons at a lower level than their initial task 3 reasons, without descriptive statistics. The three students whose responses shifted higher had their responses classified up one level each to either level 2 or 3, while two of the three students whose responses shifted lower had theirs classified down one level to level 2 or 3. The one student who provided an Idiosyncratic response wrote, “I compared the ratio of each of their Standard error to their mean scores.” Although the highest percentages of GRAD students responded at levels 3 and 4, a different trend from all other groups, a similarity to the other groups is that the frequency of responses at levels 2 and 3 also represent a majority of the GRAD responses.

Table 36.

Distribution of Level 2 and Level 3 responses, from the GRAD group, for the Pink/Black task: Without statistics and With statistics.

<u>Group</u>	<u>Response</u>	<u>Level 2</u>					<u>Level 3</u>		
		<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
GRAD (n=12)	without Stats	2	1	0	-	3	3	0	3
	with Stats	1	0	0	2	3	1	3	4

Table 36 displays the pattern of what features of the data that the GRAD students, who responded at the Transitional and Initial Distributional levels, focused on. Although the low frequency of level 2 and 3 responses make any trend observations very tentative, there were a few similarities to the reasons provided by

the 1-SE, 'GS' and 2-SE students. At level 2, after having access to the descriptive statistics, one less student each focused on Shape and on Center. At level 3, there was an observable shift away from a focus on Proportion, without descriptive statistics, and towards an Initial Global focus, with descriptive statistics. The one student who focused on Center exclusively cited the mean of the Black class as higher (as opposed to either the median or mode) with no other information. All of the Initial Global responses cited at least two or all three measures of center, mean, median and mode, for the Black class as either being "higher" or "better." Thus only one-third of the GRAD students exclusively focused on comparing measures of center, after having access to the descriptive statistics, the lowest such percentage of students out of all groups. So, after having access to the descriptive statistics, all the GRAD students decided that the Black class scored better, they provided the highest percentage of Distributional responses and the lowest percentages of Local and Transitional responses, and also had the lowest percentage of student supply reasons that were exclusively focused on measures of center. The GRAD students advanced statistical background appears to have contributed to their consistently high level responses on both Pink/Black tasks, without and with descriptive statistics.

Pink/Black task estimations: Without statistics vs. With statistics

The following discussion of estimation strategies is focused mostly on those used to quantify how much better the Black class scored. The results from each group are presented and then similarities and differences between the groups are discussed.

The number of 1-GS students who used the difference between center for their estimate for how much better the Black class scored increased from 31 without statistics to 74 with statistics, that is an increase of about 31% more of the 1-GS students. The difference between centers strategy was the only one to have increased used after the 1-GS students had access to the descriptive statistics. None switched their estimation strategy, with statistics, to the difference between proportions strategy, but most of those who found the difference between proportions, without statistics, switched to the difference between centers, with statistics. None of the 1-GS students used the ratio of centers strategy either without or with statistics. Most of the 1-GS students who switched their estimation strategy to the difference between centers, had previously used either Idiosyncratic strategies or had decided that the classes scored equally well, without statistics.

The number of 1-SE students who used the difference between center for their estimate for how much better the Black class scored increased from 18 without statistics to 24 with statistics, that is an increase of about 16% more of the 1-SE students. The only other strategy that had an increase was ratio of centers as three students used the ratio of centers for their estimation without statistics and five used that strategy with statistics. None switched their estimation strategy, with statistics, to the difference between proportions, but of the six who found the difference between proportions, without statistics, two switched to the difference between centers and three switch to the ratio of centers, with statistics. Five of the eight 1-SE students who switched their estimation strategy to the difference between centers, had previously

used either Idiosyncratic strategies, had decided that the classes scored equally well, or used the difference of sums of scores, without statistics.

The number of 2-GS students who used the difference between centers for their estimate for how much better the Black class scored increased from 20 without statistics to 44 with statistics, that is an increase of about 32% more of the 2-GS students. The difference between centers strategy was the only one to have increased used after the 2-GS students had access to the descriptive statistics. None switched their estimation strategy, with statistics, to finding the difference between proportions, but most of those who found the difference between proportions, without statistics, switch to finding the difference between centers, with statistics. The one student whose estimation strategy was the ratio of center, without statistics, switched to the difference between centers, with statistics. Most of the 2-GS students who switched their estimation strategy to the difference between centers, had previously used either Idiosyncratic strategies or had decided that the classes scored equally well, without statistics.

The number of 2-SE students who used the difference between centers for their estimate for how much better the Black class scored increased from 6 without statistics to 8 with statistics, and the number of students who used the ratio of centers strategy also increased from zero without statistics to two with statistics. The remaining 2-SE students used idiosyncratic strategies without and with statistics, except for one who found the difference between proportions without statistics but with statistics found the difference between the sample variances and also wrote “I don’t know.”

All seven of the GRAD students who used the difference between centers for their estimate for how much better the Black class scored without statistics continued to use that strategy with statistics. Of the two students who used the ratio of centers strategy without statistics, one continued to use that strategy and one switched to the difference between centers strategy, with statistics. Of the two students who used the difference of proportions strategy without statistics, one continued to use that strategy and one switched to an Idiosyncratic strategy, with statistics. The student who switched to the Idiosyncratic strategy is the same student who compared the ratio of each of their Standard error to their mean scores, and then made the estimation as the difference between those ratios.

All groups had an increased use of the difference between centers strategy and all groups had a decrease in the use of the difference between proportions strategy when estimating how much better the Black class scored, after having access to the descriptive statistics. These two trends correspond to the increased reliance on comparing centers and decreased reliance on comparing proportions to support the decision that the Black class scored better. As the measures of center are the first statistics included with task 4, these trends may be due to the convenience of referring to the centers or, particularly for the 'GS' groups, comparing measures of center could be what they are most familiar with. The 'estimation' results seemed to follow the same general trend from task 3, that is the 'GS' groups tended to have the most difficulty and the GRAD group tended to be the most successful. Yet, because the vast majority of explanations were quite minimal and exclusively reference the terms used

in the tables of statistics provided with task 4, little insight was gained from the estimation portion of this task about students' perspectives of the Pink and Black distributions.

Survey response summary: Pink/Black tasks without and with descriptive statistics

After examining the responses across the groups, for task 4, and the shifts in responses from task 3 to task 4, several trends emerged. Both the mean response level and the percentage of students who decided that the Black class scored better, increased significantly for groups 1-GS and 2-GS, from task 3 to task 4. Groups 1-SE, 2-SE and GRAD had high percentages of students responding in favor of 'Black' for task 3 and thus had no statistically significant change in their percentage of decisions favoring 'Black,' from task 3 to task 4. The students from group 1-SE did have a significant increase in their mean response level from task 3 to task 4, while groups 2-SE and GRAD had no significant change in their mean response level. Whether or not groups' mean response level changed, the ways in which the students in each group responded to the Pink/Black did change after they had access to the descriptive statistics. All groups saw an increase in students who exclusively compared measures of center to support their decision, after having access to the descriptive statistics. This increase was more than 50% for both 'GS' groups and the 1-SE group, while the increase was less than 50% for the 1-SE and GRAD groups. Another trend, across the groups, was an increase in Initial Global responses, focused on comparing multiple measures of center, and a considerable decrease in students who provided reasons that compared proportions. It is not clear why students felt compelled to abandon their

proportional arguments. It is possible that they may have felt compelled to cite the statistical measure, because they were provided, or they may have thought that citing the statistical measures was more convincing. When making their estimates for how much better the Black class scored, most students followed the trends from their decisions, that is there was a considerable increase in using the difference between centers strategy and a decrease in the difference between proportions strategy. Overall, the 1-GS students responded at lower levels of the Expanded Lattice Structure framework than any other group. It was somewhat surprising that the 2-GS students responded at only slightly higher levels than the 1-GS students and performed about as well at making estimations. Also surprising was that the 2-SE students did not respond at significantly higher levels than the 1-SE students and the 2-SE students seemed to have more difficulties at making estimations than the 1-SE students. The GRAD students advanced statistical background appears to be evident as those students consistently out performed the other groups on both Pink/Black tasks, without and with descriptive statistics. With the inclusion of descriptive statistics with the data sets, all groups recorded shifts in decisions to favor the Black class as scoring better along with corresponding shifts in reasoning that either in part or exclusively focused on comparing centers.

Survey Results: Task 5, Ambulance task

The context of the Ambulance task is similar to the Movie Wait-Time task in that for both tasks students are asked to compare wait-times. Yet, the context of the Ambulance task is considerably more crucial, as the two data sets in the Ambulance

task represent response times for ambulance services, that are essentially wait-times from the time they're called (see appendix B). The number of response times in each data set is not the same; there were more response times recorded for the Speedy ambulance service than for the Life Line ambulance service. The centers of each data set are not consistently higher or lower for one ambulance service. Both data sets are unimodal with Life Line's mode located at a lower time than Speedy's mode. Speedy's data set has several other 'peaks' located at lower times than Life Line's mode, and students may also consider those as modes. Life Line has a smaller mean while Speedy has a smaller median. Life Line has a smaller range than Speedy. Life Line has a slightly smaller minimum and a smaller maximum than Speedy. The data sets have different shapes. Although it was assumed that quicker response times are more desirable, many students also made reliability assessments, as they indicated that higher reliability was also better.

Participants' responses on this task were categorized at each level of the framework. The "GS" groups gave all the Idiosyncratic responses, except for one. Although the "GS" groups had fairly high rates of Idiosyncratic responses, it was common for those responses to exclusively address context, such as the following rather lengthy response:

Intuitively speaking, I would choose the Life Line one because Speedy Ambulance is obviously more well known since it has received more calls and, hence, busier making the possibility of it taking them longer more prominent...and which, obviously, leaves the Life Line Ambulance more readily available should people need to call. Besides, they're probably a new company and it's always good to support small businesses!

Thus, it was assumed that almost all of the survey participants understood the Ambulance survey task.

Across all the groups the most common decision was to recommend the Life Line ambulance service, although the 1-GS, 2-GS, and 2-SE groups either equally favored both services or only slightly favored Life Line, while the 1-SE and GRAD groups favored Life Line at about a 2 to 1 rate. Level 2 responses were most frequent for each of the groups, except 1-GS who provided slightly more level 1 responses than level 2 and 2-SE who provided equal amounts of level 1 and 2 responses. The Local types of responses, in favor of Life Line, frequently cited Life Line's shorter minimum response time or Speedy's longer maximum response time. The Speedy recommendations often did not account for the unequal sizes of the groups of response times. For example, the following Life Line recommendations appeared to be based on strictly comparing the ends of the distributions without consideration of the distributions' shape, center or spread and the following Speedy recommendations focused on frequencies:

Life Line: *because they don't have as many really long responces* [sic]

Life Line: *speedy had no times lower than 6 mins* [sic]

Speedy: *this one is better more dots* [sic]

Speedy: *they have more shorter response times recorded.*

On the whole, the 1-GS students appeared to experience difficulties in responding to the Ambulance task. More than 60% provided responses categorized at level 1 or level 0, with the highest percentage of responses at level 1. Many of the level 1 responses

indicated that those students either did not reason proportionally or had difficulty reasoning proportionally.

Among the four main tasks, Yellow/Brown, Movie Wait-Time, Pink/Black, and Ambulance, level 2 responses focused on center were consistently provided at lower percentages for the Ambulance task. This trend was not very surprising as the only visibly obvious difference between measures of center was between the modes.

Reasons for the decisions on the Ambulance task were coded at each of the five lattice structure levels (see table 37 and Figure 39). Similar to the Pink/Black task, reasons at the Local level were not appropriate due to the difference in sizes of the data sets. Among the responses across levels 0, 1, 2, and 3 for the Ambulance task, there was no apparent trend in recommendations. However, in the breakdown of responses at levels 2 and 3, the few that focused on Variation accurately cited the smaller spread of the response times for Life Line and hence recommended Life Line, but reasons that focused either on Center, Shape, or Proportion were used to support both ambulance services. This circumstance may be due to the difficult nature of the required comparison as the distributions had few features that could be determined as distinctly different from each other, by a cursory visual assessment. Particularly interesting was the variety of “cut-points” that students used for their proportional comparisons. Although many students arbitrarily partitioned the data at a midpoint of the range similar to some of the middle school students from the research of McClain, Cobb and Gravemeijer (2000), those researchers described that partitioning process as an initial step in reasoning about global trends in the data sets. The students who took

this survey clearly demonstrated that a proportional argument could be made to support a recommendation for either ambulance service depending on where the cut-point was chosen.

Table 37.

The distribution of framework level codes, for responses from task 5 (the Ambulance task), for all groups.

<u>Group</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>
1-GS	21 (15.3)	63 (46.0)	45 (32.8)	8 (5.8)	0 (0.0)	137 (100)
1-SE	0 (0.0)	8 (21.6)	21 (56.8)	6 (16.2)	2 (5.4)	37 (100)
2-GS	10 (13.5)	25 (33.8)	30 (40.5)	7 (9.5)	2 (2.7)	74 (100)
2-SE	1 (6.7)	6 (40.0)	6 (40.0)	1 (6.7)	1 (6.7)	15 (100)
GRAD	0 (0.0)	0 (0.0)	8 (66.7)	1 (8.3)	3 (25.0)	12 (100)
Total	32 (11.6)	102 (37.1)	110 (40.0)	23 (8.4)	8 (2.9)	275 (100)

Percent of group responses in parentheses.

Only a few responses were either Distributional or were Initial Distributional with an Initial Global focus, yet all favored the Life Line recommendation. Most of the Initial Global responses and all of the Distributional responses incorporated a comparison of centers into their overall comparisons. It was assumed that participating students generally understood the task, even though slightly more than 11% provided Idiosyncratic responses, most were from the two “GS” groups and many also were articulate yet exclusively focused on the context of the problem. Overall many students also wrote that they wanted more information about the ambulance services and “ideal” response times for ambulances. A potential for future research would be to include a secondary question that includes what would be considered an ideal maximum response time and then ask the students to reconsider their response. It

seems likely that, in that described situation, many of those who focused on comparing proportions would use that maximum time as their “cut-point” and hence produce strong agreement among those responses.

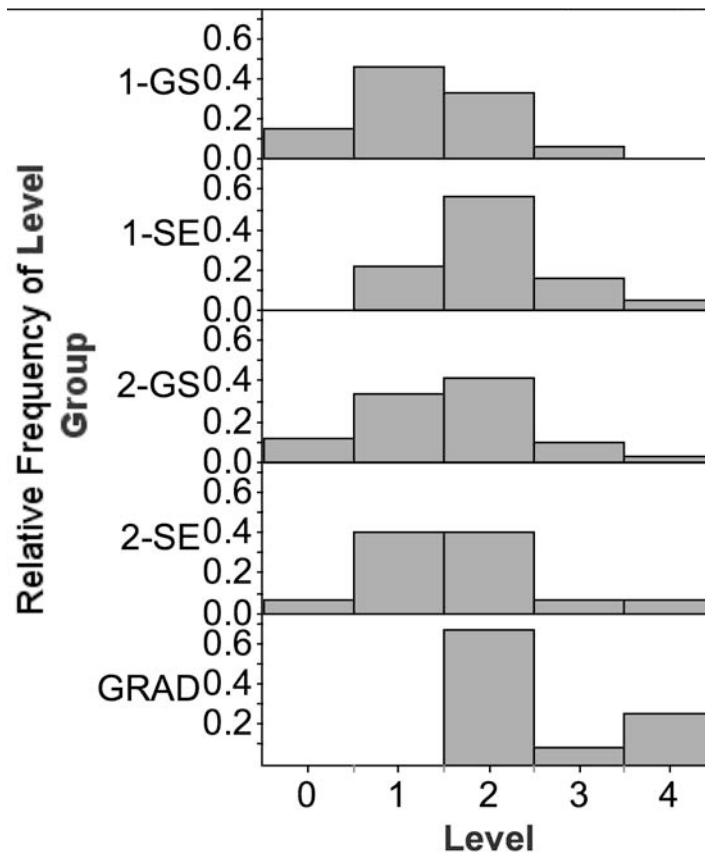


Figure 39. The distribution of response levels to the Ambulance task, separated by group.

The students from the GRAD group appeared to consistently provide responses at higher levels of the Lattice Structure Framework than all other groups. All the GRAD students responded at level 2 or higher, with the highest percentage at level 4. The 1-GS, 2-GS, and, surprisingly, the 2-SE students responded at quite similar levels of the framework, although the 1-GS was the only group to provide more level 1

responses than level 2 responses and the 2-SE provided equal frequencies of level 1 and level 2 responses. The 1-GS students provided more than 50% below level 2, while the 1-GS and 2-SE students provided more than 40% of their responses below level 2.

Although both of the “GS” groups responded at rather low levels, the mean level for the 2-GS student responses, of 1.54, was significantly higher than the mean level for the 1-GS students, of 1.29 ($t = -1.93$, $p = 0.023$). Although the difference in mean levels is significant, it does seem a bit surprising that the difference is not larger considering that the 2-GS students have had quite a bit more statistics instruction than the 1-GS students. The 1-SE students responded at a mean framework level of 2.05, also significantly higher than the mean response level of the 1-GS students ($t = -5.25$, $p < 0.01$). As students from both the 1-GS and 1-SE groups had only completed about 2 or 3 statistics classes at the time they took the survey, the 1-SE students did have a stronger mathematics background, which may be a contributing factor to the 1-SE students’ higher mean response level. When the 1-SE response levels are compared to the 2-SE response levels using Figure 39, the distribution of responses across framework levels appear quite different for each group, as the 1-SE responses appear more clustered around level 2 while the 2-SE responses are more spread out but primarily at level 2 and 1. The mean response level for the 2-SE students of 1.67 is indeed considerably lower than the 1-SE students’ mean level of 2.05. This was a rather surprising result as the 2-SE students generally have more statistics and mathematics in their background than the 1-SE students. Finally, the GRAD students

mean response level of 2.58 was considerably higher than all other groups. It appears that the advance statistics background of the GRAD group helped them to reason at higher framework levels about the challenging comparison required in the Ambulance task.

Finally, while all groups consistently had either the highest frequency of decisions in favor of ‘Life Line’ or split their decisions about equally, the 1-GS students tended to approach the task from perspectives aligned with lower framework levels, Local and Transitional, while the Grad students tended to approach the task from perspectives aligned with higher framework levels, at the Transitional level and above, and the remaining groups tended to approach the task from perspectives aligned with the Transitional framework level and Local framework level.

Survey Results: The Ambulance task –

Without descriptive statistics (Task 5) vs. With descriptive statistics (Task 6)

The results from survey task 5 and how they compare to the results from task 6 are separated into two subsections for convenience purposes. First, the decisions for recommendations of ambulance service are compared and discussed for both tasks. Second, the levels of responses that the students provided for both tasks are compared and discussed. Overall, the “with statistics” responses tended to be briefer than the “without statistics” responses as many of the “with statistics” responses merely referenced certain descriptive statistics with no further explanation.

Ambulance survey task recommendations: Without statistics vs. With statistics

On the whole, after students viewed some descriptive statistics associated with each set of response times, there appeared to be a clear shift of recommendations from ‘Speedy’ to ‘Life Line.’ Almost 95% of students who initially recommended ‘Life Line’ stayed with ‘Life Line,’ and about 56% of students who initially recommended ‘Speedy’ switched to ‘Life Line,’ but that means about 44% of students who initially recommended Speedy stayed with their recommendation after viewing the descriptive statistics. Table 38 displays the specific decision shifts, by group, for before and after viewing the descriptive statistics. For each group, a two-proportion z test was used to compare the proportion of responses that recommended the Life Line ambulance service from before and after the students had access to descriptive statistics.

The two ‘GS’ groups and the 2-SE group were the only ones to have a significant increase of the proportion of students who switched their initial decision from Speedy to Life Line. For the 1-GS group, $z = -3.51$ and $p < 0.01$, for the 2-GS group, $z = -3.23$ and $p < 0.01$, and for the 2-SE group, $z = -2.47$ and $p < 0.01$. It was not surprising that the students from the 1-SE and GRAD groups did not switch to Life Line in statistically significant proportions as both groups of those students initially had decided in favor of Life Line at about a 2 to 1 rate.

Table 38.

Decisions by group for the Ambulance tasks: Counts for
Without statistics vs. With statistics

		Ambulance Recommendation with Statistics			Group
		Life Line	Speedy	Total	
Ambulance Recommendation without Statistics	Life Line	70	3	73	1-GS (n=137)
	Speedy	31	33	64	
	Total	101	36		
	Life Line	23	2	25	1-SE (n=37)
	Speedy	7	5	12	
	Total	30	7		
	Life Line	34	3	37	2-GS (n=74)
	Speedy	22	15	37	
	Total	56	18		
	Life Line	8	0	8	2-SE (n=15)
	Speedy	6	1	7	
	Total	14	1		
	Life Line	8	0	8	GRAD (n=12)
	Speedy	3	1	4	
	Total	11	1		

Almost 96% of students from group 1-GS who initially recommended Life Line stayed with Life Line and 48% of students who initially recommended Speedy switched to Life Line, after viewing the descriptive statistics. Almost 92% of students from group 2-GS who initially recommended Life Line stayed with Life Line and almost 60% of students who initially recommended Speedy switched to Life Line, after viewing the descriptive statistics. From group 2-SE, all of the eight students who initially recommended Life Line stayed with Life Line, but six of the seven students (85.7%) who initially recommended Speedy switched to Life Line, after viewing the

descriptive statistics. Although the proportion of 1-SE and GRAD students who switched from Speedy to Life Line was not statistically significant, those proportions followed a similar trend to the other groups. The initial frequencies of students from the 1-SE and GRAD group who recommended Speedy were relatively low, but the percentage of those who switched to Life Line from Speedy were still relatively high at 58% for 1-SE and 75% for GRAD. The vast majority of students who either stayed with their Life Line decision or switched to Life Line referenced its higher mean of the response times. Most of these students exclusively focused on comparing the means, yet for those who responded at Initial Distributional or Distributional levels also incorporated a comparison of means into their support of the Life Line recommendation.

The percentage of students who switched from recommending Life Line to recommending Speedy was very low across all groups, however considerably more students initially chose Speedy and then stayed with Speedy. Most of the responses that stayed with Speedy were provided by each of the two “GS” groups at the fairly high percentages of 24% for 1-GS and 20% for 2-GS. Of those responses that stayed with Speedy after having access to the descriptive statistics, approximately $\frac{1}{6}$ were Idiosyncratic, $\frac{1}{3}$ were Local and $\frac{1}{2}$ were Transitional with slightly more of the Transitional responses focused on comparing medians than other features of the distributions.

Ambulance survey task responses: Without statistics vs. With statistics

For this follow-up task, the Ambulance task with descriptive statistics, just as with the Pink/Black task with descriptive statistics, the code ‘N/A’ was added at level 2 to account for responses that appeared to merely recite the statistical terms or refer to “the statistics” in a meaningless and uninformative way. For example, “all of the data confirms that Speedy has a better response time,” was coded N/A. Also, responses that list some terms yet the values for some are clearly not “better,” such as, “lower mean, mode and median” is in support of the Life Line recommendation, yet the median is lower for Speedy. In general, there were relatively few of these types of responses and thus were not addressed on the following discussion and analysis.

The distributions, across framework levels, of responses given by students from the 1-GS group to both tasks 5 and 6 are shown in Figure 40. From those graphs, it appears that the 1-GS students tended to provide responses at higher levels after having access to some descriptive statistics. The shift appears to be away from level 1 type responses and toward level 2 type responses. Overall, a paired two-sample t-test reveals that the mean response level, of 1.88, with statistics is significantly higher than the mean response level, of 1.29, without statistics ($t = -6.75, p < 0.01$). Thus, the 1-GS students not only decided in favor of Life Line in statistically significant greater proportions but also responded at statistically significant higher levels of the Expanded Lattice Structure Framework. The details of this shift in responses are shown in table 39.

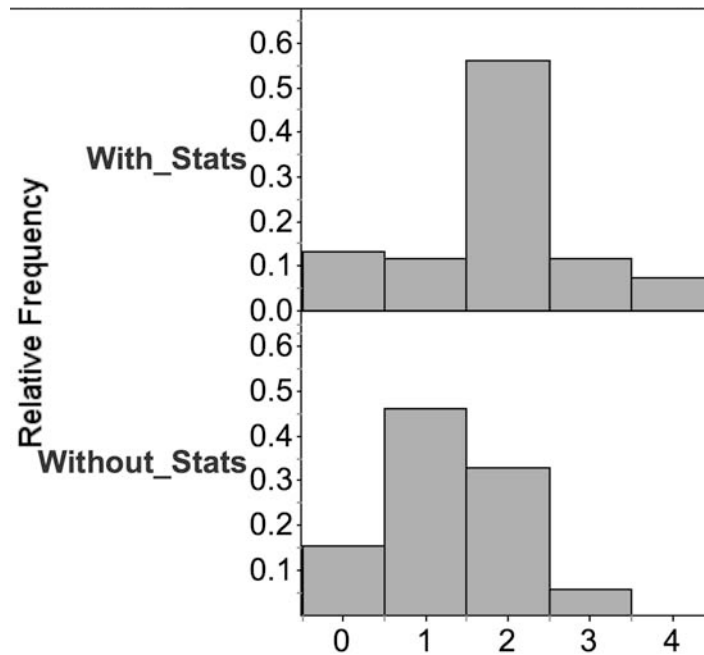


Figure 40. Response levels of the 1-GS students to both Ambulance tasks: Without statistics and With statistics.

Table 39.

Distribution of responses from the 1-GS group for the Ambulance task: Without statistics vs. With statistics.

		Response Level with stats					
		<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>Total</u>
Response Level without stats	0:	10	1	10	-	-	21
	1:	4	13	36	9	1	63
	2:	2	1	29	6	7	45
	3:	1	1	3	1	2	8
	4:	-	-	-	-	-	0
Total		17	16	78	16	10	137

The outlined cells in table 39 contain the counts of student responses that did not shift levels. Only about 39% of the 1-GS students supported their recommendation with reasons at a level equal to the reasons given before they viewed the descriptive statistics. Over 50% of the group 1-GS students, after viewing the descriptive statistics, provided reasons for their recommendation at a higher level than their initial reason. Also, after viewing the descriptive statistics, less than 10% of group 1-GS students provided reasons for their decisions that were at a lower level than their initial reasons. Each of the levels 2, 3, and 4 had considerable increases in responses after the 1-GS students had access to the descriptive statistics. In particular, about 36% of 1-GS students gave responses that shifted from other levels to level 2. The shifts in responses to higher framework levels also accompanied high success rates for recommending Life Line, as only one of the 1-GS students who responded at either level 3 or 4 also favored Speedy and about 77% of 1-GS students who responded at level 2 also favored Life Line. However, similar to when the statistics were not provided, the 1-GS students who responded at either levels 0 or 1 were about evenly split between the Life Line and Speedy recommendations.

Table 40.

Distribution of Level 2 and Level 3 responses, from the 1-GS group, for the Ambulance task: Without statistics and With statistics.

<u>Response</u>	<u>Level 2</u>					<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
without Stats	11	29	5	-	45	8	0	8
with Stats	46	5	9	18	78	0	16	16

Table 40 displays the pattern of what features of the data the 1-GS students, who responded at the Transitional and Initial Distributional levels focused on. At level 2, after having access to the descriptive statistics, considerably fewer students focused on Shape and Proportion and considerably more students focused on Center or had an Initial Global focus. Of those who focused on Center, almost all exclusively cited either the mean or “average” of the times for Life Line as lower (as opposed to the mode) with no other information. A large majority of Initial Global responses cited either two measures of center, the mean and mode, or cited a measure of center and the lower minimum and maximum times for Life Line. All of the 1-GS students who provided distributional responses after viewing the descriptive statistics also incorporated comparing the means into their responses. Thus, a majority of 1-GS students who responded at level 2 or higher, after they had access to descriptive statistics, either exclusively or in part, focused on measures of center. As the 1-GS students had limited statistical instruction, it is likely that measures of center were the statistics that they were most comfortable with and thus relied on when they had access to descriptive statistics.

The distributions, across framework levels, of responses given by students from the 1-SE group to both tasks 5 and 6 are shown in Figure 41. From those graphs, it appears that the 1-SE students tended to provide responses at higher levels after having access to some descriptive statistics. The shift appears to be away from the level 1 type responses and towards levels 3 and 4 type responses. Overall, a paired two-sample t-test reveals that the mean response level, of 2.51, with statistics is

significantly higher than the mean response level, of 2.05, without statistics ($t = -2.83$, $p < 0.01$). Thus, the 1-SE students recommended Life Line in somewhat greater proportions and responded at higher levels of the Expanded Lattice Structure Framework after they had access to descriptive statistics. The details of this shift in responses are shown in table 41.

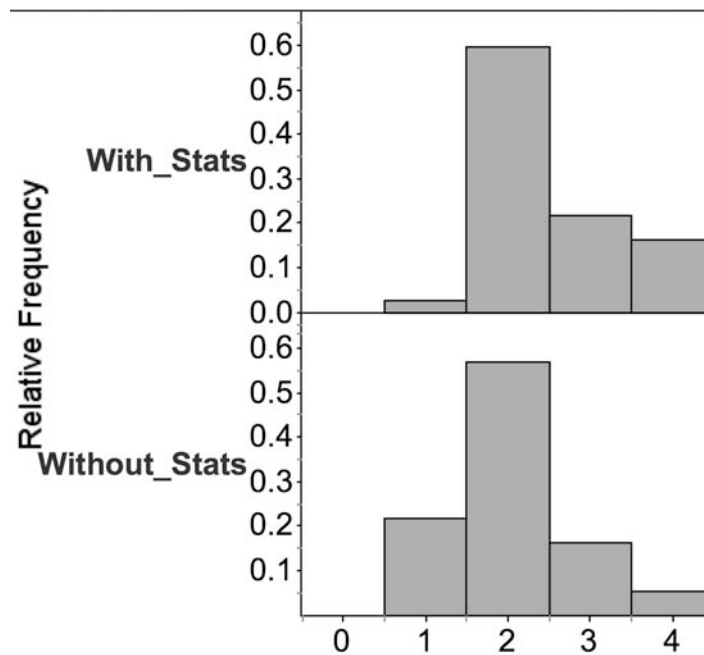


Figure 41. Response levels of the 1-SE students to both Ambulance tasks: Without statistics and With statistics

After viewing the descriptive statistics, 14 of the 37 group 1-SE students (37.8%) supported their decision with reasons at a level equal to the reasons given before they viewed the descriptive statistics, as shown in the outlined cells of table 41. Eighteen of the students (48.6%) provided reasons for their decision at a higher level than their initial reason after viewing the descriptive statistics. Also, after viewing the descriptive statistics, only 5 students (13.5%) provided reasons for their decisions that

were at a lower level than their initial reasons. Each of the levels 2, 3, and 4 had increases in responses after the 1-SE students had access to the descriptive statistics. About 24% of 1-GS students gave responses that shifted from other levels to level 2. The shifts in responses to higher framework levels also accompanied high success rates for recommending Life Line as only one of the 1-SE students who responded at either level 3 or 4 also favored Speedy and about 73% of 1-GS students who responded at level 2 also favored Life Line. After having access to the descriptive statistics only one 1-SE student provided a response that was lower than level 2. This student consistently recommended Life Line before and after viewing the statistics and also consistently cited that Life Line had no times above 24 minutes, unlike Speedy.

Table 41.

Distribution of responses from the 1-SE group for the Ambulance task: Without statistics vs. With statistics.

	<u>Response Level with stats</u>					<u>Total</u>
	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
<u>Response Level without stats</u> 0:	-	-	-	-	-	0
1:	-	1	5	1	1	8
2:	-	-	13	6	2	21
3:	-	-	3	-	3	6
4:	-	-	1	1	-	2
Total	0	1	22	8	6	37

Table 42.

Distribution of Level 2 and Level 3 responses, from the 1-SE group, for the Ambulance task: Without statistics and With statistics.

<u>Response</u>	<u>Level 2</u>					<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
without Stats	9	9	3	-	21	4	2	6
with Stats	17	0	2	3	22	0	8	8

Table 42 displays the pattern of what features of the data the 1-SE students, who responded at the Transitional and Initial Distributional levels, focused on. This pattern was quite similar to that of the 1-GS students. After having access to the descriptive statistics, considerably fewer students focused on Shape and Proportion and considerably more students focused on Center or had an Initial Global focus. A large majority (more than 80%) of 1-SE students either exclusively, or in part, focused on measures of center after they had access to descriptive statistics. As the 1-SE students also had limited statistical instruction, it is likely that measures of center were the statistics that they were most comfortable with and thus relied on when they had access to descriptive statistics.

The distributions, across framework levels, of responses given by students from the 2-GS group to both tasks 5 and 6 are shown in Figure 42. From those graphs, it appears that the 2-GS students tended to provide responses at higher levels after having access to some descriptive statistics. The shifts appear quite similar to that of the 1-GS students, which are away from levels 0 and 1 type responses and towards levels 3 and 4 types of responses. Overall, a paired two-sample t-test reveals that the mean response level, of 2.28, with statistics is significantly higher than the mean

response level, of 1.56, without statistics ($t = -5.71, p < 0.01$). Thus, the 2-GS students not only recommended Life Line in statistically significant greater proportions but also responded at statistically significant higher levels of the Expanded Lattice Structure Framework, just as the 1-GS students did. The details of this shift in responses are shown in table 43.

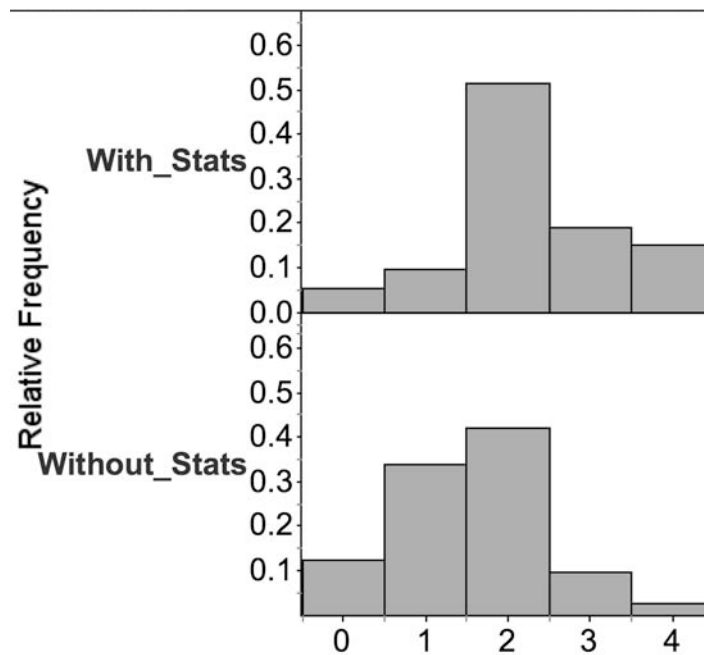


Figure 42. Response levels of the 2-GS students to both Ambulance tasks: Without statistics and With statistics

The outlined cells in table 43 contain the counts of student responses that did not shift levels. After viewing the descriptive statistics, 29 of the 74 group 2-GS students (39.2%) supported their decision with reasons at a level equal to the reasons given before they viewed the descriptive statistics. Most of the students (52.7%), after viewing the descriptive statistics, provided reasons for their decision at a higher level than their initial reason. Also, after viewing the descriptive statistics, only 6 students

(8.1%) provided reasons for their decisions that were at a lower level than their initial reasons. Those percentages are almost identical to the shift trends from group 1-GS.

Table 43.

Distribution of responses from the 2-GS group for the Ambulance task: Without statistics vs. With statistics.

		<u>Response Level with stats</u>					
		<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>Total</u>
<u>Response Level without stats</u>	0:	2	-	6	1	-	9
	1:	1	6	12	4	2	25
	2:	1	-	17	7	6	31
	3:	-	1	3	2	1	7
	4:	-	-	-	-	2	2
Total		4	7	38	14	11	74

Also similar to the 1-GS group, the 2-GS group saw considerable increases in responses at each of the levels 2, 3, and 4 had after they had access to the descriptive statistics. The shift to higher level responses for the 2-GS group was slightly different than that of the 1-GS group as the 2-GS students' increase in responses that shifted from other levels to level 2 was less than that of the 1-GS students but the shifts to levels 3 and 4 from other levels was higher for the 2-GS students than the 1-GS students. The shifts in responses to higher framework levels also accompanied high success rates for recommending Life Line as only one of the 2-GS students who responded at either level 3 or 4 also favored Speedy and about 79% of 2-GS students who responded at level 2 also favored Life Line. However, when the statistics were

not provided, the 2-GS students who responded at levels 1 were about evenly split between the Life Line and Speedy recommendations, where as after that statistics were provided all the level 1 responses from the 2-GS students favored Speedy.

Table 44.

Distribution of Level 2 and Level 3 responses, from the 2-GS group, for the Ambulance task: Without statistics and With statistics.

<u>Response</u>	<u>Level 2</u>					<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
without Stats	13	15	2	-	30	7	0	7
with Stats	24	3	2	9	38	1	13	14

Table 44 displays the pattern of what features of the data the 2-GS students, who responded at the Transitional and Initial Distributional levels, focused on. This pattern was again quite similar to that of the 1-GS and 1-SE students. After having access to the descriptive statistics, considerably fewer students focused on Shape and Proportion and considerably more students focused on Center or had an Initial Global focus. A large majority (more than 80%) of 2-GS students either exclusively, or in part, focused on measures of center after they had access to descriptive statistics. Although the 2-GS students had completed a statistics course whereas the 1-GS and 1-SE students were enrolled in their first statistics course, the 2-GS students seemed relied on measures of center when they had access to descriptive statistics, just as the 1-GS and 1-SE students did.

The distributions, across framework levels, of responses given by students from the 2-SE group to both tasks 5 and 6 are shown in Figure 43. From those graphs, it appears that the 2-SE students tended to provide responses at higher levels after

having access to some descriptive statistics, shifting away from giving responses at levels 1 and 2 and toward level 4. In a rather drastic change, the 2-SE students provided a higher mean response level, of 2.93, for task 5 (with statistics) compared to a mean response level of 1.67 for task 4 (without statistics), and a paired two-sample t -test revealed a clear significant difference between those means ($t = -4.01, p < 0.01$). Thus, as with the “GS” groups, the 2-SE students recommended Life Line in statistically significant higher proportions with descriptive statistics and also responded at a statistically significant higher mean level of the Expanded Lattice Structure Framework, with descriptive statistics. The details of those shifts in responses are shown in table 45.

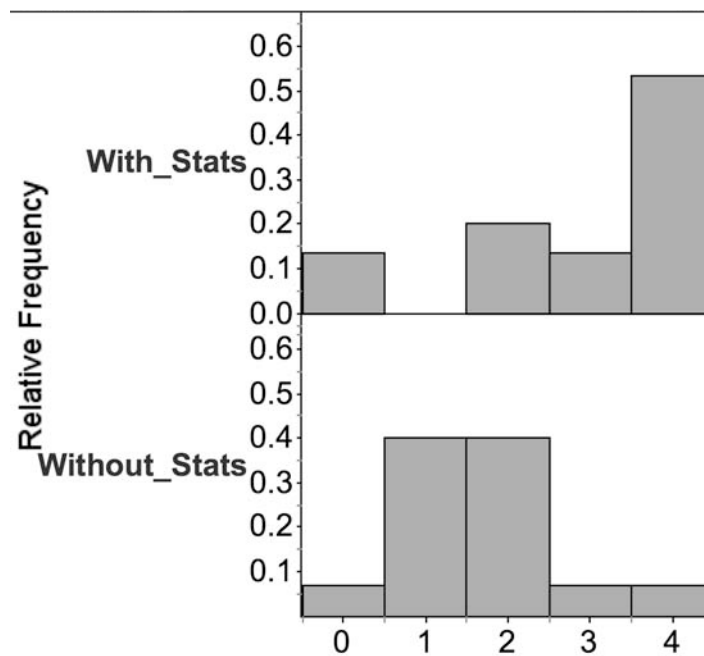


Figure 43. Response levels of the 2-SE students to both Ambulance tasks: Without statistics and With statistics

Table 45.

Distribution of responses from the 2-SE group for the Ambulance task: Without statistics vs. With statistics.

	<u>Response Level with stats</u>					<u>Total</u>
	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
<u>Response Level without stats</u> 0:	1	-	-	-	-	1
1:	1	-	1	2	2	6
2:	-	-	2	-	4	6
3:	-	-	-	-	1	1
4:	-	-	-	-	1	1
Total	2	0	3	2	8	15

After viewing the descriptive statistics, four of the 15 students from group 2-SE (27%) supported their recommendation with reasons at a level equal to the reasons given before they viewed the descriptive statistics, as shown in the outlined cells of table 45. Another two-thirds of the 2-SE students provided reasons for their decision at a higher level than their initial reason, with only one lower, after viewing the descriptive statistics. Almost half of the 2-SE students provided a response that shifted to level 4 after viewing the statistics and, as one student provided a level 4 response both before and after viewing the statistics, over half of the 2-SE students responded at level 4 to task 6. All of those level 4 responses incorporated a comparison of both the mean and a measure of variation, at least. The one 2-SE student who provided an Idiosyncratic response also recommended Life Line, but it appears that the student did

not attempt the tasks seriously as his or her written reasons were “Safety in number one.” and “Safety” for without and with statistics, respectively.

Table 46.

Distribution of Level 2 and Level 3 responses, from the 2-SE group, for the Ambulance task: Without statistics and With statistics.

<u>Response</u>	<u>Level 2</u>					<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
without Stats	3	1	2	-	6	1	0	1
with Stats	2	0	0	1	3	0	2	2

Table 46 displays the pattern of what features of the data the 2-SE students, who responded at the Transitional and Initial Distributional levels, focused on. Only five 2-SE students responded at the Transitional level, so any patterns are tentative at best. However, after having access to the descriptive statistics, none of the 2-SE students focused on Shape or Proportion, two of the three level 2 responses focused on Center, and both Initial Global responses included comparisons of the mean. Thus, similar to the responses of the “GS” and 1-SE students, 80% of the 2-SE students either exclusively, or in part, focused on measures of center after they had access to descriptive statistics.

The distributions, across framework levels, of responses given by students from the GRAD group to both tasks 5 and 6 are shown in Figure 44. From those graphs, it appears that the GRAD students also provided responses that shifted to higher framework levels, from level 2 to level 4, after having access to some descriptive statistics. The GRAD students had the highest mean response level of 3.08, out of all groups for task 6. Although the GRAD students’ mean response level

without descriptive statistics was 2.58, a paired two-sample t-test revealed no significant difference between those means without and with statistics ($t = -1.48$, $p = 0.083$). Thus, almost all of the GRAD students recommended Life Line with or without descriptive statistics and although their mean response level, of the Expanded Lattice Structure Framework, increased from without to with descriptive statistics, it was not a statistically significant increase. The details of the shifts in responses that did occur are shown in table 47.

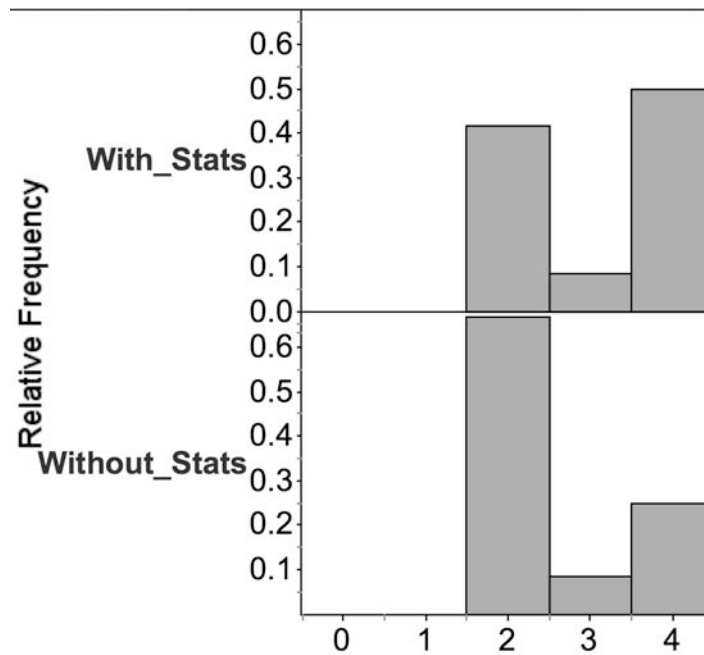


Figure 44. Response levels of the GRAD students to both Ambulance tasks: Without statistics and With statistics

After viewing the descriptive statistics, half of the 12 students from the GRAD group supported their recommendations with reasons at a level equal to the reasons given before they viewed the descriptive statistics, as shown in the outlined cells in table 47. Five of the group GRAD students (42%) provided reasons for their decision

at a higher level and only 1 provided reasons at a lower level than their initial task 5 reasons, without descriptive statistics. Four of the five students whose responses shifted higher, had their “with statistics” responses classified at level 4. All of those level 4 responses incorporated a comparison of both the mean and a measure of variation, at least.

Table 47.

Distribution of responses from the GRAD group for the Ambulance task: Without statistics vs. With statistics.

	<u>Response Level with stats</u>					<u>Total</u>
	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
<u>Response Level without stats</u>	0:	-	-	-	-	0
	1:	-	-	-	-	0
	2:	-	-	4	1	3
	3:	-	-	-	1	1
	4:	-	-	1	-	2
	Total	0	0	5	1	6

Table 48.

Distribution of Level 2 and Level 3 responses, from the GRAD group, for the Ambulance task: Without statistics and With statistics.

<u>Response</u>	<u>Level 2</u>					<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>N/A</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
without Stats	0	7	1	-	8	0	1	1
with Stats	1	1	1	2	5	0	1	1

Table 48 displays the pattern of what features of the data the GRAD students who responded at the Transitional and Initial Distributional levels focused on. Only

five GRAD students responded at the Transitional level and one at the Initial Distributional level, so any patterns are tentative at best. However, after having access to the descriptive statistics, only one of the GRAD students focused on Shape and none focused on Proportion, just one of the five level 2 responses focused on Center, and the one Initial Global responses included comparisons of the mean. Thus, only two of the GRAD students focused on measures of center at either level 2 or 3, after they had access to descriptive statistics.

So, after having access to the descriptive statistics, all the GRAD students recommended Life Line, they provided the highest percentage of Distributional responses and no Local or Idiosyncratic responses, and also had the lowest percentage of student supply reasons that were exclusively focused on measures of center. The GRAD students advanced statistical background appears to have contributed to their consistently high level responses on both Ambulance tasks, without and with descriptive statistics.

Survey response summary: Ambulance tasks without and with descriptive statistics

After examining the responses across the groups, for task 5, and the shifts in responses from task 5 to task 6, several trends emerged. Both the mean response level and the percentage of students who recommended the Life Line ambulance service increased across groups, with statistically significant increases in mean levels for all the groups except GRAD. Groups 1-SE and GRAD had high percentages of students responding who recommended Life Line for task 5 and thus the increases in their percentage of Life Line recommendations, from task 5 to task 6, was not statistically

significant. All groups saw a decrease in students who based recommendations on comparing proportions and all groups, except GRAD, had an increase in students who exclusively compared measures of center to support their recommendation, after having access to the descriptive statistics. It is not clear why students felt compelled to abandon their proportional arguments. It is possible that they may have felt there was no obvious “cut-point” to partition the distributions, or they may have thought the citing the statistical measures was more convincing or that they were supposed to use them.

Overall, for the Ambulance task with statistics, the 1-GS students responded at lower levels of the Expanded Lattice Structure framework than any other group. It was somewhat surprising that the 2-GS students responded very similarly, although at slightly higher levels than the 1-GS students. An interesting trend reversal was that the 2-SE students appeared to respond at higher levels, with the statistics, than the 1-SE students, but without the statistics the 2-SE students seemed to have more difficulties and respond at lower level than the 1-SE students. The GRAD students advanced statistical background appears to be evident as those students consistently outperformed the other groups on both Ambulance tasks, without and with descriptive statistics. With the inclusion of descriptive statistics with the data sets, all groups recorded shifts in decisions to recommend the Life Line ambulance service with corresponding shifts in reasoning that either in part or exclusively focused on comparing centers.

Survey Responses: Cross Task Numeric Codes

A *Cross Task Numeric Code* (Expanded Lattice Framework Level 0 – 4) was assigned to each student for the dominant type of reasoning that they exhibited across all the tasks, as previously described in chapter 4. The distribution of these codes, across groups, is displayed in table 49 and figure 45. Several studies, such as Gal et al (1989) and Watson (2001, 2002), noted that, frequently, students do not consistently use the same type of strategy for comparing data sets across a series of tasks. That phenomenon was true for this study as well, particularly as reasonable responses for task 1 could be classified across levels 1 to 4 and reasonable responses for tasks 2, 3 and 5 could be classified across levels 2 to 4. The researcher of this study surmised that dominant cross task code could provide insight into the students' perspective of distributions similar to those used in this study. For example, if a student's response for the Yellow/Brown task was based on sums, then exclusively on the difference in spread for the Movie-Wait time task, and then proportions and centers, together, for the Pink/Black and Ambulance tasks, the student reasoned at higher levels of the framework on the more challenging tasks that required comparisons of unequal size data sets, and thus, the student responses at lower levels on the initial tasks likely were made from a more advance perspective such as Initial Distributional or Distributional.

Table 49.

Overall reasoning levels across groups. Cross Task Numeric Codes

<u>Group</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Group Total</u>
1-GS	4(2.9)	20(14.6)	100(73.0)	13(9.5)	0(0.0)	137(100)
1-SE	0(0.0)	0(0.0)	23(62.2)	14(37.8)	0(0.0)	37(100)
2-GS	1(1.4)	6(8.1)	47(63.5)	20(27.0)	0(0.0)	74(100)
2-SE	1(6.7)	0(0.0)	10(66.7)	2(13.3)	2(13.3)	15(100)
GRAD	0(0.0)	0(0.0)	2(16.7)	7(58.3)	3(25.0)	12(100)
Level Total	6(2.2)	26(9.5)	182(66.2)	56(20.4)	5(1.8)	275(100)

Quantities in parentheses represent percent of each group total.

Figure 45 shows that the distributions of cross task numeric codes for groups 1-GS and 2-GS, the groups of students who were enrolled in either the first or second term of general statistics, appear strikingly similar. Both groups had the highest percentages of students whose cross task numeric codes were at level 1. Almost 15% of the students from group 1-GS consistently provided Local type responses and about eight percent of group 2-GS's students' cross task numeric codes were at level 1. Both groups also did not have any student who consistently responded at level 4, the Distributional level. A difference between the results for those groups is that group 1-GS had a higher percentage of cross task responses below level 2 as opposed to above level 2, whereas group 2-GS had a higher percentage of cross task responses above level 2 to as opposed to below it. Thus, it is possible that the statistics course that the 2-GS students had completed may have had some impact on the 2-GS students, promoting their slightly higher responses across tasks.

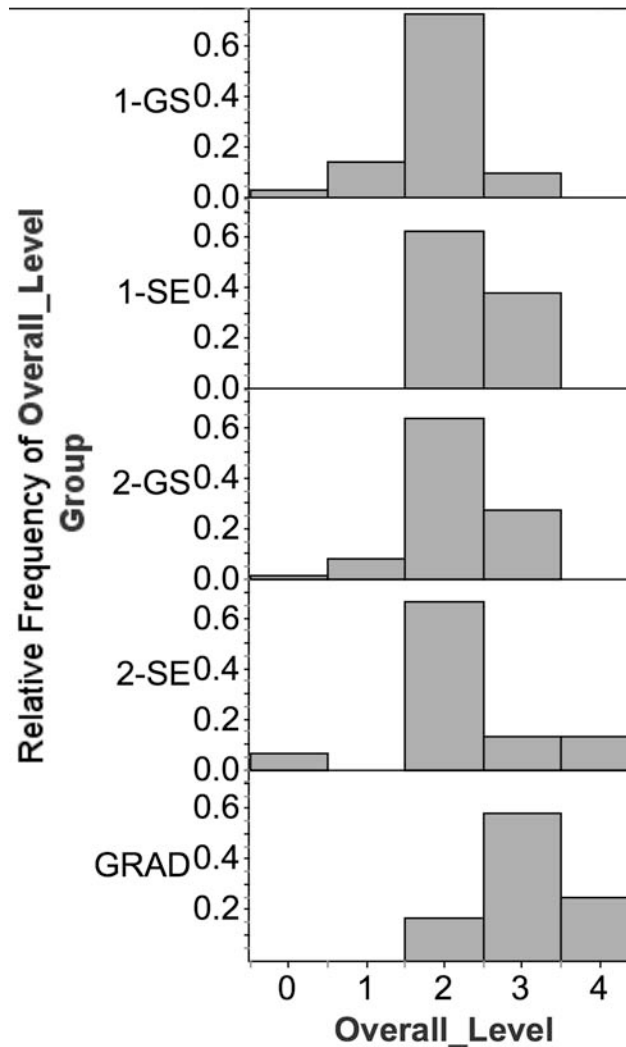


Figure 45. The distribution of Cross Task Numeric codes across groups.

From task to task, groups 1-SE and 2-SE responded quite differently in how their responses were distributed across framework levels. Neither group consistently responded at a higher mean response level from task to task. The 1-SE students were all classified, for their cross task numeric codes, at either the Transitional level, level 2 or the Initial Distributional level, level 3 with about twice as many at level 2 than at level 3. Except for one student from group 2-SE, all the students from the 2-SE group

were classified for their cross task numeric code at the Transitional level, level 2, or higher, at levels 3 and 4. The 1-SE students were beginning their first statistics class for scientists and engineers at the time they completed the survey while the 2-SE students had completed at least one prior statistics course. Thus it was rather surprising that the 2-SE students did not consistently respond at higher framework levels than the 1-SE students.

All the GRAD students were classified, for their cross task numeric codes, at the Transitional level, level 2, or higher. The 2-SE and GRAD groups were the only groups who had students who were classified, overall, at the Distributional level, level 4. Group GRAD was also the only group to have the majority of its students classified overall at level 3 as well as the only group to have the majority of its students classified at level 3 and 4 together as opposed to being classified at levels 2 and 1 or at levels 2 and 3. This trend may very well be a product of the large number of statistics courses that students from the GRAD group have completed. Thus it appears that the GRAD students' extensive statistics backgrounds has contributed to their responses, which consistently were classified at higher framework levels for each task and across tasks.

Groups 1-GS, 1-SE, 2-GS, and 2-SE all had the highest percentage of students classified for their cross task numeric codes at the Transitional level, level 2. Not all of these students responded consistently at level 2 as some provided responses that fluctuated between Local, Transitional and Initial Distributional. Yet at level 2, students who used reasoning strategies that fluctuated between levels 1 and 2 and 3

did not always refer to appropriate local reasoning strategies, particularly when comparing data sets of unequal size. This highlights the potential transitional nature of their perspective of data sets, that is, not consistently local but also not consistently global, with particular difficulties in understanding proportional reasoning and understanding statistical measures as group representatives. The GRAD students consistently responded at higher framework levels and correspondingly had the highest cross task codes and thus their perspective of data sets may tend toward a global perspective, particularly as the GRAD students consistently understood and used proportional reasoning and statistical measures as group representatives.

Interview Results

Analysis of Interviews

The analysis of interviews proceeded after the analysis of survey responses. The last four survey tasks, the Pink/Black task, the Pink/Black task with descriptive statistics, the Ambulance task, and the Ambulance task with descriptive statistics were addressed in each interview. Interviewees responded to the tasks and follow-up questions based on their responses. Additionally, interviewees were asked about their understandings of many of the statistical terms supplied in the “with descriptive statistics tasks.” The analyses of six interviewees, Jack, Amber, Eduardo, Ann, Lance, and Jill are detailed in this section along with a cross case analysis.

Background of the six interviewees

The six interviews were chosen based on the survey responses across the tasks. Jack and Amber’s responses across the survey tasks were categorized overall at

level 1, Eduardo and Ann's survey responses were categorized overall at level 2, Lance's survey responses were categorized overall at level 3, and Jill's survey responses were categorized overall at level 4. All of these six interviewees had completed their surveys during the first week of classes and then completed the interview during the third week of classes, except for Jill who was interviewed at the end of the second week of classes. All of the following background information was self-reported by the interviewees on their surveys and then confirmed at the beginning of each interview and is summarized in Table 50.

Table 50.

Background information of interviewees.

<u>Student</u>	<u>Education Level</u>	<u>Major</u>	<u>Statistics courses: Completed</u>	<u>Statistics courses: Enrolled In</u>	<u>Group</u>
Jack	Undergraduate Junior	Biology	Undergraduate: 0 Graduate: 0	Undergraduate: 1 Graduate: 0	1-GS
Amber	Post-Baccalaureate	Pre-Nursing	Undergraduate: 0 Graduate: 0	Undergraduate: 1 Graduate: 0	1-GS
Eduardo	Undergraduate Junior	Computer Science	Undergraduate: 0 Graduate: 0	Undergraduate: 1 Graduate: 0	1-SE
Ann	Undergraduate Senior	Speech and Hearing Sciences	Undergraduate: 1 Graduate: 0	Undergraduate: 1 Graduate: 0	2-GS
Lance	Graduate	Bioinformatics, Statistics	Undergraduate: 2 Graduate: 3	Undergraduate: 0 Graduate: 1	GRAD
Jill	Graduate	Statistics	Undergraduate: 6 Graduate: 14	Undergraduate: 0 Graduate: 3	GRAD

Jack and Amber were part of the 1-GS group as they both were enrolled in their first statistics class, Introduction To Probability And Statistics I for Non-Business Majors, although they had different instructors. Jack was good-natured and seemed happy to participate in the study. Jack was a junior undergraduate biology major. Amber was a post-baccalaureate who had previously earned a bachelor's degree in Liberal Arts. She had returned to school to earn a nursing degree and thus was currently enrolled in a pre-nursing program.

Eduardo was part of the 1-SE group as he was enrolled in Applied Statistics for Engineers and Scientists I. He was a junior undergraduate student who majored in Computer Science and was also working for a local high-tech company. Ann was part of the 2-GS group as she was enrolled in Introduction To Probability And Statistics II for Non-Business Majors and had previously completed Introduction To Probability And Statistics I for Non-Business Majors. Ann was a senior undergraduate student who majored in Speech and Hearing Sciences.

Lance was a graduate student who double majored in bioinformatics and statistics and was enrolled in the graduate level statistics class, Introduction to Mathematical Statistics III. Previously, he had completed two undergraduate statistics courses and 3 graduate statistics courses and thus he was part of the GRAD group.

Jill was a graduate student who majored in statistics and was enrolled in three different graduate level statistics classes, Introduction to Mathematical Statistics III, Applied Regression Analysis, and Theory of Linear Models. Previous to the current

school term she had completed six undergraduate statistics courses and 14 graduate statistics courses and thus she was part of the GRAD group.

Each of these six interviewees' responses to their surveys and interviews will be examined next. Each interview concluded with the interviewee discussing his or her understanding of the descriptive statistics included with survey task 4: Pink/Black with descriptive statistics and with survey task 6: Ambulance with descriptive statistics. These responses were analyzed first, then used to aid in informing the analysis of the responses to the tasks.

Cross case analysis of interviewees

Each of the six interviews analyzed for this research provided considerably more information about the students' reasoning than was obtained from the surveys alone. Jack and Amber were members of the 1-GS group A and their responses across the survey tasks, their cross task numeric codes, were categorized at level 1. Yet in their interviews, they both attempted to employ some of their newly acquired knowledge from their statistics class and provided responses at a generally higher levels than their survey responses. Eduardo was a member of the 1-SE group and his survey responses across the tasks were categorized at the cross task numeric code of level 2. Eduardo's interview responses, while more detailed, were quite consistent with his survey responses. Ann was a member of the 2-GS group and her survey responses were categorized at the cross task numeric code of level 2. Her interview responses were also strikingly similar to her survey responses. Lance and Jill were both members of the GRAD group but Lance's survey responses were categorized at

the cross task numeric code of level 3 while Jill's survey responses were categorized at the cross task numeric code of level 4. In the interview Lance responded similarly to his survey but added considerably more details in the interview that supported increasing his cross task numeric code to level 4 and Jill's interview responses were slightly different and while they were still global in nature they fluctuated between level 3 and level 4 and thus Jill's cross task numeric code for her interview responses was lowered to level 3. The following sections present some details behind these results and a discussion of how the students' responses support the expansion and refinement of the Lattice Structure Framework.

The interviewees' understandings of statistical terms

Each interview concluded with the interviewee discussing his or her understanding of the descriptive statistics included with survey task 4: Pink/Black with descriptive statistics and with survey task 6: Ambulance with descriptive statistics. These responses were analyzed first, then used to aid in informing the analysis of the responses to the tasks.

Jack

Jack's introductory statistics class had introduced all of the terminology used in the tasks with descriptive statistics except for *kurtosis*, *standard error of the mean*, and *sample variance*. His class had been introduced to *variance*, but not *sample variance*. He was able to recall how to calculate most of those statistical measures but had considerable difficulty expressing what they meant.

Jack was most familiar with the measures of center, i.e., mean, median, and mode. He understood how to compute those measures as he literally interpreted them as “numbers.” For example, when asked what the mean means, he replied that, “If you take all the data, add it together and divide by the amount that was entered into the equation, you get the mean.” He was then asked about the usefulness of knowing the mean. His response was somewhat algorithmic in nature. He said that finding the mean, “gives us an average. It gives us a number.” Then when he was asked why or how finding the average is helpful he responded, “You can compare it to other numbers...It’s more concrete than just looking at a graph or just raw data. It gives you a number.” Similarly, when Jack was asked about some of the measures of variation, he only discussed them in terms of computations. Jack appeared to have some recollection of how to find skewness in relative terms, i.e. “equal or skewed to one side or the other” but could not express any understanding beyond that.

Jack’s perspective of statistical measures appears to be similar to that described by Bakker and Gravemeijer (2003, 2004), that is, the data are seen as individual values that are used to calculate the numerical quantities of mean, median, range, etc. Thus for Jack measures of shape, center and spread appear to be merely the results of computations not global features of a distribution, for example, the mean is merely the result of an operation on the individual values of the data set.

Amber

Amber’s introductory statistics class had introduced all of the terminology used in the tasks with descriptive statistics except for *kurtosis*, *standard error of the*

mean, and *sample variance*. Her statistics class had introduced *variance*, but not *sample variance*. After she briefly reviewed the list of terms she indicated that she thought that she had a “pretty good idea” of what most of the terms meant, except that she only had a general idea of what standard deviation and sample variance are and had no idea what standard error of the mean and kurtosis are.

Amber was most familiar with the measures of center and understood how to calculate them. Unlike Jack, she described the mean and median, in part, as,

...a good way to start. It's a good way to summarize all the data that you have. It's like, you have all these different answers and all these different data points, but if you could summarize it to one thing [laughs] it'd be the median or the mean.

That description, in particular, is evidence that Amber potentially views the mean as a group representative. Amber did discuss standard deviation as a measure of variation from the center or from an expected center, although not necessarily an average deviation from the center. She also related sample variance to differences between pairs of data points which is conceptually quite different than assessing the differences between data points and a center. Although Amber was not specifically asked about skewness during that segment of the interview, it was addressed during the exchanges on both the Pink/Black and Ambulance tasks. In both cases she related skewness to situations when there are a few outliers on one specific side of the majority of the data and those outliers “pull” the mean toward them.

Overall, Amber indicated that she had been exposed to all of the statistical terms used in the tasks except *kurtosis*. In particular, she knew how to calculate the

measures of center and understood mean and median as global characteristics of the data. She appeared to have partial, or evolving, understandings of measures of variation and shape, that is, understanding in a transitional state.

Eduardo

Eduardo's introductory statistics for scientists and engineers course had introduced all of the terminology used in the tasks with descriptive statistics except for *kurtosis*. During this initial examination of the statistical terms, he equated the mean to average, he possibly mixed the meanings of median and mode and associated the standard deviation with "how much variance there is." Eduardo had trouble articulating his understanding of the meanings of these terms.

Eduardo understood the algorithm for calculating the mean and tentatively described how the mean is different from the median as,

It is different because, I guess, because the middle of the data basically gives you, it doesn't really give you an accurate reflection of the entire set. Where as the mean, you know, you get the entire picture and the median is, you know, whatever is in between.

Thus, he potentially understood the mean as a group representative, but not the median. Eduardo's employment background in the high tech industry is evident from his description of variation as he initially described standard deviation as, "how much variance there is," but when asked to explain his description, he sketched Figure 46 and said,

The standard deviation is, I guess, um, so you have a target and how much it deviates would be your standard deviation...Let's say you are doing an experiment and you want to get some kind of output to be, let's say some x-value and you want that deviation to be within that x-

value or pretty close...so I guess to get the deviation you could just sort of see how far apart they are?

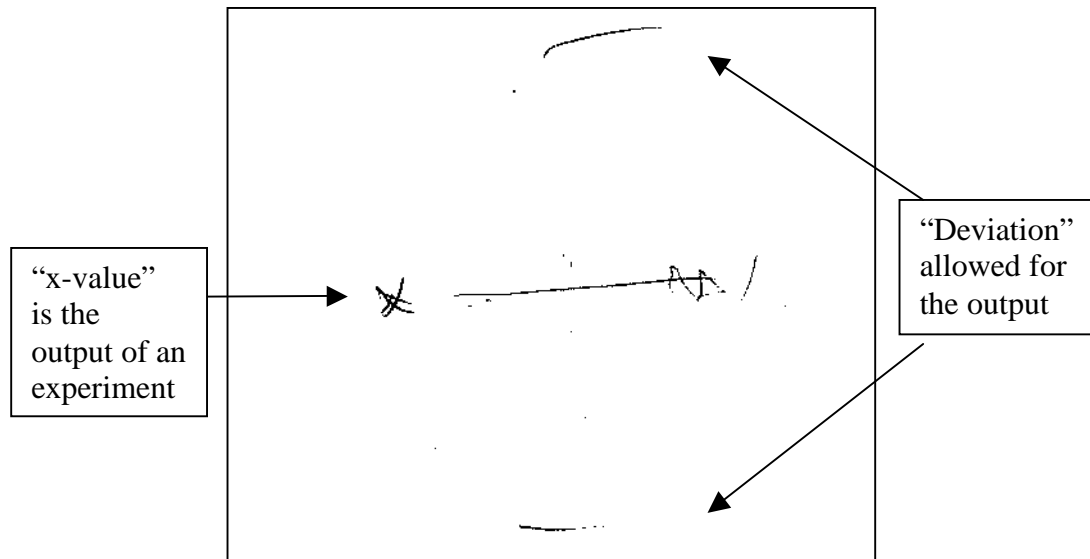


Figure 46. Eduardo's sketch of standard deviation.

Eduardo went on to relate standard deviation to “consistency,” and range to “a picture of the entire spectrum,” and sample variance to the difference in the size of two data sets. Thus while it’s possible that he had correct intuitions about the standard deviation and the range, he was confused about the sample variance. Eduardo was not specifically asked about his understanding of measures of shape.

From his responses it appears that Eduardo may not have had a solid understanding of the computations of many of the descriptive statistics nor had a clear understanding of their meanings. He did attempt to associate a naïve global meaning to mean, standard deviation and possibly range. This is an indication that Eduardo reasons transitionally about the statistical measures.

Ann

Ann was enrolled in her second statistics course, and thus likely had previous experience with statistical problems that employed all or some of the terminology used in the tasks with descriptive statistics. Ann's responses concerning the meanings of the statistical terminology used in the interview indicate that she had been introduced to these terms but she appeared to have a limited and confused working knowledge about them.

Ann was able to describe the algorithm for calculating the mean, but described characteristics of the mean only as related to the "middle" and "most" for a normal or slightly skewed distribution and thus it was not clear that she differentiated the mean from the median from the mode. She related the usefulness of the mean to using it to compare a single test score to all the test scores of a particular group. She said,

... If you have the mean and you performed higher or below you can know where you are in comparison to everybody else.

This interpretation, of how the mean is used, appears to be focused on intra-group comparison as opposed to inter-group comparison. She described the use of standard deviation as related to how far a specific value is from "zero," such as in her following explanation with her accompanying sketch in Figure 47,

OK, so you have a bell curve and you have zero and say I'm giving you the standard deviation right here...That's my first one, and then I have my second one and it goes on from there...If I'm too far over that's a bad thing. So over would be bad. But the closer I am is good.

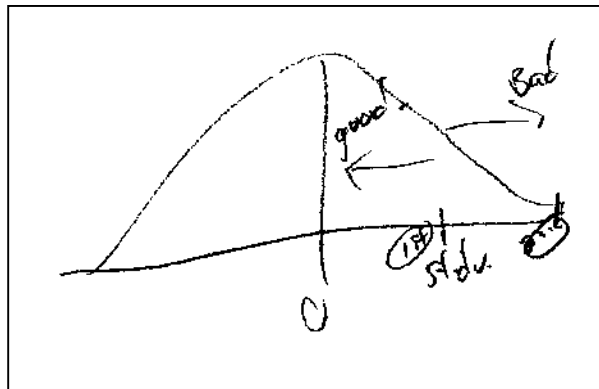


Figure 47. Ann's sketch of standard deviation.

Ann also frequently referred to “zero” as an idealized score or value for the data, not necessarily the mean of the data. Concerning the range, Ann’s description was connected to how many distinct outcomes there are between the extremes, a distinctive Local perspective. Ann was not specifically asked about her understanding of measures of shape.

Ann’s responses concerning the meaning and usefulness of the mean were initially algorithmic but then showed that she understood the mean, tenuously, as a group characteristic as she also conflated the mean with the median and mode and only discussed any of the measures of center in the context of bell shaped or slightly skewed distributions. When discussing the terms related to variation, she seemed to know the algorithms but could not articulate the normative uses or meanings. For the range, she possibly held a local view, that is, when describing range she related it to how many discrete outcomes there were in the data. When discussing the meaning of standard deviation she seemed to relate it to an assessment of how far away a specific

data point is from an idealized value. It was not clear that Ann understood standard deviation as a global characteristic of a distribution.

Lance

Lance had previously completed several undergraduate and graduate statistics courses and thus likely had experience, from several classes, using all of the terminology in the tasks. Lance had no trouble correctly defining the algorithmic calculations of most of the statistical measures used in the tasks. He did admit that he would need to “to review the details of skewness and kurtosis to use them accurately.” Lance also had no trouble providing meanings and examples of the statistical terms in ways that a novice statistics student could understand, which was the context that Lance was asked to place his responses in.

Lance not only knew how to calculate the various measures of center but his description of their uses indicated that he had a sophisticated understanding of the mean and median as group representatives. During the conversation, Lance provided the following example related to the meaning of the mean:

How tall are people? Eh, people are about five and a half feet. It doesn't say how far they're spread but if you need a good ballpark figure of how tall people are, you know, people are about five and a half feet.

Lance continued on and differentiated the measures of mean and median by describing them as capturing different places of a distribution and also implying that they are group representatives. Concerning standard deviation, Lance provided a definition of in the context of explaining it to a novice statistics student. He said that,

That's a measure of how many of your observations fall within a distance of the mean. Of the, ah, whatever units you are measuring in, you can say how many of them will be within a distance of that mean using the standard deviation.

As this explanation is intended for a novice statistics student, it is difficult to assess if Lance was considering standard deviation as a global characteristic of the distribution, yet it does seem likely that he was attempting to explain it that way. Lance was not asked specifically about skewness and kurtosis.

Lance's responses were evidence that he understood the measures of center and measures of variation from a global perspective, as group characteristics; however, his description of standard deviation did not necessarily provide specific evidence that he understood it as a global characteristic.

Jill

Jill had previously completed many undergraduate and graduate statistics courses and thus likely had experience, from several classes, in using all of the terminology in the tasks. She said that she understood all the statistical terms presented in the tasks, but then admitted that her understanding of skewness and kurtosis was minimal. Jill also had no trouble providing meanings and examples of the statistical terms in ways that a novice statistics student could understand, which was the context that she was asked to place her responses in.

Jill understood the details of the computations for the mean and median and her responses somewhat indicated that she understood them as global characteristics. She provided a description of the mean that was similar to the "fair share"

interpretation as described by Mokros and Russell (1995). Jill also appeared to understand calculations for finding standard deviation, sample variance and interquartile range; for example, Jill described the standard deviation as “The average deviation from the center.”

Jill’s explanations for the meanings of the descriptive statistical terminology showed that she has detailed understandings of most of the measures and can perceive them as global representatives. Jill also had no trouble providing meanings and examples of the terms in ways that a novice statistics student could understand.

Summary of interviewees’ understandings of statistical terms

Jack, Amber and Eduardo all were enrolled in their first statistics course and had been attending classes for about three weeks at the time of the interview, yet they all had rather different understandings of statistical measures. Jack’s understanding focused on computations. Amber gave some specific explanations that related measures of center to group representatives, but did not clearly have a similar understanding of other measures. Eduardo, although confused about some of the calculations, appeared to have a tenuous understanding of the mean, not the median, as a group representative as well as a tenuous understanding of measures of variation as group features.

Ann had completed one general statistics course and was enrolled in her second. Her explanations for the meanings of various measures were often confusing. She mixed up terms and tended to describe measures tentatively as group features, but only in the context of intra-group comparisons as opposed to inter-group comparisons.

Lance and Jill were both graduate students and had completed many undergraduate and graduate level statistics courses. Although their descriptions were slightly different, they both tended to describe various measures of center and variations as group features.

Responses to task 1: the Yellow/Brown task

The data sets in Yellow/Brown task were of equal size and had equal centers and similar uni-modal shapes (see appendix B). The range of the scores for the Yellow class was slightly smaller than the range of scores for the Brown class. Because the classes had the same number of scores comparisons of sums of frequencies would produce the same conclusions as comparing the means or proportions. This task was only addressed in the interview if there was extra time at the conclusion.

Jack

On the survey Jack wrote that the Yellow class scored better because “the average is higher for the yellow.” This response was coded at level 1 because it was deemed similar to other responses that indicated that the mode for the Yellow class had a higher frequency than the mode for the Brown class. This type of reason, that compares frequencies of specific values, is indicative of a Local view of data sets.

There was time at the end of Jack’s interview to review Jack’s written survey response to this task. Jack did confirm that he meant, “The mode was taller.” When asked how he would decide now, he indicated that he would still decide that the Yellow class scored better, but his reason changed. His new reason was, “Yellow, because it has more scores at five and above,” similar to the “cut-point” reasoning

previously described. That statement is true, the Yellow class has seven scores at five or more correct, while the Brown class has six scores at five or more correct. Also, given that the Yellow class and the Brown class are of equal size, this reason is appropriate. The response is based on a comparison of frequencies, but there is potential for this type of reason to be proportional because equally sized data sets imply that comparing the frequencies above the score of five and comparing the proportions above five will produce results that are only different by a scalar. The lower level code was assigned because of the absence of evidence to support assigning a higher code.

Amber

Amber decided that the classes scored equally well because “Though the Brown class had one high and one low score, they essentially cancel each other out.” This response was coded at level 1, local, because of its focus on the location of specific values. This type of response does show potential for more sophisticated reasoning but without follow-up questions a level higher than level 1, Local, cannot be inferred.

Fortunately, there was time at the end of the interview to address Amber’s survey response to task 1. In the follow-up, Amber was shown the task and her survey response, then asked to describe why she had “canceled” the one high score and one low score from the Brown class. Her explanation was that, “there is always like an exception to the rule and these are the exceptions.” Then she continued by explaining how she made her decision, “In general, the [Brown] class did pretty good, but why

should you have, like, one low scoring student and one exceptional one effect the overall assessment?" Thus, Amber did not average the two end points she removed them from consideration. Amber's reference to an "overall assessment" is possibly based on a global assessment rather than a local one but it was not clear exactly how she made her overall assessment. Her interview response confirmed the categorization of her survey response at level 1.

Eduardo

Eduardo made the same decision as Amber, that the classes scored equally well, but Eduardo's written reason for his decision referred to comparing the averages of each class. He also made a separate comment that a different conclusion could be reached if variation was considered instead of averages. He wrote,

A mental addition indicates that both classes score an equivalency in value, meaning that both classes have the same average score. Although the variation in score might perceive an entirely different perspective.

Eduardo seemed to have considered his two reasoning strategies separately and felt that he could only use one or the other, not both. Thus Eduardo's response was Transitional, level 2 focused on center. Although he did consider comparing variation, that comparison was not integrated with his comparison of centers. Eduardo's written responses to the Yellow/Brown task were not reviewed in the interview.

Ann

On the survey, Ann decided that the Yellow class scored better than the Brown class. Ann implicitly noted that the centers were in the same location and based her

decision primarily on the frequency or height of the centers, that is, the yellow class had more fives so it was better. It is not clear if Ann is referring to medians or modes because she uses the term “medial.” Ann’s written response follows.

I choose this answer [Yellow] because in the dot plot for the yellow class there were 5 students who all received the same score of 5, and in the dot plot for the brown class there were 3 students who received the same score of 5. Even though the score of 5 is a medial score, not an exceptional one, that is what the majority of students in both classes received, hence why I choose the answer. I reasoned, what is the effectiveness of having one student receive a high grade when all the other students receive the same medial grade.

Her last statement was rather unclear and appeared to address the context of what is ‘better’ for a classroom environment. Unfortunately the researcher was unable to make sense of Ann’s last statement as it related to the graphs. Although Ann could have been thinking about the shapes of the graphs, overall her response indicated that she focused on absolute frequencies; in particular she compared the frequencies of the centers of the two groups of scores.

The data sets presented for comparison in the Yellow/Brown task were of equal size, had equal measures of center and similar shapes but differed slightly in variation. Ann’s response to this task was primarily focused on comparing absolute frequencies of the centers, i.e. their heights. Her use of the language “medial” highlights her conflation in her understanding of measures of center. Although she implied that the data sets had equal centers and focused on comparing the heights of those centers, which could be an intuitive assessment of shapes, her focus on

comparing frequencies is indicative of a Local-type, level 1 comparison strategy.

Ann's written responses to the Yellow/Brown task were not reviewed in the interview.

Lance

On the survey, Lance decided that the classes scored equally well. He cited, "same average, same total correct, different variance. Lacking more information on how to judge results, they scored equally well." Although this response was quite minimal, it was coded at level 4-distributional, because it incorporated comparing both center and variation. Lance's written responses to the Yellow/Brown task were not reviewed in the interview.

Jill

On the survey, Jill decided that the Yellow class scored better. In her written response, she described a comparison of highest scores and noted that both modes were at five, but her explanation primarily focused on the consistency of the Yellow class's scores around the score of five. She wrote,

While the brown class did have the highest score of the two classes, it was only one score. Both classes had the highest frequency at a score of 5, but the yellow class had less variation in those scores. The yellow class was more consistently near 5 than the brown class.

The primary focus of Jill's response appeared to be on consistency, with attention to modes, but without mention of the modes' alignment with the medians and means. Thus, she made no specific assessment about how the data was distributed on each side of the mode. Of course, because the shapes of the distributions were

obviously similar with their measures of center aligned, Jill could have been comparing all the centers in her thinking process but only expressed the comparison of modes. Her response was coded at the Initial Distributional level with an Initial Global focus because of the primary focus on consistency with additional focus on frequencies not explicitly on multiple, global characteristics.

Summary for the Yellow/Brown task

All the interviewees' decisions and response levels for the Yellow/Brown task are summarized in table 51.

Table 51.

Interviewees' decisions and response levels for task 1:
the Yellow/Brown task.

<u>Student (Group)</u>	<u>Format</u>	<u>Decision</u>	<u>Response Level</u>
Jack (1-GS)	Survey	Yellow	1
Amber (1-GS)	Survey	Equal	1
Eduardo (1-SE)	Survey	Equal	2-C
Ann (2-GS)	Survey	Yellow	1
Lance (GRAD)	Survey	Equal	4
Jill (GRAD)	Survey	Yellow	3-IG

Jack and Ann each decided that the Yellow class scored better based on Local, level 1, reasoning strategies. Amber also described a Local reasoning strategy but concluded that the classes scored equally well. Both Jack and Ann compared the frequencies of the centers and observed that the Yellow class had the taller mode, but Ann also considered, then discounted, making the decision based on comparing the

highest scores. Each of the reasoning strategies described by Jack, Ann and Amber were considered Local and were potentially not valid from a statistical perspective. Eduardo decided that the classes scored equally well, but unlike Amber's reasoning strategy, Eduardo based his decision on the observation that the classes had the same average and he specifically excluded considering any difference in variation. So Eduardo's reasoning strategy was classified as Transitional focused on centers and while his strategy was potentially valid, it was not necessarily made from a global perspective. Jill decided that the Yellow class scored better based on an Initial Distributional, level 3 reasoning strategy with an Initial Global focus. Jill assessed one distributional characteristic, consistency, along with local features such as comparing endpoints and modes and did not specifically integrate more than one global characteristic. Lance decided that the classes scored equally well based on a reasoning strategy classified as Distributional, level 4, although the evidence for that classification was somewhat weak. He noted that the averages and the sums were the same and that the "variances" were different. Those assessments were also not necessarily separated, as were Eduardo's comparisons, so his strategy was considered Distributional.

Jack and Amber were the only interviewees to have their Yellow/Brown survey task responses reviewed during the interview. In both cases they provided local type responses on the survey and then confirmed the interpretation of those responses during the interview. Both interviewees provided greater details of their thinking, with Jack revising his comparison strategy, and while the extra explanations showed that

the interviewees had potential for making the required comparison from a perspective at a higher framework level, their responses continued to warrant a level 1 code.

Responses to Survey task 2: the Movie-Wait-Time task

The data sets in Movie Wait-Time task were of equal size and had equal means and medians and both were bi-modal (see appendix B). The range of the wait-times for the Royal Theater was considerably smaller than the range of wait-times for the Maximum Theater. Because the theaters had the same number of wait-times recorded, comparisons of sums of frequencies would produce the same conclusions as comparing the means or proportions. Reasoning about the difference in the variation of each distribution was expected to produce claims that there is a difference in wait times while reasoning about the centers without considering the variation was expected to produce claims that there was no difference in wait times. This task was only addressed in the interview if there was extra time at the conclusion.

Jack

Jack agreed with the assertion that there was no difference in the wait-times between the Royal and Maximum Movie Theaters, and then he decided that, given the choice, he would attend the Maximum Theater. Jack's written explanation for why he agreed that there was no difference in wait-times was difficult to assess. He wrote that the student who made the claim that there was no difference in wait times must have experienced a 10 minute wait-time for each theater, a focus on the specific wait-time. His reason for choosing to go to the Maximum Theater was that there was a chance to get a shorter wait-time, a comparison of the low ends of the distributions. So, while it

is possible that Jack did not understand the first part of the question, his decision was apparently based on a comparison of specific, individual points and did not compare the data sets as wholes.

Jack's survey response to the Movie-Wait-Time task was also reviewed off camera at the end of the interview. He was shown his decision and explanation and asked if he could recall what he meant when he wrote it. He indicated that he meant that "Eddy must have been one of the 10 minute wait times at Royal and Eddy must have been one of the 10 minute wait times at Maximum." Thus, Jack's survey response was verified and considered to be a Local, level 1 type.

Amber

Amber disagreed with the assertion that there was no difference in the wait-times between the Royal and Maximum Movie Theaters. The reason Amber provided was that, "The range of the wait times are quite different." For her choice on which theater to attend, Amber chose the Royal Theater because, "Though there is a chance I would wait a short period of time at Maximum, it is not very likely. Also, there is a greater chance I would wait longer at Maximum." This response was difficult to interpret, but it is possible that Amber first addressed the positive value of waiting a short time versus the negative value of waiting a long time and thus focused on an implied comparison of the variation in the data from the two theaters. Hence her "choice" response was also focused on comparing variation. Amber's responses exclusively addressed the variation of the data. Her survey response was hence categorized at level 2 focused on variation.

Amber's survey responses to the Movie Wait Time question were briefly addressed at the conclusion of the interview. She was shown the task and her survey response. She confirmed that she disagreed with Eddy because the range was different for each theater's wait times and thus did not conclude that the wait times were about the same. Then the following exchange took place that addressed her responses to both the Yellow/Brown task and to the Movie Wait Time task:

Int: ...*you said these were about the same* [points to the Yellow and Brown graphs] *because these cancel out* [points to the high and low scores] *but you are not canceling out the Maximum minutes to say, well it's about the same as the Royal. What about those two situations are different?*

A: *Ah, well for these it was only one* [points to Brown graph] *and here there is kind of more cluster for the low wait times and the high wait times* [points to Maximum graph].

Her response above may be evidence that Amber views the data from the Maximum Theater as separated into three clusters, low, middle and high, similar to the partitioning described by Makar and Confrey (2005) as "distributional chunks." Also Amber's assessment of the data from the Royal Theater as just one cluster whose location corresponds to the middle cluster from Maximum's data is similar to a modal clump as described by Konold et al. (2002). Both of these descriptions focus on partial distributions and are potential evidence that Amber was utilizing an Informal Global view of the data, but the inarticulate nature of her response makes that type of conclusion a questionable one.

Eduardo

Eduardo disagreed with the assertion that there was no difference in the wait-times between the Royal and Maximum Movie Theaters. Eduardo's explanation cited the range for Maximum's times, "from 5 to 14 minutes" and then concluded that Royal's times were more consistent and "within a smaller percentage of the actual mean." Given the choice, Eduardo decided that he would rather attend the Royal Theater because of the consistency of Royal's times. Eduardo's response was clearly focused on variation and "consistency" but his additional observation that the data for the Royal Theater was "within a smaller percentage of the mean" is global observation and is considered an assessment of the whole distribution of Royal's wait times. Although Eduardo's reasoning strategies focused on variation for Maximum's wait-times, he extended his variation assessment to include variation about a center and thus his response was categorized at the Distributional level, level 4. Eduardo's written responses to the Movie Wait-Time task were not reviewed in the interview.

Ann

On the survey, Ann disagreed with the assertion that there was no difference in the wait-times between the Royal and Maximum Movie Theaters, that is, she believed that there was a difference in wait-times. Ann also chose to attend the Royal Theater. She observed that the times for the Royal Theatre were "closer together" with no outliers while the times for the Maximum Theater were "diverse" and "at all levels with peaks only at 2 different times." Given Ann's previous described understanding of range as closely related to 'variety,' it is likely that Ann's reasoning on this task is

intuitively focused on number different wait times and on the frequencies of individual wait times. This is particularly evident in her comment about the data for the Maximum Theater having only two peaks, yet both data sets are bi-modal.

Ann wrote a final comment that was a bit unclear and appeared to be highly contextual as she tried to understand Eddy's assertion from his perspective. She hypothesized that Eddy made his decision based on the wait time he experienced (a local view) not on the data as a whole (a global view). She wrote,

I concluded that for Eddy to say this he must have been one of the students who did not have an outlier type number. It seems more students would report otherwise because they are more likely to have had an outlier number in the first theatre Maximum than a similar number.

It is not clear whether or not Ann would also claim the wait-times were the same if she was a member of the class and experienced similar wait times at each theater. Ann's reasoning strategy appeared to be more focused on shapes than variation and thus was categorized at the Transitional level, level 2 focused on shape.

At the conclusion of the interview, there was time to review Ann's written response to the Movie-Wait-Time task. In the follow-up, Ann confirmed that she was referring to variation in her survey response, yet she also focused on comparing only the ends of the distributions. A portion of her response is given below.

I looked at the waiting time for the Maximum theater and I thought well jeez they got 5 minutes and 5 and a half minutes and that's really good, but then what if I got a 14 minute wait?...I would say that, um, the Royal waiting time is better just because it makes me think that they're trying to stay within a certain bracket. So the variation, I guess variance, on this [points to Maximum], um, it's a big range, it's a large range, like, I could get either 5 or 14 and this one [points to Royal] is so small...

In this follow-up response Ann started by considering the possibility of experiencing the shortest or longest wait-times from Maximum, a Local perspective, then considered the wait-times from Royal as a group, i.e., staying “within a certain bracket,” that bracket being the shortest and longest times, a potential global perspective. Her response was not exclusively focus on individual times or on comparing variation but did appear to be focused more on comparing variation than shapes. This response supported the categorization of her survey response at Level 2, although with a focus on variation as opposed to shape.

Lance

On the survey, Lance disagreed with the assertion that there was no difference in the wait-times between the Royal and Maximum Movie Theaters based on the difference in the “variances” of the distributions. He chose to attend the Royal Theater and inferred that smaller variance for Royal made it possible to time his arrival at the theater to minimize wait time. With out more information Lance’s overall responses were coded level 2-variation, but given the extra information from the interview about Lance’s global perspective of distributions, it is quite possible he was thinking distributionally when he referred to “variances.” Lance’s written responses to the Movie Wait-Time task were not reviewed in the interview.

Jill

Jill disagreed with the assertion that there was no difference in the wait-times between the Royal and Maximum Movie Theaters. Jill explained that while the centers were the same there were also differences in variation and shape between the two

distributions, specifically citing, “Royal theatre's wait times are nearly evenly distributed.” Similar to Lance, Jill chose to attend the Royal Theater based on being better able to predict when a movie at the Royal Theater would begin. As Jill’s reasoning strategies included comparisons of the two sets of wait times as whole units, as she attended to center, variation and shape. Her response was categorized at the Distributional level, level 4. Jill’s written responses to the Movie Wait-Time task were not reviewed in the interview.

Summary of the Movie Wait-Time task

All the interviewees’ decisions and response levels for the Movie Wait-Time task are summarized in table 52.

Table 52.

Interviewees’ decisions and response levels for task 2:
the Movie Wait-Time task.

<u>Student (Group)</u>	<u>Format</u>	<u>Decision</u>	<u>Choice</u>	<u>Response Level</u>
Jack (1-GS)	Survey	Agree	Maximum	1
Amber (1-GS)	Survey	Disagree	Royal	2-V
Eduardo (1-SE)	Survey	Disagree	Royal	4
Ann (2-GS)	Survey	Disagree	Royal	2-V
Lance (GRAD)	Survey	Disagree	Royal	2-V
Jill (GRAD)	Survey	Disagree	Royal	4

Jack was the only one of the six interviewees who agreed with the assertion that the wait times for each theater were the same. It was not clear that Jack actually understood the task as was intended, so his response was potentially idiosyncratic, yet

his explanation focused on the actual wait-times of 10 minutes, frequencies of specific times, and thus a Local type strategy. Jack also chose to attend the theater that had the larger range of times, the Maximum Theater, based on a comparison of the shortest times, a local type reasoning strategy. So, overall for the Movie Wait-Time task, Jack responses were classified as Local, level 1.

Amber, Ann, and Lance each disagreed with the assertion that there was no difference in wait times, chose to attend the Royal theater, the theater with a smaller range of wait times, and primarily used reasoning strategies focused on comparing the variation in the distributions. Amber specifically cited the difference in ranges of wait times and when explaining why she chose to attend the Royal Theater she noted the possibilities of experiencing either a very short or a very long wait time at the Maximum Theater. Ann seemed to relate the diversity of times and their frequencies to variation. Both responses, from Amber and Ann, appeared to be exclusively focused on variation, with little to no evidence that their comparisons were made globally. Lance's minimal response exclusively cited different "variances" and he chose to attend the Royal theater because of the predictability of its times. Because of his previously described understanding of standard deviation, it is likely that Lance's comparison on this task was from a global perspective.

Eduardo and Jill also disagreed with the assertion that there was no difference in wait times and chose to attend the Royal theater, the theater with a smaller range of wait times. While they both used reasoning strategies that compared the variation in the distributions, their described strategies also included assessments of other

characteristics and thus were Distributional. Eduardo cited the difference in the ranges and also assessed the consistency of times around the mean. Jill cited the difference in ranges but she included an acknowledgement that the centers were indeed the same with an assessment of the shapes, specifically that the times for the Royal Theater were evenly distributed. So Jill's response, in particular, appeared to be made from a global perspective.

Jack, Amber and Ann were the only interviewees to have their Movie Wait-Time survey task responses reviewed during the interview. Both Amber and Ann wrote responses coded at level 2 focused on variation while Jack's responses was coded at level 1. In each case those interviewees confirmed the interpretation of their written survey responses during the interview. All three provided greater details of their thinking, with only Amber showing potential to make, and understand, the required comparison from a perspective at a higher level, specifically from a proportional reasoning perspective.

Responses to task 3 and task 4: the Pink/Black task –

Without descriptive statistics and With descriptive statistics

The data sets in the Pink/Black task were of different sizes. The Pink class recorded 36 scores while the Black class recorded 21 scores. There is a clear difference in centers between these distributions. Each of the mean, median, and mode are higher for the Black class. The ranges are the same but the shapes are different. While both distributions are uni-modal, the distribution for the Pink class is symmetric and approximately "bell shaped" and the distribution for the Black class is skewed

(see appendix B). Due to the size difference, reasoning additively by summing the scores was expected to result in claims that the Pink class scored better. Comparisons of centers or reasoning proportionally or distributionally were expected to result in claims that the Black class scored better. Those students who decided that one of the classes did score better were also asked to quantify how much better. Those estimates, for how much better a class scored, with their supporting explanations potentially provided either confirming or contradictory evidence of the students' reasoning on the first part of the task. For example, if a student decided that the Black class scored better and his reasoning strategy focused on comparing the means, then made an erroneous estimation for how much better the Black class scored that did not incorporate the difference between the centers, the student may not truly understand the mean as a group representative. For task 3, students made their assessments based on their intuitions without descriptive statistics, while for the follow up task, task 4, students were asked to rethink their decision in light of being provided with some descriptive statistics.

Jack

Jack's survey decisions for the Pink/Black task without and with the descriptive statistics was that the classes scored equally well. When he did not have access to descriptive statistics, Jack considered the averages of each class, but decided that they were about equal due to the difference in the class sizes. This is an indication that Jack did not completely understand the mean in a global way. He did not have to estimate how much better one of the classes scored because he decided they scored

equally well. For the follow up task, with descriptive statistics, Jack wrote that he did not understand the meaning of the terms so he did not change his original responses.

During the interview, before Jack responded to the Pink/Black task, he first interpreted the numbers on the horizontal axes of the graphs as percentage of questions answered correctly. After a brief discussion with the interviewer, he saw that the numbers represent the actual number of questions answered correctly. As these two interpretations are similar, the first interpretation was not considered problematic. When asked what the dots in the graphs represented, Jack said that they were “people.” Finally when Jack was asked to explain what he thought the task was asking him to do, he responded, “Out of the two classes, I believe the question is asking me, what to decide if the classes scored equally well or one scored better than the other.” Then when Jack was asked what he interpreted the term ‘better’ to mean, he indicated that to him ‘better’ meant higher scores. Thus, it was assumed that Jack understood the task as it was intended.

Jack’s decisions switched from the survey as in the interview he argued that the Black class scored better, for both tasks, without and with descriptive statistics. Without descriptive statistics, Jack decided that the Black class scored better based on comparing frequencies with an informal consideration of how the difference in size impacts that comparison. He explained,

...I know there are less students [in the Black class] and I’m sure that this plays a big part of this, but where I’m at right now in statistics, I really can’t explain it...It just seems like, although, for 7, 8, and 9 the values seem to match, but below 7 it seems like the Pink class didn’t do as well, had more values in the 2, 3, 4, 5, 6.

If Jack had only compared frequencies of individual scores, it would be some evidence that he was comparing the data sets as collections of individual scores. He was aware, from his experiences in his statistics class, that the size difference between the classes should influence his comparison of frequencies but he could not explain how or why. His attempt to account for the difference in class size, appeared to be the beginning of a transition away from a local view of the data to a global view, as he partitioned the scores at 7 and then made comparisons of scores in each partition and although he did not explicitly compare proportions of scores it appeared that his thinking was leading in that direction.

In an effort to verify that Jack was in the process of transitioning his comparison strategy away from a locally oriented one, he was asked about a comparison made by one of his classmates who decided in favor of the Pink class based on comparing the sums of scores for each class. In assessing that argument, Jack continued to struggle with how to deal with the difference in class size. He responded as follows:

That's a good argument but then you'd have to say there were more students in that Pink class than the Black class and that's the part that I really can't think of a mathematical equation to cipher the difference between the 36 and the 21. So just adding up the sum of one class compared to the other, um, I wouldn't hold that to be true because there's a big difference, um, 15 students more in the Pink class.

This response is more evidence that Jack views the comparison of the data sets from neither a completely local perspective nor a completely global perspective.

Particularly, his desire to use “a mathematical equation to cipher the difference” is an

indication that he could rely on computations of measures to make a comparison, such as the mean, without understanding them as group representatives. Next Jack was asked to assess arguments for 'Black' based on comparing the average scores, and then based on comparing proportions of scores above 6. Jack thought that comparing averages "world work" but felt the proportional argument was more convincing as he said, "I think that's what I was trying to say, that there's a higher proportion above 6 compared to the proportion of what's below it." So, although Jack had trouble articulating a proportional type argument he did agree with it when it was presented to him.

Although Jack indicated that comparing means or proportions were "convincing," he attempted to quantify how much better the Black class scored by estimating the difference in frequency of scores below seven. He estimated that Pink class had about double or triple the number of scores below seven so that meant the Black class scored more than 50% better than the Pink class. Thus, he fell back to using a strategy that was similar to viewing the data as a collection of individual scores to aid him in making an estimation of how much better the Black class scored.

With access to the descriptive statistics Jack also decided that the Black class scored better but he made a change in his reasoning as he based his decision on the fact that the Black class had a higher mean, median, and mode. That reasoning strategy is potentially at the initial distributional level as the locations of the mean, median and mode are related to the shape of a distribution. However, based on Jack's responses to some follow up questions it was clear that he still struggled with

accounting for the difference in the sizes of the distributions and he potentially cited the measures of center in a way similar to that described by Ben-Zvi (2004) who characterized similar group comparisons made by junior high students as “insignificant and monotonous use of statistical measures.” Jack specifically admitted that he was confused as to how the difference in size “played out in determining the mean, median, and mode.” When he was asked to quantify how much better the Black class scored, Jack considered the difference of the means but was, again, unsure of how the difference in the classes’ size would effect the validity of that quantification. After addressing the follow up questions Jack reconsidered using the difference of means for his estimation and returned to using his strategy from when he did not have access to the descriptive statistics, a comparison of frequencies below seven.

On both the survey and interview for the Pink/Black tasks, with and without the descriptive statistics, Jack did not understand how to account for the difference in the size of each class. On the survey this resulted in an erroneous comparison of centers. In the interview Jack integrated a consideration of the different sizes into his rough shape assessment that actually was consistent with informal proportional reasoning. When he had the descriptive statistics available on the survey Jack wrote that he did not understand the terminology, but in the interview, Jack mentioned that all three measures of center were higher but focused on the mean and was unsure of how the differences in the size of each class effected that comparison. He also did not understand that he could use a measure of center, particularly the mean, to quantify the difference between the classes’ scores. Thus his reasoning strategies on the survey and

interview were coded at the Transitional level, level 2. For the survey he focused on centers without the descriptive statistics, but with the descriptive statistics gave an idiosyncratic response. For the interview he focused on shape without the descriptive statistics but focused on centers with the descriptive statistics.

Amber

Amber's survey decision without descriptive statistics was in favor of the Pink class and was based on the observation that Pink had more correct answers as she specifically wrote that, "More students got correct answers in the Pink class." Her quantification for how much better the Pink class scored was consistent with her decision reason as she estimated that Pink scored better by 37%, based on the difference of sums of scores. This type of reasoning was described as 'additive' and was used by some middle and high school students to compare distributions in a variety of contexts, as reported by Petrosino, Lehrer, and Schable (2003), Cobb (1999) and Shaughnessy, Ciancetta, and Canada (2004).

When provided with the descriptive statistics on the survey, Amber switched her decision to 'the Black class scored better.' Her explanation was contradictory as it referred to the sum for the Black class as better, yet for her estimation she wrote that she compared the medians and means to determine that the Black class scored "15% better." So both of Amber's responses, without and with descriptive statistics, were classified as Local, however when she had access to some statistics this classification was much more tentative as she attempted to quantify how much better the Black class scored by comparing centers.

During the interview, before Amber responded to the Pink/Black task, she first interpreted the numbers on the horizontal axes of the graphs as, “the number of correct answers.” When asked what the dots in the graphs represented, Amber said that they were “the number of students.” This was interpreted as meaning all the dots represented all the students. Finally, when Amber was asked to explain what she thought the task was asking her to do, she responded, “Assess the graphs and figure which class did better overall.” Then when asked what she interpreted the term ‘better’ to mean, she said that “Better means more correct answers.” Then when she was asked to clarify that statement she said it meant more correct answers for “the class as the whole.” So, it appeared that Amber’s criteria, although not specifically articulated in a proportional way, tends toward the proportional as she required a comparison of the correct answers of the classes on the whole. Thus, it was assumed that Amber understood the task as it was intended.

Amber’s interview decision for the Pink/Black task without the descriptive statistics was different than her survey decision. She decided that the Black class scored better. In part, Amber said,

I’m looking at more the shape of the graph and the first one [Pink] has a traditional, like, bell shaped curve but the Black class had fewer incidences of lower numbers...

Her reasoning strategy was actually quite similar to Jack’s, that is, it was based on the shapes of the graphs with a specific observation that the Black class had lower frequencies for the lower scores. Amber’s responses implied that she was informally accounting for the differences in class sizes but later when asked to assess a

proportional argument in support of the black class she was unconvinced. Amber's quantification for how much better the Black class scored was also similar to Jack's. For her estimation, Amber relied on her assessment of shapes, yet her estimation did not incorporate a comparison of averages, it focused solely on the shapes. Amber estimated that the Black class scored 30% better and she explained that she looked at the shapes of the graphs to make her estimate. When asked to provide some details of how she looked at the shapes she said,

I'm looking at it like this [Amber sketched a curve over the Black graph, shown in Figure 48], and then this shaded area I estimated to be about 30%.... If I put this [Black] on top of that [Pink] there's this much [points to shaded area above Black graph]

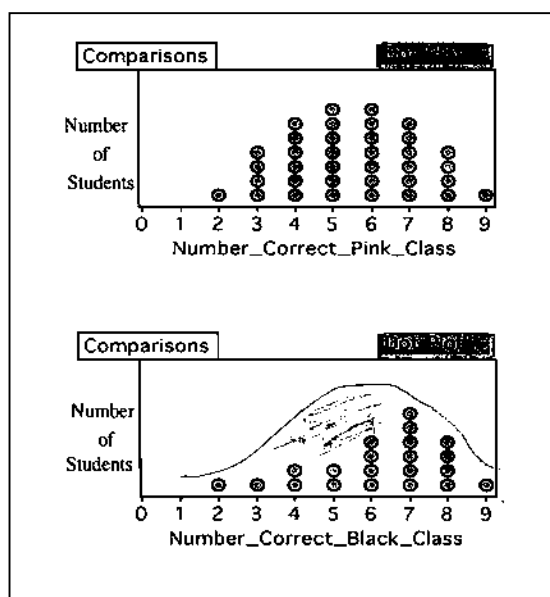


Figure 48. Amber's comparison of shapes.

From Amber's explanation of her estimation it appears that she constructed an interesting, however inappropriate way to compare the shapes of the data sets. She seems to have estimated how much data might be added to the Black graph to make it

look like the Pink graph. Amber was also asked a follow up question about assessing another student who made an estimation based on comparing the sums of scores of the two classes. Amber was unsure of this method because the Pink class had more students than the Black class. When asked about a way to account for that difference Amber said, “There is a way. I’m not real certain about what that way is but, um, if given some time I bet I could figure it out.” This may indicate that Amber does not yet completely understand using the mean as a group representative or how other proportional comparisons account for differences in group size.

After Amber examined the descriptive statistics associated with each data set, she still decided that the Black class scored better but her reasoning strategy changed. Again, similar to Jack, she supported her decision by citing that the mean and median were higher for Black. She then quantified how much better the Black class scored by estimating the difference of those measures of center at about 20%. The actual difference between the means is 0.69 and between the medians is 1.5 so it is not clear how Amber concluded 20%. Similar to Jack, Amber may have cited the measures of center in a way similar to Ben-Zvi’s (2004) characterization of “insignificant and monotonous use of statistical measures.” In a follow up question Amber was asked why she changed from her original method of comparing the shapes to comparing the means and medians. She explained that the mean was a “quantitative value [sic]” and that by comparing the means she did not have to “use intuition or gut instinct to defend.” That explanation was not helpful in assessing if Amber was comparing the means as global characteristics, yet it had elements of Bakker and Gravemeijer’s

(2003, 2004) assertion about students who view data as individual values who, in turn, understand measures such as mean, median, range, etc, as merely the results numerical calculations, not group features.

So prior to viewing the descriptive statistics, on the survey, Amber decided that the Pink class scored better based on a reasoning strategy was at the Local level, specifically with an additive focus. After Amber viewed the descriptive statistics she changed her decision to the Black class but her reasoning strategy on the survey was difficult to categorize as it was inconsistent, part Local and part Transitional. It is likely that she did not fully understand how to compare data sets of different sizes. In the interview Amber decided that the Black class scored better both before viewing the descriptive statistics and after viewing the descriptive statistics. Before she had access to the descriptive statistics she compared the classes' scores based on shapes and frequencies and estimated how much better the Black class scored by comparing the frequencies. After Amber examined the descriptive statistics she made her decision in favor of the Black class based on comparing the means and medians and also estimated the difference between the means and medians to quantify how much better the black class scored. Although Amber's reasoning appeared to be at a higher framework level in the interview, was not clear that Amber considered her comparison of the centers from a global perspective.

Eduardo

Eduardo's survey decision without descriptive statistics was in favor of the Pink class. His explanation was contradictory. He noted that the Pink class had more

students with “most averaging in the 5 – 6 range” and that the Pink class’s scores formed a bell curve, whereas the Black class had scores shifted higher. He then concluded the “bell curve system indicates that the Pink class is more successful since most students will fall within the allowed percentage range.” It appears that Eduardo believed that the apparent bell shape of the Pink class’s data is “better” even though he observed that the Black class’s scores are shifted higher. Although Eduardo’s explanation mentions a comparison of frequencies, it appears to be mostly based on a comparison of shapes, although a misinterpretation of what the ‘Bell Shape’ implies.

Eduardo’s quantification for how much better the Pink class scored was less than 3% better. His written explanation was, again, rather contradictory. He claimed to have compared the shapes of the graphs and then wrote that if an infinite amount of data was collected then “the Black might essentially exceed the average of the Pink.” Although Eduardo’s earlier statements and decision may be evidence that he was focused on assessing shapes and variation and may not have interpreted the idea of equating a class scoring better with higher scores, he seemed to be claiming that the average for the Pink class is higher than the average for the Black class, which is not true whether comparing location or frequency. Eduardo did have an intuition that the skewness of the data for the Black class is related to its higher average but he only understood this if the size of both data sets was infinite.

After examining the descriptive statistics on the survey, Eduardo did not change his decision. He decided that the Pink class scored better. His written explanation for his decision was based on a comparison of standard deviations for

each class, that is, the Pink class's scores had a slightly smaller standard deviation. Eduardo's estimated that the Pink class scored "~2%" better. His explanation contradicted his quantification in favor of the Pink class as it referred to making a "mental comparison" of the means.

On both survey tasks, without and with descriptive statistics, Eduardo's responses indicated that he had difficulty making a comparison because of the difference in class size. His main focus appeared to be on comparing variation yet when it came to estimating the difference between the classes' scores he referred to the means, which actually provide evidence contrary to his decision.

On the Interview, Eduardo interpreted the numbers on the horizontal axes of the graphs as 'grades' related scores out of 100 points. For example he interpreted 3 as 30 and 4 as 40. Although the numbers on the horizontal axes were scores out of 9 not out of 10, Eduardo was not corrected, as his interpretation was not different enough to cause potentially different conclusions. Eduardo interpreted the 'dots' as the amount of students at each grade. Finally, when Eduardo was asked to explain what he thought the task was asking him to do, he responded, "It's basically asking which class did better based on the results of the given data and knowing the amount of students that each class had." Then when asked what he interpreted the term 'better' to mean, he indicated that to him 'better' meant "consistency." Thus, it was assumed that while Eduardo could read the graphs correctly, he understood the task differently than was intended, as 'better' was intended to be interpreted as higher grades.

Eduardo's interview responses for each Pink/Black task, without and with the descriptive statistics, were similar to his survey responses. While, throughout his reasoning on each of the Pink/Black tasks, Eduardo relied on his interpretation of equating 'better' with being more consistent, he still had difficulty accounting for the difference in class sizes and had difficulty resolving the fact that the Black class had higher overall scores. Without access to the descriptive statistics Eduardo decided that the Pink class scored better and, as he did on the survey, described that the 'bell shape' of the Pink class's scores implied that they were more consistent and thus better. Along with his 'consistency' statement Eduardo also made the following speculation similar to what he wrote on his survey:

...but I guess if we had more students, I guess over a certain period of time, this class [points to the Black graph] would be achieving a lot more because of the fact that you have more students getting grades on the top end of the line.

Eduardo seemed to be wrestling with his stated interpretation of better as more consistent and the observation that although the Black class has fewer students, it also has more students who scored higher, i.e., "at the top end of the line," an initial type of proportional observation.

When Eduardo was asked some follow up questions about other possible reasoning strategies, he always returned to his original assessment. When specifically asked about comparing means, he said,

I put this [Pink average] just a little lower than this one here [Black average]. I would give this a higher average [Points to Black]...It contradicts my answer because now the average is a lot higher [on

Black] but going back to the answer I gave you, there is more consistency in the first one [Pink].

As with his original response Eduardo wrestled with the observation that the Black class had higher scores but in the end did not change from his belief that the Pink class was more consistent and thus better. Eduardo was asked another follow up question about a proportional type argument. The argument he assessed was that the Pink class had about half of its students score above five while the Black class had more than half of its students score above five, so the Black class scored better. Eduardo's response was,

... it's definitely a fact so, I don't think I could argue against that one [laughs]... You can see that the Black actually did better than the Pink, but [pause] so I think over a period of time I could see this one [Black] did better, but as far as what we have, I still want to stay with Pink.

So, Eduardo continued to struggle with the fact that there is a size difference between the classes. He understood that the mean for the Black class is higher and that the scores for the Black class are shifted higher, but misinterpreted that the bell shape of the Pink class's scores and their greater class size imply Pink's scores were more consistent and thus better despite Pink's lower mean.

Eduardo's attempt to quantify how much better the Pink class scored was not successful. He initially switched his decision to favor the Black class because of his estimation of the means for each class, but then had trouble reasoning about the difference in the size of each class, and finally decided not to switch his decision. He ultimately did not make an estimation for how much better the Pink class scored but returned to referring to his understanding that the Pink class was more consistent.

When Eduardo was provided with some descriptive statistics associated with the scores for each class, he stayed with his decision that the Pink class scored better and cited the lower standard deviation of the Pink class's scores. He did address the fact that the mean for the Pink class was lower than the mean for the Black class, but ultimately felt that a lower standard deviation was the better indicator. The difference in standard deviation between the classes was 0.08 and it was not clear that Eduardo understood how small of a difference that is as related to the data sets. This was evident when Eduardo attempted to quantify how much better the Pink class scored. Once again he did not provided an estimate but did refer to the difference in standard deviations and noted that, "... just looking at the standard deviation I can see it's a lot smaller and that counts a lot more than having a bigger deviation in the data."

Overall, for the Pink/Black without and with descriptive statistics, Eduardo related higher consistency to scoring better. He believed that the closer a distribution is to a bell shape, the better or more ideal it is, as it will have more consistency. It appeared that Eduardo understood that the Black class higher scores than the Pink class and that the Black class had a higher mean, but because of the difference in the number of scores for each class he ultimately did not believe those comparisons were valid. Thus as Eduardo's reasoning strategies were specifically focused on one feature of the distributions, his responses were classified at the Transitional level, level 2 alternately focused on shape and variation.

Ann

Ann's survey decision without descriptive statistics was in favor of the Black class and was based on an observation that was intuitively proportional. She considered that the highest scores of seven, eight, and nine had equal frequencies for both classes but the Pink class had higher frequencies for the lower scores of four, five and six, so the Pink class did not score as well. Ann's estimation for how much better the Black class scored was 40% to 50% better. She did not explain how she determined that estimate and only wrote that, "the Black class only exceeded the Pink class by a small margin," which is rather inconsistent with the quantification she provided.

After having access to the descriptive statistics on the survey, Ann decided again that the Black class scored better. Her explanation showed that she did not understand the statistical terminology used and that she did not understand that comparing statistics, such as the mean, would account for a difference in the size of data sets to be compared. In her explanation supporting her decision for the Black class, she almost came to the conclusion that the Pink class scored better. She wrote,

The mean, median and mode all reflect higher scores in the Black class in comparison to the Pink, but only by a small margin. However, this question becomes difficult in that the sum is higher for the Pink class and the standard deviation isn't as far away from zero. Thus it makes me conclude that perhaps the Pink class really did do better than the Black class.

Ann compared several of the statistical measures, but because of the different sizes of the data sets, she was unsure about which comparisons were valid and thus potentially

understands the measure only as computations, not as group features. Her assertion that the Pink class has a higher sum and is therefore ‘better’ is further evidence that she had difficulty reasoning proportionally. Her previous description of the meaning of standard deviation and her confused understanding of that meaning is also highlighted in this response.

On the survey, Ann estimated that the black class scored 5% to 10% better. She explained that, “If I maintain my answer above, that the black class did do better, and the new data provided shows that they only did better by a few standard deviations.” This is an idiosyncratic response because doing better by a few standard deviations would be doing significantly better.

In the interview, Ann interpreted the numbers on the horizontal axes of the graphs as the number of correct answers. When asked what the dots in the graphs represented, Ann indicated that they were ‘students.’ Finally when Ann was asked to explain what she thought the task was asking her to do, she responded, “OK, when I see it, it says, um, which class scored better. So which class got the higher grade in comparison to the other class?” Then when asked what she interpreted the term ‘better’ to mean, she indicated that to her ‘better’ meant “higher grades.” Thus, it was assumed that Ann understood the task as it was intended.

In the interview, Ann decided that the Black class scored better for both Pink/Black tasks, without and with descriptive statistics. Without descriptive statistics, she initially decided that the Pink class scored better based on a comparison of frequencies, that is, each class had equal frequencies of scores at seven and above but

the Pink class had higher frequencies for the scores below seven. Then she quickly rethought her argument and concluded that it actually supported the decision that the Black class scored better. Her final explanation focused on comparing frequencies with an informal consideration of how the difference in size impacts that comparison. She summarized her initial observation as follows:

So, more students got higher grades in the Black class and there were less students that got poorer grades, like a C, D, F average. But in the Pink class there's a lot that got a middle average and lower. So, even though there are more students, you know I don't have a calculator or anything, but it seems like the Black class did better.

Each of Ann's final explanations, from the survey and interview, were almost identical as they both supported the Black class scoring better and both employed an informal proportional comparison. Yet, in follow up questioning in the interview Ann was unsure about the argument that the sum of the Pink class was higher so the Pink class scored better, an indication that she did not completely understand proportional comparisons.

Similar to the survey, in the interview Ann had difficulty quantifying how much better the Black class scored. Her explained her estimate as follows:

So, my decision was the Black class scored better. Um, I couldn't say how much better. I'd probably guess, say, half as well, half as much better. I mean it doesn't seem like they did significantly better, just maybe, on average say the Pink class scored, there mean was 70%, they [the Black class] probably got an 85 for their mean.

It appears that Ann might have been using the means as group representatives. That conclusion would be tenuous at best given Ann's earlier description of the meaning of the mean and her estimation that the Black class scored better by "half as much." Her

estimation also indicates potential proportional reasoning difficulties as the difference between Ann's mean estimations of 70 and 85 is 15, so it is quite unclear how she estimates "half as much better."

After she viewed the descriptive statistics in the interview, Ann also decided that the Black class scored better. Ann based her decision primarily on a comparison of the means, but not necessarily the medians and modes. She also compared the standard deviations and sums. In particular she felt that the smaller sum of the Black class was problematic which is an indication that she did not fully understand how to account for the difference in the number of scores for each class. Ann's quantification for how much better the Black class scored was consistent with her decision. She estimated that the Black class scored one point to half a point better based on the difference between the means of the classes' scores.

For both Pink/Black tasks, without and with descriptive statistics, Ann consistently decided that the Black class scored better but had difficulty quantifying how much better and also had difficulty understanding how to make comparisons that account for the difference in number of scores, i.e. proportional reasoning. On the survey, without descriptive statistics, her reasoning strategy employed an implied proportional argument and was categorized at the Transitional level, focused on shape. On the survey, without descriptive statistics, Ann's reasoning strategy was also Transitional but focused on center. Ann's responses in the interview were strikingly similar to her survey responses. Without descriptive statistics she used an implied proportional strategy categorized at the Transitional level focused on shape and with

descriptive statistics she compared the means, a Transitional strategy focused on center. Ann consistently was unable to form explicit proportional arguments and had considerable difficulty using the means to quantify how much better the Black class scored.

Lance

Lance's survey decision without descriptive statistics was in favor of the Black class and his explanation only mentioned a comparison of means. His quantification for how much better the Black class scored was consistent with his decision as he used the means to determine his estimate. Although he did not estimate the difference between the means, he did estimate the ratio of the means for his quantification.

On the survey, Lance decided that the Black class scored better. His explanation was quite brief, "higher average score in Black than Pink." Without more information this response alone would be categorized at level 2-center. Next, Lance estimated that the Black class scored about 12% better. He determined this estimation by a "rough calculation of mean in black class, observed mean in pink class, rough calculation of $(\text{black avg})/(\text{pink avg}) - 1$." So, Lance produced a sophisticated estimation that relied on a ratio of mean scores for each class. So, even though Lance exclusively referred to the means throughout his explanations, given the sophisticated nature of his estimation and the extra information from the interview about Lance's global perspective of the mean, it is quite probable he was comparing the means from a distributional perspective.

Lance's survey decision and estimate with the descriptive statistics was also in favor of the Black class but his explanation only referred to the "descriptive statistics support observations and rough calculations on previous page," meaning his responses to the Pink/Black task without descriptive statistics. This type of response was difficult to assess, particularly because Lance's explanations "on the previous page," for Task 3, the Pink/Black task, were quite minimal and focused on the mean. It was unclear which statistics, other than the mean, Lance referred to and how he used them to decide that the Black class scored better. Also, Lance did not change his estimate as he just referred to his previous estimate. Thus the supplemental code of level 2-N/A was assigned to Lance's response. In general, this code was used in this situation, that is, when the "statistics" are referenced without any more explanatory information.

On the interview, Lance interpreted the numbers on the horizontal axes of the graphs as, "the number of correct answers." When asked what the dots in the graphs represented, Lance indicated that, "Each of the dots is a student." Finally when Lance was asked to explain what he thought the task was asking him to do, he responded, "It's asking me to say if one class performed better than the other." Then when Lance was asked what he interpreted the term 'better' to mean, he indicated that to him 'better' meant "higher score." Thus, it was assumed that Lance understood the task as it was intended.

Lance's responses in the interview were mostly consistent with his survey responses. For both Pink/Black tasks, without and with descriptive statistics, he decided that the Black class scored better. His explanations were considerably more

detailed that his explanation on his survey. Without access to the descriptive statistics Lance incorporated both shape and center characteristics to give a strong distributional response. In part, he said,

I'm looking at getting a visual feel for what the mean of the distribution would be. Um, mean, mode. Um, I'm looking at how they are spread, evenly or unevenly and it appears that although that the Black class has a lot longer tail on the lower end that it would have a, um, higher over all score.

He also provided sophisticated, global assessments of other possible arguments. One such argument that was given by another student asserted that the bell shape of the Pink class is the nearest to perfect, the most desirable shape and thus the Pink class scored better. After Lance stopped looking at the interviewer as if the interviewer had just sprouted three heads, Lance rejected this response and provided a distributional type explanation as to why that response was nonsense. His first reaction was to say, "I don't see any particular reason to say that, um, we would prefer that shape." Then when asked to explain why he said,

Well, they, um, they have the same range. They have the same low end. They have the same high end. Um, the Black class, within that range, is shifted more towards the upper end of that range. The body of its score is shifted more upward. So, in order to, on that same range to achieve that symmetrical shape, more people would have to get a lower score. It doesn't seem like the better measure of the class's performance.

His estimate in the interview was a bit different than on the survey, as he quantified how much better the Black class scored by estimating the difference between the means.

Lance's interview response, with access to the statistics, was fairly consistent with his survey response. He cited the list of statistics and specifically compared the mean, median and mode and then connected those comparisons with shape when he said, "those [centers] describe the shift to the higher score that I see as being the higher score for the [Black] class." Lance was not asked any follow-up questions. He quantified how much better the Black class scored by considering the difference between each of the centers citing that "the mean, median, and mode are about a point higher."

For the Pink/Black tasks, without and with descriptive statistics, Lance consistently decided that the Black class scored better and quantified how much better by comparing the measures of center. On the survey, Lance's explanations were rather minimal and were coded at the Transitional level, level 2 focused on centers. Lance's interview responses were considerably more detailed and were clearly made from a global perspective with a categorization at the distributional level, level 4.

Jill

On the survey, Jill's survey decision without descriptive statistics was in favor of the Black class and was based on a comparison of the proportion of scores above seven, a comparison of the medians, and skewness assessment. Jill estimated that the black class scored two points better and indicated that she determined this by finding the difference of the medians. Her comparison appeared to of the distributions as whole units.

When provided with the descriptive statistics on the survey, Jill also decided that the Black class scored better. Her written explanation was, “Same reasons as last time, but now I have evidence from the data that the median and the mean score for the black class is higher.” She then estimated that the Black class scored better by about one point. She wrote that her estimate was, “Kind of a mix between 1.5 for the differences in the medians and 1 (about) for the difference in means.” Jill appeared to use some of the descriptive statistics as confirming evidence for her original response.

In the interview, Jill interpreted the numbers on the horizontal axes of the graphs as the number of questions answered correctly. When asked what the dots in the graphs represented, Jill indicated that they were ‘students.’ Finally when Jill was asked to explain what she thought the task was asking her to do, she responded, “We are supposed to look at all the frequencies and decide which class did better.” Then when Jill was asked what she interpreted the term ‘better’ to mean, she indicated that to her ‘better’ meant higher scores in a probabilistic sense (see the transcript below).

J: *The probability that if you picked a person [points to Black] they would probably land in the upper grades, than if you picked any person [points to Pink] they could land almost anywhere, it's fairly symmetric. So you'd land more of the lower scores [for Pink].*

Int: *You're pointing to the middle of the [Pink] graph?*

J: *Yes, fives and sixes.*

Thus, it was assumed that Jill understood the task. Although the task was not originally designed for students to overlay a sampling environment over it, the interviewer felt that there was no need to further explore this interpretation.

Jill's decisions in the interview were the same as on her survey, the Black class scored better for each task, without and with descriptive statistics. Jill's explanation was quite similar to her explanation on her survey, although, surprisingly, her interview explanation was not quite as sophisticated as her survey explanation (see the transcript below):

The first class [Pink] has a lot more students, than the second class [Black]. So we get a lot higher frequencies for the number correct for the first class than we do in the second class. Um, to decide which one did better, I look at the frequencies and I notice that in the Pink class they're higher up in the middle than they are in the Black class. And in the Black class they tended to do better, more of the 7s and 8s, even though there was the exact same number of 7s and 8s, um, there were fewer people who did worse in the second class [Black] than in the first class [Pink]. So, I may say, well maybe there is not enough data compared to the first class [Pink], maybe there could potentially be a larger class [Black], this class [Pink] would do as well, but just from looking at where most of the data lies, it tends to be higher up in the second class [Black].

Her interview explanation informally compared proportions of scores at seven and above, but she never actually said that is what she did and thus was closer to a focus on shape. She also mentioned the difference in class size but it was unclear whether or not she found that problematic.

Jill then quantified how much better the Black class scored as "one or two points." She explained how she determined her quantification by, again, assessing the shapes and then comparing the modes and "centers." It was not clear if she meant means or medians when she said "centers." She again mentioned the difference in class size and again it was unclear if she found that problematic.

After Jill viewed the descriptive statistics for the Pink and Black classes, she still thought that the Black class scored better. She initially scanned the list of the statistics and noted which were higher for Black and which were higher for Pink. She then claimed that “everything is higher” for the Black class except standard deviation and later noted that the difference in standard deviation was small, so it was not too much different. Also, her estimation of “a point or two” for how much better the Black class scored was based on the difference between the measures of center, the means, medians, and modes, just as she did on her survey.

For both Pink/Black tasks, without and with descriptive statistics, Jill consistently decided that the Black class scored better and quantified how much better by comparing several of the measures of center. On the survey, without the descriptive statistics, Jill’s explanations were a little more formal as she specifically referenced comparisons of proportions, centers and shapes. While in the interview, Jill’s explanations were a little more informal as she focused on shapes and centers while mentioning the difference in class size. With the descriptive statistics, for both survey and interview, Jill compared almost all of the measures in support of her previous decision. Overall for the survey, Jill explicitly reasoned proportionally and also consistently responded at the Distributional level, level 4. In the interview, without the descriptive statistics, she responded at the Initial Distributional level, level 3 with an Initial Global focus, and with the descriptive statistics she responded at the Distributional level, level 4.

Summary of the Pink/Black task – Without and With descriptive statistics

The results of the responses, for all six interviewees, to both Pink/Black tasks are summarized in table 53. Each of the first four students, Jack, Amber, Eduardo, and Ann, experienced trouble in accounting for the difference in the size of the data sets. Reasoning about proportions appeared to be a significant conceptual obstacle for all four of those interviewees. Irrespective of the decision for which class scored better, all four had difficulty at estimating the difference between the classes' scores and contributed to the evidence that these students generally understood statistical measures, such as the mean, more as computations with the data as opposed to group characteristics.

All the students provided more detailed responses in the interview as opposed to the survey and all of the students, except for Jill, provided interview responses that were categorized at an equal or greater framework level than their survey responses. This was not surprising as the interviewees had more opportunities to explain their thinking in the interview and had more accountability for their interview responses, as they were not typing their responses into their computer, on line. Jill's survey and interview responses were similar, yet one distinct difference was that she explicitly reasoned proportionally in her survey explanation as she compared proportions of scores above seven, while in the interview she compared frequencies of scores above and below seven with a only qualification that the Pink class had more students.

Table 53.

Interviewees' decisions, estimates and response levels for tasks 3 and 4: the Pink/Black task (without statistics) and the Pink/Black task with statistics.

Student (group)	Format	Without	Without	Without	With	With	With
		stats	stats	stats	stats	stats	stats
		<u>Decision</u>	<u>Estimate</u>	<u>Response</u> <u>Level</u>	<u>Decision</u>	<u>Estimate</u>	<u>Response</u> <u>Level</u>
Jack (1-GS)	Survey	Equal	None	2-C	Equal	None	0
	Interview	Black	Diff. of shapes	2-S	Black	Diff. of shapes	2-C
Amber (1-GS)	Survey	Pink	Diff. of sums	1	Black	Diff. of centers	1
	Interview	Black	Diff. of shapes	2-S	Black	Diff. of centers	2-C
Eduardo (1-SE)	Survey	Pink	Diff. of shapes	2-S	Pink	Diff. of st. dev.	2-V
	Interview	Pink	Diff. of var.	2-V	Pink	Diff. of st. dev.	2-V
Ann (2-GS)	Survey	Black	Unclear	2-S	Black	Diff. of centers	2-C
	Interview	Black	Diff. of centers	2-S	Black	Diff. of centers	2-C
Lance (GRAD)	Survey	Black	Ratio of centers	2-C	Black	Ratio of centers	2-N/A
	Interview	Black	Diff. of centers	4	Black	Diff. of centers	4
Jill (GRAD)	Survey	Black	Diff. of centers	4	Black	Diff. of centers	4
	Interview	Black	Diff. of centers	3-IG	Black	Diff. of centers	4

All the students provided equal or higher level framework responses after they had access to the descriptive statistics. Yet it was not clear from the “with statistics” responses that students were viewing the data sets differently. Eduardo and Ann

responded nearly the same before and after having access to the descriptive statistics. Lance and Jill tended to refer to the descriptive statistics as supporting evidence for their initial decisions without the statistics. Jack and Amber's responses changed the most after they had access to the descriptive statistics. However, Jack, Amber, Eduardo, and Ann all made at least some contradictory comparisons by using the descriptive statistics and their responses that referred to statistics often showed no evidence that those students understood the statistical measures as group representatives.

Lance and Jill's responses each provided some evidence that they were able to view and compare the Pink/Black data sets as whole units. They both demonstrated that they could reason proportionally and understood that representative measures, the means in the case of these data sets, could also be used to compare groups of scores and estimate the difference between the groups of scores. Both students had considerable experiences using statistics as they had each completed numerous statistics courses and particularly Lance frequently performed statistical analyses on data obtained for classes related to his Bioinformatics major. Lance and Jill appeared to have global perspectives of the Pink and Black data sets.

Jack, Amber, Eduardo, and Ann all had very different educational and work backgrounds yet all appeared to, at times, view the Pink and Black data sets from a local perspective as a collection of individual scores and other times from a somewhat global perspective as they attempted to compare statistical measures and intuitive proportions but often came to contradictory or incorrect conclusions. Those students'

responses contained evidence that their perspectives of the Pink and Black data sets were in transition from a local perspective and towards a global perspective. Whether or not students were provided descriptive statistics, understanding proportional reasoning appeared to be strongly correlated with students who compared the centers as group representatives and consequently used those centers to quantify the difference between the groups.

Responses to task 5 and task 6: the Ambulance task –

Without descriptive statistics and With descriptive statistics

The data sets in the Ambulance task were of different sizes. There were 36 recorded response times for the Life Line ambulance service and 74 response times recorded for the Speedy ambulance service. There is not a clear difference in centers between these distributions. It is not easy to visually determine the mean of each distribution, although the mean for Life Line is slightly less than Speedy's. The median for Speedy is slightly smaller than Life Line's, but the mode for Life Line is clearly at a shorter time. Life Line also has a smaller range. While both distributions are uni-modal, the distribution for the Speedy has several "spikes." Due to the size difference, reasoning additively by summing the times was expected to result in claims that Life Line was better. Comparisons of only the endpoints of the distributions was also expected to produce a recommendation for Life Line as its shortest and longest times were less than Speedy's shortest and longest times, respectively. Comparisons of proportions could result in recommendations for either service depending on which proportions were being compared. The differences between the Ambulance

distributions were not as clear as the differences between the Pink/Black classes' distributions and thus students were expected to potentially approach this task differently than the previous ones. For task 5, students made their assessments based on their intuitions without descriptive statistics, while for the follow up task, task 6, students were asked to rethink their decision in light of being provided with some descriptive statistics.

Jack

On the survey, before and after viewing the descriptive statistics, Jack decided to recommend Speedy. When he did not have access to the descriptive statistics he wrote, "Speedy is mostly there in 20 minutes" which was interpreted as a naïve shape assessment, as he informally compared the times below 20 minutes to above 20 minutes. Jack's written response to the Ambulance task with descriptive statistics was to stay with his original survey recommendation of Speedy because, "again I just started this class and do not understand the data. I just feel by looking at the other set that Speedy is faster." This explanation was not helpful in categorizing Jack's response.

During the interview, before Jack responded to the Ambulance task, he interpreted the numbers on the horizontal axes of the graphs as response time in minutes. When asked what the dots in the graphs represented, Jack said that they represented each ambulance response. Finally, when Jack was asked to explain what he thought the task was asking him to do, he responded that "the question is asking to compare response time to determine which company is better." Then when Jack was

asked what he interpreted the term ‘better’ to mean, he indicated that to him ‘better’ meant “quicker.” Thus, it was assumed that Jack understood the task as it was intended.

In the interview, Jack experienced similar difficulties in making the Ambulance recommendation as he did making the Pink/Black decision, that is, he was unsure of how to account for the difference in the size of each data set. Jack’s interview decision, without descriptive statistics, was the same as his survey decision. He experienced similar difficulties as he did with the Pink/Black task in dealing with the difference in the size of each data set. His reason for this decision, again, included a slightly more articulate proportional argument than his survey response, in an attempt to reconcile the difference in number of response times recorded. He said,

...It just seems that you’d want a quicker response time and again we have 36 responses and 74, um so we have double the amount of calls for Speedy. But even if we double the lower response times [for Life Line] they’re still not as many as Speedy [below 20 minutes]...

Jack was then asked what he meant by “double the lower response times.” He responded by describing a process of doubling each frequency for each of Life Line’s response times. Using this reasoning strategy, Jack possibly compared rough proportions of individual scores, not proportions of partial distributions, and thus is closer to a comparison of shapes than of proportions. For a follow up question, Jack was asked to assess a decision that was based on a comparison of means, Jack thought that it would be helpful to compare the means but concluded that the decision could not be based solely on the means because of the difference in the number of response

times. That conclusion was evidence that Jack did not understand the mean proportionally, as a group propensity (Konold & Pollatsek, 2002).

After Jack viewed the descriptive statistics he switched his recommendation to Life Line specifically because Life Line's mean was less than Speedy's mean. Jack was then asked about his previous recommendation that was based on a rough comparison after doubling the number of response times for Life Line below 20 minutes. Jack said that comparing the means was more convincing for him. When asked why he responded,

Um, statistics, I mean that's, um there's a science to it here and if they tell me the mean is lower, 15.56 is lower than 16.45, I interpret that as a better response time on average. Although we don't have anything in the higher range here, the higher response times [points to high end of Life Line], now that I look at it, it seems like Life Line has a statistically better response time.

Jack's comment about the comparison of the means being convincing because there is a 'science' to statistics is particularly illustrative of a Transitional level response. That is, Jack relied on comparing the means without being able to articulate why it is appropriate.

Thus, Jack seems to have an inclination that comparing the means would be helpful and when given the mean calculation he deferred to comparing them. Yet the difference in the sizes of the data sets was problematic for him, a problem that remained unresolved from the survey through the interview. On these Ambulance tasks, Jack seems to have used a similar strategy to what he used on the Pink/Black tasks, that is, he compared the frequencies with an informal attempt to compensate for

the differences in the sizes of the data sets, and when provided with the means defer to comparing them. His attempt to base his comparison on more than the frequencies implied that his reasoning strategy was higher than level 1, but he did not understand how to correctly utilize proportional type arguments and he did not understand the means from a proportional perspective. Thus his responses were coded at the Transitional level, level 2, but without the statistics he focused on shape whereas with statistics he focused on center.

Amber

On the survey, without the descriptive statistics, Amber recommended Speedy with the vague explanation that “Speedy is more likely to respond quickly.” Then, after Amber had access to the descriptive statistics for the ambulance response times, she switched to recommend Life Line. Her written explanation referred to Speedy’s times as “less predictable.” Fortunately, there was time at the conclusion of the interview to review Amber’s survey responses to both of the Ambulance tasks. Concerning the Ambulance task without statistics, she was asked if she could remember what she meant by “more likely.” Amber said that she did recall what she was thinking and went on to describe that she had compared the absolute frequencies of the response times at the low ends of the distributions. She did not compare relative frequencies. For the Ambulance task with statistics, she said that she remembered that by “less predictable” she was saying that the data for Speedy was “more spread out.” It is unclear if she based that assessment on a comparison of ranges or on something else.

During the interview, before Amber responded to the Ambulance task, she first interpreted the numbers on the horizontal axes of the graphs as “Minutes until the ambulance gets to where they are called.” When asked what the dots in the graphs represented, Amber indicated that they represented, “how many times the ambulance went to a destination in that many minutes.” Finally when Amber was asked to explain what she thought the task was asking her to do, she responded, “it’s asking you to just look these graphs over and determine which is the best ambulance company to go with.” Later during the discourse on this task, the idea of which company was ‘better’ came up. When Amber was asked what she interpreted the term ‘better’ to mean for this task, she indicated that to her ‘better’ meant “more reliable.” As ‘more reliable’ is a reasonable interpretation to make for the Ambulance task, it was assumed that Ann understood the task as it was intended.

In the interview, Amber recommended the Life Line ambulance service. Along with changing her recommendation from what she decided on the survey, she also changed how she examined the data. She made some very rough comparisons of shapes and compared the relative frequencies for the “mid-range” times. She described her estimation process in the following exchange:

A: Speedy is very erratic. You don’t really, you never know what you are going to get, because we have high numbers all across the board. Where as Life Line, um, even though some of their, they basically have this kind of mid-range with the most instances, the most ambulances coming within probably this right here 12 to 20 [Amber’s mid-range marks for Life Line are in Figure 49]... Speedy has, let’s see, lots of incidences of 6, 9, 13, 18, 23. They’re more all across the board, but this one [Life Line] seems more generally reliable.

Int: So there is more in the 12 to 20 in the Life Line than there is in the Speedy? Is that what you are saying?

A: Relatively for each one of these, yes.

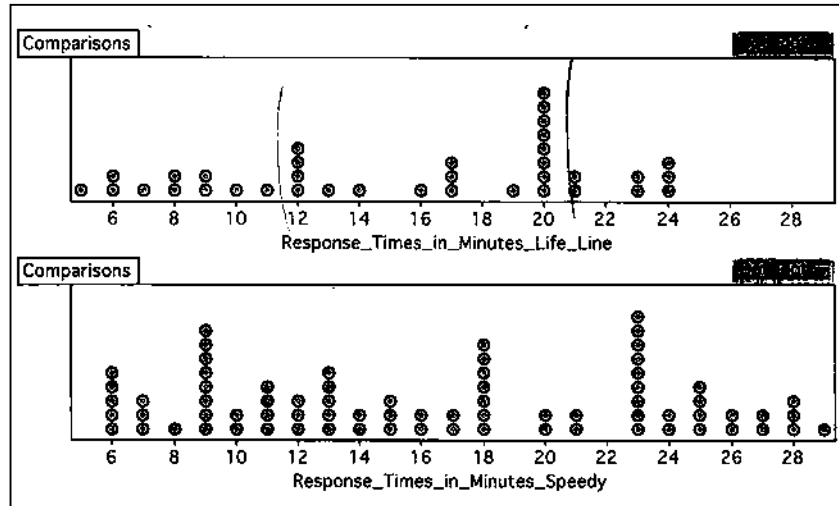


Figure 49. Amber's estimation of "mid-range" times.

When Amber was asked why it is necessary to deal in "relative" terms, she said that,

Ah, because you have to compare, because we only have 36 responses for Life Line and 74 for Speedy. So like, if you just look at the numbers, it looks like, oh Speedy has all these instances of less than 12 minutes and Life Line doesn't have that many, but then you also have to consider that Speedy has almost double the amount of, um, recorded ambulance arrival times.

From the above discourse it appears that Amber compared rough assessments of the proportion of "mid-range" times for each ambulance service. Her partitioning of the data into the three groups of low, middle and high, was similar to the informal reasoning about distribution of middle school students as described by Bakker and Gravemeijer (2004).

In the interview, after Amber examined the descriptive statistics, she again recommended Life Line. She said that she compared the means, medians and modes as well as the interquartile ranges (IQR). She observed that the mean and mode are lower in Life Line and that the IQR for Speedy is 12.75 and for Life Line it is 9.25. Then Amber was asked what the IQR meant to her. She explained, “It means there’s a bigger variance in the times...the range of times is larger in Speedy than it is in Life Line.” She was then asked how that information helped her to make her recommendation. She said,

The sample variance for Speedy is a lot higher than for Life Line and again the descriptive statistics reassert the erratic nature of Speedy and even though the median is like 1 minute higher in Life Line than it is in Speedy, overall it’s, um, you’re going to get a pretty reliable ambulance...

Amber’s statistics class had not yet been introduced to sampling distributions so her reference to the sample variation had to be based on an assumption of what she thought it meant. From her previous description of the meaning of variance as related to the pair-wise differences between data points and her description of the data associated with Speedy as “erratic,” it appears that she did not consider variation beyond range and focused more on assessment of shape. Similar to the Pink/Black task with statistics, Amber listed several of the statistics so support her conclusion yet it was not clear what meanings she attached to them. It was clear that Amber attempted to make global type assessments when she said, “overall... you’re going to get a pretty reliable ambulance [Life Line],” but could not articulate her comparisons well and had some lingering misconceptions about the meaning of variance.

For both Ambulance tasks, Amber's brief explanations on the survey did not provide much information. Before having access to the descriptive statistics, Amber's survey response was so minimal that it was coded as idiosyncratic, level 0. With the extra information about her survey response that was obtained in the interview, Amber's survey response would be classified at level 1. After the inclusion of the descriptive statistics, Amber's survey response focused mostly on variation and was categorized at the Transitional level, level 2 with a focus on variation. That survey response was later confirmed at the conclusion of the interview. In the interview situation, Amber gave considerably more detailed explanations for both tasks. Without the descriptive statistics she examined the shapes and compared the proportion of "mid range" times. With the descriptive statistics she compared the centers with an inarticulate inclusion of shape. Thus Amber's interview responses would be classified at the Initial Distributional level, level 3 with an Initial Global focus.

Eduardo

As with the previous survey tasks, Eduardo continued to focus on variation for both Ambulance tasks, specifically that less variation is better, irrespective of center location. Before he viewed the descriptive statistics, Eduardo decided to recommend the Life Line ambulance service based on his observation that the times for Life Line had "more consistence and the range value is smaller than Speedy Ambulance service." It is not clear if Eduardo's assertion that "Life Line has more consistence" refers to shape or variation.

On the survey, after examining the descriptive statistics, Eduardo continued to recommend Life Line and based his decision on comparisons of the range and standard deviation for each service. He specifically claimed that both measures were “much smaller” for Life Line. The difference between each range is 4 minutes and the difference between each standard deviation is 0.98. Eduardo gave no indication as to how he determined that those measure for Life Line were “much smaller” and it does seem possible that Eduardo could interpret any difference in variation as a significant difference.

In the interview, Eduardo interpreted the numbers on the horizontal axes of the graphs as “The amount of time that it took for the ambulance to respond.” When asked what the dots in the graphs represented, Eduardo said that they represented, “the amount of data points.” Finally when Eduardo was asked to explain what he thought the task was asking him to do, he responded, “If I had the option to call either one, which one would I call?” Later during the discourse on this task, the idea of which company was ‘better’ came up. When Eduardo was asked how he interpreted the term ‘better’ for this task, he indicated that he was looking for less variation to him as well as for individual times ‘better’ meant “faster.” Thus, it was assumed that Eduardo understood the task as it was intended.

Eduardo’s recommendations and explanations in the interview were consistent for both Ambulance tasks and were fairly consistent with his survey responses. When addressing the Ambulance task in the interview, Eduardo acknowledged that the shorter response times were more desirable, but continued his focus on assessment of

variation. He approached the Ambulance task with a slightly different focus than the Pink/Black task where his focus was on the Pink class as better because he perceived it as more consistent. On the Ambulance task, he decided to recommend the Life Line Ambulance service because he perceived the data for the Speedy Ambulance service as having too much variation. Part of his explanation was that,

... Even though there is a lot of calls better, there is a lot of data points that are pretty impressive on the lower end [of Speedy] but you never know when you will have a critical situation... There [points to Speedy] you are really not certain. There is a lot of variations, I mean it can go anywhere from 6 minutes all the way to 29 or 30, it's a big variation. Whereas here, [points to Life Line] the max time is only 24 [sic].

So Eduardo attempted to weigh the possible benefits of short response times versus the detriment of the long response times, but this comparison was frequency oriented. He noted that Speedy has short response times but also has longer response times than Life Line. Eduardo appeared to base his final decision on a comparison of the longer response times for each ambulance service and a comparison of each range of response times. In a particular follow up question Eduardo was asked to assess a proportional type argument in favor of Speedy. That argument was made by another student who described doubling the data at each of Life Line's response times. This student claimed that this process would make the total number of response times about equal for both services and then when the times were compared it would appear that Speedy had more lower times. So the student recommended Speedy. This argument did not change Eduardo's decision. He reiterated that Life Line was more consistent. He said,

Well one of the things that I am seeing by looking at this data, is there's a lot spikes going on in this guy here [points to Speedy]. Where as here [Life Line] ... even if you double the amount you are pretty much [makes horizontal back and forth motion in the air with his hand], you don't have the big spikes that you would with this guy [makes up and down motion with hand then points to Speedy]. So, I still see that as being more, if I have something that is reacting more consistent [repeats horizontal motion with hand], I will stick with that as opposed to something that really varies a lot [repeats up and down motion with hand].

Eduardo's description of consistency was related to shape in that he discussed, with hand motions, comparing the frequencies or heights of the response times. He assessed Speedy as having "spikes" and motioned up and down with his hand, while he assessed Life Line as "more consistent" and motioned horizontally with his hand. This type of shape assessment is a pair-wise comparison of frequencies of response times. So it appears that Eduardo related shape to consistency more than he related variation to consistency.

After Eduardo examined the descriptive statistics, he continued to recommend Life Line primarily because of its lower standard deviation. He also supported his recommendation with the observation that the mean is lower, although it appeared that Eduardo did not necessarily use the comparison of means to contribute to making his decision, he merely cited the comparison as supporting his recommendation.

Eduardo's explanation was:

... I still think the Life Line did better. Of course I'm going back to the deviation, it's a lot smaller, and in this case I have something else to argue about that the mean is better. So here I do have two things. But on the previous one [Pink/Black task with stats] it was a little different. But yes, I would definitely agree that this [Life Line] is better. The deviation is a lot smaller which means it's more consistent.

Eduardo's previously described understanding of standard deviation was tied to consistency, thus his responses from before and after he had access to the statistics coincide as they are both intuitively focused primarily on comparisons of shape in the context of consistency.

On both Ambulance tasks, for the survey and interview, Eduardo's main focus was on assessing and comparing variation and consistency. Eduardo related consistency to shape in terms of comparing pair-wise frequencies of the response times, although at times he also seemed to relate variation and consistency. Eduardo largely did not attend to the issue of the different size of each data set, and thus an assessment concerning his ability to reason proportionally about the Speedy and Life Line data sets is inappropriate. However, Eduardo assessment of the variation for the Life Line times as being "a lot smaller" appeared to be a nominal comparison, not a proportional one. For both tasks on the survey and in the interview setting, Eduardo's explanations were classified as transitional, level 2 primarily addressed comparisons of variation and shape, alternately.

Ann

On the survey, without the descriptive statistics, Ann recommended Speedy but her explanation was rather inconsistent with that recommendation. She wrote, "Speedy had almost half the amount of trails in comparison to Life Line. Yet, Speedy seemed to still yield more instances where the response time was less [sic]." It is possible that she might have simply confused "half" and "twice," and she really meant

to write that Speedy had twice the instances. She also could have mistakenly switched Life Line and Speedy. If that was the case then she would have made a naïve proportional comparison as she would have cited more low time frequencies but fewer overall times for Life Line. Unfortunately without more information Ann's response at it is could only be classified as Idiosyncratic.

After examining the descriptive statistics for the ambulance response times, Ann switched her recommendation on the survey to Life Line. Her written explanation was that, "Life Line has a smaller mean time, and the minimum and maximum were both smaller in comparison to speedy. However I still question the skewness in Life Line." Although this response does not address proportions it does indicate that she is tentatively examining the data sets as wholes by comparing the means and ends. Her reservation about the skewness does highlight her mis-understanding of the meaning of skewness or potential problems understanding proportions.

On the interview, Ann interpreted the numbers on the horizontal axes of the graphs as "The minutes it takes to get to their location." When asked what the dots in the graphs represented, Ann indicated that they represented, "the number of times it took them to get to somewhere." Finally when Ann was asked to explain what she thought the task was asking her to do, she responded,

OK, so basically, before making a decision they wanted to see the data. So they wanted to see how many times did they respond in 6 minutes, how many times did they respond in 7, 8, etc...and then, um, then make a decision on which company to go with based on their times that they normally get.

Later during the discourse on this task, the idea of which company was ‘better’ came up. When Ann was asked how she interpreted the term ‘better,’ she indicated that to her ‘better’ meant “faster.” Thus, it was assumed that Ann understood the task as it was intended. There was also another cue in her response that she may have viewed the data sets as a collection of individual times and not as a whole unit as part of her description of the task was to examine the frequencies of individual response times.

Ann’s interview decision, without the descriptive statistics, was to recommend Speedy. Her initial response was extraordinarily long and a bit difficult to follow. Paraphrasing the major points of her response, she noted that Speedy had more “trials,” and it had “more dots” on the low end of its graph as compared to Life Line. Ann also compared the modes (although she did not use the term “mode”) and noted that Speedy’s mode of 23 is only 3 minutes more than Life Line’s mode of 20. She claimed that 23 and 20 were about the same and then discounted the response times above the modes in influencing her choice. Ann’s final decision, in favor of Speedy, was primarily based on a comparison of frequencies of the shorter response times. She did attempt to informally account for the size difference with a qualifier noting that Speedy had more trials, but she then made no attempt to adjust the frequencies accordingly.

In an attempt to follow up on Ann’s potential to make or understand a proportional comparison, she was asked to assess a proportional type argument that claimed that Life Line had a higher proportion of times below 12 minutes, so Life

Line should be recommended. Ann initially said she would switch to that response but then questioned how valid that type of reasoning was in light of the difference in sizes of the data sets. This was evidence of Ann's incomplete understanding of proportional reasoning.

After examining the descriptive statistics, Ann switched her interview response from recommending Speedy to recommending Life Line, just as she did on the survey. Her switch was primarily based on comparing the means, yet she still had difficulty accounting for the difference in the number of response times. She said,

I don't know, it's kind of hard to say because the mean is higher [for Speedy] so that means they took longer in the end. So Speedy is not really so speedy in comparison to Life Line, but then if I look at the count they [Speedy] had 74 instances and they [Life Line] only had 36, so that's half as much. So, I'm wondering if the count had something to do with it or not. Like if they had given Life Line 74 trials maybe their sum and statistics would end up different. So, hmm, I don't know and then looking at Life Line their minimum is 5, Speedy is 6...

Specific evidence that Ann understood the descriptive statistics as results of calculations on the data was her speculation that changing the count may change the mean, the sum and "statistics." So although Ann's initial statement implying that Speedy's mean was higher so "in the end" Speedy took longer, was a potential use of the means to make a global comparison, she didn't fully understand that comparison globally. Later, Ann continued to discuss the implication of the different size and range of each data set and for Ann they are related. She said,

See that's the other thing with the count, the range is larger [Points to Speedy's data]. So again it seems like, well that'd be great if I got the 5 minute ambulance but then what if I ended up getting the 29 minute ambulance? It's, it seems like I have more variation, like I have more

chances of not getting a fast response versus their [Life Line] range is only 19.

She seemed to relate the range to the number of “chances” of getting specific times. This relationship is similar to her previously described understanding of variation as number of distinct outcomes, which can imply an understanding of data from a local perspective.

For both the survey and interview, Ann initially recommended Speedy, when she did not have access to the descriptive statistics, but then after examining the descriptive statistics she switch to recommend Life Line. Her survey response without the descriptive statistics was idiosyncratic although with further explanation had the potential to be informally proportional. Ann’s reasoning strategy on the survey, after viewing the descriptive statistics, was primarily focused on the means and endpoints, an initial distributional level response with an initial global focus. In the interview Ann had difficulty making comparisons because of the difference in number of response times, a cue that she has difficulty with proportional reasoning. Her reasoning strategy primarily focused on comparing the modes and comparing the ends of the distributions and thus were categorized at the transitional level, level 2 focused on shape. After examining the descriptive statistics, Ann based her recommendation switch to Life Line on a comparison of means and a weak accounting for variation. She continued to have difficulty accounting for the difference in count of each data set and appeared to relate variation to the number of distinct outcomes. Her reasoning strategy was primarily based on comparing the means, yet it was not clear she

understood her comparison from a global perspective and thus was categorized at the Transitional level focused on center.

Lance

On the survey, Lance decided to recommend Life Line. His explanation integrated proportional characteristics with center characteristics, and a comparison the high ends of the distributions. He wrote,

Both arrive within 10 minutes in 1/4 of all cases. Speedy arrives at 23 minutes or later in 1/4 of all cases, while Life Line arrives 20 minutes or later in 1/4 of all cases. Speedy has a 1 minute lower median, but ~ 1 minute higher mean, 3 minute higher mode, and 5 minute higher maximum time than Life Line. They are close on several statistics, but Life Line appears to give a quicker response where they differ.

Lance estimated and compared the proportion of times less than 10 minutes, then he estimated and compared the proportions of times above the mode for each ambulance service. Lance then compared the centers and the longest times for each. Particularly because Lance attended to proportional aspects and centers his response was likely a describing his comparison from a global perspective, of whole units of response times, not a comparison of collections of individual times.

When provided with a list of descriptive statistics, on the survey, Lance gave a minimal response that referenced the descriptive statistics in general, and hence his written response was difficult to categorize. Lance did not change his recommendation, that is, he stayed with Life Line and gave the explanation that “descriptive statistics support observations and rough calculations from previous page.” Lance’s response to this task was similar to response from survey Task 4, the

Pink/Black task with statistics. His responses to each “with statistics” task were coded at level 2-N/A. In these cases Lance appeared to be citing “the statistics” as confirming evidence of his original response, not just meaninglessly quoting the statistical terms.

In the interview, Lance interpreted the numbers on the horizontal axes of the graphs as, “the minutes from, presumably, the time called to when the ambulance arrived.” When asked what the dots in the graphs represented, Lance indicated that each dot represented each response time. Finally when Lance was asked to explain what he thought the task was asking him to do, he responded, “based on the data, which ambulance company one should choose.” Then when Lance probed more on what he meant, he replied, “who ever has the shortest, the fastest measure of response from the data.” Thus, it was assumed that Lance understood the task as it was intended.

In the interview, as with the survey, Lance approached the Ambulance task from a global perspective. Lance spent a rather long time examining both graphs. As he worked through his response, he made some shape comparisons, range comparisons and also noted some specific response times, such as modes and other “peaks.”

Lance’s explanation for his tentative recommendation of Life Line is provided below.

It sure looks like Life Line has, overall, um, shorter, faster response times than Speedy...The first thing I see is that the range goes for a longer time on Speedy and this is a service that quick response time is often very important...better is going to be who ever has the shortest, the fastest measure of response from the data and the first thing I see is that the data is spread over a longer time in Speedy... Now, the spread of response times seems a little bit more even on Speedy than on Life

Line. There is an awful lot on 20 there [points to Life Line graph], but it looks about the same, an awful lot on 23 on Speedy, then an awful lot at 9 on Speedy. So, all that becomes, it gets to be, it gets to be close, with out kind of processing the data more.

It is not clear that he formally considered and compared two global features of the distributions, but by saying that “overall” it appears that Life Line has shorter times, based on range and an comparison of the frequencies of some specific times, Lance approached the comparison from an informal global perspective. For some follow up questions, Lance thought that a recommendation based on the comparison of the high ends of the distributions was possibly acceptable only if a “more analytical” approach was not possible. He also discounted a recommendation based only on comparing the number of response times, i.e., Speedy had more responses times so it is ‘better,’ reaffirming that he was likely considering the distributions as whole units.

After he examined the descriptive statistics, Lance provided a detailed explanation, focused on not only which service was faster but also which was more reliable. He confidently and accurately compared measures of center and variation to form his explanation for Life Line as follows:

To start with the mean median and mode: Life Line is lower on two of them and higher on the median. So it seems to do a little bit better there. Again the range, I hadn’t noticed that this one [Speedy] had a little bit lower low end, but that seems almost, almost trivial... the variance was lower on this one, on Life Line, and they need to be fast but I want them to be, I want to know better what to expect... So, it [Life Line] looks a little bit shorter, shorter time, shorter range, smaller variance, ah, yea so the mean and mode are a little bit lower, the range is noticeably lower, the variance is lower. Ah, I prefer it to be even better, but it’s well enough to choose it [Life Line] over Speedy with a degree of confidence.

Lance compared a variety of features, some local and some global, yet a key in his responses were informal assessments of how significant the differences were. For example, he assessed the difference between the shortest times was “almost insignificant” while differences between some of the measures of variation, range and variance, were “well enough” although he preferred bigger differences. Lance clearly attached meanings to his comparisons, as he was not just monotonously citing various measures and appeared to compare the statistical measures as group representatives.

For both Ambulance tasks, without and with statistics, on the survey and interview, Lance recommended the Life Line ambulance service. On the survey, without the descriptive statistics Lance attended to proportional aspects and centers, globally, and thus his reasoning strategy was characterized at level 4, Distributional. Although after being supplied with the descriptive statistics, Lance provided a minimal explanation that reference the “statistics” to support his previous conclusion and so, without more information, that response was categorized at Level 2 – N/A. Lance’s interview responses without the descriptive statistics, specifically relied on comparisons of variation while informally assessing the shapes and were categorized at the Initial Distributional level, level 3 with an Initial Global focus. After examining the descriptive statistics Lance’s provided a sophisticated response that included comparisons of the centers and variation from a global perspective and was categorized at the Distributional level, level 4. In particular, his interview responses, after examining the descriptive statistics, provide potential evidence that his level 2-

N/A response from the survey would likely be categorized at level 4 with opportunities to further explain and address follow-up questions.

Jill

On the survey, before she had access to the descriptive statistics, Jill decided to recommend Speedy. She cited a comparison of frequencies of specific times but also attended to the difference in sample sizes. Jill's complete written response was,

I like the larger sample size for Speedy (more representative of the population), even though there are [for Speedy] larger frequencies for longer times. There are two lower numbers that have high frequencies [for Speedy] (and higher than Life Line), which is promising to the person in the ambulance.

Jill's response was difficult to interpret, as it is not clear if she compared raw frequencies or informally compared relative frequencies. Jill's assessment of the data set for Speedy as a "more representative" sample was an indication that her comparisons of specific peaks in the data sets were more related to an informal shape assessment than to a comparison of isolated response times.

On the survey, after having access to the descriptive statistics, Jill switched her recommendation from Speedy to Life Line, as a result of comparing many of the statistics for center and variation. Her complete written response was,

I suppose I would choose Life Line because most of the important numbers are less than that of Speedy (mean, mode, SD, range, IQR). While the median is more (17 v. 16) 16 and 17 are awfully close numbers anyway and this difference may be negligible.

As Jill specifically noted that even though the median is higher for Speedy, the amount that it is higher is negligible. Similar to Lance, Jill made an informal assessment of

how significant the difference was between the medians. This was some evidence that she attached meanings to her comparisons and was not just monotonously citing various measures. Jill's comparison of measures of center and variation appear to have been made about the data as a whole global unit.

In the interview, Jill interpreted the numbers on the horizontal axes of the graphs as response times in minutes. When asked what the dots in the graphs represented, Jill indicated that each dot represented, "a sample data point, so I guess it's one ambulance." Finally when Jill was asked to explain what she thought the task was asking her to do, she responded, "to figure out which one was better." Then when Jill was asked what she interpreted the term 'better' to mean, she indicated that to her 'better' meant "shorter time." Thus, it was assumed that Jill understood the task as it was intended.

On the interview, without statistics, Jill spent a rather long time examining both graphs and noting the difference in "sample size" and also noting some specific response times, she summarized her decision to recommend Speedy by using a proportional type reason. She said,

I will go with Speedy because there is a lot of data that is more towards lower response times. Um, a lot at the 9 and a lot at the 18, compared to the proportion of data for the other ambulance.

Jill then estimated that about two-thirds of Speedy's data was between 9 and 18 compared to about half for Life Line. This was rather similar to her survey response in that she cited a comparison of frequencies, however for the survey, her response was rather vague as to if she was considering her comparison proportionally, whereas in

the interview she added an explicit proportional comparison. For one of the follow-up questions for Jill, she assessed another student's recommendation for Life Line that was based on a comparison of both ends of the distributions, i.e. Life Line had the shortest response time and its longest response time was also shorter than Speedy's. Jill rejected that method of comparison as she cited the need to also have additional information about the distributions, such as the shapes, in order to make valid comparisons. She specifically said,

I like to see where it's going to peak, because you don't know. The data could be stacked on the two end points or it could be evenly distributed through the middle if you just looked at the range.

This response reaffirmed that Jill likely did not exclusively compare raw frequencies on the survey, but focused more on comparing shapes.

After Jill examined the descriptive statistics in the interview, she observed that most are in favor of Life Line, but then decided her original reason was more convincing and did not switch her recommendation, that is and she continued to recommend Speedy. The transcript of the exchange in which she describes the follows:

J: *...Life Line was the faster one when you compare the averages [means]...the median was a little bit higher for Life Line, the mode was lower for Life Line. Here I was picking Speedy. Standard error for Speedy is a bit smaller, but the standard deviation is bigger, as well as the variance and IQR and the range. So, I think I picked the one that had higher for everything except the median and the standard error. So while all the numbers seem to be pointing towards Life Line as better, I still pick Speedy.*

Int: *And that's because?*

J: *Looking at more amounts of data down towards lower. So just strictly how many ambulances had these lower times and I guess having that larger sample size makes it more reliable and that shows in that the standard error is smaller.*

Int: *So, is this a proportional argument or a frequency argument or ?*

J: *I think you can turn the frequencies into proportions. I guess I haven't really counted so I don't know exactly, I could be off, but I just see more of the data down lower.*

Jill appeared to begin in a strikingly similar way to her survey responses as she chose the same measures to compare for both responses and cited the same differences between those measures. Yet her conclusions were opposite. Surprisingly, Jill acknowledged the 'better' statistics for Life Line but still decided that her original argument, that was focused on shape and informally focused on proportion, was more convincing. It was not clear how to interpret Jill statement that Speedy's "larger sample size makes it more reliable and that shows in that the standard error is smaller." At the end of the interview, when Jill was specifically asked about her understanding of the standard error of the mean she correctly described it as the standard deviation of a sampling distribution, yet the researcher does not understand how she has applied that information to this task.

Jill's responses to the Ambulance task, without the descriptive statistics, were similar in that she made comparisons of frequencies at some specific values that led to a shape assessment on the survey and a shape and proportions assessment on the interview, all in favor of Speedy. So, on the survey, Jill responded at the transitional level but because of her explicit comparison of proportions, between two 'cut-points,'

her interview response was Initial Distributional. After Jill examined the descriptive statistics, for both the survey and interview, she compared various measures of center and variation and determined most favored Life Line, yet came to different recommendations. On the survey she switched her recommendation to Life Line, but in the interview she stayed with her Speedy recommendation. Apparently in the interview she felt that her proportional argument was more convincing. Thus her with statistics survey response was categorized as Distributional but her interview response remained classified at the Initial Distributional level with a focus on proportion, because although she compared several features of the distributions, she fell back to rely on her initial observations.

Summary of the Ambulance task without and with descriptive statistics

The results of the responses, for all six interviewees, to both Ambulance tasks are summarized in table 54. Similar to the Pink/Black task, each of the first four students, Jack, Amber, Eduardo, and Ann, experienced trouble in accounting for the difference in the size of the data sets and thus reasoning about proportions appeared to be a significant conceptual obstacle for all four of those interviewees. Unlike the Pink/Black task, reasoning proportionally about partial distributions for the Ambulance task did not lead to one recommendation over the other. Both Lance and Jill cited proportional comparisons of pieces of each distribution. Each student compared different pieces and contributed to Lance's recommendation of Life Line and Jill's recommendation of Speedy.

Table 54.

Interviewees' decisions and response levels for tasks 5 and 6: the Ambulance task (without statistics), and the Ambulance task with statistics.

Student (Group)	Format	Without stats		With stats	
		Without stats Decision	Response Level	With stats Decision	Response Level
Jack (1-GS)	Survey	Speedy	2-SH	Speedy	0
	Interview	Speedy	2-SH	Life Line	2-C
Amber (1-GS)	Survey	Speedy	1	Life Line	2-V
	Interview	Life Line	3-IG	Life Line	3-IG
Eduardo (1-SE)	Survey	Life Line	2-V	Life Line	2-V
	Interview	Life Line	2-SH	Life Line	2-SH
Ann (2-GS)	Survey	Speedy	0	Life Line	3-IG
	Interview	Speedy	2-SH	Life Line	2-C
Lance (GRAD)	Survey	Life Line	4	Life Line	2-N/A
	Interview	Life Line	3-IG	Life Line	4
Jill (GRAD)	Survey	Speedy	2-SH	Life Line	4
	Interview	Speedy	3-P	Speedy	3-P

Without access to the descriptive statistics, most of the interviewees focused on comparing shapes. This was not surprising as the differences between most of the distributional features were not clearly evident through visual inspection. With access to the descriptive statistics, all the interviewees focused on comparisons of measures of center and variation, however Lance and Jill were the only ones who clearly and consistently used the measures as global representatives. Although Jack, Amber, Eduardo, and Ann each gave responses that contained elements of viewing the data sets as collections of individual points, a Local perspective, they also attempted to

make comparisons the characteristics of either shape, center or spread, but also did not fully understand those characteristics as global representatives particularly after they were provided with some statistics, their responses contained evidence that these students generally understood statistical measures, such as the mean, more as computations with the data as opposed to group characteristics. Thus, their perspectives of the Ambulance data sets were in a transition out of a purely Local view towards a more global view. Although Lance's survey responses were minimal, his interview responses were considerably more detailed, confirmed his interview responses, and showed it was very likely that his perspective of the Ambulance data sets was global. Jill's responses were the most difficult to interpret. As she seemed most comfortable basing her decisions on informal shape assessments and proportional assessments about the partial distributions, her perspective of the Ambulance data sets did not appear to be consistently global. In general, when students were provided descriptive statistics, those who made meaningful comparisons and meaningful assessments of differences between statistical measures, provided some evidence that they understood the measures, such as the mean, as group representatives, and thus were likely making their comparison from a global perspective.

Summary of Cross Task Numeric Code assignment

The Cross Task Numeric codes based on their survey response and then on their interview responses are shown in Table 55.

Table 55.

Interviewees' cross task numeric framework levels.

<u>Student</u> <u>(Group)</u>	<u>Format</u>	<u>Cross-Task Numeric Code:</u> <u>Framework Level</u>
Jack	Survey	1
(1-GS)	Interview	2
Amber	Survey	1
(1-GS)	Interview	2
Eduardo	Survey	2
(1-SE)	Interview	2
Ann	Survey	2
(2-GS)	Interview	2
Lance	Survey	3
(GRAD)	Interview	4
Jill	Survey	4
(GRAD)	Interview	3

Not surprisingly, Jack and Amber saw the most dramatic increases on a task-by-task basis from survey to interview, and consequently their cross task numeric codes also increased. They had attended several more statistics classes in-between the survey and interview and often explicitly stated in the interview that they were attempting to use some of what they had learned in class. Jack's and Amber's survey responses across the tasks were fairly consistent at the Local level, level 1, with occasional level 0 and level 2 responses. Thus, at the time of the survey it is possible that they typically viewed data sets from a local perspective. Both Jack and Amber provided higher level responses in their interviews. It was clear that Jack had attempted to use some of statistical knowledge that he was trying to learn in class, but did not yet understand their meaning and uses. Thus, in the interview Jack consistently

responded at the Transitional level. In the interview Amber also attempted to use some of statistical knowledge that she was trying to learn in class. Her explanations were somewhat closer to global than Jack's explanations, yet, Amber's were still a bit inarticulate. Both Jack and Amber encountered difficulty in making valid comparisons of data sets with unequal counts and thus had considerable difficulty understanding proportional reasoning, although, in the interview, both students appeared to be on the verge of grasping proportional arguments.

Eduardo and Ann were strikingly consistent in providing level 2 responses on the survey and in the interview. Thus, both students' cross task numeric codes also remained consistent at level 2. This was not surprising for Ann as she was enrolled in her second statistics class, but Eduardo was in the same situation as Jack and Amber, thus it was surprising that the extra time that he spent in statistics class, from survey to interview, did not apparently effect how he viewed the data sets and approach the tasks. Eduardo was employed in the high tech industry and it did appear that his work experiences using statistics had a considerably larger impact on his comparisons than his limited time he had spent in statistics class. Eduardo's responses across the survey tasks and in the interview were consistently categorized at level 2 and exclusively focused on either centers or shapes or variation. Thus, at the time of the survey and at the time of the interview it is possible that Eduardo's perspective of data sets was in transition from local to global. On the survey and in the interview, Ann's responses across the tasks fluctuated between local and transitional and global. Ann fairly consistently considered comparisons of frequencies then extended her reasoning to

either a transitional or initial global level. Thus, at the time of the survey it is possible that Ann's perspective of data sets was also in transition from local to global. Both Eduardo and Ann also encountered difficulty in making valid comparisons of data sets with unequal counts, a cue that they have difficulty with proportional reasoning.

Lance's responses across the survey tasks fluctuated between Transitional and Distributional. So although his perspective of data sets, at the time of the survey, was potentially global, the fluctuation in his response levels is reflected in his cross task numeric code of level 3. In the interview Lance provided considerably more information about his reasoning strategies and those strategies were all at the Initial Distributional level and Distributional level. This considerable increase in detailed sophistication was reflected in his cross task numeric code being increased to level 4. Thus, Lance had likely made the various comparisons in the tasks from a global perspective, in the interview, and likely made his level 2 comparison on the survey from a global perspective as well.

Finally, at the time of the survey, Jill's responses across the survey tasks fluctuated between Transitional and Distributional, but were mostly Distributional, particularly on the tasks that required comparisons of unequal sized data sets. Thus, her cross task numeric code for her survey responses was at level 4 so it is possible that her perspective of data sets at the time of the interview was global. Her interview responses also fluctuated between Transitional and Distributional, however, those responses, although more detailed, were more consistently at the level 3, Initial Global level, than at the Distributional level. Hence, Jill's focus on comparing partial

distributions, particularly on the Ambulance task in the interview, is reflected in the cross task numeric code of level 3 for her interview responses. So her perspective of the Ambulance data sets is apparently not clearly global, although nearly so.

The results from the interviews generally validated the coding categorizations of the survey responses as lower bounds of students' reasoning. The interview responses also provided further evidence of the separation of statistical reasoning abilities between students who understood proportional reasoning and students who did not. The four students who showed clear difficulties in understanding proportional arguments also had limited understandings of the statistical measures and how they can be used as group representatives to make valid comparisons. Those four students' arguments tended to make their comparisons from perspectives that fluctuated between being aligned with the Local framework level, to the Transitional level, to the Initial Distributional level. The two students who showed proficiency in using and understanding proportional arguments tended to make their comparisons from perspectives aligned with the Initial Distributional and Distributional framework levels. Those two students also regularly used statistical measures as group representatives.

Chapter 5

Discussion and Conclusion

This research study was designed to investigate students' reasoning strategies as they engaged in tasks that required making informal statistical inferences about pairs of data sets. It was a descriptive study with a major component focused on building and then refining an interpretive framework. In order to observe a wide spectrum of responses and reasoning on the tasks, this research involved a large number of participants from a diverse group of university students at the undergraduate, post baccalaureate, and graduate levels.

Data were initially collected from a task-based web survey completed by 275 undergraduate and graduate students who were enrolled in statistics courses. Additional data were collected through follow up interviews, six of which were analyzed for this study. The survey consisted of six tasks; each task contained two data sets of quasi-real data derived from the context of each task. The data sets were presented in graphical form and students were asked to make decisions based on the data and explain their decisions.

Research Goal: Expand and refine the interpretive framework

An important goal of this research was to further expand and refine the conceptual framework, originally developed by Shaughnessy, Ciancetta, Best, and Noll (2005), for describing middle and high school students' statistical reasoning in a sampling environment. Based on an extensive review of the relevant literature, the researcher hypothesized that the framework by Shaughnessy et al. (2005), could be

expanded for use in describing the statistical reasoning of students with more mature educational backgrounds and possibly more sophisticated statistical backgrounds, as they reason about data that were not explicitly set in a sampling environment. Figure 50 displays the framework by Shaughnessy et al. (2005) with the corresponded final refined version of the Expanded Lattice Structure Framework that resulted from this research.

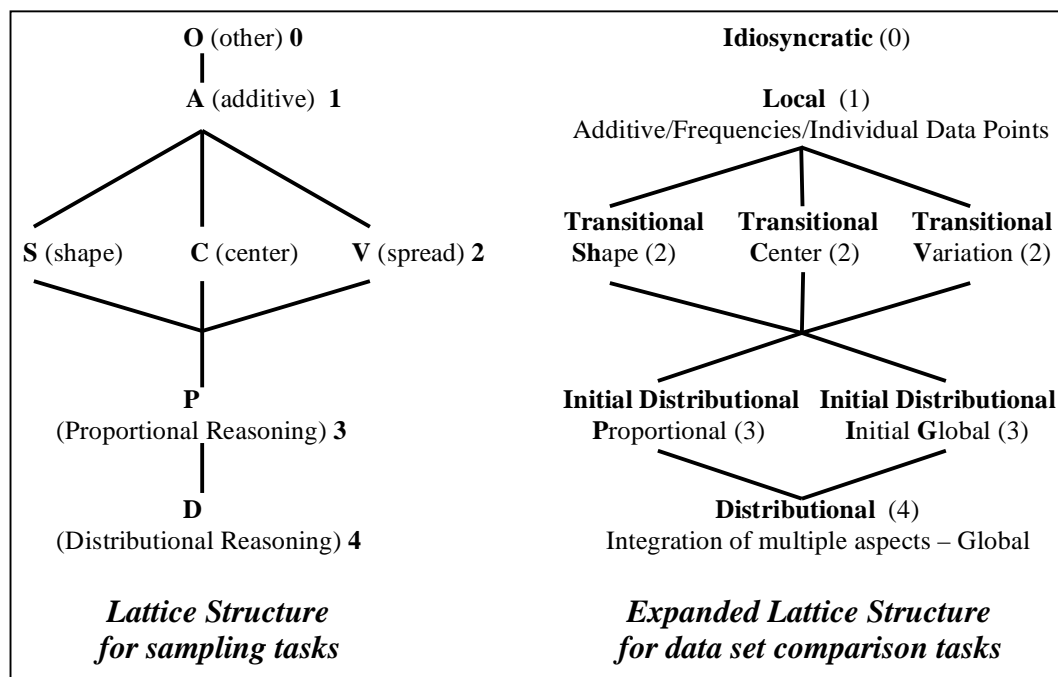


Figure 50. Partial evolution of the Lattice Structure Framework used to describe students' statistical reasoning.

Through the process of constant comparison (Dey, 1999; Strauss & Corbin, 1990) as part of the survey analysis, the Expanded Lattice Structure Framework was refined to the form shown in figure 45. The substances of these refinements are summarized next.

At level 0, the new descriptor “Idiosyncratic” was employed to highlight that the students lacked a focus on reasoning about the data and thus provided no information about the students’ statistical reasoning, other than a potential inability to reason about data. This level was also disconnected from the lattice to further emphasize that the response was not helpful in characterizing the students reasoning about the data.

At level 1, the new descriptor “Local” was employed to more accurately capture the types of responses in which the student potentially reasoned about the data from a local perspective. These types of responses include those that were similar to the “additive” description in the former framework, but also included responses that indicated that the student viewed the data sets as groups of individual points not as whole units.

At level 2, the descriptors were largely unchanged, as those responses that focused on one particular feature of the distributions were distinctive in both sampling and non-sampling contexts. The responses coded at level 2 focused on only one feature of a distribution, as there was little evidence that those students considered that single feature as a group representative. Yet by using a group feature to make a comparison or description of data, those students have taken a step away from considering the data as an amalgam of individuals and towards considering the data as a whole unit. A potential limitation for using the framework to describe student reasoning emerged on the two tasks that had descriptive statistics included with the data sets. Between 10% and 12% of all students provided explanations for their

decisions on those two tasks by either making a reference such as, “because of the stats” or listing many names of the statistical measures, not all of which, when compared, supported their decision. Those responses were coded at level 2 with an “N/A” descriptor, to imply that the response was not explicitly associated with one of the three characteristics.

Level 3 was expanded and given the overall descriptor of “Initial Distributional” to capture responses that focused on comparing the proportions of partial distributions or focused on a single feature of the distribution with an informal integration of other group features or local features that together provide an informal global “picture” of the data. The responses coded at level 3 showed more sophistication than the level 2 responses, yet it was not clear that they were entirely focused on the distributions as whole units.

Level 4 was slightly expanded to capture the possibility of students comparing the proportions of partial distributions with another group feature, such as center. Thus students whose responses integrated comparisons of multiple group features, or integrated a proportional comparison with a comparison of a group feature, were likely focused on the distributions, for a particular task, as whole units, i.e., from a global perspective.

A variety of reasoning strategies, at any of the lattice levels 1 – 4, were potentially valid to support the possible decisions, for each task comparison, although this does not necessarily imply that all reasoning strategies were valid. Codes were assigned conservatively to responses, based on the framework levels, and thus

represent a potential lower bound for the reasoning of each student on each task. Initial steps toward assessing the reliability and validity for using the framework to code the responses to the tasks were also taken.

Reliability was addressed through the four independent coders, who were each trained in using the Expanded Lattice Structure Framework and each coded about 5% of all the responses. The two coders who had experience in using similar frameworks consistently had at least 70% agreement with the researcher. The two other coders did not meet that standard, however one had less initial training and the other had no prior experience using similar frameworks to code task responses. Thus there is potential for reproducibility of code assignments to similar tasks using the Expanded Lattice Structure Framework as refined from this research.

Validity was addressed through triangulation of the survey responses and interview responses of six participants. Students who were interviewed not only responded to the survey tasks, but also responded to follow up questions designed to investigate their understanding of responses classified at different framework levels. At the conclusion of each interview, students described their understanding of some common descriptive statistics measures and that information was used to aid in interpreting their task responses. Although interviewees did not always provide the same responses on the survey and in the interview, when responses differed from survey to interview those differences were addressed near the conclusion of the interview, and in each case the student confirmed that the interpretation of the task response was an accurate reflection of their thinking at the time they took the survey.

For all interviewees, except for one student on one task, their interview responses were coded at equal to or higher framework levels than their survey responses, thus confirming the lower bound nature of the survey codes.

Research Questions

Students who participated in this research study were separated into five distinct groups, with varying statistical backgrounds. The research questions for this study will be answered in the context of comparing and contrasting the responses from each group. Responses of all the participants, on the whole, will not be addressed as the 1-GS group accounted for half of all participants, and that group had the least statistical experience with the least mathematics in their backgrounds. Thus any assessment of all the participants, on the whole, would tend to heavily reflect the assessment of that largest group. The statistics backgrounds of the students in each group are: 1-GS, students who were beginning their first general statistics class; 1-SE, students who were beginning their first statistics class specifically designed for engineers and scientists; 2-GS, students who were beginning their second general statistics course; 2-SE, students who were science and engineering majors and who had completed more than one statistics course; and GRAD, graduate students and senior level undergraduate students who had completed at least five statistics courses. In general students with less statistical experience provided responses at lower levels than students with more statistical experience. The details of the results are addressed as the research questions are explored.

Research Question 1:

What types of reasoning strategies do students use when making comparisons of data sets? Specifically, are the strategies global or local or in transition from local to global?

A goal of statistics education is to help students develop a global perspective of data and then use that perspective to make global comparisons of data sets. This will lead to more fully understanding statistical inference, in particular tests, such as a t-test, as a global comparison of distributions. The cross task numeric codes were one indicator used to provide some evidence as to the dominant perspective that the students, in this study, made their comparisons from.

The “GS” groups of students, who were enrolled in either the first or second term of general statistics, were the only groups to have some of their students receive cross task numeric codes at level 1. Those “level 1 students” likely made their comparisons from a local perspective of the data sets, as they fairly consistently responded in ways that indicated they considered the data as an amalgam of points and not a whole unit.

Groups 2-SE and GRAD were the only groups who had students who received cross task numeric codes at level 4. Those “level 4 students” likely made comparisons of the data sets from a global perspective, on a consistent basis. They also tended to provide responses that effectively employed proportional reasoning and/or had indications that statistical measures, such as the mean, were group representative and could be used to compare groups of different size.

Groups 1-GS, 1-SE, 2-GS, and 2-SE all had the highest percentage of their students classified at cross task numeric codes of level 2. Those “level 2 students” likely made comparisons of the data sets from a perspective that was in transition away from local and toward global. They either consistently made their comparisons based on a single feature of the distributions or they made some comparisons from a global perspective and others from a local perspective, but not always appropriately so. Additionally, they tended to use statistical measures to make comparisons without necessarily understanding their meaning beyond how they are calculated. Many of the “level 2 students” also had considerable difficulties in making comparisons of unequal sized data sets, an indication that they also had difficulty understanding proportional arguments.

The GRAD group was also the only group to have the majority of its students classified at a cross task numeric code of level 3. Those “level 3 students” likely made comparisons of the data sets from a perspective that was also transitioning away from local and toward global, but this type of perspective is close to being global. Their comparisons focused on partial distributions or focused on one group feature with an informal or incomplete incorporation of other features. They also may have provided responses that effectively employed proportional reasoning and/or had indications that statistical measures, such as the mean, were group representatives and could be used to compare groups of different size.

The results of this study also provide some evidence of the separation of statistical reasoning abilities between students who understood proportional reasoning

and students who did not. Students who showed clear difficulties in understanding proportional arguments also had limited understandings of the statistical measures and how they can be used as group representatives to make valid comparisons, particularly when comparing groups of different size. Students who showed proficiency in using and understanding proportional arguments tended to make their comparisons from perspectives aligned with the Initial Distributional and Distributional framework levels. Those students also regularly used statistical measures as group representatives.

The students from the GRAD group consistently provided responses at higher framework levels than any other group and the 1-GS students consistently provided responses at lower framework levels than any other group. Those trends are not surprising because of the large number of statistics courses that students from the GRAD group have completed and the minimal mathematics and statistics backgrounds of the 1-GS students. It was rather surprising that while the 2-GS students responded consistently at higher levels than the 1-GS students, the differences, although significant, were minimal. Apparently, the one general statistics course that the 2-GS students completed did have some positive impact on the reasoning of the “GS” students. It was also surprising that the 2-SE students did not consistently respond at higher framework levels than the 1-SE students. Although the comparison may not be completely accurate because of the very small size of the 2-SE group, the 2-SE students had more statistics in their backgrounds with equal or more mathematics. It is not clear what factors may have lead to this trend.

Research Question 2:

What aspects of distribution (i.e. center, shape, spread) do students attend to when comparing data sets?

Making group comparisons is at the heart of statistics as it leads to the most basic questions in statistics, that is, to examine differences between two data sets in order to ascertain if some factor has produced a difference or differences between them (Konold & Higgins, 2003; Konold & Pollatsek, 2002). Thus, the educational research community has given increased attention to research, instruction, and assessment of students' understanding of data sets as distributions that are understood and can be examined and described in terms of shape, center, and spread, among other features (Garfield & Ben-Zvi, 2004). By investigating what students attend to as they make comparisons of data sets, we can gain insight into how they understand the most basic of statistical questions and about their understanding of data sets as distributions.

Comparing equal sized data sets

The first two survey tasks required students to compare pairs of small data sets of equal size. Each of the data sets in the Yellow/Brown task were both approximately mound shaped with all three centers coinciding within each distribution. The equality of their centers between the distributions was assumed to be visually evident along with the slight difference in variation. The data sets in the Movie Wait-Time task had their means and medians coincide within each distribution and between the distributions. This information was provided to the respondents, as it may not have

been visually evident. Both Movie Wait-Time data sets were bi-modal and it was assumed that the distinct difference in range was visually evident.

The trends of decisions made about the data set comparisons in these first two tasks were quite different. For the Yellow/Brown data sets, the two 'GS' groups were about evenly split between deciding that the data sets were equal versus deciding that one was 'better' than the other. The rest of the groups decided largely in favor of the groups being equal. However, for the Movie Wait-Time data sets, all the groups tended to favor the decision that the data sets were not equal as opposed to equal.

Irrespective of which of the first two tasks that students were reasoning about, undergraduate students in this study, who saw the data sets as 'equal,' tended to focus on comparing centers to support their decision. They often focused exclusively on centers, and a few informally considered other aspects of the distributions as well. The group of graduate and senior undergraduate students who had completed many statistics courses was the only group who had a considerable portion of their students, who decided 'equal' and supported that decision by comparing and relating multiple aspects of the distributions.

Students who decided that the pairs of data sets from the first two tasks were unequal tended to focus on different aspects of the distributions than those students who decided the data sets were equal. The undergraduate students, particularly the 1-GS group, made more comparisons of local features to support the 'unequal' decision as opposed to the 'equal' decision. However, similar to those who decided that the data sets were equal, the majority of undergraduates who decided that the data sets

were unequal also focused on comparing a single feature of the distributions. This single feature was generally either variation or shape. Also, students who decided specifically that the Movie Wait-Time data sets were unequal tended to provide more responses that incorporated comparisons of multiple features, distributional type responses, as opposed to those students who decided that the Yellow/Brown data sets were unequal.

For the first two data set comparisons, where the data sets were small and of equal size, all students rarely made comparisons based on proportional type strategies. The students from the GRAD group consistently provided responses corresponding to the higher framework levels and the 1-GS students consistently provided responses corresponding to the lower framework levels.

Comparing unequal sized data sets

The next two tasks, along with both of their follow up tasks, required students to make comparisons of unequal sized data sets. For the Pink/Black task, the 'Black' data set was smaller than the 'Pink' data set. Many similarities and differences between the features of those two data sets were assumed to be visually evident. Both data sets were unimodal with the 'Pink' data set symmetric and the 'Black' data set skewed. The ranges were equal, but the 'Black' data set clearly had higher measures of center. For the Ambulance task, the 'Life Line' data set was smaller than the 'Speedy' data set. Similarities and differences between the features of the data sets were assumed to be not easily determined through visual inspection. The one exception was that it was clear that the range of the 'Life Line' data set was smaller

than the range of the 'Speedy' data set. Both data sets were unimodal with the 'Life Line' mode located at a lower time than the 'Speedy' mode, however, the 'Speedy' data set had several other 'peaks' located at lower times than the 'Life Line' mode. The 'Life Line' data had a smaller mean while the 'Speedy' data had a smaller median. Thus, it was assumed that comparing these data sets would be more challenging than any of the other comparisons.

For the Pink/Black task, all students who provided responses at level 3 or level 4 successfully decided that the Black class scored 'better.' A key to making this decision from an advanced perspective was understanding that while the 'Black' data sets was smaller, it had proportionally more data shifted higher. Confirming evidence that students who made comparisons of the Pink/Black data sets at levels 3 or 4 also were viewing the data as whole distributions is that 85% of those students also made reasonable estimations for how much better the Black class scored, by using either the centers or proportions of partial distributions as group representatives to estimate differences. Students who responded at level 4 for the Ambulance task all decided in favor of the 'Life Line' data set mostly by incorporating comparisons of the mean and variation. However students who responded at level 3 for the Ambulance task did not have similar success as they were about evenly split between favoring each data set. This may be because there was no clearly visible proportional shift between the data sets that might lead to different conclusions, depending on which piece of the distributions were compared. The GRAD group provided the highest percentage of level 4 responses for each task and also was the only group to have almost all of its

students consistently provide responses at level 2 and higher. Thus, based on the results of this research, the Expanded Lattice Structure Framework seems to be well suited for capturing the levels of sophistication that the students from the GRAD group worked at, and it seems to be a useful tool for capturing important distinctions across students of varying levels of educational/statistical experiences.

For the Pink/Black task, the ‘GS’ undergraduate students who decided that the Black class scored better primarily provided responses at levels 2 and 3, with more at level 2. The 1-SE undergraduate students also primarily provided responses at levels 2 and 3, but they provided slightly more at level 3. The 2-SE students provided responses at levels 2 and 4 with considerably more at level 2. Across all groups the students who responded at level 2 overwhelmingly focused on comparing centers. Yet almost 60% of those students had difficulty with estimating how much better the Black class scored, as they did not use the centers as group representatives to find the differences, nor did they use any other reasonable group representatives to find the difference between the groups. This is evidence of the Transitional nature of these students perspective of the Pink/Black data sets as these students made comparisons based on group features but they appeared to not understand them as group representatives.

For the Ambulance task, both groups of ‘GS’ students and both groups of ‘SE’ students tended to provide their responses at levels 1 and 2, irrespective of their recommendation. Most of the students who responded at level 2 relied on informal shape assessments or on comparisons of one particular measure of center, i.e., they

compared only means, only medians or only modes. Depending on which feature was focused on, different conclusions could be reached. Thus it is unlikely that those students were making those comparisons from a global perspective as they did not seem to view the distributional as whole units.

Students who provided level 1 responses for the Pink/Black task primarily came from the ‘GS’ groups, while for the Ambulance task between 20% and 45% of each groups’ students provided level 1 responses (except for the GRAD group). Many of the level 1 responses indicated that the students had difficulty resolving how to account for the different sizes of the data sets. Others ignored the size difference and based their decision on finding the sum to support either the ‘Pink’ or ‘Life Line’ decisions. In either case, those students had clear difficulties with understanding proportional reasoning in the context of the research tasks. Still other students based their decision solely on comparing individual data points, a sign that they view those data sets as collections of individual points, not as whole units.

As with the comparisons of equal size data sets, the GRAD students reasoned about the comparisons of the unequal size data sets at consistently higher framework levels than any other group and the 1-GS reasoned about the comparisons at consistently lower framework levels.

Comparison without vs. with descriptive statistics

Both of the Pink/Black and Ambulance tasks had follow-up tasks where students had access to some descriptive statistics associated with each data set. On these tasks students were allowed to revise or change their original responses. For both

tasks groups provided responses at higher framework levels and tended to shift their decisions to largely favor ‘Black’ and ‘Life Line.’ For both tasks, students from all groups tended to abandon proportional arguments in favor of arguments that used the statistical measures that were provided. For the Pink/Black task, there was also a considerable increase in the number of students who estimated the difference between the groups by finding the difference between the centers. Overall, it was not clear that the Expanded Lattice Structure Framework captured that students reasoned at higher levels. That is, it is likely that students felt that that task was specifically asking them to use the provided statistics to make their decision and consequently their revised response was coded at a higher level. However, the results for the “with statistics” tasks did highlight that some students clearly did not understand the meaning of the measures as they monotonously and meaninglessly cited measures to support their decisions, even if some of the comparisons of the measures contradicted that decision.

Limitations of the research

There are several limitations regarding this research that I will mention. One is concerned with how the participants were obtained, another concerns the contexts of the tasks, and the last concerns the applicability of the lattice levels to advanced statistical reasoning.

Participant Selection

All the participants for this research were self-selected. All volunteered and most, but not all, of the undergraduate students received extra credit applied to their

statistics course for completing the survey. Thus, extending the results of statistical comparisons between the five groups to larger populations of students is not appropriate. Also, the surveys were completed over a two-week period and some students potentially took the survey after only one instructional session while others may have taken the survey after several instructional sessions. For those students who were beginning their first course, the difference between having one versus several instructional sessions in statistics could have a significant impact on their thinking about the data set comparisons. For example, both Amber and Jack were just beginning their first statistics course and they each were interviewed several weeks after they completed their surveys. Both Amber and Jack gave responses in the interview that were at a higher lattice level from their survey responses, and while some of the change can be attributed to the interview environment where more detailed explanations can be given, some of the change is also attributed to their experiences in their statistics class.

Task Contexts

The tasks were set in two main contexts of comparing test scores and comparing wait/response times. In each task, the students were required to make some interpretations. For example, students had to form their own interpretation for what criteria to assess ‘better’ test scores. A large majority of the students interpreted ‘better’ scores as higher scores, however a few students, such as Eduardo, focused on consistency and variation for assessments of ‘better.’ Because there were few of these

students, their alternate interpretation of ‘better’ was not accounted for in the analysis. Also, knowledge of the context, such as with the Ambulance response time task, may promote the addition of contextual reasons and explanations for the decisions. Thus it is possible that tasks using similar data sets but embedded in different contexts may yield different responses and different categorizations of lattice levels.

Finally these tasks were designed to elicit informal comparisons, that is, without the use of statistical tests or other investigative techniques. Thus the Lattice Framework constructed as a result of this research only applies when students are making informal inferences and making intuitive assessments of the data sets. The framework has no structure to describe how students may respond when using more advanced statistical techniques.

Implication for future research and teaching

There are two areas for which I recommend future research relating to understanding informal conceptions of distribution. The first area concerns the refinement and testing of the Lattice Framework. The second area concerns curriculum development.

Refinement

Refinement of the levels and their branches within the framework should include investigations with research tasks crafted to explore possible expansion and

clarification of the levels. For example, it is possible that refinement of the Local level would include reorganization into branches such as a branch for a focus on sums and a branch for a focus on absolute frequencies. Conducting interviews with students before they begin their first statistics course may aid in clarifying and teasing out possible branches of the Local level. Refinement of the branches at the Transitional level and Initial Distributional level might address the issue that sometimes students do not separate their meaning for variation, shape, range and consistency. New tasks or new lines of questioning could be designed to investigate these problematic situations. These new tasks could address similar types of questions but be embedded in an Exploratory Data Analysis environment that allows students to change the graphical representations or construct their own representations and decide whether or not to use descriptive statistics.

Curriculum Development

An immediate use for the tasks from the data set comparison survey and the results associated with students' responses to those tasks is to design an assessment tool. This tool could employ tasks used in this study with a group of representative responses from students in a multiple-choice format. Initial steps toward the reliability and validity of such an assessment could be claimed, based on the results of the research study. Once proven reliable and valid, this assessment would have great potential for use by teachers as it could be quickly completed and graded. It could be

used as either a pre- and post-assessment tool or to gain a “snap shot” of student thinking and understanding of distributions and data set comparisons.

Based on the results of this research study, it appears that for undergraduate university students, taking one statistics course is sufficient to expand their perspective, of many types of distributions, to a transitional state that is broader than a local perspective yet not completely global. Many traditional statistics courses investigate, separately, the individual group features of distributions of data, such as shape, center and spread, yet those features are intertwined as none exist without all the others. This type of course design may tend to promote reasoning about features of distributions in isolation of each other. A curriculum designed to specifically focus on the inseparable connections between the group features of distributions could be organized around reference to the Expanded Lattice Structure Framework. Instruction could specifically address the visible separation and then joining of the branches of the lattice structure, at various levels, to illustrate of how various features of a distribution can be initially considered separately but then must be integrated to form a whole complete distribution. This type of course may effectively promote reasoning about features of distributions in the context of the distribution as a whole, as opposed to individual data points or unconnected individual features.

Use of the Data Set Comparison tasks could be an integral piece of the proposed curriculum. The data set comparison tasks promoted reasoning at all levels

of the framework. Sharing various strategies and thinking about these tasks could be an effective launching point for student discussions and debate about valid comparisons of data. The sequence of tasks could be particularly useful for comparing and contrasting various strategies for comparing distributions of equal size versus comparing distributions of unequal size and the integral part that understanding proportional reasoning plays in comparing distributions of unequal size. After students' informal comparisons of distributions, instruction could be designed to evolve into investigating comparisons with descriptive statistics and then with formal statistical test, each of which could be related back to the students' initial informal comparisons. Thus promoting a deeper understanding of data sets as distributions and providing a solid foundation from which to understand statistical inference.

Conclusion

Statistics students' informal conceptions of distribution are vital to understand because of the integral role that the concept of distribution plays in data analysis and understanding statistical inference. Making and understanding valid, yet informal, comparisons of data sets lays the foundation for understanding formal statistical inference. This research adds to the literature in the area of statistical education by offering an in-depth exploration of university-level statistics students' informal comparisons of data sets and provided insight into those students' conceptions of distribution. This study is possibly the first of its kind to investigate graduate and

senior level undergraduate statistics students' conceptions of distributions. It also is one of hardly any studies that have investigated the conceptions of science and engineering students. In general, the graduate students consistently provided responses at the higher levels of the framework and thus appeared to perceive the data sets globally. The undergraduate students, whether or not they were science and engineering majors, largely responded at the middle levels of the framework and thus likely perceived the data sets from a perspective that was in transition from local to global.

The other significant contribution is the development and use of The Lattice Framework, a five-tiered interpretive framework for statistical reasoning that was constructed based on an integration of existing frameworks and the survey and interview data. This research has contributed to the groundwork for understanding students' informal conceptions of distribution. In doing so it points to the need to improve instruction at the college level, in ways that might result in better experiences for students and their increased success in understanding and applying statistics.

References

- Albert, J. H., & Rossman, A. J. (2001). *Workshop Statistics: Discovery with data, a bayesian approach*. Emeryville, CA: Key College Publishing.
- Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, 3(2), 64-83.
- Bakker, A., & Gravemeijer, K. (2003). Planning for teaching statistics through problem solving. In H. L. Schoen (Ed.), *Teaching mathematics through problem solving: Grades 6-12* (pp. 105-117). Reston, VA: National Council of Teachers of Mathematics.
- Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 147-168). Boston, MA: Kluwer Academic Publishers.
- Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (1994). Errors and Difficulties in Understanding Elementary Statistical Concepts. *International Journal of Mathematics Education in Science and Technology*, 25(4), 527-547.
- Batanero, C., Tauber, L. M., & Sánchez, V. (2004). Students' reasoning about the normal distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 257-276). Dordrecht: Kluwer Academic Publishers.
- Ben-Zvi, D. (2002). Seventh grade students' sense making of data and data representations. In B. Phillips (Ed.), *Developing a Statistically Literate Society: Proceedings of the Sixth International Conference on Teaching Statistics [CD-ROM]*. Voorburg, The Netherlands: International Statistical Institute.
- Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, 3(2), 42-63.
- Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics*, 45(1/3), 35-65.

- Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 3-16). Boston, MA: Kluwer Academic Publishers.
- Biggs, J. B. (1992). Modes of learning, forms of knowing, and ways of schooling. In M. Shayer & A. Efklides (Eds.), *Neo-Piagetian theories of cognitive development: Implications and applications for education* (pp. 31-51). Routledge, London.
- Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York, NY: Academic Press.
- Bright, G. W., & Friel, S. N. (1998). Graphical representations: Helping students interpret data. In S. J. Lajoie (Ed.), *Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12*. (pp. 63-88). Mahwah, NJ: Lawrence Erlbaum Associates.
- CEEB. (2007). <http://apcentral.collegeboard.com>: College Entrance Examination Board (CEEB).
- Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 295-323). Dordrecht: Kluwer Academic Publishers.
- Cobb, G. (1992). Teaching statistics. In L. Steen (Ed.), *Heeding the call for change* (pp. 3-43). Washington, DC: Mathematical Association of America.
- Cobb, P. A. (1999). Individual and collective mathematics development: The case of statistical data analysis. *Mathematical Thinking and Learning*, 1(1), 5-44.
- delMas, R. C., & Liu, Y. (Eds.). (2003). *Exploring students' understanding of statistical variation*. Mt. Pleasant, MI: Central Michigan University.
- Dey, I. (1999). Coding. In *Grounding grounded theory: Guidelines for qualitative inquiry*. San Diego, CA: Academic Press.
- Estepa, A., Batanero, C., & Sánchez, F. T. (1999). Students' intuitive strategies in judging association when comparing two samples. *Hiroshima Journal of Mathematics Education*, 7, 17-30.

- Fendel, D., & Doyle, D. (1999). Welcome to our Focus on Statistics. *The Mathematics Teacher*, 92(8), 658-659.
- Fischbein, E., & Schnarch, D. (1997). The Evolution with Age of Probabilistic, Intuitively Based Misconceptions. *Journal for Research in Mathematics Education*, 28(1), 96-105.
- Gal, I. (2004). Statistical literacy: Meanings, components, responsibilities. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 47-78). Boston, MA: Kluwer Academic Publishers.
- Gal, I., Rothschild, K., & Wagner, D. A. (1989). *Which Group Is Better?: The Development of Statistical Reasoning in Elementary School Children*. Paper presented at the Meeting of the Society for Research in Child Development, Kansas City, MO.
- Gal, I., Rothschild, K., & Wagner, D. A. (1990). *Statistical Concepts and Statistical Reasoning in School Children: Convergence or Divergence?* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.
- Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In J. Garfield & D. Ben-Zvi (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 397 - 409). Dordrecht: Kluwer Academic Publishers.
- Garfield, J., & Gal, I. (1999). Teaching and assessing statistical reasoning. In L. Stiff & F. R. Curcio (Eds.), *Developing mathematical reasoning in grades K-12* (pp. 207 - 219). Reston, VA: National Council of Teachers of Mathematics.
- Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 517 - 545). Mahwah, NJ: Lawrence Erlbaum.
- Jones, G. A., Langrall, C. W., Thornton, C. A., & Mogill, A. T. (1999). Student's probabilistic thinking in instruction. *Journal for Research in Mathematics Education*, 30(5), 487-519.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430-451.

- Konold, C. (1989). Informal Concepts of Probability. *Cognition and Instruction*, 6(1), 59-98.
- Konold, C., & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 193-215). Reston, VA: National Council of Teachers of Mathematics.
- Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, 33(4), 259-289.
- Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). *Students analyzing data: Research of critical barriers*. Paper presented at the 1996 IASE Round Table Conference: Research on the Role of Technology in Teaching and Learning Statistics.
- Konold, C., Robinson, A., Khalimahtul, K., Pollatsek, A., Well, A., Wing, R., et al. (2002, July). *Students' use of modal clumps to summarize data*. Paper presented at the Sixth International Conference on Teaching Statistics: Developing a statistically literate society, Cape Town, South Africa.
- Lann, A., & Falk, R. (2003). What are the clues for intuitive assessment of variability? In C. Lee & A. Satterlee (Eds.), *Reasoning about variability: Proceedings of the The Third International Research Forum on Statistical Reasoning, Thinking and Literacy [CD-ROM]*. Mt. Pleasant, MI: Central Michigan University.
- Lee, C. (1998). *An assessment of the PACE strategy for an introductory statistics course*. Paper presented at the Fifth International Conference on Teaching Statistics, Singapore.
- Lee, C. (1999). Computer-assisted approach for teaching statistical concepts. *Computers in the Schools*, v16 n1 p193-208 1999, 16(1), 193-208.
- Lee, C., Zeleke, A., & Wachtel, H. (2002, July). *Where do students get lost: The concept of variation?* Paper presented at the Sixth International Conference on Teaching Statistics: Developing a statistically literate society, Cape Town, South Africa.
- Loosen, F., Lioen, M., & Lacante, M. (1985). The Standard Deviation: Some Drawbacks of an intuitive approach. *Teaching Statistics*, 7(1), 2-5.

- Lovie, P., & Lovie, A. D. (1976). Teaching Intuitive Statistics I: Estimating Means and Variances. *International Journal of Mathematics Education in Science and Technology*, 7(1), 29-39.
- Lutzer, D. J., Maxwell, J. W., & Rodi, S. B. (2002). *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States: Fall 2000 CBMS Survey*: American Mathematical Society.
- Makar, K., & Confrey, J. (2002, July). *Comparing two distributions: Investigating secondary teachers' statistical thinking*. Paper presented at the Sixth International Conference on Teaching Statistics: Developing a statistically literate society, Cape Town, South Africa.
- Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of developing Statistical Literacy, Reasoning and Thinking* (pp. 353-373). Dordrecht: Kluwer Academic Publishers.
- Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal*, 4(1), 27-54.
- McClain, K. (2003). *Supporting teacher change: A case from statistics*. Paper presented at the the 27th annual meeting of the International Group for the Psychology of Mathematics Education, Oahu, HI.
- McClain, K., Cobb, P. A., & Gravemeijer, K. (2000). Supporting Students' Ways of Reasoning about Data. In J. M. Burke & F. R. Curcio (Eds.), *Learning Mathematics for a New Century* (pp. 174-187). Reston, VA: National Council of Teachers of Mathematics.
- McClave, J. T., & Sincich, T. (2003). *Statistics* (9th ed.). UpperSaddle River, NJ: Pearson Prentice Hall.
- Meletiou, M. (2000). *Developing students' conceptions of variation: An untapped well in statistical reasoning*. Unpublished Dissertation, The University of Texas at Austin, Austin, TX.
- Meletiou, M., & Lee, C. (2002, July). *Student understanding of histograms: A stumbling stone to the development of intuitions about variability*. Paper presented at the Sixth International Conference on Teaching Statistics: Developing a statistically literate society, Cape Town, South Africa.

- Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 26(1), 20-39.
- Moore, D. S. (1990). Uncertainty. In L. Steen (Ed.), *On the Shoulders of Giants* (pp. 95-137). Washington, DC: Academy Press.
- Moore, D. S. (1991). *Statistics: Concepts and controversies* (3rd ed.). New York: W. H. Freeman and Company.
- Moore, D. S. (2004). Foreword. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. ix-x). Boston, MA: Kluwer Academic Publishers.
- NCTM. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- NCTM. (2000). *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- Onwuegbuzie, A. J., & Teddlie, C. (2003). A framework for analyzing data in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 351 - 383). Thousand Oaks, CA: Sage.
- Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods* (Third ed.). Thousand Oaks, CA: Sage Publications.
- Petrosino, A. J., Lehrer, R., & Schable, L. (2003). Structuring error and experimental variation as distribution in 4th grade. *Mathematical Thinking and Learning*, 5.
- Pfannkuch, M. (1997). Statistical thinking: One statistician's perspective. In F. Biddulph & K. Carr (Eds.), *People in mathematics education* (Vol. 2, pp. 406-413). Waikato, New Zealand: Mathematics Education Research Group of Australasia.
- Pfannkuch, M., & Wild, C. (2000). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. *Statistical Science*, 15(2), 132-152.
- Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 17-46). Boston, MA: Kluwer Academic Publishers.

- Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Resnick, L. B. (1988). Treating mathematics as an ill-structured discipline. In R. I. Charles & E. A. Silver (Eds.), *The Teaching and Assessing of Mathematical Problem Solving* (Vol. 3, pp. 32-60). Reston, VA: National Council of Teachers of Mathematics.
- Sanders, W. J. (1981). Statistical inference in junior high and middle school. In A. P. Shulte & J. R. Smart (Eds.), *Teaching statistics and probability: National council of teachers of mathematics 1981 yearbook* (pp. 194-202). Reston, VA: The national council of teachers of mathematics, Inc.
- Scheaffer, R. L. (2000). Statistics for a new century. In M. J. Burke & F. R. Curcio (Eds.), *Learning mathematics for a new century* (pp. 158-173). Reston, VA: National Council of Teachers of Mathematics.
- Schoenfeld, A. H. (1988). Problem solving in context(s). In R. I. Charles & E. A. Silver (Eds.), *The Teaching and Assessing of Mathematical Problem Solving* (Vol. 3, pp. 82-92). Reston, VA: National Council of Teachers of Mathematics.
- Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning: A project of the National Council of Teachers of Mathematics* (pp. 334-370). New York: Simon & Schuster Macmillan.
- Shaughnessy, J. M. (1997). *Missed opportunities in research on the teaching and learning of data and chance*. Paper presented at the People in mathematics education, the 20th annual conference of the Mathematics Education Research Group of Australasia (MERGA) Inc., Waikato, New Zealand.
- Shaughnessy, J. M. (2003). *The development of secondary school students' conceptions of variability* (NSF Grant No. REC 0207842). Portland, OR: Portland State University.
- Shaughnessy, J. M. (2006). Research on students' understanding of some big concepts in statistics. In G. Burrill (Ed.), *National Council of Teachers of Mathematics 2006 yearbook*. Reston, VA: National Council of Teachers of Mathematics.

- Shaughnessy, J. M., Ciancetta, M., Best, K., & Canada, D. (2004). *Students' attention to variability when comparing distribution*. Paper presented at the Research Presession of the 82nd Annual Meeting of the National Council of Teachers of Mathematics, Philadelphia, PA.
- Shaughnessy, J. M., Ciancetta, M., Best, K., & Noll, J. (2005). *Secondary and middle school students' attention to variability when comparing data sets*. Paper presented at the Research Presession of the 82nd Annual Meeting of the National Council of Teachers of Mathematics, Anaheim, CA.
- Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004). *Types of student reasoning on sampling tasks*. Paper presented at the The 28th meeting of the International Group for Psychology and Mathematics Education, Bergen, Norway.
- Shaughnessy, J. M., & Pfannkuch, M. (2002). How faithful is old faithful? Statistical Thinking: A story of variation and prediction. *Mathematics Teacher*, 95(4), 252-259.
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved January 18, 2007 from <http://PAREonline.net/getvn.asp?v=9&n=4>.
- Strauss, A., & Corbin, J. (1990). *Basic qualitative research: Grounded theory procedures and techniques*. London: Sage.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4), 293-315.
- Utts, J. M. (1999). *Seeing Through Statistics* (2nd ed.). Pacific Grove, CA: Brooks/Cole Publishing.
- Watson, J. M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics*, 47, 337-372.
- Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics*, 51, 225-256.

- Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematics Education in Science and Technology*, 34(1), 1-29.
- Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, 37, 145-168.
- Wirt, J., Choy, S., Rooney, P., Provasnik, S., Sen, A., & Tobin, R. (2004). *The condition of education (NCES 2004-077)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Appendix A

Informed Consent forms

Survey Participation – Informed Consent

You are invited to participate in a doctoral research project entitled “Undergraduate and Graduate Statistics Students’ Strategies for Comparing Data Sets and Conceptions of Distribution”, being conducted by Matthew Cianceceta from the department of Mathematics and Statistics at Portland State University. The researcher hopes to develop a characterization of how undergraduate and graduate students reason as they compare data sets. You were selected as a possible participant by virtue of your enrollment in a statistics class at Portland State University.

By giving your consent to take part in this study, you are agreeing to complete a web survey. The survey will take from 20 to 40 minutes to complete. In the survey you will be shown pairs of data sets, asked to make a decision about the sets, and required to provide your reasoning for your decisions. Your responses on survey will be used as data.

It is possible that your instructor may offer extra credit for taking this survey. The extra credit will be given whether or not you participate in the study. Other than that, you may not receive any direct benefit from taking part in this portion of the study, but the results of this study may help to increase knowledge concerning the teaching and learning of statistics, which may help others in the future. Potential risks include the possibility that an unauthorized person may view the data, or that your actual name may inadvertently become associated with the data. To minimize this risk, all written responses, notes, audio or video tapes, and transcripts will be kept confidential, and will be kept locked up in the researcher’s office in the Department of Mathematics and Statistics at PSU. After three years, these records will be destroyed. In writing any results of this study, pseudonyms will be used so that your identity cannot be matched with the responses provided.

Your participation in this study is voluntary and you are completely free to withdraw from the study at any time. Your decision whether or not to participate will not affect your relationship with the researcher, your instructor, or with any academic program at PSU in any way. If you have concerns about your participation in this study or your rights as a research subject, please contact the Human Subjects Research Review Committee, Office of Research and Sponsored Projects, 111 Cramer Hall, Portland State University, (503) 725-8182. If you have any questions about the study itself, please contact Matthew Cianceceta, at the Department of Mathematics and Statistics, 334 Neuberger Hall, Portland State University, (503) 725-XXXX.

By checking the box titled ‘I agree to participate in the study’ indicates that you have read and understand the above information and agree to take part in this study. Please remember that you may withdraw your consent at any time without penalty. Also, by checking the box, you are not waiving any legal rights or remedies. Please print this page for a copy of this form for your records.

Interview Participation – Informed Consent

You are invited to participate in a doctoral research project entitled “Undergraduate and Graduate Statistics Students’ Strategies for Comparing Data Sets and Conceptions of Distribution”, being conducted by Matthew Ciancetta from the department of Mathematics and Statistics at Portland State University. The researcher hopes to develop a characterization of how undergraduate and graduate students reason as they compare data sets. You were selected as a possible participant by virtue of your enrollment in a statistics class at Portland State University.

You have previously taken part in a portion of this study by completing a web survey. In the survey you were shown pairs of data sets, asked to make a decision about the sets, and required to provide your reasoning for your decisions. Your responses on survey will be used as data. Now, your signature on this form indicates that you consent to be interviewed about your responses on your survey. Only a portion of those surveyed will be interviewed. The interview will be scheduled at a mutually convenient time and place; it will be audio and video recorded, and will last approximately 30 minutes. The recordings and transcripts from this interview can be used as data.

Those participating in interview will receive the direct benefit of \$10, while indirect benefits include the possibility of helping to increase knowledge concerning teaching and learning statistics, which may help others in the future. Potential risks include the possibility that an unauthorized person may view the data, or that your actual name may inadvertently become associated with the data. To minimize this risk, all written responses, notes, audio or video tapes, and transcripts will be kept confidential, and will be kept locked up in the researcher’s office in the Department of Mathematics and Statistics at PSU. After three years, these records will be destroyed. In writing any results of this study, pseudonyms will be used so that your identity cannot be matched with the responses provided.

Your participation in this study is voluntary and you are completely free to withdraw from the study at any time. Your decision whether or not to participate will not affect your relationship with the researcher, your instructor, or with any academic program at PSU in any way. If you have concerns about your participation in this study or your rights as a research subject, please contact the Human Subjects Research Review Committee, Office of Research and Sponsored Projects, 111 Cramer Hall, Portland State University, (503) 725-8182. If you have any questions about the study itself, please contact Matthew Ciancetta, at the Department of Mathematics and Statistics, 334 Neuberger Hall, Portland State University, (503) 725-XXXX.

Your signature indicates that you have read and understand the above information and agree to take part in the interview portion of this study. Please remember that you may withdraw your consent at any time without penalty. Also, by signing, you are not waiving any legal rights or remedies. The researcher has provided you with a copy of this form for your records.

Name of Participant (Please print clearly.)

Matthew Ciancetta, Researcher Date
Department of Mathematics and Statistics
Portland State University
(503) 725-XXXX

Signature of Participant Date

Appendix B

Text version of survey tasks

Task 1, the Yellow/Brown task, from the Data Set Comparison Survey:

Two teachers are comparing classes to see which is better at quick recall of 9 math facts. Please help the teachers with their comparison of the Yellow and Brown classes.

The scores for the Yellow and Brown classes are shown in the charts to the right. Both classes contain 9 students.

1a) Examine the scores for all the students in each class, then decide:

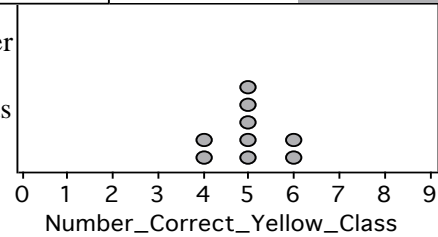
The classes scored equally well
or
The Yellow scored better
or
The Brown scored better

1b) Explain how you decided.

Comparisons

Dot Plot ▼

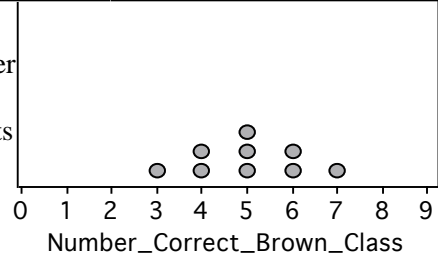
Number
of
Students



Comparisons

Dot Plot ▼

Number
of
Students



Task 2, the Movie Wait-Time task, from the Data Set Comparison Survey:

A recent trend in movie theaters is to show commercials along with previews before the movie begins. The *wait-time* for a movie is the difference between the ADVERTISED start time (like in the paper) and the ACTUAL start time for the movie.

A class of 21 students investigated the wait-times at two popular movie theater chains: Maximum Movie Theaters and Royal Movie Theaters. Each student attended two movies, a different movie in each theater. The class's results are shown in the chart below. (Times were rounded to the nearest half-minute.)

The students in the class found the median wait-time for both of the theaters to be 10 minutes. The students also calculated the mean wait-time for each theater to be 10 minutes.

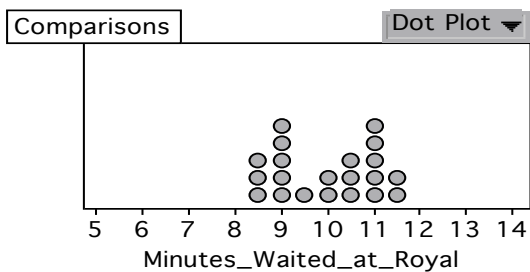
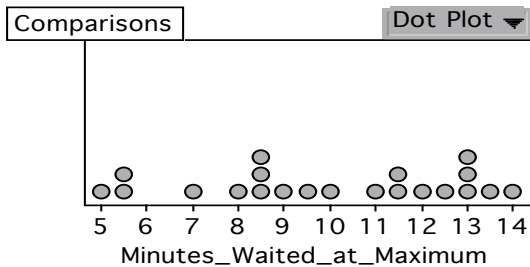
2a) One student in the class, Eddy, concluded that there was no difference in wait-times for the theaters because they both were about 10 minutes.

Do you agree or disagree with Eddy?

2b) Explain all of your reasoning:

2c) Suppose a movie that you wanted to see was playing at each theater at the same times. If both theaters were of equal quality and equally convenient to attend, then which theater would you choose to go to, to see the movie?

2d) Explain your reasoning for your choice:



Task 3, the Pink/Black survey task, from the Data Set Comparison Survey:

Two teachers are comparing classes to see which is better at quick recall of 9 math facts. Please help these teachers with their comparison of the scores for the Pink and Black classes.

The scores for the Pink and Black classes are shown in the charts to the right. The Pink class contains 36 students and the Black class contains 21 students.

3a) Examine the scores for all the students in each class, and then decide:

The classes scored equally well.

or

The Pink class scored better.

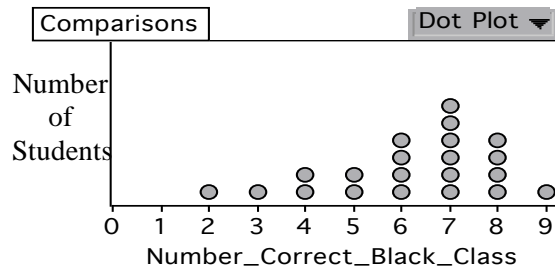
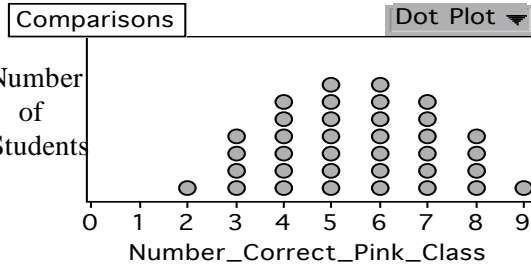
or

The Black class scored better.

3b) Explain how you decided.

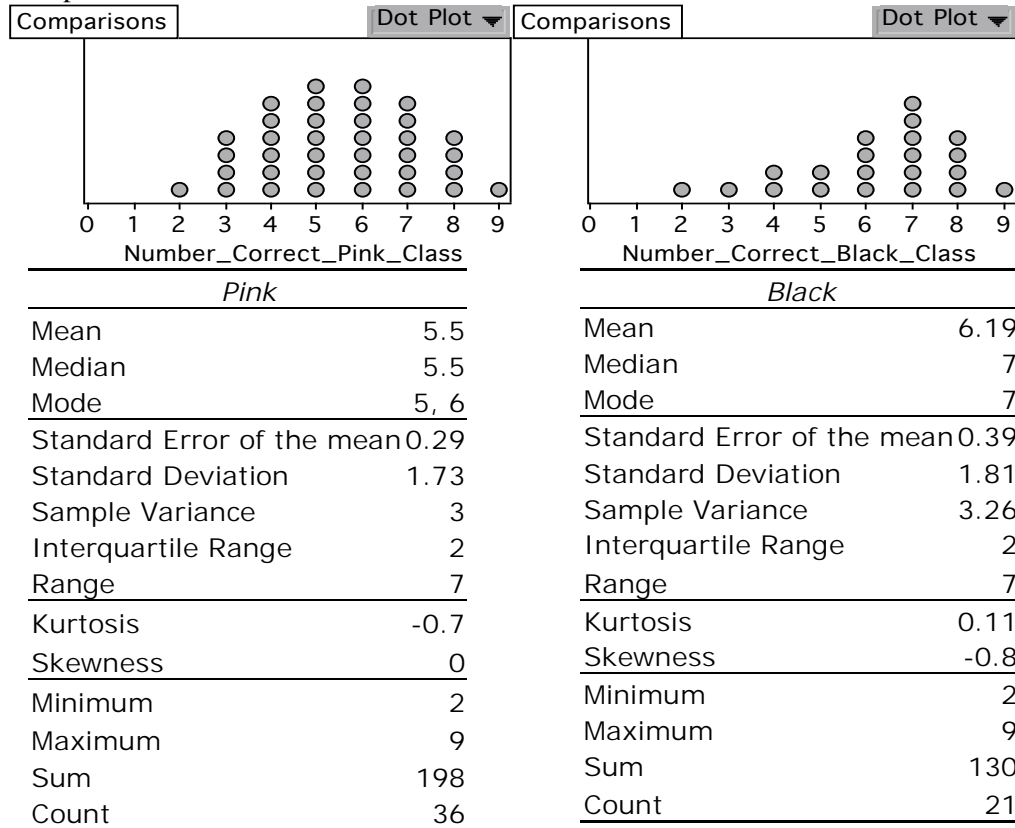
3c) If you decided that one of the classes scored better, then estimate how much better. (If you decided the classes scored equally well please enter 0.)

3d) Explain how you determined your estimation. (If your estimation in part 3c was 0, then enter the word 'equal' for your estimation.)



Task 4, the Pink/Black task with statistics, from the Data Set Comparison Survey:

Now you have the opportunity to re-examine the Pink and Black class's scores. The same charts containing the data are shown below along with some of their corresponding descriptive statistics.



4a) Re-examine the data sets along with their corresponding descriptive statistics. You now have the opportunity to change or amend your responses to the previous question.

In light of these descriptive statistics, decide:

The two classes scored equally well.

or

The Pink class scored better.

or

The Black class scored better.

4b) Explain how you decided:

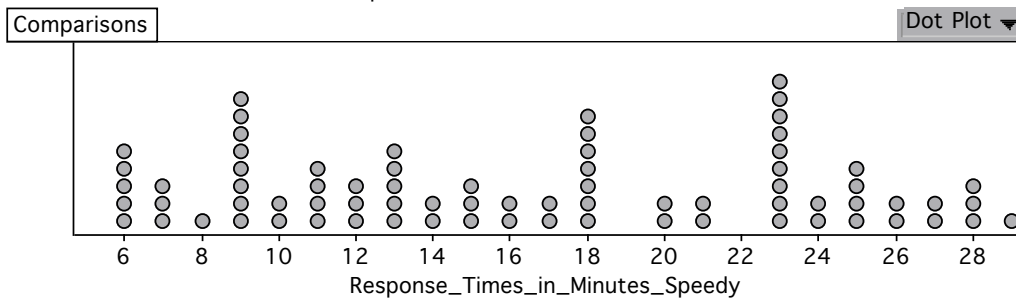
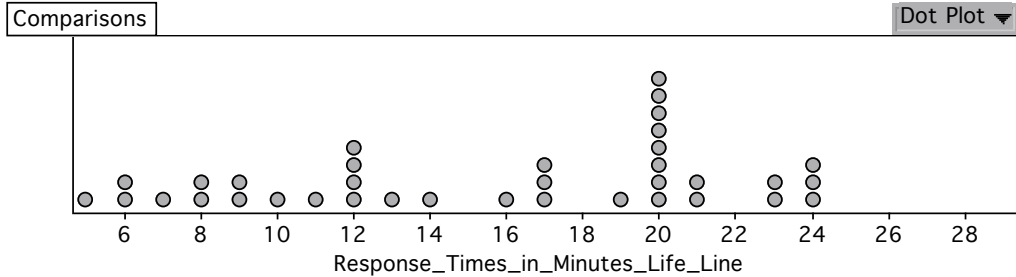
4c) If you decided that one of the classes scored better, then estimate how much better. (If you decided the classes scored equally well please enter 0.)

4d) Explain how you determined your estimation. (If your estimation in part 3c was 0, then enter the word 'equal' for your estimation.)

Task 5, the Ambulance task, from the Data Set Comparison Survey:

The school board for BIG School had to make a decision about which one of two ambulance service companies to call when emergencies arise at their school. The two ambulance companies in the area of the school are Life Line Ambulance Service and Speedy Ambulance Service.

The school board members obtained the most recent 36 response times for Life Line and the most recent 74 response times for Speedy. These response times are shown in the charts below. (Times are rounded to the nearest minute.)



5a) Before the school board members began their debate as to which ambulance service to use, they requested that you look at the data and give your intuitive opinion as to which ambulance service they should choose. Examine the data, then decide:

Recommend Life Line Ambulance Service.

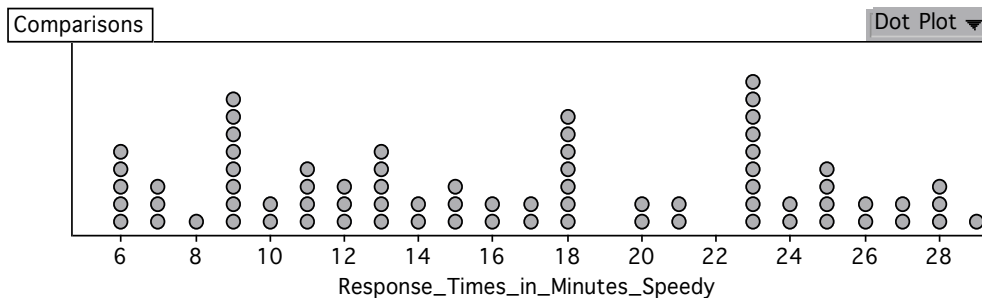
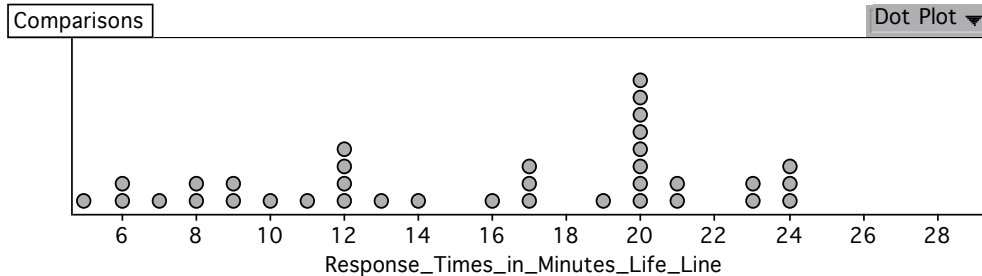
or

Recommend Speedy Ambulance Service.

5b) Explain your reasoning on your choice:

Task 6: the Ambulance task with statistics, from the Data Set Comparison Survey:

After you gave the BIG School board members your intuitive opinion, one member calculated some descriptive statistics and asked you to re-examine the response times for both ambulance services. The same charts displaying the response time data for both Life Line and Speedy ambulance services are shown below along with the corresponding descriptive statistics.



<i>Life Line</i>		<i>Speedy</i>	
Mean	15.56	Mean	16.45
Median	17	Median	16
Mode	20	Mode	23
Standard Error of the mean	0.992	Standard Error of the mean	0.806
Standard Deviation	5.95	Standard Deviation	6.93
Sample Variance	35.4	Sample Variance	48.1
Interquartile Range	9.25	Interquartile Range	12.75
Range	19	Range	23
Kurtosis	-1.295	Kurtosis	-1.276
Skewness	-0.249	Skewness	0.1304
Minimum	5	Minimum	6
Maximum	24	Maximum	29
Sum	560	Sum	1217
Count	36	Count	74

6a) Re-examine the data and the corresponding descriptive statistics, then determine which ambulance service you would recommend.

Recommend Life Line Ambulance Service OR Recommend Speedy Ambulance Service.

6b) Explain your reasoning on your choice:

Appendix C

Detailed survey results

Group 1-GS survey results: Yellow/Brown Task

Table 56 displays the results for the decisions and coded reasons provided by the 1-GS students. Almost half of these students (48.9%) decided that the classes scored equally well, closely followed by those who decided that the Yellow class scored better at 42.3%, with 8.8% deciding that the Brown class scored better. Overall, most of the 1-GS students either provided local, level 1 type reasons (40.1%) or transitional, level 2 type reasons (38%), with only 9.5% providing reasons coded at either level 3 or 4. Across all the groups, the 1-GS students provided the highest percentage of level 1 responses and the lowest percentage of levels 3 and 4 responses. Several factors may have combined to produce this result, such as, the relatively small size of the data sets and equal size of the data sets. The statistically naïve background of the 1-GS students could also contribute to those students finding non-proportional type arguments more accessible and thus more convincing. Finally, 1-GS students gave a high percentage of idiosyncratic responses for task 1. Many of these responses simply provided no reasonable information to base a code on. For example, explanations similar to “Just by looking at the numbers.” were common for 1-GS students and coded at level 0. The inarticulate nature of these explanations did not give any insight into students’ thinking and may be due to the lack of statistical experience that these students had at the time of the survey.

Table 56.

The distribution of responses from task 1 (the Yellow/Brown task), coded across framework levels, for group 1-GS.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Brown	2	6	3	1	0	12
Yellow	10	33	13	1	1	58
Equal	5	16	36	7	3	67
Level	17	55	52	9	4	137
Total	(12.4)	(40.1)	(38.0)	(6.6)	(2.9)	(100)

Percentage of total count of participants in group 1-GS (n = 137) in parentheses.

Table 57.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 1, for group 1-GS.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Brown	0	3	0	3	1	0	1
Yellow	1	6	6	13	1	0	1
Equal	29	6	1	36	0	7	7
Total	30	15	7	52	2	7	9
	(21.9)	(10.9)	(5.1)	(38.0)	(1.5)	(5.1)	(6.6)

Percentage of total count of participants group 1-GS (n = 137) in parentheses.

Of the 67 students who decided ‘equal,’ 29 of them (43.3%) cited a level 2 comparison of centers, i.e., their reasons focused on means or medians or modes, without mentioning other characteristics such as shapes or spread (see table 57). Sixteen of the students who decided ‘equal’ supported their decision with local type responses, particularly citing a comparison of sums, such as, “The same number of correct answers was recorded for both classes” (see table 56). For each of the groups

in this study, of the students who decided ‘equal,’ the 1-GS students had the lowest percentage of comparisons of centers and the highest percentage of local type comparisons. Again, because of these students limited statistical background, strategies such as comparing sums may be more accessible than comparing features such as centers.

More than half (55.7%) of the 70 students who decided that either the Yellow or Brown class scored better supported their decision with local type reasons. Some of these responses could be considered appropriate because of the equal sizes of the data sets. Some of these appropriate responses focused on comparing the number of scores at 5 and higher or at 6 and higher. Yet, other non-appropriate responses, in support of ‘Yellow,’ focused on comparing the heights of the columns of scores of 5, i.e., the larger number of scores at 5 for the yellow class as compared to the Brown class. The prevalence of these types of local responses is somewhat surprising, even for 1-GS students, as strategies such as focusing on whichever group had the “tallest” column is also a common strategy for elementary grade students (Gal et al., 1989, Watson & Moritz, 1999).

Group 1-SE survey results: Yellow/Brown Task

Table 58 displays the results for the decisions and coded reasons provided by the 1-SE students. A large majority of the students from group 1-SE (70.3%) decided that the classes scored equally well, followed by those who decided that the Yellow class scored better at 24.3%, with only two students (5.4%) deciding in favor of ‘Brown.’ Overall, a majority (54.1%) of the 1-SE students provided transitional, level

2, type reasons while slightly more students provided responses at levels 3 and 4 compared to levels 0 and 1. The majority of those responses at level 2 could be due to the 1-SE students' limited experiences in statistics along with their non-statistics mathematics backgrounds, as they were required to have completed or be enrolled in a first course in calculus. Thus 1-SE students could be at ease with performing computations and hence find computational arguments compelling in support of their decisions, such as deciding 'equal' based on a comparison of means.

Table 58.

The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for group 1-SE.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Brown	0	2	0	0	0	2
Yellow	0	1	5	1	2	9
Equal	1	4	15	3	3	26
Level	1	7	20	4	5	37
Total	(2.7)	(18.9)	(54.1)	(10.8)	(13.5)	(100)

Percentage of total count of participants in group 1-SE (n = 37) in parentheses.

Table 59.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 1, for group 1-SE.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Brown	0	0	0	0	0	0	1
Yellow	0	4	1	5	0	1	1
Equal	15	0	0	15	0	3	3
Total	15	4	1	20	0	4	4
	(40.5)	(10.8)	(2.7)	(54.1)	(0.0)	(10.8)	(10.8)

Percentage of total count of participants group 1-SE (n = 37) in parentheses.

More than half of the 1-SE students who decided ‘equal’ provided a reason categorized at level 2, all of which cited a comparison of centers (see table 59). As the centers of the data sets from task 1 are equal and the shape and variation features differ only slightly, these students may have found an argument base solely on the center sufficient and convincing. Of the remaining 1-SE students who decided ‘equal’ none of those students who provided responses at levels higher than level 2 included proportional arguments in their responses. This, again, could be due to the simplistic nature of the data sets, that is, small size, equal sizes, and similar, unimodal shapes.

The two students who decided that the Brown class scored better provided local type responses. One compared the number of scores above 5 and the other focused on the frequencies of specific scores, i.e., “The brown class had a higher grade overall and an equal number of scores in the sixes...” Five of the nine students who decided in favor of ‘Yellow’ provided level 2 responses with four of the five focused on shape (see table 59). Shape was also typically addressed in the level 3 and 4 responses that decided in favor of ‘Yellow.’ Most of the shape arguments related that a favorable feature of the Yellow class was the “consistency” of its scores. This specific contextual interpretation would appear to induce reasoning about the data sets at level 2 and higher.

Group 2-GS results: Yellow/Brown Task

Table 60 displays the results for the decisions and coded reasons provided by the 2-GS students. Half of those students decided that the classes scored equally well, followed by just over one-third who decided that the Yellow class scored better, with

just less than one-sixth who decided that the Brown class scored better. A high percentage (39.2%) of 2-GS students provided responses at the transitional level with less at levels 1 and 3, although approximately the same number of students gave responses at levels 1 and 3 at slightly more than 20% each. The fewest number responses were categorized at levels 0 and 4, with about the same number of responses in each level. Overall, group 2-GS students appear to have responded to task 1 at higher framework levels than their 1-GS counterparts as 2-GS students provided proportionally fewer level 0 and 1 responses and proportionally more level 3 and 4 responses. The 2-GS students had completed one semester of a general statistics course and thus it is not surprising that, proportionally, more of the 2-GS students reasoned at higher framework levels.

Table 60.

The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for group 2-GS.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Brown	3	3	4	2	0	12
Yellow	2	11	8	2	2	25
Equal	0	4	17	12	4	37
Level	5	18	29	16	6	74
Total	(6.8)	(24.3)	(39.2)	(21.6)	(8.1)	(100)

Percentage of total count of participants in group 2-GS (n = 74) in parentheses.

As with groups 1-GS and 1-SE most of the students from group 2-GS, who made the ‘equal’ decision, also provided reasons that were at level 2 and similarly almost all of those transitional responses, in support of ‘equal,’ focused on comparing

centers (see table 61). Yet, considerably more 2-GS students provided reasons at level 3 than at level 1 to support a decision of ‘equal’ where as in groups 1-GS and 1-SE more of the students who decided ‘equal’ provided reasons at level 1 than at level 3. Of the level 3 responses, from 2-GS students, that supported an equal decision, none included proportional arguments. This may be a result of the specific features of the data sets, as previously mentioned.

Table 61.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 1, for group 2-GS.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Brown	2	1	1	4	1	1	2
Yellow	1	2	5	8	0	2	2
Equal	15	2	0	17	0	12	12
Total	18 (24.3)	5 (6.8)	6 (8.1)	29 (39.2)	1 (1.4)	15 (20.3)	16 (21.6)

Percentage of total count of participants group 2-GS (n = 74) in parentheses.

Students from group 2-GS who decided in favor of the Brown class provided responses that were almost equally distributed across levels 0,1, 2, and 3 while the responses in favor of the Yellow class were primarily at levels 1 and 2 with a few each at levels 0, 3, and 4. Several of the Level 1 reasons for ‘Brown’ were based on the feature that the Brown class had the highest score. Level 1 reasons in favor of ‘Yellow’ were similar to those from 1-GS students in that they tended to focus either on the higher number of scores at 5 and above or on the feature that the column height for 5 was higher for the Yellow class than for the Brown class. Also, similar to the 1-

GS and 1-SE students, most of the 2-GS students who favored ‘Yellow’ and provided level 2 reasons focused either on the narrower range of the Yellow scores or the higher consistency of the Yellow scores as positive features. The one student who provided a proportional response supporting the Brown class wrote, “A higher percentage of students in the brown class "passed" the recall with at least a 60%.” This response is considered by the researcher as a prototypical level 3, proportional, type response yet it was uncommon among all the groups for their responses to the Yellow/Brown task.

Group 2-SE results: Yellow/Brown Task

Table 62 shows the distribution of student responses from group 2-SE. Of the 15 students from that group, 10 (two-thirds) decided that the classes scored equally well, four (26.7%) decided that Yellow scored better and only one (6.7%) decided that Brown scored better. The two 2-SE students whose responses were coded as idiosyncratic appear to have misread either the question or the graphs or both. Most of the group 2-SE students (60%) provided transitional, level 2 type reasons. Although no 2-SE student wrote a level 4 response, three did provide level 3 responses and only one provided a level 1 response. The concentration of level 2 and 3 responses seems reasonable given that these students have completed at least 1 semester of a statistics for engineers course, however it was a bit surprising that there were no distributional responses.

Six of the nine transitional type responses supported the ‘equal’ decision and four of those six focused on a comparison of centers. Again, because of the obvious ‘bell shape’ and equal centers, these students may have felt that it was sufficiently

convincing to only compare the centers (see table 63). While no 2-SE students provided reasons that could be categorized at level 4, distributional, three (20%) students provided responses at the initial distributional level, all of which supported the ‘equal’ decision and were categorized as Initial Global, not proportional. Those Initial Global responses also tended to include a strong focus on centers, hence it is apparent that the comparing the centers of the data sets for task 1 is rather compelling for many of the 2-SE students.

Table 62.

The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for group 2-SE.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Brown	1	0	0	0	0	1
Yellow	0	1	3	0	0	4
Equal	1	0	6	3	0	10
Level	2	1	9	3	0	15
Total	(13.3)	(6.7)	(60.0)	(20.0)	(0.0)	(100)

Percentage of total count of participants in group 2-GS (n = 15) in parentheses.

Three of the four 2-SE students who decided that the Yellow class scored better provided transitional reasons, two of which argued that the narrower range of the Yellow class’s scores was better (see table 63). This similar contextual interpretation was observed in each of groups 1-GS, 1-SE, and 2-GS.

Table 63.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 1, for group 2-SE.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Brown	0	0	0	0	0	0	0
Yellow	1	0	2	3	0	0	0
Equal	4	2	0	6	0	3	3
Total	5 (33.3)	2 (13.3)	2 (13.3)	9 (60.0)	0 (0.0)	3 (20.0)	3 (20.0)

Percentage of total count of participants group 2-SE (n = 15) in parentheses.

Group GRAD results: Yellow/Brown Task

The distribution of the GRAD students' responses is displayed in table 64. Of the 12 students from the GRAD group, none decided that the Brown class scored better, seven decided that the classes scored equally well and five decided that the Yellow class scored better. Ten of the 12 GRAD students supported their decisions with reasons that were classified at either level 3, initial-distributional or level 4, distributional. Only two students provided reasons classified at level 2 with no students giving reasons at level 1 or level 0. This is a rather dramatic shift higher in response level compared to all the other groups and likely could be the result of these students' extensive statistical backgrounds.

The seven students who decided 'equal' all integrated a comparison of centers into their explanations (see table 65). Four of these students provided reasons at the distributional level, two provided reasons categorized as initial global and one provided a transitional reason. This pattern is strikingly different than all the other

groups whose highest percentage of responses in support of ‘equal’ are at level 2 focused on centers. Thus, for many of the GRAD students it is apparently not sufficient to make a comparison of centers of the data sets in task 1, in isolation, without a larger, global picture.

Table 64.

The distribution of responses coded across framework levels from task 1 (the Yellow/Brown task), for group GRAD.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Brown	0	0	0	0	0	0
Yellow	0	0	1	1	3	5
Equal	0	0	1	2	4	7
Level	0	0	2	3	7	12
Total	(0.0)	(0.0)	(16.7)	(25.0)	(58.3)	(100)

Percentage of total count of participants in group GRAD (n = 12) in parentheses.

Table 65.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 1, for group GRAD.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Brown	0	0	0	0	0	0	0
Yellow	0	0	1	1	0	1	1
Equal	1	0	0	1	0	2	2
Total	1 (8.3)	0 (0.0)	1 (8.3)	2 (16.7)	0 (0.0)	3 (25.0)	3 (25.0)

Percentage of total count of participants group GRAD (n = 12) in parentheses.

Of the five students who decided ‘Yellow,’ three provided reasons at the distributional level, one provided a reason categorized as initial global and one

provided a reason focused on variation (see table 65). All of these students included in their explanations the interpretation that the lesser variation in the scores for the Yellow class was better. As with most of the other students who decided in favor of ‘Yellow,’ these GRAD students’ interpretation of “better test scores” for a class includes more consistency as apart of “better.”

Group 1-GS results: Movie Wait-Time Task

Table 66 displays the results for the decisions and coded reasons provided by the 1-GS students. Just over two-thirds (67.9%) of group 1-GS respondents disagreed with Eddy, that is, they decided that there was a difference in wait times between the two theaters, while just less than one-third (32.1%) agreed with Eddy, that is, they decided that there was no difference in wait times between the two theaters. Irrespective of their decision, students chose to go to the Royal Theater over the Maximum Theater at about a 2 to 1 rate.

Table 66.

The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group 1-GS.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Agree	6	1	20	11	6	44
Disagree	6	12	57	7	11	93
Level	12	13	77	18	17	137
Total	(8.8)	(9.5)	(56.2)	(13.1)	(12.4)	(100)

Percentage of total count of participants in group 1-GS (n = 137) in parentheses.

Students from group 1-GS provided the highest percentage of responses at level 2, the Transitional level (see table 66). Of the level 2 responses in support of

‘Disagree,’ more than half cited the difference in variation between the two data sets and more than one-third cited differences in shape (see table 67). Both of those strategies seem reasonable, particularly because of the clear difference in spread between the data sets. Of the level 2 responses in support of ‘Agree,’ most focused on centers (see table 67) as they paraphrased Eddy’s statement implying that the means are equal so there is no difference in wait-times.

Table 67.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 2, for group 1-GS.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Agree	16	4	0	20	0	11	11
Disagree	2	16	39	57	2	5	7
Total	18 (13.1)	20 (14.6)	39 (28.5)	77 (56.2)	2 (1.5)	16 (11.7)	2 (1.5)

Percentage of total count of participants group 1-GS (n = 137) in parentheses.

Concerning 1-GS students who provided responses at the Local level, only one “Agreed” while 12 “Disagreed.” Reasoning that the data sets are different, as opposed to the same, may have been considerably easier to do at the Local level because so many of the local features between the data sets were not the same, such as different maximums and minimums, which were frequently referenced in these responses.

Responses at levels 3 and 4 showed an interesting trend in that more level 3 responses supported the ‘Agree’ decision while more level 4 responses supported the ‘Disagree’ decision (see table 66). All of the level 3 – “Agree” responses were classified as Initial Global (see table 67) with most citing the given information that

the mean and median for each data set is the same at 10 minutes. Although this type of response does go beyond the argument given by “Eddy,” it still is only repeating information given in the task description and thus is accessible to a wide variety of students. Only two of the level 3 – “Disagree” responses were Proportional as the remaining were Initial Global. The Initial Global, ‘Disagree’ responses appeared to be more clearly a step beyond Transitional than the Initial Global ‘Agree’ responses, but still not completely Distributional. Many of those ‘Disagree’ responses also cited the equal means and medians but additionally noted that the endpoints were different. They only referred to specific values not spread. Also, other level 3 – ‘Disagree’ responses cited the equal means and medians but additionally noted that modes were different. All the level 4 responses from the 1-GS students were surprisingly articulate and included the given information about the equal means and medians with the ‘Agree’ responses generally adding the data appears to be “evenly distributed on each side” of the mean for each data set and the ‘Disagree’ responses often included assessments about how different the spreads (or shapes) are. The given information that the mean and median are equal, along with the clear difference between the ranges of the data sets are both easily accessible and thus may have been a significant factor as to why more 1-GS students disagreed at level 4 than agreed at level 4.

Group 1-SE results: Movie Wait-Time Task

Table 68 displays the results for the decisions and coded reasons provided by the 1-SE students. Over 80% disagreed with Eddy, that is, they decided that there was a difference in wait times between the two theaters, while less than 20% disagreed

with Eddy, that is, they decided that there was no difference in wait times between the two theaters. Irrespective of their decision, students chose to go to the Royal Theater over the Maximum Theater at a rate of about 2.7 to 1.

The 1-SE students provided slightly more responses at level 4 than at level 2 and they were the only group to provide the most responses at level 4. Only one of those level 4 responses favored ‘Agree.’ That student cited the equal means and medians as well as indicating that for each data set the data was distributed equally on each side of the mean. The remaining students who provided level 4 responses all favored ‘Disagree.’ These students also cited the equality of the centers but additionally assessed the differences in the spreads (or shapes). The percentage of Distributional responses was highest in the 1-SE group and it is not clear as to what factors may have caused this phenomena.

Table 68.

The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group 1-SE.

<u>Decision</u>						Decision
	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>
Agree	1	1	3	1	1	7
Disagree	1	3	9	4	13	30
Level	2	4	12	5	14	37
Total	(5.4)	(10.8)	(32.4)	(13.5)	(37.8)	(100)

Percentage of total count of participants in group 1-SE (n = 37) in parentheses.

As with the level 4 responses, the level 2 responses from the 1-SE students largely favored ‘Disagree’ (see table 69). All of the ‘Agree’ reasons at level 2 focused on the equal centers while most of the ‘Disagree’ reasons focused on the difference in

spreads with a few making an assessment that the shapes were different (see table 69). Although these students have limited statistical backgrounds, the focus on the difference in the spreads seems reasonable, as that difference visibly stands out.

Table 69.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 2, for group 1-SE.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Agree	3	0	0	3	0	1	1
Disagree	0	2	7	9	0	4	4
Total	3 (8.1)	2 (5.4)	7 (18.9)	12 (32.4)	0 (0.0)	5 (13.5)	5 (13.5)

Percentage of total count of participants in group 1-SE (n = 37) in parentheses.

Only five of the 37 students from group 1-SE provided level 3 responses, all of which were Initial Global with only one in support of ‘Agree.’ That one student referenced the equal means and medians while the remaining four students who disagreed either referenced the equal centers along with citing that the longest and/or shortest wait-times were not the same (they did not explicitly refer to spread) or they cited the difference in spreads and the difference in “the most frequent times,” i.e., the different modes.

Students from group 1-SE provided the fewest number of level 1 responses (except or level 0), almost all of which supported the ‘Disagree’ choice (see table 68). Most of these responses focused on comparing individual times, such as noting that the shortest times are different. The shift of responses toward higher levels could be

because of the accessibility of the given information about the means and medians or because of the visible difference in the spreads.

Group 2-GS results: Movie Wait-Time Task

Table 70 displays the results for the decisions and coded reasons provided by the 2-GS students. Almost 70% of group 2-GS respondents disagreed with Eddy, that is, they decided that there was a difference in wait times between the two theaters, while just more than 30% agreed with Eddy, that is, they decided that there was no difference in wait times between the two theaters. Irrespective of their decision, more students chose to go to the Royal Theater over the Maximum Theater at a consistent rate of slightly less than 2 to 1. Overall, the 2-GS and 1-GS students responded in fairly similar ways for both decisions and response levels for the Movie Wait-Time task.

Table 70.

The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group 2-GS.

						Decision
<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>
Agree	0	1	15	4	3	23
Disagree	2	4	27	2	16	51
Level	2	5	42	6	19	74
Total	(2.7)	(6.8)	(56.8)	(8.1)	(25.7)	(100)

Percentage of total count of participants in group 2-GS (n = 74) in parentheses.

More than half of the group 2-GS students provided responses at the Transitional level and a majority of those responses favored ‘Disagree’ (see table 70). As with the previous groups, the majority of level 2 – ‘Agree’ responses focused on

the given information about the equal centers and the majority of level 2 – ‘Disagree’ responses focused on the differences in variation or spread (see table 71). However, the 2-GS students gave a rather large number of level 2 – ‘Disagree’ responses that focused on center. After a closer inspection of these responses, it was found that there was no common reasoning trend in support of the decision. For example, some students recalculated the means and/or medians and made errors doing so, some misinterpreted the meaning of the median, and some correctly recalculated the means and found that there was indeed a two second difference and decided that the wait times were different because of that small difference.

Table 71.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 2, for group 2-GS.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Agree	15	0	0	15	0	4	4
Disagree	7	9	11	27	0	2	2
Total	22 (29.7)	9 (12.2)	11 (14.9)	42 (56.8)	0 (0.0)	6 (8.1)	6 (8.1)

Percentage of total count of participants in group 2-GS (n = 74) in parentheses.

The second most frequent responses from the 2-GS students were at level 4 and the least frequent responses were at level 1 (except level 0 responses). The trends of responses at these levels was similar to those from the 1-GS and 1-SE groups, that is, most decided that there was a difference in wait times, they disagreed, the level 1 responses focused on comparing individual times and the level 4 responses included the given information about the equal means and medians with the ‘Agree’ responses

adding that the data appears to be “balanced” or “equal” on each side of the mean, for each data set, and the ‘Disagree’ responses included additional assessments about the different spreads (or shapes).

In a different trend from the 1-GS and 1-SE students, the 2-GS students provided more level 3 responses in support of ‘Agree’ than in support of ‘Disagree,’ although the reasons for each decision were similar to the previous groups. Those who agreed cited the equal means and medians and those who disagreed referenced the equal centers along with citing that the longest and/or shortest wait-times were not the same (they did not explicitly refer to spread).

Group 2-SE results: Movie Wait-Time Task

Table 72 displays the results for the decisions and coded reasons provided by the 2-SE students. Of the 15 students in group 2-SE, 11 (73.3%) disagreed with Eddy, that is, they decided that there was a difference in wait times between the two theaters, while only 4 (26.7%) agreed with Eddy, that is, they decided that there was no difference in wait times between the two theaters. Students chose to go to the Royal Theater over the Maximum Theater at a rate that was slightly less than 3 to 1.

Seven of the 15 group 2-SE students responded at level 2, more than any other level. Unlike all the other groups, the 2-SE students almost evenly split their Transitional responses between ‘Agree’ and ‘Disagree’ (see table 73). The three students who agreed, all cited the equal centers and the four who disagreed all cited the difference in variation (or shape). Those responses were very similar to the Transitional responses given by the 1-SE group.

Table 72.

The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group 2-SE.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Agree	0	0	3	1	0	4
Disagree	2	2	4	0	3	11
Level	2	2	7	1	3	15
Total	(13.3)	(13.3)	(46.7)	(6.7)	(20.0)	(100)

Percentage of total count of participants in group 2-SE (n = 15) in parentheses.

Table 73.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 2, for group 2-SE.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Agree	3	0	0	3	0	1	1
Disagree	0	1	3	4	0	0	0
Total	3	1	3	7	0	1	1
	(20.0)	(6.7)	(20.0)	(46.7)	(0.0)	(6.7)	(6.7)

Percentage of total count of participants group 2-SE (n = 15) in parentheses.

Except for one level 3 response in support of ‘Agree,’ the remaining responses favored ‘Disagree,’ three at level 4 and two at level 1. The level 3 and 4 responses were similar to other groups responses in that the level 3 – ‘Agree’ response cited the equal means and medians and the level 4 – ‘Disagree’ responses also cited the equal centers but added an extra assessment about the different amount of variation in each distribution. The Local responses were a bit different than the previously described level 1 responses from other groups. These two students tended to assess the difference between the data sets by comparing each data point. For example one

student wrote, “Certainly there is a difference. The data is not identical.” This student seems to be viewing these particular data sets as an amalgam on points, not as whole units.

Group GRAD results: Movie Wait-Time Task

Table 74 displays the results for the decisions and coded reasons provided by the 2-GS students. Of the 12 students in group GRAD, 9 (75%) disagreed with Eddy, that is, they decided that there was a difference in wait times between the two theaters, while only 3 (25%) agreed with Eddy, that is, they decided that there was no difference in wait times between the two theaters. Students chose to go to the Royal Theater over the Maximum Theater at a rate of 5 to 1.

Table 74.

The distribution of responses coded across framework levels from task 2 (the Movie Wait-Time task), for group GRAD.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Agree	0	0	1	1	1	3
Disagree	0	0	6	0	3	9
Level	0	0	7	1	4	12
Total	(0.0)	(0.0)	(58.3)	(8.3)	(33.3)	(100)

Percentage of total count of participants in group GRAD (n = 12) in parentheses.

All of the GRAD students provided responses at level 2 or higher. A majority, seven of the 12 students (58.3%), provided responses at the transitional level. Similar to all the other groups, the students from the GRAD group who provided Transitional responses also largely favored ‘Disagree.’ All of the reasons provided for the ‘Disagree’ decision cited the difference in variation (see table 75), and most of those

responses specifically referred to “standard deviation” or “variance.” There was one student who did write that the difference in variance is “sufficient” to determine that the data sets are different. It is quite possible that this student was considering the data sets from a global perspective, yet his or her give response was coded at level 2. This situation highlights the nature of the coding process for this research, that is, codes were assigned conservatively and thus may represent a lower bound for the students’ reasoning. A similar situation may have also arisen with the one student who agreed and gave a Transitional reason. The student wrote, “Without performing a statistical test, it does appear that the difference between the two average wait times is not significant.” Although the task did not ask about statistical test, this student appears to have interpreted the problem as a prediction about a test for significantly different means. It is not entirely clear that the Expanded Lattice Structure Framework sufficiently captures this response and thus points to a potential limitation of the framework.

Table 75.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 2, for group GRAD.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Agree	1	0	0	1	0	1	1
Disagree	0	0	6	6	0	0	0
Total	1 (8.3)	0 (0.0)	6 (50.0)	7 (58.3)	0 (0.0)	1 (8.3)	1 (8.3)

Percentage of total count of participants group GRAD (n = 12) in parentheses.

The students from the GRAD group were, again, the only ones to provide no level 0 and no level 1 responses. They provided one level 3 response and four level 4 responses. The level 3 – ‘Agree’ responses was similar to those from other groups as its focus was on only the equality of the means and medians. The three level 4 – ‘Disagree’ responses were also similar to those from other groups in that they cited the equal centers and added an extra assessment about the different amount of variation in each distribution. The one level 4 – ‘Agree’ response was, “While the measures of central tendency are similar (identical), the dispersion is quite different. Thus, I would agree that both distributions are centered around the same value.” While it does appear that this student is considering the data sets from a global perspective, it does seem a bit contradictory in the acknowledgment that the dispersion is different but the wait-times are the same because they are centered around the same value. This may be an issue to further explore in future research.

Group 1-GS results: Pink/Black Task

Table 76 displays the results for the decisions and coded reasons provided by the 1-GS students. The majority, 61.3%, decided that ‘the Black class scored better,’ followed by ‘the classes scored equally well’ at 25.5% and then ‘the Pink class scored better’ at 13.1%. Most of the reasons supporting ‘Black’ were classified at level 2 with slightly more reasons that addressed comparing shapes than centers. Just over one-quarter of the ‘Black’ decisions were supported by reasons at levels 3 or 4, and almost all of those reasons specifically relied on proportional reasoning. Most of the reasons supporting the ‘equal’ decision were either Local or Transitional focused on shape or

Transitional focused on centers. A large majority of the reasons supporting the ‘Pink’ decision were Local.

Table 76.

The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group 1-GS.

Decision						Decision
	Level 0	Level 1	Level 2	Level 3	Level 4	Total
Black	2	0	59	21	2	84
equal	3	13	19	0	0	35
Pink	1	14	3	0	0	18
Level	6	27	81	21	2	137
Total	(4.4)	(19.7)	(59.1)	(15.3)	(1.5)	(100)

Percentage of total count of participants in group 1-GS (n = 137) in parentheses.

The researcher expected most of the Transitional responses for the Pink/Black task to focus on center, but the 1-GS students had more students who focused on shape (see table 77). These shape responses generally described a “shift” or “slide” of the ‘Black’ scores, or a “curve,” to the right. Others explained that the Black class had fewer total scores but more scores to the right. Although the previous example is a preliminary proportional type argument, it was classified as level 2 – shape because of its inarticulateness. For statistically naïve students, such as the 1-GS students, the prominent difference between the shapes may have been easier to describe than reasoning about the centers. The Transitional – Center responses mostly cited the higher “average” or “mean” of the Black class, however a few cited the median and some cited the mode, such as, “...the greatest frequency in Black class is 7 while the Pink class seems to be between 5 and 6.” When these students who focused on center

and the students who focused on shape were asked to quantify how much better the Black class scored, they had very different success rates. Of the 22 students who focused on center, 19 of them estimated the difference between the means, medians, or modes, while only 3 provided idiosyncratic estimations. Yet, 33 of the 36 students who focused on Shape could not provide a reasonable estimation for how much better the Black class scored, such as estimating the difference between the centers. The students who compared centers to decide that the Black class scored better, and then made reasonable estimations, such as finding the difference of the center, are potentially considering the center as a group representative.

The amount of level 2 responses, from the 1-GS students, in support of ‘equal’ was a bit surprising (see table 76). Most cited the difference in shapes or means but then referenced the difference in class size and claimed that because the Black class had fewer students the classes’ shapes (or the averages) were equal (or about equal). For example, “The graphs were both following the same kind of pattern, one with just fewer students or sample size. If they had been equal it looks like they would appear the same.” was categorized at level 2, focused on shape and “On avg [sic] each class scored the same, due to the difference in class size.” was categorized at level 2, focused on center. It appears that these students may not understood the need to reason proportionally or were unable to reason proportionally about the data sets.

Table 77.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 3, for group 1-GS.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Black	22	36	1	59	19	2	21
equal	7	11	1	19	0	0	0
Pink	0	3	0	3	0	0	0
Total	29 (21.2)	50 (36.5)	2 (1.5)	81 (59.1)	19 (13.9)	2 (1.5)	21 (15.3)

Percentage of total count of participants group 1-GS (n = 137) in parentheses.

The level 1 responses were almost evenly split between favoring ‘equal’ and favoring ‘Pink.’ Most of the level 1 responses that favored ‘equal’ cited the difference in class size as an obstacle to making a comparison of scores. Additionally many noted the equal frequencies of the top scores between the classes and claimed that if more students were added to the Black class, then the scores would be equal. Many of these students did attempt to account for the difference in class size but could not accomplish this in a proportional sense and thus seemed to result in the ‘equal’ decision. The students who decided in favor of ‘Pink’ generally did not attempt to compensate for the unequal class sizes and focused on sums and frequencies such as, “More students got correct answers in the pink class.” None of the students who provided level 1 responses could also provide a reasonable estimate for the difference between the classes’ scores, although 6 of the 14 students who favored ‘Pink’ did find the difference in the sums of the scored and use that as their estimate for how much

better the Pink class scored. This strategy does at least seem to have potential to build on, to eventually make proportional comparisons between the classes.

All of the level 3 and 4 responses supported the Black class as scoring better. The large majority of these responses were at level 3 and of those, almost all focused on comparing proportions (see tables 76 and 77). These responses were similar to those report by McClain, Cobb and Gravemeijer (2000) and McClain (2003) where some middle school students reasoned proportionally about partial distributions and middle school teachers formed partial distributions with “cut-points” and reasoned proportionally about the piece of the distribution either above or below the “cut-point.” The “cut-point” used by the 1-GS students was a score of 5, 6, or 7. Those who reasoned proportionally then compared the proportion of points above their “cut-point,” between the two classes. This strategy is contextually appropriate as scores below a “cut-point” may represent failure on the test or merely undesirable scores, thus comparing the proportion of passing or desirable scores for each class would seem to be appropriate. Most students who used proportions to make their decision appeared to also view the described proportions as groups representatives as more than half also found the difference between the relative frequencies above the “cut-point” for their estimation of how much better the Black class scored. Only three 1-GS students who provided a level 3 response also gave an Idiosyncratic estimation, the rest found the difference between either centers or proportions. Although there were only two Distributional type responses that cited comparisons of proportions and centers, both subsequently made reasonable estimations for how much better the Black

class scores; one found the difference between centers and the other found a difference between proportions.

The responses at levels 1, 3, and 4, from the 1-GS students, show a clear separation. Responses at level 1 only supported either ‘equal’ or ‘Pink’ while the Levels 3 and 4 responses only supported ‘Black.’ Thus reasoning about these data sets at higher framework levels did appear to elicit the normative conclusion that the Black class scored better.

Group 1-SE results: Pink/Black Task

Table 78 displays the results for the decisions and coded reasons provided by the 1-SE students. Almost all of the students from group 1-SE (86.5%) decided that ‘the Black class scored better’, with only three students (8.1%) deciding that ‘the classes scored equally well’ and just two students deciding ‘the Pink class scored better.’ Most (46.9%) of the reasons supporting ‘Black’ were classified at level 3 and most of those relied on proportional reasoning with only one reason classified as distributional. Many students supported their ‘Black’ decision with transitional type responses (40.6%). Of the transitional reasons supporting the ‘Black’ decision practically all referred to a comparison of centers. None of the reasons in favor of ‘Pink’ or ‘equal’ were categorized at either of the upper levels.

The 1-SE students performed considerably better than the 1-GS students as a higher percentage of 1-SE students decided in favor of ‘Black’ with slightly more supporting the ‘Black’ decision with level 3 reasons than with level 2 reasons (see table 79). All of the level 3 and 4 responses supported ‘Black.’ Although the 1-SE

level 3 responses were distributed similarly to those from the 1-GS group, with considerably more Proportional responses, the 1-SE level 2 responses were largely focused on comparing centers where as there were slightly more 1-GS responses focused on shape as opposed to centers.

Table 78.

The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group 1-SE.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Black	2	1	13	15	1	32
Equal	0	1	2	0	0	3
Pink	0	1	1	0	0	2
Level	2	3	16	15	1	37
Total	(5.4)	(8.1)	(43.2)	(40.5)	(2.7)	(100)

Percentage of total count of participants in group 1-SE (n = 37) in parentheses.

Table 79.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 3, for group 1-SE.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Black	11	2	0	13	11	4	15
Equal	0	0	2	2	0	0	0
Pink	0	1	0	1	0	0	0
Total	11	3	2	16	11	4	15
	(29.7)	(8.1)	(5.4)	(43.2)	(29.7)	(10.8)	(40.5)

Percentage of total count of participants group 1-SE (n = 37) in parentheses.

Only one student from the 1-SE group provided a level 4 response. This student related the shapes and location of the centers to support the ‘Black’ decision. The Initial Global responses at level 3 cited the mean, median, and mode as higher for

the Black class, thus the Black class scored better. The Proportional responses at level 3 mostly compared the proportion of scores above a “cut-point,” although a few compare the proportion below a “cut-point.” Except for one student, all those who provided either level 3 or 4 responses also made a reasonable estimation for how much better the Black class scored. Two students found the ratio of the means for their estimations and the remaining either found the difference between the centers or computed the difference of the proportions they had previously determined.

The 1-SE students who responded at the Transitional level had a fairly high success rate of deciding in favor of the Black class and of providing reasonable estimations for how much better the Black class scored. Of the 16 level 2 responses, 13 favored ‘Black’ and 11 of those 13 focused on comparing a center with two focused on comparing shapes (see table 79). Of the students who focused on centers, only one could not provide an estimation for how much better the Black class scored, while nine extended their strategy to find the difference between centers and one found the ratio of means. Both of the students who provided level 2 responses and supported ‘equal’ cited the equal ranges of scores for each class. The one student who provided a level 2 response and supported ‘Pink’ seemed to mis-interpret implications about the shapes of the distributions as the student wrote, in part, “...the bell curve system indicates that the pink class is more successful [sic] since most students will fall within the allow percentage range.” This student appears to have moved beyond a local perspective to the data as he or she attempted to compare the distributions a ‘units’ yet the student is clearly still in the process of building a normative

understanding of what shapes, such as a ‘bell shape’ imply about how the data is distributed.

Only three 1-SE students provided Local type responses as they either compared specific frequencies of scores or sums of scores. Thus, the 1-SE group generally performed quite well on the Pink/Black task as most provided responses at either level 2 or 3, choosing ‘Black’ and with most of those also making reasonable estimations for how much better the Black class scored.

Group 2-GS results: Pink/Black Task

Table 80 displays the results for the decisions and coded reasons provided by the 2-GS students. Students from the 2-GS group appeared to perform similarly to the 1-GS students, with the 2-GS students providing a slightly higher percentage of level 3 and 4 responses and a slightly lower percentage of level 1 responses. Almost 70% of students from group 2-GS decided that ‘the Black class scored better,’ followed by about 15% each deciding either ‘the classes scored equally well’ or ‘the Pink class scored better.’ Most (54.9%) of the reasons supporting ‘Black’ were classified at level 2 and were about equally split among comparing centers and comparing shapes, with slightly more reasons that addressed shapes. Except for one idiosyncratic reason, the rest of the ‘Black’ reasons (43.1%) were at levels 3 or 4, and almost all of those reasons specifically relied on proportional reasoning. Most of the reasons supporting the ‘equal’ decision were either local or transitional focused on shape or transitional focused on centers. All of the reasons supporting the ‘Pink’ decision were either local or transitional.

Table 80.

The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group 2-GS.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Black	1	0	28	18	4	51
equal	1	4	8	0	0	13
Pink	0	6	4	0	0	10
Level	2	10	40	18	4	74
Total	(2.7)	(13.5)	(54.1)	(24.3)	(5.4)	(100)

Percentage of total count of participants in group 2-GS (n = 74) in parentheses.

For the Transitional responses in favor of ‘Black,’ the 2-GS and 1-GS groups were the only ones to provide more of those responses focused on shape than on a center (see tables 81 and 77). Most of the 2 –GS students who provided level 2 – shape responses that supported ‘Black’ described the Black class’s scores as “skewed” or “pushed to the right.” Except for one student who compared modes, all the level 2 – center responses cited a comparison of either “average” or “mean.” As with the 1-GS students, the 2-GS students who decided in favor of ‘Black’ and who focused on center at level 2 were largely successful and those who focused on shape at level 2 were largely not successful at making a reasonable estimate for how much better the Black class scored. For the level – center responses, 10 estimated the difference between the centers and only three provided idiosyncratic estimations, but for the level 2 – shape responses 12 of 15 provided idiosyncratic estimations with only two who found the difference between centers and one who found a difference between proportions above a ‘cut-point.’

Table 81.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 3, for group 2-GS.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Black	13	15	0	28	16	2	18
equal	3	5	0	8	0	0	0
Pink	3	1	0	4	0	0	0
Total	19 (25.7)	21 (28.4)	0 (0.0)	40 (54.1)	16 (21.6)	2 (2.7)	18 (24.3)

Percentage of total count of participants group 2-GS (n = 74) in parentheses.

For the level 2 responses not in support of the Black class scoring better, more supported ‘equal’ than ‘Pink.’ For the three ‘equal’ responses that focused on center, they all described the means as “about equal,” but the three ‘Pink’ responses that focused on center, each seem to have relied on miscalculations of the mean for each class as they cited a greater mean for ‘Pink’ (see table 81). The responses focused on shape that supported equal, all essentially noted the difference in class size and noted the different shapes made it appear that the Black class scored better, but then if more students were added to the Black class, its scores would become similar to the Pink class’s scores. The one student who focused on shape and supported ‘Pink’ appeared to be confused about the implication of a “bell shape” and “skewed” as the student correctly used both terms to describe the shape of the distribution of scores for the Pink and Black classes, however this student then claimed that those shapes implied that the Pink class had scored better. All of the 2-GS students who supported ‘equal’ or ‘Pink,’ with reasons focused on shape or center, attempted to compare the data sets

not as individual data points but as groups, yet their understanding of how to do this contained serious flaws. For example, those who chose ‘Pink’ based on their miscalculations of the means appear to have relied solely on their calculations without regard to the shapes and those who decided in favor of ‘equal’ or ‘Pink,’ in an attempt to account for the difference in class size, appear to have difficulty reasoning proportionally about the required comparison.

Although, the 1-GS students gave the highest percentage of level 1 responses, the 2-GS students gave the second highest percentage of level 1 responses (see tables 76 and 80). All the 2-GS students who responded at level 1 and chose ‘equal’ compared the frequencies of the highest scores with a few students also citing the difference in class size as problematic for making a comparison, similar to the following response:

I feel that classes scored equally because although the pink class has more students, both classes had the same amount of students who scored 7,8,9. It is hard to say that one class scored higher than the other because the sample size is not the same.

Those who decided in favor of ‘Pink,’ at level 1, did not address the difference in class size. All of those responses either compared sums of scores or frequencies, such as “In Pink class [sic], there are more dots in each range.” Yet 5 of the 6 level 1 – ‘Pink’ responses then went on to estimate how much better the Pink class scored by finding the difference of the sums of scores. As with the 1-GS students who responded in this way, there does appear to be potential with these students to build on this strategy to eventually make proportional comparisons between the classes.

Group 2-SE results: Pink/Black Task

Except for one student who decided that ‘the Pink class scored better’ and gave a Local type reason to support that decision, all the 14 other students from group 2-SE decided that ‘the Black class scored better.’ Nine of these 14 students supported their ‘Black’ decision with Transitional reasons, and although none responded at level 3, there were four who provided Distributional type responses (see table 82).

Table 82.

The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group 2-SE.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Black	1	0	9	0	4	14
equal	0	0	0	0	0	0
Pink	0	1	0	0	0	1
Level	1	1	9	0	4	15
Total	(6.7)	(6.7)	(60.0)	(0.0)	(26.7)	(100)

Percentage of total count of participants in group 2-SE (n = 15) in parentheses.

Table 83.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 3, for group 2-SE.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Black	7	2	0	9	0	0	0
equal	0	0	0	0	0	0	0
Pink	0	0	0	0	0	0	0
Total	7	2	0	9	0	0	0
	(46.7)	(13.3)	(0.0)	(60.0)	(0.0)	(0.0)	(0.0)

Percentage of total count of participants group 2-SE (n = 15) in parentheses.

Of the level 2 responses that the 2-SE students provided, most focused on comparing a measure of center while only two compared shapes (see table 83). Except for two students, one who compared medians and one who compared modes, all of those who compared center, specifically compared means. The students, who compared medians or modes, then went on to estimate how much better the Black class scored by finding the difference between the medians or modes, respectively. The students who compared the means did not have the same success, on the whole, at making estimations. Only two went on to find the difference between the means while three could not give a reasonable estimation for how much better the Black class scored.

All of the 2-SE students who provided Distributional responses relied on comparing both centers and shapes to support their conclusion that the Black class scored better. Two of these students went on to find the difference between centers and one went on to find the difference between the proportion of scores at 7 and higher for their estimations of how much better the black class scored. The remaining student summed the values of the mean and standard deviation for each class then found the difference between those sums. That estimation was classified as idiosyncratic.

Although the 2-SE students performed very well in that almost all decided that the Black class scored better, half of those who decided 'Black' also were unable to provide reasonable estimates for how much better the Black class scored. This was a

somewhat surprising result as all the 2-SE students had previously completed at least one statistics course.

Group GRAD results: Pink/Black Task

Except for one student who decided that ‘the classes scored equally well’ and gave a local type reason to support that decision, all of the 11 other students from the GRAD group decided that ‘the Black class scored better.’ Almost half, that is, five of these 11 students supported their ‘Black’ decision with Distributional type reasons and three students each supported their ‘Black’ decision with Initial Distributional and Transitional type reasons, respectively (see table 84).

Table 84.

The distribution of responses coded across framework levels from task 3 (the Pink/Black task), for group GRAD.

						Decision
<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>
Black	0	0	3	3	5	11
Equal	0	1	0	0	0	1
Pink	0	0	0	0	0	0
Level	0	1	3	3	5	12
Total	(0.0)	(8.3)	(25.0)	(25.0)	(41.7)	(100)

Percentage of total count of participants in group GRAD (n = 12) in parentheses.

The five GRAD students, who provided level 4 responses, all supported their decision on favor of ‘Black’ by comparing center with three additionally integrating shape comparisons and two additionally comparing proportions of scores above a “cut-point.” All of these students went on to make reasonable estimations for how much better the Black class scored as four found the difference between center

measurements and one computed the ratio of the two means. All of these students appear to have compared the distributions as whole units, i.e., from a global perspective.

Table 85.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 3, for group GRAD.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Black	2	1	0	3	3	0	3
Equal	0	0	0	0	0	0	0
Pink	0	0	0	0	0	0	0
Total	2 (16.7)	1 (8.3)	0 (0.0)	3 (25.0)	3 (25.0)	0 (0.0)	3 (25.0)

Percentage of total count of participants group GRAD (n = 12) in parentheses.

All of the students from the GRAD group, who provided level 3 responses, decided in favor of ‘Black’ and supported their decision by comparing proportions of scores above either a score of 5 or 7 (see table 85). Two of these students went on to estimate the difference between the proportions the reference and one found the difference between the means to estimate how much better the Black class scored. In particular, the two students who compared proportions and then made estimates based on those proportions likely viewed the partial distributions, which they compared, as group representatives.

Of the three GRAD students who supported their ‘Black’ decision with level 2 reasons, two focused on comparing a measure of center, one each one mean and median, and one focused on comparing shapes (see table 85). In making an estimation

for how much better the Black class scores, the level 2 – center responses were consistent as the student who initially compared means, then used the ratio of the means for an estimation and the student who initially compared medians found the difference between medians for an estimation. The student who compared shapes made his or her estimation by initially finding the difference between the modes, then adjusting that value smaller because, for the Black class, “the mean will be less than the mode.” All three of these students successfully responded to this task, and although they may have focused on a single feature of the data, potentially made their comparisons from a global perspective.

The one GRAD student who decided that the classes scored equally well wrote, “no way to tell because you can't compare with different values of n.” This student reported that he or she had previously completed five statistics course and was currently enrolled in three more courses, thus it is a bit surprising that he or she felt that a comparison could not be made merely because the size of each distribution was different. Despite this one student, the rest of students from the GRAD group chose ‘Black’ using reasonable strategies and then made reasonable estimates, using group representatives, for how much better the Black class scored.

Responses for Group 1-GS: Ambulance Task

Table 86 and Figure 51 display the results for the decisions and coded reasons provided by the 1-GS students. Slightly more than half of all the 1-GS students recommended Life Line. Irrespective of their recommendation, students from group 1-GS most frequently provided level 1, local type reasons (46%), with the next most

frequent type of reasons at level 2 (32.8%) followed by level 3 reasons (5.8%) with no students providing distributional type reasons. Figure 51 also shows that the pattern of response levels was quite similar between the two choices of ambulance service.

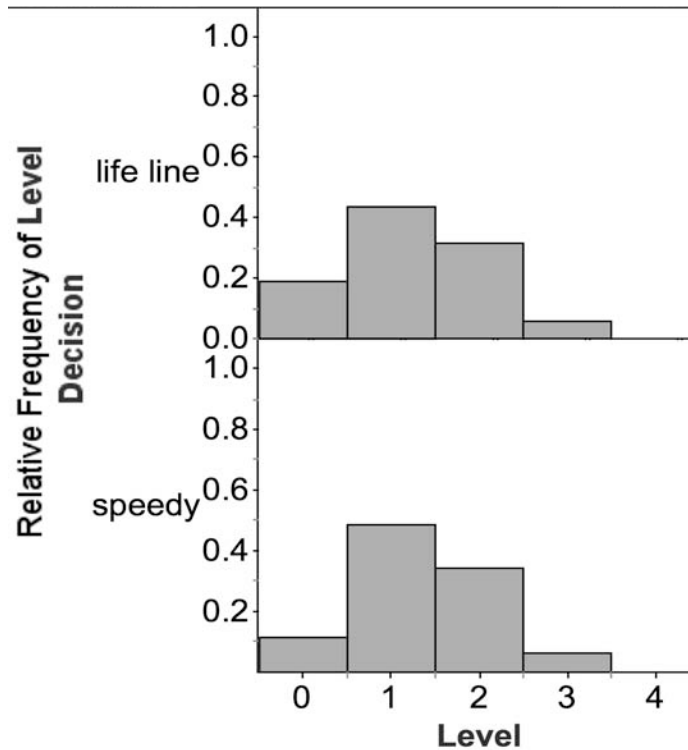


Figure 51. The distribution of response levels of the 1-GS students to the Ambulance task, separated by recommendation.

Table 86.

The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group 1-GS.

						Decision
<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Total</u>
Life Line	14	32	23	4	0	73
Speedy	7	31	22	4	0	64
Level	21	63	45	8	0	137
Total	(15.3)	(46.0)	(32.9)	(5.8)	(0.0)	(100)

Percentage of total count of participants in group 1-GS (n = 137) in parentheses.

The Local types of responses, in favor of Life Line, frequently cited Life Line's shorter minimum response time or Speedy's longer maximum response time. The Speedy recommendations often did not account for the unequal sizes of the groups of response times. For example, the following Life Line recommendations appeared to be based on strictly comparing the ends of the distributions without consideration of the distributions' shape, center or spread and the following Speedy recommendations focused on frequencies:

Life Line: *because they don't have as many really long responses* [sic]

Life Line: *speedy had no times lower than 6 mins* [sic]

Speedy: *this one is better more dots* [sic]

Speedy: *they have more shorter response times recorded.*

The 1-GS students provided a noticeably high frequency of responses similar to the first example in favor of Speedy. Although some of those responses referred to "response times" instead of "dots," many of these level 1 responses included no explanation as to why "more response times" was better, and a few stated that more times meant that Speedy could be "trusted" more or was more reliable.

Table 87.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 5, for group 1-GS.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Life Line	4	14	5	23	4	0	4
Speedy	7	15	0	22	4	0	4
Total	11 (8.0)	29 (21.2)	5 (3.7)	45 (32.9)	8 (5.8)	0 (0.0)	8 (5.8)

Percentage of total count of participants group 1-GS (n = 137) in parentheses.

Almost two-thirds of the Transitional responses provided by the 1-GS students focused on comparing shapes (see table 87). There were no apparent trends in the Shape arguments used to support each recommendation. Both of the Life Line and Speedy distributions were described by some students as “grouped more lower” and described by other students as “grouped higher.” The Life Line distribution was described several times as “concentrated around 20” and that description was used to support both recommendations. One student did decide to recommend Life Line because it “looks skewed to the left.” Some students who focused on comparing a measure of center also encountered difficulties with the Ambulance task. Half of the students who compared means recommended Life Line and half recommended Speedy, although two of the students who claimed the mean was lower for Speedy specifically stated that they made estimates, not calculations. The remaining students who cited a comparison of medians all recommended Speedy.

None of the 1-GS students gave Distributional responses, but eight did provide Initial Distributional responses, all focused on Proportion and evenly split between the recommendations. These students compared proportions of times below a variety of “cut-points,” that are: 10 minutes, 12 minutes, 14 minutes, 18 minutes, and 20 minutes. One of the students who recommended Speedy compared the proportion of times above 20 minutes. Not all of the proportions referenced in these responses were computed accurately.

On the whole, the 1-GS students appeared to experience difficulties in responding to the Ambulance task. More than 60% provided responses categorized at

level 1 or level 0, with the highest percentage of responses at level 1. Many of the level 1 and responses indicated that those students either did not reason proportionally or had difficulty reasoning proportionally.

Responses for Group 1-SE: Ambulance Task

Table 88 and Figure 52 display the results for the decisions and coded reasons provided by the 1-SE students. These students recommended Life Line at about a 2 to 1 rate over Speedy. More than half of the students from group 1-SE supported their recommendations with level 2 type reasons. About half of the responses supporting each recommendation were at level 2, and each recommendation had about the same frequency of level 1 responses as levels 3 and 4 responses combined, although the only two distributional responses supported Life Line. No students from the 1-SE group gave idiosyncratic reasons.

Table 88.

The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group 1-SE.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Life Line	0	6	14	3	2	25
Speedy	0	2	7	3	0	12
Level	0	8	21	6	2	37
Total	(0.0)	(21.6)	(56.8)	(16.2)	(5.4)	(100)

Percentage of total count of participants in group 1-SE (n = 37) in parentheses.

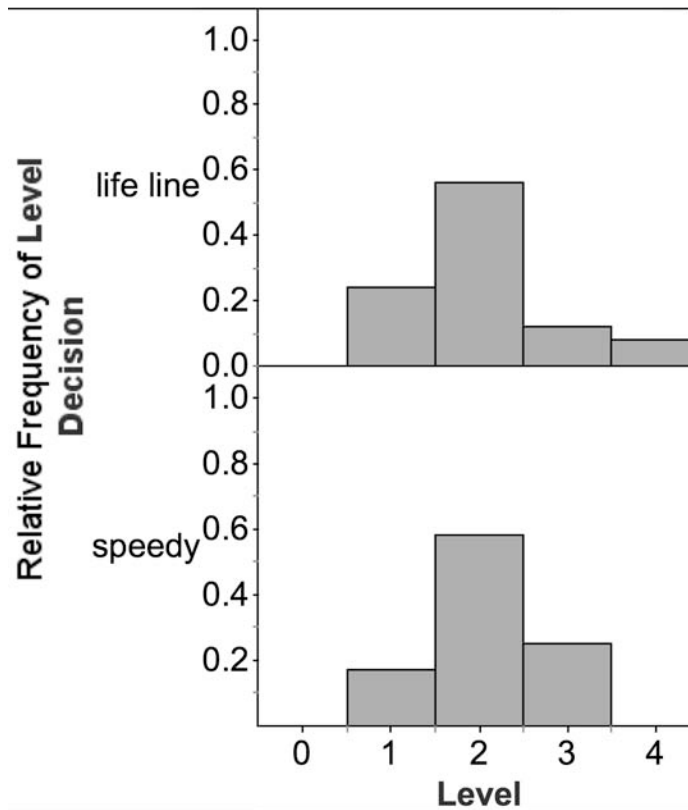


Figure 52. The distribution of response levels of the 1-SE students to the Ambulance task, separated by recommendation.

All of the 1-SE students who provided level 1 responses, in favor of Life Line, compared the longest response times for each Ambulance service, most of them noting that Life Line had no times above 24 minutes. Both of the students who provided level 1 responses in favor of Speedy compared the frequency of “calls.” One student even noted that Speedy had “more short response times” and “more calls.”

Table 89.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 5, for group 1-SE.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Life Line	5	6	3	14	1	2	3
Speedy	4	3	0	7	3	0	3
Total	9 (24.3)	9 (24.3)	3 (8.1)	21 (56.8)	4 (10.8)	2 (5.4)	6 (16.2)

Percentage of total count of participants group 1-SE (n = 37) in parentheses.

Most of the level 2 responses that the 1-SE students provided were equally split between focusing on Shape and focusing on Center (see table 89). Although the Transitional responses from the 1-SE students were distributed differently than those from the 1-GS students, the content of level 2 – Shape and Level 2 – Center responses, from each group, were strikingly similar. The Shape responses described both Life Line and Speedy as “grouped more lower” and as “grouped higher” and the Life Line distribution was described as “concentrated around 20.” Students who focused on comparing means recommended Life Line and Speedy, and the students who cited a comparison of medians all recommended Speedy. The few students who compared spreads all favored Life Line.

Of the two 1-SE students who gave Distributional responses supported Life Line, one compared centers and proportion, and the other is particularly worth noting, as it was unique:

well I split up the charts in half... and then I determined the average of each half and my results were that life lines average was quicker on the first half and second half of the chart.

The Initial Distributional responses were evenly split between the recommendations. All that supported Speedy focused on Proportion, but only one that supported Life Line focused on Proportion. These students compared proportions of times below “cut-points” of 15 minutes and 20 minutes, and above 20 minutes and 24 minutes. The two Initial Global responses, which supported Life Line, focused on one distributional feature and a local feature, such as, “...I’d tend to select the one that seems to have fewer high-end times and more consistent response times.”

Both groups, 1-GS and 1-SE, had some similarities in their responses, such as both groups showed no pattern of responses that supported one recommendation over the other. Yet, a considerably higher proportion of 1-SE students recommended Life Line and the 1-SE group also appeared to provide their responses at a higher level. Their most frequent response level was Transitional as opposed to Local and they provided a few Distributional responses and no Idiosyncratic responses.

Responses for Group 2-GS: Ambulance Task

Table 90 and Figure 53 display the results for the decisions and coded reasons provided by the 2-GS students. Exactly half of all the 2-GS students recommended each ambulance service. Most responses from the 2-GS students were at either level 2 or 1. The highest percentage of responses supporting the Life Line recommendation were at level 2, but when supporting the Speedy recommendation there was one more level 1 response than level 2 responses. Figure 56 also shows that the pattern of response levels was a bit different between the two choices of ambulance service, particularly as there were fewer Life Line responses at level 0 and more at level 4.

Table 90.

The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group 2-GS.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Life Line	2	12	18	3	2	37
Speedy	8	13	12	4	0	37
Level	10	25	30	7	2	74
Total	(13.5)	(33.8)	(40.5)	(9.5)	(2.7)	(100)

Percentage of total count of participants in group 2-GS (n = 74) in parentheses.

The 2-GS students gave almost as many Local responses as Transitional responses. The Local responses were very similar to those given by the 1-GS and 1-SE students. These responses tended to focus on comparing either the high ends or low ends of the distributions or focused on comparing frequencies without regard for the difference in the amount of response times for each ambulance service. One student use a somewhat different strategy in that he or she compared the times for when the first “peak” occurred for each ambulance service. This student concluded that the first peak for Life Line was at 12 minutes but the first peak for Speedy was at 6 minutes, so Speedy was the recommended service.

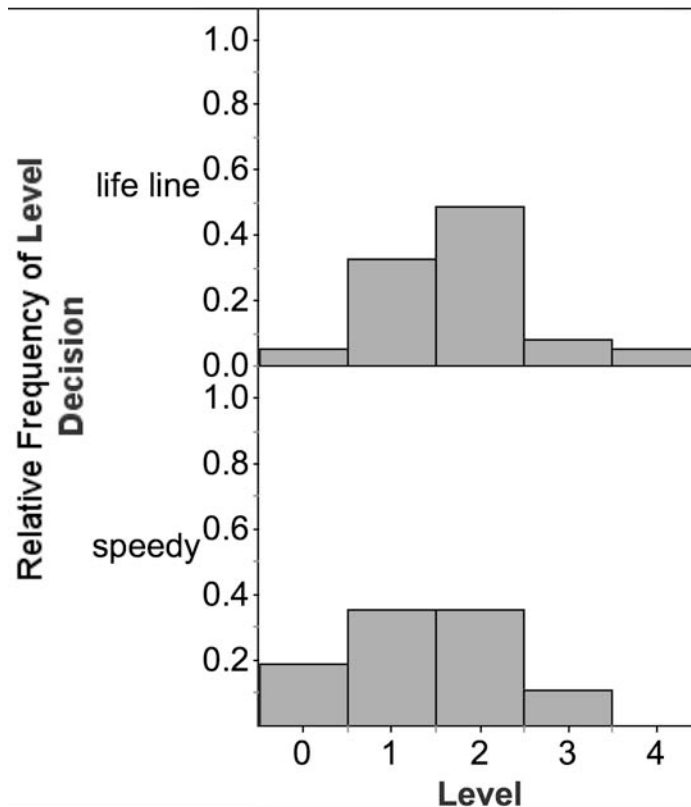


Figure 53. The distribution of response levels of the 2-GS students to the Ambulance task, separated by recommendation.

Overall, for the Ambulance task recommendations, the 2-GS students most frequently gave Transitional responses. Those responses were distributed differently than those from the 1-GS students as most of the level 2 responses that the 2-GS students provided were equally split between focusing on Shape and focusing on Center, as opposed to more that focused on Shape (see table 91). Those 2-GS students who did focus on Shape were about evenly split between the two recommendations. The Shape responses, although similar to those given by the 1-GS and 1-SE students, also seemed to be a bit less articulate as descriptive words such as “grouped” and

“concentrated” were used less descriptive trend description, such as, “I would choose speedy because their data seems to be mostly towards the left side” were used more. Of the students who focused on comparing centers, a large majority referenced the “mean” or “average” with most making estimation statements similar to, “[the] mean for speedy appears lower.” A few compared modes to support Life Line and one compared medians to support Speedy.

Table 91.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 5, for group 2-GS.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Life Line	9	7	2	18	3	0	3
Speedy	4	8	0	12	4	0	4
Total	13 (17.8)	15 (20.3)	2 (2.7)	30 (40.5)	7 (9.5)	0 (0.0)	7 (9.5)

Percentage of total count of participants group 2-GS (n = 74) in parentheses.

Only about 10% of the 2-GS students focused on comparing proportions to make their recommendation. Those responses were about evenly split between the recommendations. As with the previously described Proportional type responses the students used “cut-points” but there was no particular trend in which points were used to support particular recommendations. Most of these students used a “cut-point” of 20 minutes and although a majority of those supported Speedy, not all supported Speedy. This may be due to inarticulate descriptions of their reasons or from miscalculations.

The only two Distributional responses supported the Life Line recommendation. Both compared estimations of the means and variation, with one

also comparing skewness. Both appeared to be making comparisons of whole distributions.

Both groups, 2-GS and 1-GS, had some similarities in their responses, such as relatively high percentages of responses at level 1 and level 0 and the inability of many 2-GS students to reason proportionally. While the 1-GS students had no pattern of responses that supported one recommendation over the other, the 2-GS students who gave Idiosyncratic responses, most purely contextual, favored Speedy over Life Line, and another slight trend was that students who focused on comparing center favored Life Line over Speedy. Overall, the 2-GS students appeared to respond only at slightly higher levels than the 1-GS students, as the 2-GS students provided a higher percentage of level 2 responses with a few distributional responses and a lower percentage of idiosyncratic responses. The statistics course that all the 2-GS students completed possibly had a small impact on the 2-GS students when responding to the Ambulance task.

Responses for Group 2-SE: Ambulance Task

Table 92 and Figure 54 display the results for the decisions and coded reasons provided by the 2-SE students. Almost half of all the 2-SE students recommended each ambulance service. The majority of responses from the 2-SE students were at either level 2 or level 1, with only one response at each of the other levels. The highest percentage of responses supporting the Life Line recommendation were at level 2, but when supporting the Speedy recommendation the highest percentage of responses was

at level 1. Figure 54 also shows that the pattern of response levels was quite different between the two choices of ambulance service.

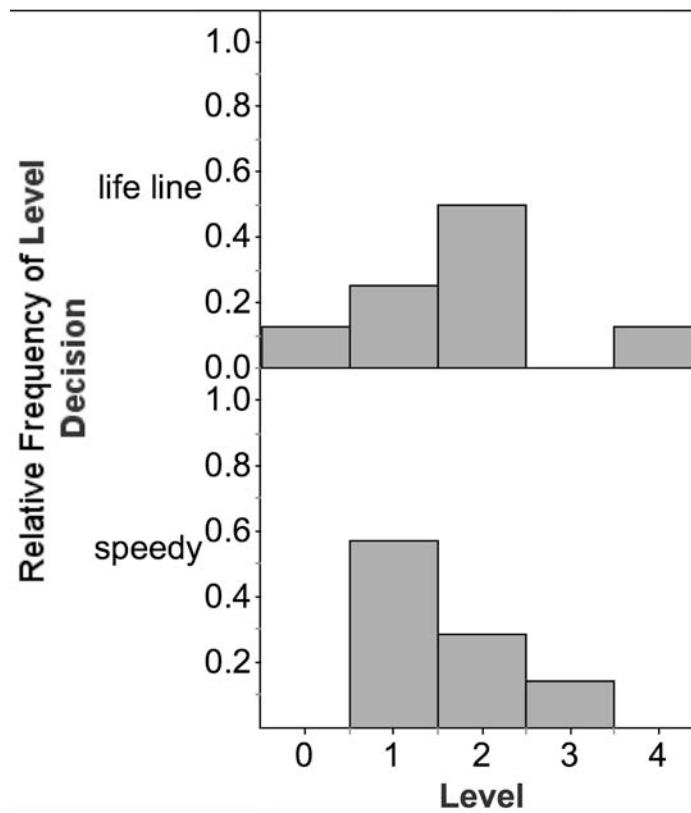


Figure 54. The distribution of response levels of the 2-SE students to the Ambulance task, separated by recommendation.

Table 92.

The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group 2-SE.

Decision	Decision					Total
	Level 0	Level 1	Level 2	Level 3	Level 4	
Life Line	1	2	4	0	1	8
Speedy	0	4	2	1	0	7
Level	1	6	6	1	1	15
Total	(6.7)	(40.0)	(40.0)	(6.7)	(6.7)	(100)

Percentage of total count of participants in group 2-SE (n = 15) in parentheses.

Two of the 2-SE students provided Local responses that favored Life Line while four provided local responses that favored Speedy. The Local responses were very similar to those given by the “GS” and 1-SE students. These responses, in favor of Life Line, focused on comparing the high ends of the distributions, such as “In the event of an emergency, seconds can mean life or death. Life Line never had a wait time of more than 24 min.” Those that favored Speedy compared frequencies without regard for the difference in the amount of response times for each ambulance service, such as, “While both companies have several high data points, Speedy has a lot of examples of arriving quickly.” All of these students did not appear to be comparing the response times as whole units and also did not reason proportionally about the distributions.

While the 2-SE students provided the same frequency of Local and Transitional responses, more level 2 responses favored Life Line whereas more level 1 responses favored Speedy. There were not enough level 2 responses from the 2-SE students to cite specific trends in focus (see table 93), but there were, again, similarities with the “GS” and 1-SE students. All the responses that focused on Center, cited comparisons of mean, although they did not all cite that the mean was lower for Life Line. The two that focused on variation cited the smaller spread of the times for Life Line and the one that focused on Shape claimed that more of the Speedy times looked lower as opposed to higher.

Table 93.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 5, for group 2-SE.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Life Line	2	0	2	4	0	0	0
Speedy	1	1	0	2	1	0	1
Total	3 (20.0)	1 (6.7)	2 (13.3)	6 (40.0)	1 (6.7)	0 (0.0)	1 (6.7)

Percentage of total count of participants group 2-SE (n = 15) in parentheses.

Only two of the 2-SE students responded to the Ambulance task at a level higher than level 2. The one response that focused on Proportion compared the proportion of times at 18 minutes and lower to favor Speedy. The one Distributional response chose Life Line based on its lower mean and smaller variation.

The 2-SE and 1-SE groups seemed to have more difference than similarities in their responses to the Ambulance task. Although both gave relatively high percentages of responses at level 1 and level 2, more 1-SE students provided level 2 responses than level 1 responses whereas the 2-SE students provided level 1 and 2 responses at equal frequencies. The 1-SE students also gave a higher percentage of level 3 responses than the 2-GS students did, even though the 2-SE students reported that they had previously completed at least one statistics course. For both groups of “SE” students, those who recommended Life Line appeared to provide responses at a slightly higher level than those who recommended Speedy.

Responses for Group GRAD: Ambulance Task

Table 94 and Figure 55 display the results for the decisions and coded reasons provided by the GRAD students. Twice as many students from the GRAD group recommended Life Line as recommended Speedy. All of the responses from the GRAD students were at level 2 or higher. While all of the responses supporting the Speedy recommendation were at level 2, the responses that supported the Life Line recommendation were mostly at level 2 and level 4. Figure 55 also shows that the pattern of response levels was quite different between the two choices of ambulance service.

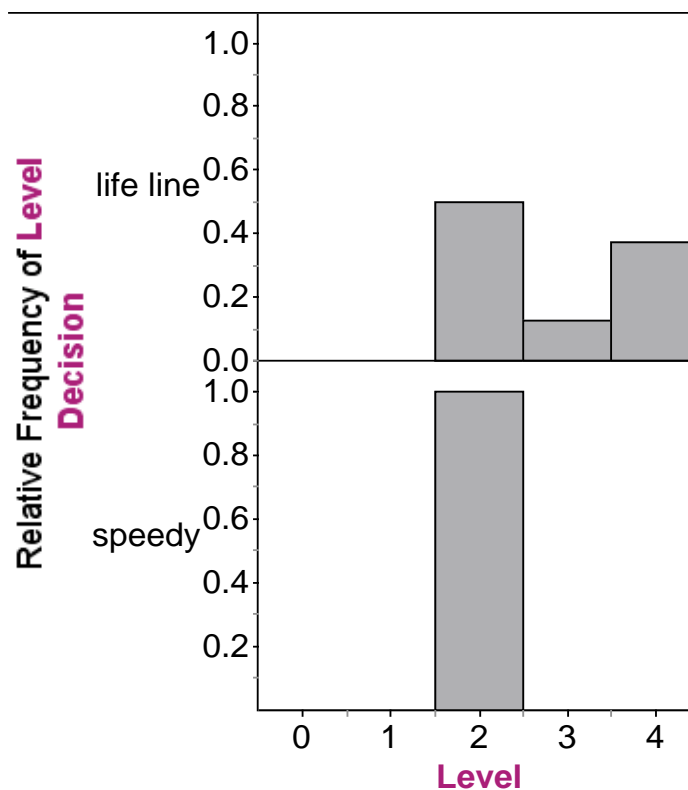


Figure 55. The distribution of response levels of the GRAD students to the Ambulance task, separated by recommendation.

Table 94.

The distribution of responses coded across framework levels from task 5 (the Ambulance task), for group GRAD.

<u>Decision</u>	<u>Level 0</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Decision Total</u>
Life Line	0	0	4	1	3	8
Speedy	0	0	4	0	0	4
Level	0	0	8	1	3	12
Total	(0.0)	(0.0)	(66.7)	(8.3)	(25.0)	(100)

Percentage of total count of participants in group GRAD (n = 12) in parentheses.

All but one of the Transitional responses provided by the GRAD students focused on Shape. For most of the level 2 responses that favored Speedy, those GRAD students described comparisons of their visual estimations of skewness (see table 95). Two of those who favored of Life Line, cited its lower mode time and lower minimum and maximum times. The other who recommended Life Line its lower maximum time and that more of its times were “close” to 20 minutes, although it is not clear how or why times closer to 20 minutes are better.

Table 95.

The distribution of responses coded at level 2 (*transitional*) and level 3 (*initial distributional*) from survey task 5, for group GRAD.

<u>Decision</u>	<u>Level 2</u>				<u>Level 3</u>		
	<u>C</u>	<u>SH</u>	<u>V</u>	<u>Total</u>	<u>P</u>	<u>IG</u>	<u>Total</u>
Life Line	0	3	1	4	0	1	1
Speedy	0	4	0	4	0	0	0
Total	0	7	1	8	0	1	1
	(0.0)	(58.3)	(8.3)	(66.7)	(0.0)	(8.3)	(8.3)

Percentage of total count of participants group GRAD (n = 12) in parentheses.

The one GRAD student who provided an Initial Distributional response recommended Life Line with an Initial Global focus. The student cited Life Line's lower median, mode, maximum, and minimum times. The three who provided Distributional responses also recommended Life Line. Two of those considered proportions and centers while the other compared means and shape.

The GRAD students, recommended Life Line at a 2 to 1 rate. Generally they responded at higher levels of the framework than any other group, with no responses below level 2 and the highest percentage at level 4. All of the responses from the GRAD group were at level 3 or level 4 and supported the Life Line recommendation.