DEVELOPMENT OF PRESERVICE MIDDLE SCHOOL MATHEMATICS TEACHERS' REASONING ABOUT DISTRIBUTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

NAZLI AKAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
MATHEMATICS EDUCATION IN MATHEMATICS AND SCIENCE EDUCATION

MAY 2024

Approval of the thesis:

DEVELOPMENT OF PRESERVICE MIDDLE SCHOOL MATHEMATICS TEACHERS' REASONING ABOUT DISTRIBUTION

submitted by **NAZLI AKAR** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** i**n Mathematics Education in Mathematics and Science Education, Middle East Technical University** by,

Prof. Dr. Emre Altun
Dean, **Graduate School of Natural and Applied Sciences** _____

Prof. Dr. Mine Işıksal Bostan
Head of the Department, **MSE** _____

Prof. Dr. Mine Işıksal Bostan
Supervisor, **Mathematics and Science Education, METU** _____

Examining Committee Members:

Prof. Dr. Erdinç Çakıroğlu
Mathematics and Science Education, TED University _____

Prof. Dr. Mine Işıksal Bostan
Mathematics and Science Education, METU _____

Assoc. Prof. Dr. Sibel Kazak
Mathematics and Science Education, METU _____

Assoc. Prof. Dr. Gülseren Karagöz Akar
Mathematics and Science Education, Boğaziçi University _____

Assoc. Prof. Dr. Şerife Sevinç
Mathematics and Science Education, METU _____

Date: 14.05.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Nazlı Akar

Signature :

ABSTRACT

DEVELOPMENT OF PRESERVICE MIDDLE SCHOOL MATHEMATICS TEACHERS' REASONING ABOUT DISTRIBUTION

Akar, Nazlı
Doctor of Philosophy, Mathematics Education in Mathematics and Science Education
Supervisor : Prof. Dr. Mine Işıksal Bostan

May 2024, 330 pages

The study aims to investigate the development of preservice mathematics teachers' reasoning about distribution as they participated in the classroom teaching experiment based on Realistic Mathematics Education and enriched by dynamic statistics software. To this end, 12 compelling learning activities were implemented in the classroom teaching experiment with 14 preservice middle school mathematics teachers. The activities in the teaching experiment were designed to promote classroom discussion and collaboration. The study's data comprised the pretest, posttest, semi-structured interviews, video recordings of classroom and debriefing sessions, the participants' written artifacts, and field notes. Qualitative methods were employed to analyze the data. The findings revealed that the preservice mathematics teachers' reasoning levels improved considerably in reasoning about statistical measures, data representations, shape and shifts in data, and proportional reasoning after the teaching experiment. Classroom statistical practices emerging from taken-as-shared ideas based on communal reasoning in the classroom showed how the preservice mathematics teachers reasoned throughout the teaching experiment enriched by dynamic statistical software. This study indicated that classroom

teaching that gradually progressed by making sense of central tendency and variability, interpreting density in the data, and reasoning about the shape of the distribution in an integrated manner in the social learning environment contributed to the improvement of preservice mathematics teachers' reasoning about distribution. In that respect, the study has valuable implications for teacher education in developing preservice mathematics teachers' reasoning about distribution.

Keywords: Distribution, Dynamic Statistics Software, Preservice Mathematics Teachers, Realistic Mathematics Education, Statistical Reasoning

ÖZ

İLKÖĞRETİM MATEMATİK ÖĞRETMEN ADAYLARININ DAĞILIM HAKKINDA AKIL YÜRÜTMELERİNİN GELİŞTİRİLMESİ

Akar, Nazlı
Doktora, Matematik Eğitimi, Matematik ve Fen Bilimleri Eğitimi
Tez Yöneticisi: Prof. Dr. Mine Işıksal Bostan

Mayıs 2024, 330 sayfa

Bu çalışmanın amacı, Gerçekçi Matematik eğitimine dayanan ve dinamik istatistik yazılımı ile zenginleştirilmiş sınıf öğretim deneyine katılan matematik öğretmen adaylarının dağılım hakkındaki akıl yürütmelerinin gelişimini araştırmaktır. Bu amaçla, 14 ortaokul matematik öğretmeni adayının katıldığı sınıf öğretim deneyinde 12 zorlayıcı öğrenme etkinliği uygulanmıştır. Öğretim deneyindeki etkinlikler sınıf içi tartışma ve işbirliğini teşvik edecek şekilde tasarlanmıştır. Çalışmanın verileri ön test, son test, yarı yapılandırılmış görüşmeler, sınıf ve bilgilendirme oturumlarının video kayıtları, katılımcıların yazılı eserleri ve saha notlarından oluşmaktadır. Verileri analiz etmek için nitel yöntemler kullanılmıştır. Bulgular, matematik öğretmen adaylarının öğretim deneyinden sonra istatistiksel ölçümler, veri temsilleri, verilerdeki şekil ve kaymalar ve orantısal akıl yürütme konularında akıl yürütme düzeylerinin önemli ölçüde geliştiğini ortaya koymuştur. Sınıfta ortak akıl yürütmeye dayalı olarak paylaşılan fikirlerden ortaya çıkan sınıf istatistik uygulamaları, matematik öğretmen adaylarının dinamik istatistik yazılımı ile zenginleştirilen sınıf öğretim deneyi boyunca nasıl akıl yürüttüklerini göstermiştir. Sosyal öğrenme ortamında merkezi eğilim ve değişkenliği anlamlandırma,

verilerdeki yoğunluğu yorumlama ve dağılımın şekli hakkında akıl yürütme ile bütünleşik bir şekilde aşamalı olarak ilerleyen sınıf öğretim deneyi, matematik öğretmen adaylarının dağılım hakkında akıl yürütmelerinin gelişimine katkıda bulunmuştur. Bu açıdan çalışma, matematik öğretmeni adaylarının dağılımla ilgili akıl yürütmelerini geliştirmede öğretmen eğitimi için değerli çıkarımlara sahiptir.

Anahtar Kelimeler: Gerçekçi Matematik Eğitimi, Dağılım, Dinamik İstatistik Yazılımı, Matematik Öğretmen Adayları, İstatistiksel Akıl Yürütme

To my Family

ACKNOWLEDGMENTS

First and foremost, I wish to express my deepest gratitude to my supervisor, Prof. Dr. Mine Işıksal Bostan, for her patience, exceptional support, constant guidance, and valuable encouragement. Without her immense mentorship, I would have been lost in this long, challenging research process. She is my inspiration in my academic journey. Beyond research and teaching, I have gained a great deal of experience about life from her and will continue to learn to do so.

Also, I would like to thank the rest of my thesis committee: Prof. Dr. Erdinç ÇAKIROĞLU, Assoc. Prof. Dr. Sibel KAZAK, Assoc. Prof. Dr. Gülseren KARAGÖZ AKAR, and Assoc. Prof. Dr. Şerife SEVİNÇ. Their valuable knowledge, suggestions, and feedback have greatly enriched my dissertation.

I would especially like to thank my dear friend Emine Çatman Aksoy, who touched every step of my thesis. With her support, I took more confident steps forward. Following that, I am fortunate to have my close friends: Nurbanu, Öznur, Emine A., Cansu, Aniş, Ceren, Mine, Beyza, Sema Nur, Gözde, Merve and Sevgi. Whenever I felt down, I got up with their motivation and support. Thank you very much.

I want to express my gratitude to my large family: my parents and siblings. They have always loved me and believed in me. It's great to see that they’re behind me whenever I need them.

My dear husband, Erhan, I can not thank you enough. Your continuous love and encouragement, unwavering support, understanding, patience, tolerance, compassion, etc., are the most incredible power behind me. You fulfilled with your poetry and food in this journey. I feel alive again in your loving eyes. You always give me the hope. My prayer has been answered, we did it together. Let's continue to realize our dreams one by one…

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xix

CHAPTERS

1 INTRODUCTION 1

1.1 Research Questions 5

1.2 Significance of the Study 6

1.3 Definition of Important Terms 9

2 LITERATURE REVIEW 11

2.1 Statistical Reasoning 11

2.2 The Concept of Distribution 16

2.2.1 Center 18

2.2.2 Spread 21

2.2.3 Density 24

2.2.4 Skewness 26

2.2.5 The Studies on Reasoning About Distribution 28

2.3 Theoretical Framework 40
2.3.1 Realistic Mathematics Education 41
2.3.2 Emergent Perspective 44
3 METHODOLOGY 47
3.1 Design of the Study 47
3.2 Participants of the Study 51
3.3 Data Collection Process 53
3.3.1 Preparing for the Experiment 53
3.3.2 The Design Experiment 75
3.3.3 The Retrospective Analysis 94
3.4 Data Analysis 96
3.5 Trustworthiness 112
3.6 Researcher Role 114
3.7 Ethical Considerations 115
4 FINDINGS 117
4.1 The PMTs' Statistical Reasoning Levels About Distribution Before and After the Classroom Teaching Experiment 118
4.1.1 Reasoning about Statistical Measures 121
4.1.2 Reasoning about Representations of Data 125
4.1.3 Reasoning about Shape and Shifts in Data 129
4.1.4 Proportional Reasoning 141
4.2 The Emerged Classroom Statistical Practices Throughout the Teaching Experiment 150
4.2.1 Classroom Statistical Practice 1: Reasoning about the average as the signal in a noisy process 152

4.2.2 Classroom Statistical Practice 2: Measuring of variability in data . 177

4.2.3 Classroom Statistical Practice 3: Selecting appropriate statistics for the context 189

4.2.4 Classroom Statistical Practice 4: Exploring qualitative characteristics of collections of data points 200

4.2.5 Classroom Statistical Practice 5: Exploring qualitative characteristics of distributions 223

5 CONCLUSIONS AND DISCUSSION 267

5.1 The Reasoning Skills to Reason about Distribution..... 267

5.2 The Classroom Statistical Practices That Emerged Throughout the Classroom Teaching Experiment 272

5.3 Implications of the Study 281

5.4 Limitations of the Study and Suggestions for Future Research 284

REFERENCES 287

APPENDICES

A. Elementary Mathematics Education Curriculum 309

B. Ethical Approval 311

C. Permission for Activities..... 312

D. Consent Form (Turkish Version) 313

E. Pretest..... 314

F. Posttest..... 322

CURRICULUM VITAE.....329

LIST OF TABLES

TABLES

Table 2.1 The strategies in proportional reasoning 33
Table 2.2 The interpretative framework for analyzing individual and collective activity at the classroom level (Cobb & Yackel, 1996, p. 6) 44
Table 3.1 The order of the activities in the initial instructional sequence and the HLT of this study 56
Table 3.2 The last version of the HLT implemented in the classroom teaching experiment 73
Table 3.3 The schedule was followed in the course 75
Table 3.4 The skills that the questions measure in the tests 91
Table 3.5 The adapted version of the model of statistical reasoning (Garfield, 2002; Jones et al., 2004) to code the PMT's reasoning 97
Table 3.6 Research questions, corresponding data collection tools, and data analysis 109
Table 4.1 The reasoning levels of PMTs while reasoning about statistical measures before and after the teaching experiment 122
Table 4.2 The reasoning levels of PMTs while reasoning about representations of data before and after the teaching experiment 127
Table 4.3 The reasoning levels of PMTs while reasoning about shape and shifts in data before and after the teaching experiment 132
Table 4.4 The reasoning levels of PMTs for proportional reasoning before and after the teaching experiment 144
Table 4.5 The CSPs and TAS ideas that emerged throughout the classroom teaching experiment 151
Table 4.6 Groups' ranking of the factors from most important to least important 153
Table 4.7 The emerged CSP, TAS ideas, and the PMTs' reasoning levels while dealing with these ideas 262
Table 5.1 The CSPs and the taken-as-shared interest/activities in the HLT 273

LIST OF FIGURES

FIGURES

Figure 2.1. The perspectives for statistical reasoning, thinking, and literacy 13
Figure 2.2. The relationship between data and distribution (Bakker & Gravemeijer, 2004, p. 148) 18
Figure 2.3. The hypothetical data in dot and box plots 25
Figure 2.4. Different shapes of distributions (Gravetter & Wallnau, 2015, p. 48) . 26
Figure 2.5. Ratio table for the example 36
Figure 2.6. The graphs developed from AIDS data and the teacher's argument (Cobb, 1999, p. 26) 37
Figure 3.1. Teaching cycle in the classroom teaching experiment 49
Figure 3.2. The aerial photograph of the elephant herd 58
Figure 3.3. The book page whose middle is empty in the third activity of the HLT 60
Figure 3.4. The points scored by the two teams in each of nine games (I) and two games (II) (Kader & Mamer, 2008, pp. 39-40) 62
Figure 3.5. The data of the activity related to resting heart rate 67
Figure 3.6. The data of the activities related to backpack weights of the same (a) and different (b) numbers of students 68
Figure 3.7. The data of anorexia patients using traditional therapy or family therapy 70
Figure 3.8. The data on response times of ambulances 71
Figure 3.9. Question 1 in the pretest and posttest 83
Figure 3.10. The data of the employees who got a flu shot and did not get a flu shot in Question 5 in the pretest 87
Figure 3.11. The data of the headache patients using new or old medication in Question 5 in the posttest 89
Figure 3.12. The skeleton proposed by Toulmin (2003, p. 92) 101

Figure 3.13. The graphs structured by the group in TinkerPlots for the activity related to anorexia 103
Figure 3.14. The argumentation scheme of the excerpt 104
Figure 3.15. The outline of the data analysis process 111
Figure 4.1. The stacked bar graphs of the PMTs' reasoning levels based on skills before and after the teaching experiment 119
Figure 4.2. Question 2 in the pretest (a) and posttest (b) 121
Figure 4.3. Question 3 in the pretest (a) and posttest (b) 126
Figure 4.4. Question 4 in the pretest (a) and posttest (b) 131
Figure 4.5. The graphs that PMT13 drew in TinkerPlots in Part 2 of the posttest 135
Figure 4.6. The graphs that PMT8 drew in TinkerPlots in the posttest 137
Figure 4.7.The graphs that PMT3 drew in TinkerPlots in the posttest 139
Figure 4.8. The graphs that PMT12 drew in TinkerPlots in the posttest 140
Figure 4.9. The graphs that PMT9 drew in TinkerPlots in Part 2 of the posttest .. 147
Figure 4.10.The graphs that PMT7 drew in TinkerPlots in Part 2 of the posttest . 148
Figure 4.11. The argumentation scheme related to ranking the data 156
Figure 4.12. Group 3's grouping of the elephants 160
Figure 4.13. The picture is divided into ten equal parts 161
Figure 4.14. The argumentation scheme for the typical value representing the data 163
Figure 4.15. The table was drawn by Group 3 on page 166
Figure 4.16. The argumentation scheme regarding the equal distribution of words and lines 167
Figure 4.17. The solution for complete and incomplete lines 169
Figure 4.18. Normal (a) and left-skewed (b) distributions plotted by PMT5 172
Figure 4.19. The argumentation scheme regarding the effect of the outliers on the mean 173
Figure 4.20. The argumentation scheme regarding the balance point interpretation of the mean 174
Figure 4.21. The distributions for balance point interpretation of the mean 175

Figure 4.22. The points scored by two teams in each of the nine games (Kader & Mamer, 2008, p. 39) 178
Figure 4.23. The argumentation scheme regarding the relation between normality and variability 181
Figure 4.24. The points scored by the two teams (Kader & Mamer, 2008, p. 40) 182
Figure 4.25. The argumentation scheme regarding the use of SAD and MAD 184
Figure 4.26. The argumentation scheme regarding the relation between MAD and the standard deviation 186
Figure 4.27. The plots drawn by PMTs for the two battery brands in TinkerPlots 188
Figure 4.28. The solution of Group 3 for the activity related to batteries 192
Figure 4.29. The argumentation scheme for the use of measures of variability for the activity related to batteries 193
Figure 4.30. The argumentation scheme where measures of center and variability were connected for the activity related to batteries 196
Figure 4.31. The horizontal bar graphs of the three battery brands drawn by the PMT2-PMT4 pair in TinkerPlots 198
Figure 4.32. The argumentation scheme regarding the balloon can safely carry 12 students 204
Figure 4.33. The graph is drawn by PMT2 for the weights of students and adults 205
Figure 4.34. The graph drawn by Group 1 for the weights of 12 students in TinkerPlots 206
Figure 4.35. Different distributions for weight context (Bakker, 2004, pp. 219-222) 210
Figure 4.36. The argumentation scheme regarding the watermelon activity 215
Figure 4.37. The heart rate distribution of the class represented by PMT7 in TinkerPlots 216
Figure 4.38. The heart rate graphs partitioned into fixed-width intervals in TinkerPlots by Group 2 218

Figure 4.39. The argumentation scheme regarding the effect of smoking on resting heart rate 221
Figure 4.40. The partitioned backpack distributions into equal-width intervals ... 225
Figure 4.41. The partitioned backpack 1 distribution of Group 1 (a) and Group 2 (b) in TinkerPlots 228
Figure 4.42. The partitioned backpack 1 distribution of Group 3 in TinkerPlots . 228
Figure 4.43. The graph is partitioned into four groups with equal sizes 230
Figure 4.44. The Backpack 2's data partitioned into equal-width intervals 233
Figure 4.45. The inscription with 30 pounds of the cut-point 234
Figure 4.46. The data partitioned into four equal sizes (a) and the box plots (b) for the backpacks 2 in TinkerPlots 235
Figure 4.47. The argumentation scheme regarding the effect of lockers on the weights carried in the backpacks 238
Figure 4.48. Dot (a) and box (b) plots and sample student responses for Question 1 regarding memorizing words in the last assignment 240
Figure 4.49. The anorexia distributions partitioned by Group 1 in TinkerPlots ... 245
Figure 4.50. The anorexia distributions partitioned by Group 3 in TinkerPlots ... 246
Figure 4.51. The box plots drawn by Group 1 in TinkerPlots for the activity related to anorexia 248
Figure 4.52. The argumentation scheme regarding the effect of the therapies on weight gain in the "anorexia" activity 251
Figure 4.53. The graphs drawn by Group 2 (a) and Group 1(b) in TinkerPlots for the activity related to ambulance 254
Figure 4.54. The data that Group 2 partitioned into equal-width intervals in TinkerPlots for the activity related to ambulance 257
Figure 4.55. The argumentation scheme regarding the response time of the ambulances 259

LIST OF ABBREVIATIONS

ABBREVIATIONS

CSP: Classroom Statistical Practice

HLT: Hypothetical Learning Trajectory

MAD: Mean of Absolute Deviations

PMT: Preservice Mathematics Teacher

SAD: Sum of Absolute Deviations

TAS: Taken-as-shared

CHAPTER 1

INTRODUCTION

Statistical ideas are necessary for posing evidence-based arguments, critically evaluating data-based claims, and making logical decisions in an environment full of quantitative information (Bargagliotti et al., 2020; Franklin et al., 2007; Garfield & Ben-Zvi, 2008). To raise a conscious citizen who can make logical decisions, the interest in statistics education in school has grown. Therefore, statistics has become an essential component of mathematics curricula in many countries (e.g., Australia, Estonia, Singapore, Turkiye, the U.S., and so on). For instance, Common Core Standards for Mathematics aims that students should be able to formulate research questions, collect, organize, and display relevant data, select and use appropriate statistical methods, make inferences and predictions, and understand and apply the basic probability concepts to answer research questions within the scope of Data Analysis and Probability Standard (National Council of Teachers of Mathematics, [NCTM], 2000). Similarly, the middle school mathematics curriculum of Turkiye (Ministry of National Education, [MoNE], 2018a) includes this aim in the content of the data processing learning area. The curriculum summarizes the content as follows:

> In the 5th grade level, students are expected to form research questions that require data collection and to display and interpret the data suitable for these questions with appropriate tables, frequency tables, and bar graphs. At the 6th grade level, they are expected to obtain data on two data groups and organize and analyze these data. The students' use of arithmetic mean and range in comparing and interpreting data belonging to two groups is among the targeted objectives at this level. In the 7th grade, the concepts of pie charts and line graphs are addressed, and students are expected to interpret these graphs. In addition to these, it is expected that the concepts of mean, median, and mode are understood, calculated, and interpreted by students. Furthermore, they are asked to present the data with appropriate representations. In the 8th grade, it is expected that they interpret line and bar graphs containing at most three data groups and display the data related to the research questions with bar graphs,

> pie charts, and line graphs according to their suitability and make appropriate conversions between these representations (MoNE, 2018a, pp. 13-14).

As can be seen from the content, there is an emphasis on generating researchable questions, data collection, organizing relevant data in appropriate tables and graphs, analyzing them, making inferences and predictions to answer the questions, and interpreting results. That is, students should be able to use appropriate strategies to solve a statistical problem and check the meaningfulness of the results achieved by applying these strategies. Garfield (2002, p.1) calls this process statistical reasoning, in which "people reason with statistical ideas and make sense of statistical information."

Statistical reasoning is essential in daily life, research, and many disciplines, such as medicine, politics, social sciences, economics, etc., to solve statistical problems and make critical decisions through inferences and generalizations based on data (Garfield, 2002). It requires interpreting data sets, representations, or statistical summaries (Ben-Zvi & Garfield, 2004). It also includes understanding and explaining why and how statistical processes occur while making inferences based on statistical results. For this, educational researchers aim to develop reasoning skills about data, data representations, statistical measures, uncertainty, samples, and association (Garfield, 2002). All these reasoning skills are undoubtedly essential to developing statistical reasoning. In this context, this study focuses on the development of reasoning about distribution involving these skills because distribution is an overarching concept that includes interrelated concepts of center, variability, shape, density, skewness, relative frequency, probability, proportionality, and causality (Bakker & Gravemeijer, 2004; Pfannkuch & Reading, 2006).

Reasoning about distribution requires extracting information from the data in the scope of exploratory data analysis (Ben-Zvi, 2004; Friel et al., 2001). It includes data analysis by describing and predicting aggregate characteristics of data sets instead of a collection of individual data values (Bakker & Gravemeijer, 2004; Cobb, 1999). In other words, reasoning about distribution can be described as a method for comprehending statistical information that characterizes the entirety of the data and

interpreting its meaning. Based on the definition of distribution as a conceptual entity (Bakker & Gravemeijer, 2004), reasoning about distribution refers to finding an overall pattern and deviations from the pattern in the data (Reading & Reid, 2006). Accordingly, it includes reasoning about center, spread, and shape in an interrelated way. This way of thinking, which corresponds to a global understanding of the data (Ben-Zvi & Arcavi, 2001), requires the combination of the center, spread, density, and skewness aspects of distribution in the bilateral orientation between data and distribution (Bakker & Gravemeijer, 2004). In other words, distribution serves as an aggregator concept that incorporates the key statistical subjects, such as center, spread, and shape, to recognize the general characteristics of the data set. In this context, a person who interprets distribution reasonably can get information from the data by integrating the aspects of the distribution.

Although the notion of distribution is a central idea in statistics education, many studies indicate that many difficulties are experienced in understanding the characteristics of distribution. Like most students (Aridor & Ben-Zvi, 2018; Bakker & Hoffman, 2005; Konold et al., 2015; Wild, 2006), some preservice teachers and teachers could not see the data values in distribution as an aggregate (Frischemeier & Biehler, 2018; Garfield & Ben-Zvi, 2008; Gökçe, 2019; Madden, 2008). In other words, they exhibited a local understanding of the distribution rather than a global understanding (Ben-Zvi & Arcavi, 2001). When interpreting quantities varying in different ranges, they focused only on specific values, such as maximum or minimum values, instead of the distribution of the data. They had problems defining and interpreting the aspects of distribution, such as center and spread, as representative of the whole data based on a context (delMas & Liu, 2005; Dubreil-Frémont et al., 2014; Pfannkuch, 2007; Turegun, 2011; Watson, 2006; Watson & Callingham, 2003). For example, they calculated statistical measures that summarize the data, such as the mean and standard deviation, without developing a conceptual understanding (Ben-Zvi & Garfield, 2004). In addition, most preservice mathematics teachers (PMTs) performed poorly in statistical justification and reasoning about distribution when interpreting statistical findings in a context (Frischemeier &

Biehler, 2018; Karatoprak, 2011; Kurt, 2016; Mercimek, 2013; Sorto, 2004; Yılmaz, 2019).

The previous studies cited above indicate that there are challenges in reasoning about distributions. However, to teach statistics effectively, mathematics teachers teaching statistics in schools need a well-developed understanding of distribution (Reading & Canada, 2011). Teachers require profound and sophisticated reasoning methods on distribution to guide and help students learn statistics (Mickelson & Heaton, 2004). Since it is through the guidance of teachers that students can develop this skill, teachers need to have developed reasoning about distribution. Therefore, it is important that teachers have this competence before they start their professional life. In this context, there is a need to develop PMTs' statistical reasoning about distribution.

When investigating developmental studies on improving reasoning about distribution, it was observed that these studies are predominantly classroom teaching experiments conducted with students (Bakker, 2004; Bakker & Gravemeijer, 2004; Cobb, 1999). These studies emphasize that reasoning about distribution can be developed by interpreting the distribution's shape in the classroom's social environment. For this, in the studies, the activities in which students had a history with data were designed to promote classroom discussion and collaboration. In the classroom setting where students experience data analysis within the framework of Realistic Mathematics Education (RME), their classroom practices are monitored to rely on their taken-as-shared reasoning patterns. Based on the emergent perspective which focuses on the collective learning of the class, classroom mathematical practices (CSPs) serve as indicators of how students' reasoning develops. For example, in Cobb's study (1999), the students' reasoning about distribution progressed from *exploring qualitative characteristics of collections of data points* to *exploring qualitative characteristics of distributions*.

The classroom teaching experiments conducted within the framework of RME (Bakker, 2004; Bakker & Gravemeijer, 2004; Cobb, 1999) underline the structuring

of distribution in different ways, including multiplicative structuring, to develop reasoning about distribution. In this context, the studies emphasize the use of technological computer tools that allow different ordering, partitioning, and organizing of the data. In these studies, the *minitool* developed specifically for reasoning about the distribution enabled the development of reasoning with different configurations, such as partitioning the data into groups of a specified size (e.g., 10 data points in each group), partitioning the data into groups with a specified interval width, partitioning the data into two equal groups, and partitioning the data into four equal groups. On the other hand, studies using TinkerPlots with similar and more features have found that dynamic statistical software supports statistical reasoning (Fielding-Wells, 2018; Frischemeier & Biehler, 2018; Hammerman & Terc, 2004; Koparan, 2019; Makar et al., 2011; Meletiou-Mavrotheris et al., 2009; Watson et al., 2020).

In light of these studies on the development of reasoning about distribution, we offer a classroom teaching experiment based on RME using TinkerPlots. In this regard, the current study aims to investigate the development of PMTs' reasoning about distribution as they participated in the classroom teaching experiment based on RME and enriched by dynamic statistics software. Concerning the purpose, the study seeks to answer the following research questions:

1.1 Research Questions

1. How did the PMTs' reasoning about distribution develop throughout the classroom teaching experiment based on RME and enriched by dynamic statistics software?
2. What classroom statistical practices emerged throughout the teaching experiment based on RME and enriched by dynamic statistics software?

1.2 Significance of the Study

The study aims to investigate the development of PMTs' reasoning about distribution as they participated in the classroom teaching experiment based on RME and enriched by dynamic statistics software. Regarding the purpose, this study holds many significant aspects. I highlight below these aspects that are important for the improvement of statistics education.

The focus on the distribution concept in the study contributes to its significance. According to Gravemeijer and Cobb (2006), the notion of distribution should be a central idea in statistics education as it encompasses the interrelationship of statistical topics such as mean, median, and mode, and statistical displays rather than their separate collection as the statistical content. Since distribution is an overarching concept that includes big statistical ideas (Bakker & Gravemeijer, 2004; Cobb, 1999; Pfannkuch & Reading, 2006), it allows for thinking and reasoning in the process of doing statistics holistically. For example, understanding how data is distributed provides a sense of the variability with complex structures among measures, as distribution is a lens through which to look at the variation in the data (Wild, 2006). Considering that statistics has its own territory as a discipline due to the *omnipresence of variability* (Cobb & Moore, 1997), the concept of distribution appears as an indispensable concept in statistics education.

The distribution concept is included in mathematics curriculum standards and guidelines worldwide, such as Common Core State Standards for Mathematics (CCSSM, 2010), Ministry of Education Singapore (2012), etc. As Gravemeijer and Cobb (2006) emphasize, the notion of distribution can eliminate the difficulties of learning statistics mentioned above. In addition, since multiplicative reasoning is done based on relative frequencies while reasoning about distribution (Cobb, 1999), the concept of distribution is also related to different mathematical topics, such as multiplication, division, fractions, ratios, and proportion in the multiplicative conceptual field (Ezaki et al., 2023). Accordingly, the development of reasoning about distribution can contribute to the progress in different subject matters of

mathematics. Therefore, it should be included in the content of statistics education. However, the notion of distribution is not explicitly introduced in the middle school mathematics curriculum in Turkiye (MoNE, 2018a). In this context, the investigation of the position of the distribution in the school curriculum can open further research areas. For example, curriculum design research can be conducted about distribution (Clements, 2007; Van den Akker, 2010). In this sense, this study may lead to the inclusion of the concept of distribution as a key concept, such as the mean, median, mode, or different graphical representations in the curriculum. In this respect, the study's findings can provide politicians, mathematicians, mathematics educators, and all stakeholders of the education system with different perspectives regarding statistics education.

There are promising studies that contribute to making sense of the data thanks to classroom discussions based on tasks (Bakker & Gravemeijer, 2004; Cobb, 1999; Fielding-Wells, 2018; Gravemeijer & Cobb, 2006; Selmer et al., 2012; Watson et al., 2020) using technology (Fielding-Wells, 2018; Frischemeier & Biehler, 2018; Hammerman & Terc 2004; Koparan, 2019; Makar et al., 2011; Meletiou-Mavrotheris et al., 2009; Watson et al., 2020) guided by the teacher or researchers. By addressing sociocultural learning and the use of technology, the study offers a classroom teaching experiment in which collective learning of the class focuses. In this sense, the current study presents valuable insight into promoting PMTs' reasoning about distribution as they participated in the classroom teaching experiment based on RME and enriched by dynamic statistics software. In this study, the PMTs can develop their reasoning about distribution by studying compelling activities that allow the use of technology in the social classroom environment. Regarding this process, the study presents a route that mathematics educators can use to develop PMTs' statistical reasoning on the distribution concept in teacher education. In this sense, the study offers some considerable insights into teacher education in developing PMTs' statistical reasoning about distribution.

In addition to the characteristics of the activities mentioned above, another aspect to be considered is how teachers orient these activities. Teachers tend to teach the way

that they are taught (Chicoine, 2004; Hall et al., 2006), "the way they prefer to be taught," or "the way they think students learn best" (Cox, 2014, p. 39). They are expected to lead their students to develop an understanding of the key statistical ideas and the relationships between them by finding ways to go beyond individual concepts and skills within the classrooms (Garfield & Ben-Zvi, 2004). Since statistical reasoning is rarely taught effectively (Sedlmeier, 1999), developing teachers' understanding of distribution as an overarching concept is important. In this context, the study would be beneficial for future statistics teachers to graduate from teacher education programs with a robust understanding of distribution before entering professional life as they can follow the way in this study to teach the concept of distribution to their students.

There are few attempts to investigate and improve the knowledge and reasoning regarding the key statistical concepts of PMTs (Ben-Zvi & Garfield, 2004; Groth & Meletiou-Mavrotheris, 2018). Although limited developmental studies have been conducted, those that exist focus on assessing the participants' statistical knowledge and reasoning while comparing distributions rather than on how they developed their reasoning on distribution (Ben-Zvi & Garfield, 2004; Frischemeier & Biehler, 2018; Madden, 2008). The subjects being studied in the studies focusing on the development of reasoning about distribution are students (Bakker, 2004; Bakker & Gravemeijer, 2004; Cobb, 1999). This study is an adaptation of these studies with different participants, PMTs, in the Turkiye context. Since teachers may be trained in statistics, in the same way, students are, this study not only ensures the transferability of previous studies (Gravemeijer & Cobb, 2006; Lincoln & Guba, 1985) but also makes original contributions to teacher education programs. Compared to studies on the development of students' reasoning, it is necessary to examine how the distribution concept is given to PMTs. Thereupon, the current study fills a gap in the literature by examining the development of PMTs' reasoning about distribution. Considering all of these, the study aimed to investigate the development of PMTs' reasoning about distribution as they participated in the classroom teaching experiment based on RME and enriched by dynamic statistics software.

1.3 Definition of Important Terms

Statistical Reasoning: Statistical reasoning is one of the three domains necessary for learning statistics (Ben-Zvi & Garfield, 2004). Although it can be used interchangeably with the other domains, that is, statistical thinking and statistical literacy, in this study, statistical reasoning refers to "the way people reason with statistical ideas and make sense of statistical information" (Ben-Zvi & Garfield, 2004, p. 7). It includes interpreting data sets, representations, or summaries and connecting concepts and ideas such as the center, spread, chance, and data to understand and explain statistical processes and results comprehensively.

Distribution: Researchers define distribution in different ways. According to Leavy (2006) and Moore and McCabe (2003), a distribution is the arrangement of the values of a variable and their frequencies in data representations. In this study, since the focus is on the conceptual meaning of the distribution rather than its representation, the distribution is addressed as a conceptual entity consisting of an overall pattern and deviations from the pattern (Reading & Reid, 2006). It refers to an aggregate structure with specific characteristics (i.e., center spread, density, and skewness) rather than the collection of individual values (Bakker & Gravemeijer, 2004; Cobb, 1999).

Classroom Teaching Experiment: The classroom teaching experiment employed in this study is design-based research aiming to improve education by designing, developing, and assessing the influence of an educational intervention or innovation (Bannan, 2010; Plomp, 2010). It is conducted in the social learning ecology of the classroom where students participate in learning, and one of the researchers in the team has the role of the teacher leading the instruction in the study (Cobb, 2000). In the classroom teaching experiment in the context of this study, the instructor is my supervisor. The educational intervention in the study is the implementation of instructional activities designed according to RME and the use of dynamic statistical software to develop PMTs' reasoning about distribution.

Realistic Mathematics Education (RME): RME is a theoretical approach specific to mathematics education, developed by Hans Freudenthal. It is based on understanding and dealing with problem situations that are experientially real to students. In this study, the activities that the PMTs work on in the classroom are closely related to contexts from real situations, which allows them to engage in statistics as a human activity (Freudenthal, 1991).

Dynamic Statistical Software: It is a statistics program that enables interactive analysis and visualization of data. The adjective dynamic reflects the characteristics of the program that allow for ordering, partitioning, and organizing the data set and reconstructing statistical representations in different ways. In this study, we specifically use TinkerPlots, which is a computer tool designed for data exploration (Konold & Miller, 2011). In other words, the dynamic statistics software in this study refers to TinkerPlots.

Classroom Statistical Practices (CSPs): CSPs refer to "the taken-as-shared ways of reasoning, arguing, and symbolizing established while discussing particular *statistical* ideas" (Cobb et al., 2001, p. 126). In other words, they are the established patterns of behaving, communicating, and symbolizing statistical concepts at a particular time without requiring any justification (Cobb & Yackel, 1996; Gravemeijer & Cobb, 2006). In this study, CSPs are indicators of "collective learning of the classroom community" that participates in classroom activities (Cobb et al., 2001, p. 113).

Hypothetical Learning Trajectory (HLT): The HLT is a route of learning that might progress in the way that is anticipated by the teacher. It includes the learning goals that students are expected to learn, the learning activities they might engage in, and conjectures of the learning process - how students' thinking and understanding about what to learn can evolve in this process (Simon, 2020). The HLT in this study is adapted from the instructional sequence in previous research focusing on improving 7th-grade students' reasoning about distribution (Cobb, 1999).

CHAPTER 2

LITERATURE REVIEW

The purpose of the study is to investigate the development of PMTs' reasoning about distribution as they participated in the classroom teaching experiment enriched by dynamic statistics software. Concerning the purpose, this study clarifies the PMTs' statistical reasoning before and after the classroom teaching experiment and documents their collective development in reasoning about distribution by analyzing classroom statistical practice emerging throughout the teaching experiment. Corresponding to the scope of the study, this chapter comprises three sections: statistical reasoning, the distribution concept, and the theoretical framework of the study. The first section puts forward the nature of statistical reasoning with its definition and types. Since the study addresses the development of statistical reasoning about distribution, the next section presents a brief review of what distribution is, aspects of distribution, and related studies investigating the development of reasoning about distribution. It also includes literature on proportional reasoning and the use of technology that is considered necessary for the development of reasoning about distribution. The focal point of the study is the development of the PMTs' reasoning about distribution as collective learning of the classroom community through their discoveries in the social and interactive classroom environment. In this regard, the last section introduces Realistic Mathematics Education (RME) and the Emergent Perspective, the theoretical frameworks on which the study is based.

2.1 Statistical Reasoning

The main goal of statistics education is to lead students to reason and think statistically beyond skills, procedures, and computations (Ben-Zvi & Garfield,

2004). Specifically, merely calculating measures such as the mean, range, and standard deviation, along with some operations, is insufficient for learning statistics. This approach involves only applying mathematical procedures to statistical problems. On the contrary, it is necessary to understand statistical information, different representations, and language, to explain, interpret, and evaluate the statistical findings, and to reason and grasp the way behind statistical investigation (Ben-Zvi & Garfield, 2004). Accordingly, three cognitive domains need to be achieved to ensure statistical learning. These are statistical reasoning, statistical thinking, and statistical literacy. According to Ben-Zvi and Garfield (2004, p. 7),

> ***Statistical literacy*** includes basic and important skills that may be used in understanding statistical information or research results. These skills include being able to organize data, construct and display tables, and work with different representations of data. Statistical literacy also includes an understanding of concepts, vocabulary, and symbols, and includes an understanding of probability as a measure of uncertainty.
>
> ***Statistical reasoning*** may be defined as the way people reason with statistical ideas and make sense of statistical information. This involves making interpretations based on sets of data, representations of data, or statistical summaries of data. Statistical reasoning may involve connecting one concept to another (e.g., center and spread), or it may combine ideas about data and chance. Reasoning means understanding and being able to explain statistical processes and being able to fully interpret statistical results.
>
> ***Statistical thinking*** involves an understanding of why and how statistical investigations are conducted and the "big ideas" that underlie statistical investigations. These ideas include the omnipresent nature of variation and when and how to use appropriate methods of data analysis such as numerical summaries and visual displays of data. Statistical thinking involves an understanding of the nature of sampling, how we make inferences from samples to populations, and why designed experiments are needed in order to establish causation. It includes an understanding of how models are used to simulate random phenomena, how data are produced to estimate probabilities, and how, when, and why existing inferential tools can be used to aid an investigative process. Statistical thinking also includes being able to understand and utilize the context of a problem in forming investigations and drawing conclusions, and recognizing and understanding the entire process (from question posing to data collection to choosing analyses to testing assumptions, etc.). Finally, statistical thinkers are able to critique and evaluate results of a problem solved or a statistical study.

Statistical literacy concerns understanding and interpreting statistical information in daily life (Chance, 2002). Statistical reasoning also refers to making inferences from data by understanding statistical information (Garfield & Chance, 2000). Similarly, statistical thinking is defined as an attempt to clarify "the relation of quantitative data to a real-world problem, often in the presence of variability and uncertainty" (Mallows, 1998, p. 3). As can be seen from the definitions based on their use in doing statistics, the distinctions among them are not clear-cut, and they are likely to be used interchangeably (delMas, 2002). However, delMas (2002) separates them by discussing their relationship and the points where they diverge.

delMas (2002) reveals two different perspectives, focusing on the relationship between statistical literacy, reasoning, and thinking. In the first perspective, he demonstrates the independence of literacy, thinking, and reasoning with some overlaps based on the definitions of Chance (2002), Garfield (1998), and Rumsey (2002). They describe statistical literacy as knowing how to use data to make decisions, statistical thinking as knowing how to apply data in contexts, and statistical reasoning as making sense of statistical ideas and information. From another perspective, since literacy is considered to be the purpose of statistics education, it includes both thinking and reasoning.

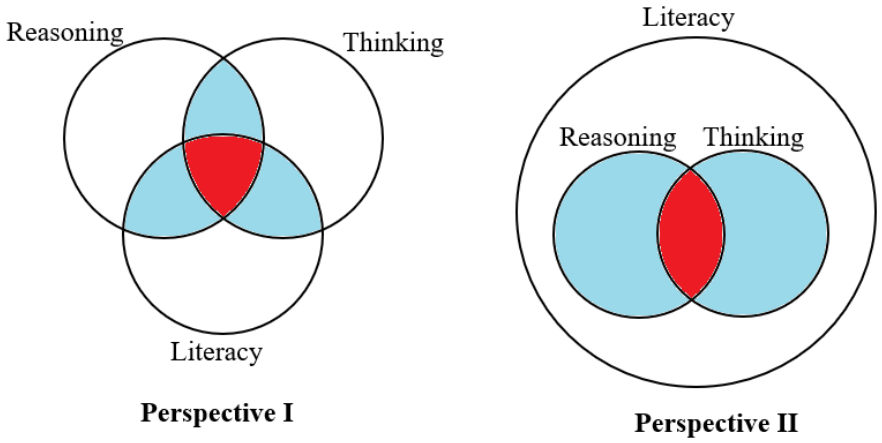


Figure 2.1. The perspectives for statistical reasoning, thinking, and literacy

In both perspectives in Figure 2.1, there are common areas, colored parts, where thinking, reasoning, and literacy overlap. It means that the same content can direct each of them. According to delMas (2002), the distinguishing factor is the nature of the task based on statistical content. The word in the task determines which domain the focus is. The task is oriented toward statistical literacy if it requires the actions of *identifying, describing, rephrasing, translating*, or *interpreting* the results in a statistical procedure. When it asks to *explain* the *why* and *how* of the statistical results reached in the process, it aims to develop or measure statistical reasoning. Statistical thinking is used while *applying* the understanding to real-life problems, *critiquing* and *evaluating* the design and findings, and *generalizing* the results to new situations. For example, while defining the mean of a data set reflects statistical literacy, applying the arithmetic mean formula to solve a statistical problem is within the scope of statistical thinking. When one explains how the mean acts as a balancing point and why the mean is over-affected by outliers, they reason statistically (delMas, 2002). Although delMas (2002) distinguishes domains with descriptive verbs, an action defined for one can also apply to another. For example, *interpreting* is included in statistical literacy. However, according to BenZvi and Garfield (2004), interpreting statistical results is the cornerstone of statistical reasoning. In this study, the main idea is to obtain statistical results reasonably and make sense of these results. Therefore, the study is addressed in the context of statistical reasoning.

Statistical reasoning is "the way people reason with statistical ideas and make sense of statistical information" (Ben-Zvi & Garfield, 2004, p. 7). It requires a deeper and more comprehensive thinking about statistical ideas and concepts. It includes representations of data, reducing data through concepts of central tendency and variability, making judgments and inferences based on sets of data, and interpreting statistical results (Garfield, 1998). In this context, statistical reasoning is necessary to conceptually understand the ideas of center, variability, association, sampling, stochastic/randomness, chance, and inference. Garfield (2002) introduces the following types of statistical reasoning based on these statistical ideas.

Reasoning about data: Recognizing the type of data, e.g., qualitative, quantitative, discrete, or continuous, and knowing the effect of the data type on the choice of data representation (table and graphs) and statistical measures.

Reasoning about representations of data: Understanding the ways of reading, interpreting data representations, transforming graphs into better representations of data, and seeing beyond random artifacts of distribution to realize shape, center, and variability.

Reasoning about statistical measures: Comprehending the meaning of shape, center, and variability of the data, knowing which statistical measures (e.g., mean, median, mode, range, percentile, and standard deviation) are the most appropriate to represent the data, understanding the importance of big data- the law of large sample- in prediction, and recognizing the necessity of using center and variability measures together for a good summary when comparing data sets.

Reasoning about uncertainty: Understanding and using randomness, chance, and likelihood for uncertain events, knowing the differences among all outcomes of an event, and determining the likelihood of different events with different methods.

Reasoning about samples: Knowing how to make inferences about the population with the help of a well-chosen representative sample, understanding the way of sampling to choose representative and unrepresentative samples, and noticing skeptical inferences for the population, which are due to small and biased samples.

Reasoning about the association: Knowing to judge and interpret the bivariate relationship of two variables represented in a two-way table or a scatterplot and comprehending that a strong correlation does not mean that one variable causes the other.

The types of reasoning described above by Garfield differ in terms of the data set of interest and the meaning of the statistical process. Reasoning about data, representations of data, and statistical measures primarily focus on describing the dataset and fall within the scope of descriptive statistics (Gravetter & Wallnau,

2015). While descriptive statistics involves summarizing, organizing, and reducing data without discriminating between sample and population (Cobb & Moore, 1997), inferential statistics aims to make generalizations about the corresponding population based on the work done with samples. In this manner, inferential statistics encompasses reasoning about uncertainty, samples, and association. On the other hand, according to Curcio (1987), reasoning about data, representation of data, and statistical measures are in *reading the data*, reasoning about association is in *reading between the data*, and reasoning about uncertainty and samples is in *reading beyond the data*. Friel et al. (2001) update this triad as *extracting information from the data, finding relationships in the data, and moving beyond the data*, respectively. Ben-Zvi (2004) stresses exploratory data analysis, which puts aside formal statistical inference but focuses on informal statistical inference to make sense of data by organizing, describing, representing, and analyzing the data. In the context of the exploratory data analysis, the current study focuses on the reasoning about distribution. Since the primary concern of this study is to examine the development of reasoning about distribution, the study deals with extracting information from the data in the framework of Friel et al. (2001), including the first three reasoning types of Garfield (2002). Depending on the concern of the study, since the ways of reasoning about distribution would be better understood with the nature of the distribution concept, the following section presents a brief review of distribution as the basic concept of the study and how to reason about distribution.

2.2 The Concept of Distribution

Distribution is considered a fundamental concept for statistical thinking since it provides the conceptualization of variation in the data (Pfannkuch & Wild, 2004; Pfannkuch & Reading, 2006; Wild, 2006). In addition, it is an overarching concept that includes interrelated concepts of center, variability, shape, density, skewness, relative frequency, probability, proportionality, and causality (Bakker & Gravemeijer, 2004; Pfannkuch & Reading, 2006). Based on these concepts, the

distribution is the basis for sampling distribution, statistical confidence, statistical inference, estimation, probability, and establishing causation with designed experiments. Leavy (2006) defines distribution as “the arrangement of values of a variable along a scale of measurement resulting in a representation of the observed or theoretical frequency of an event” (p. 90). This definition requires a scaled representation to be a distribution. However, according to Moore and McCabe's (2003) definition, a distribution represents variable values and their frequencies. In the light of the later definition, unscaled representations such as stem and leaf plots or tables also indicate distribution. Based on the meaning of the concept rather than the representation of distribution, Bakker and Gravemeijer (2004) define distribution as a conceptual entity. Distribution substitutes for an aggregate structure with specific characteristics rather than the collection of individual values (Bakker & Gravemeijer, 2004; Cobb, 1999). Accordingly, distribution is a conceptual entity consisting of an overall pattern and deviations from the pattern (Reading & Reid, 2006).

Distinguishing the data and distribution provides insight into the meaning of distribution. According to Bakker and Gravemeijer (2004), the relationship between data that are individual values and distribution are four aspects. These are center, spread, density, and skewness. Although Reading and Reid (2006) attach outliers to these aspects as key elements in their framework, this study considers four aspects to examine the characteristics of distribution because center, spread, density, and skewness already reflect outliers. Figure 2.2 shows the relationship between data and distribution.

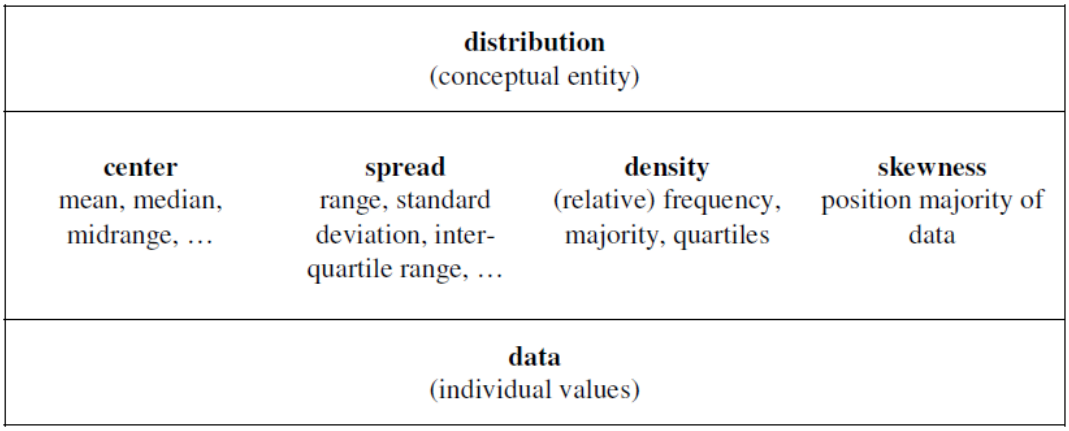


Figure 2.2. The relationship between data and distribution (Bakker & Gravemeijer, 2004, p. 148)

The relationship can be approached in two ways: upward and downward. From an upward perspective, it is concerned with individual data values. On the other hand, from the downward perspective, distribution is perceived as an organized structure through the conceptualization of distribution. Bakker and Gravemeijer (2004) argue that experts in statistics easily combine these two perspectives. The bilateral orientation between the perspectives can be accomplished by understanding the meaning of center, spread, density, and skewness. For this, it would be useful to explain what aspects mean. These aspects are explained in the next sections.

2.2.1 Center

As can be seen from Figure 2.2, the first aspect of distribution is the center. In a data set, values are directed to the center. For the central tendency of data, the term average is often used in the literature. It summarizes the data to a single value but with different meanings. For example, the typical value that can be generalized for everyone, the most common value, the middle value, the value that is equidistant to all values in the data, etc., can be the average of the data. Konold and Pollatsek (2002) categorize the average meanings as data reduction, fair share, typical value, and signal in noise because of the variability in the data. The ultimate goal of all

average meanings is to understand the center of the data or central tendency, albeit in different meanings. The measures of central tendency are indicators of the center of the data. Of these measures, the mean is the most commonly used one compared to the median, which is the value in the middle of the ordered form of a data set, and the mode, which is the value with the most frequency in data. When defining the mean, it is specified how it is calculated procedurally (Van de Walle et al., 2019), as in the definition of "A measure of center in a set of numerical data, computed by adding the values in a list and then dividing by the number of values in the list" (CCSSM, 2010, p. 86). Nevertheless, the underlying meanings of this operation are important to understand the mean completely. For example, the mean balances the data values above and below it (Hardiman et al., 1984; Uccellini, 1996; Van de Walle et al., 2019) and shares the data evenly by compensating what is less with what is more (Hardiman et al., 1984; Leavy & O'Loughlin, 2006; Uccellini, 1996; Van de Walle et al., 2019).

As Cobb and Moore (1997) state, the conclusions obtained with statistical analysis depend crucially on how to produce data. The data analyzed in statistics learning environments generally consist of variables collected from students such as height, weight, sleep time, etc., as in the task of "Lunch," designed by Selmer et al. (2012) to develop statistical reasoning. Students mostly encounter the mean in the classroom as average or center unless there are unexpected situations. Therefore, they think that the mean is average without considering the mode and median (Bakker, 2004; Friel et al., 2006; Garfield & Gal, 1999). In other words, they tend to find arithmetic mean by carrying out the procedure (Mokros & Russel, 1995). In addition to the algorithm approach, Mokros and Russel (1995) classified younger students' reasoning on average as mode, reasonable, midpoint, and mathematical point of balance. The students with the mode approach appreciate average as the most common value that occurs more often than the others. In the reasonable approach, they consider that average is a tool for making sense of data. For average as a midpoint, they searched the middle of the data. The students with the mathematical point of balance approach try to find a balance point that represents data by considering all the values in the

data. According to Mokros and Russel (1995), while the reasonable, midpoint and mathematical point of balance approaches represent data, others are not representative. Although these appear to be difficulties (Batanero et al., 1994), each of them may be a correct interpretation according to the conditions of a statistical problem. For instance, for a data set consisting of qualitative variables, the mode value that occurs more than the others makes sense. On the other hand, for quantitative data sets with outliers, the midpoint of the data is more reasonable.

2.2.1.1 Studies on the Center Concept

Research on understanding the concept of average concurred that students focused too much on calculating measures of center without developing conceptual understanding (Ben-Zvi & Garfield, 2004; Gresalfi, 2015; Leavy & O'Loughlin, 2006; Makar, 2014; Konold & Pollatsek, 2002; Saidi & Siew, 2019). For example, Yolcu and Haser (2013) exhibited that 1074 8th-grade students tended to implement the "add them up and divide" algorithm as average. Most could not interpret the average in the given context by recognizing outliers. In another study (Saidi & Siew, 2019) assessing 148 10th-grade students' understanding of central tendency measures, it was found that most high school students did not know mean, median, and mode as average types and could not choose the best representative one for the given data. Similarly, some PMTs and even some teachers teaching statistics apply algorithms without conceptual understanding and interpreting the average in the context (Engledowl & Tarr, 2020; Groth & Bergner, 2006; Jacobbe, 2012). For example, when comparing two distributions, secondary teachers (Makar & Confrey, 2004) and PMTs (Huey, 2011) draw on the mean to make decisions. They rely on computing arithmetic mean disregarding the median and mode (Landtblom & Sumpter, 2021; Leavy & O'Loughlin, 2006) for measures of center despite their various meanings.

Although the mean is an important concept in statistics with its function of representing the data, reducing a data set to a single value causes a lot of information

about the data to be lost. Konold and Pollatsek (2002) criticize the situation as the data reduction interpretation of average. The average removes the uncertainty caused by many values in the dataset. However, since the data does not consist of a single value, other aspects also provide information about the data. In light of this, Konold and Pollatsek (2002) prefer to use the central tendency phrase for the average. Conceptually, central tendency has two components: stable and variable. While the stable component is summarized by measures, such as the mean, median, and mode, the variability component focuses on the spread of data values around the center (Konold & Pollatsek, 2002). In that case, the center of the data can make sense entirely depending on the spread. The meaning of the average already takes into account the spreadoutness in the data. Spread is explained in the following section.

2.2.2 Spread

Spread is the second aspect of the distribution shown in Figure 2.2. It is related to the variability in a data set. Since statistics is full of uncertainties derived from the omnipresence of variability (Cobb & Moore, 1997), it is important to conceptualize variability to make sense of how data is distributed. The concept of variability refers to differences among the values that vary from one individual to another in the distribution and deals with how far the values are spread. The following quotation from Gravetter and Wallnau (2015, pp. 101-102) demonstrates the definition of variability:

> Variability provides a quantitative measure of the differences between scores in a distribution and describes the degree to which the scores are spread out or clustered together (…) variability is defined in terms of distance. It tells how much distance to expect between one score another; or how much distance to expect between an individual score and the mean.

It is realized from the definition that variability is related to the position of the values in the distribution in two ways. The first is the distance between values, and the other is between a value and the center. Based on their definitions, the range, which is the distance spread from the minimum value to the maximum value in a distribution and

interquartile range (IQR) that is the distance between the 25th percentile (Q1) and the 75th percentile (Q3) of all values in the distribution involve in the first category. In the other category, standard deviation and variance are related to the distances between data values and the center. While the standard deviation is the square root of the squared distance between each value in the distribution and the mean, the variance is the square of the standard deviation ($\sigma^2 = [\sum_{i=1}^{n}(x_i\text{-}\bar{x})^2]/n$). Based on the absolute deviations as outlined in the formulas, the mean of absolute deviations (MAD) and the sum of absolute deviations (SAD) are also used to measure the degree of variability (Kader & Mamer, 2008; Van de Walle et al., 2019). Since MAD is the mean of the total distance from the mean, it is obtained by dividing the SAD by the number of data.

There are many calculations, such as the one above, to measure variability. Abstract mathematical calculations only measure the degree of the variability. What is essential in descriptive statistics is to discover the story hidden by variability in distribution (Shaughnessy, 2007). Since variability is often discussed in the context of sampling (Aridor & Ben-Zvi, 2018; Chance et al., 1999; delMas et al., 2007) or comparing distributions (Ben-Zvi, 2004; Biehler et al., 2018; Shin, 2019), there is need to understand the variability inherent in statistical inquiries (Garfield & Gal, 1999) for reasoning about distribution. According to Wild and Pfannkuch (1999) and Reading and Shaughnessy (2004), variability can be understood by noticing and acknowledging, measuring and modeling, explaining and dealing with, developing investigative strategies about variation, describing, and representing.

2.2.2.1 Studies on the Variability Concept

Research focusing specifically on variability and measures reveals that students at different levels have difficulty understanding variability-related concepts (Çatman Aksoy & Işıksal Bostan, 2020; Shaughnessy, 1997; Reading & Shaughnessy, 2004). For example, delMas and Lui (2005) investigated the development of the standard deviation concept of twelve university students. The students experienced how the

data size, mean, and range affected standard deviation through a computer program. Unfortunately, some of them could not realize that the standard deviation would be smaller for the data with a small range while other factors were the same. The students in high school (Rosidah et al., 2018) and territory level (delMas & Liu, 2005; Dubreil-Frémont et al., 2014) could calculate the standard deviation of a distribution, but they had a problem defining and evaluating it (Watson, 2006; Watson & Callingham, 2003). For example, the undergraduate students perceived the standard deviation as the range (Dubreil-Frémont et al., 2014).

In another study (Turegun, 2011), it was confirmed that college students had very limited pre-existing knowledge of measures of spread. None of the 33 students could choose the appropriate measures of spread and center for a skewed distribution. There were conceptual errors in definitions of range, IQR, and standard deviation. Exemplarily, the description of a student for measures of spread is as follows:

> Range is the difference between maximum and minimum. Interquartile range between the upper quartile in other word interquartile range is the whole. IQR is the middle of data. Standard deviation observation from mean.

Although Turegun (2011) coded this student's definitions at the uni-structural level in terms of associating the measures, there are ambiguities in the definitions. The student specified the spread with a single value. However, the mean of IQR and standard deviation are not clear. The student seems to have reduced the IQR covering the middle 50% of the data to the median. Although s/he attempted to relate the standard deviation to the mean, s/he failed to define it.

Limited studies on teachers who teach statistics (Groth & Meletiou-Mavrotheris, 2018; Huey et al., 2018; Shaughnessy, 2001) have shown that teachers and preservice teachers have similar thinking and understanding of variability with students (Makar & Confrey, 2005; Peters, 2014). For example, secondary mathematics teachers perceived the standard deviation as a measure of sample homogeneity (Silva & Coutinho, 2008). Although they started to elaborate on the standard deviation around the mean, their explanations about what it meant remained ambiguous. That is, they could not construct a robust understanding of variation and

oriented toward measures rather than interpreting the spread in the data (Peters, 2014; Silva & Coutinho, 2008).

In the conceptual analysis of variation in the data (Reading & Shaughnessy, 2004), primary and secondary school students defined the variation based on middle values, extreme values, deviations from any anchor, or deviations from the central anchor. Although they were not asked about the sources, they mentioned causes when explaining variation. They stated the causes of variation as extraneous variables in the sampling process (e.g., the location of the sample, the way of sampling, the characteristics of the person who selected the sample, etc.), the frequencies or proportions of data values, and likelihood based on the proportions. Similarly, the preservice teachers also expressed variability with different terminology. In the variation talk of preservice teachers, which Makar and Confrey (2005) categorized as standard and nonstandard, the preservice teachers used the range and standard deviation for the measures of variability. In addition to expressing variation as a "spread, low-middle-high, modal clump, and distribution chunks," they described spread in nonstandard language- "clustered, clumped, grouped, bunched, gathered, spread out, evenly distributed, scattered, dispersed" (Makar & Confrey, 2005, pp.47-48). The use of nonstandard terminology is essential for developing an understanding of variability. The language also shows that variability is interpreted in terms of the spatial aspects of distribution (Makar & Confrey, 2005). The aspects of density and skewness are included in the next sections.

2.2.3 Density

The third aspect of the distribution in Figure 2.2 is density. Bakker and Gravemeijer (2004) express density in terms of (relative) frequency, majority, and quartiles. Based on the number and position, density can be defined as an indicator of the quantity of data values in a space. Accordingly, it depends on two dimensions: the size of the space and the number of data in the space. In that case, more data values in a narrower space mean more density. Even if the space is narrow, the density is

low when there is no or very little data in this space. Comparing the components, density is often directly related to the number of data and inversely related to the size of the space, provided that one dimension remains constant. However, this relationship depends on how data values are distributed in the space. For example, if 20 data values are distributed evenly over ten units, the density is the same in each partial space as in Figure 2.3a because the number of data per unit space is equal. When these data values are clustered in a unit space, density increases in favor of frequency. In Figure 2.3b, 50% of the data whose values are two is concentrated on two.

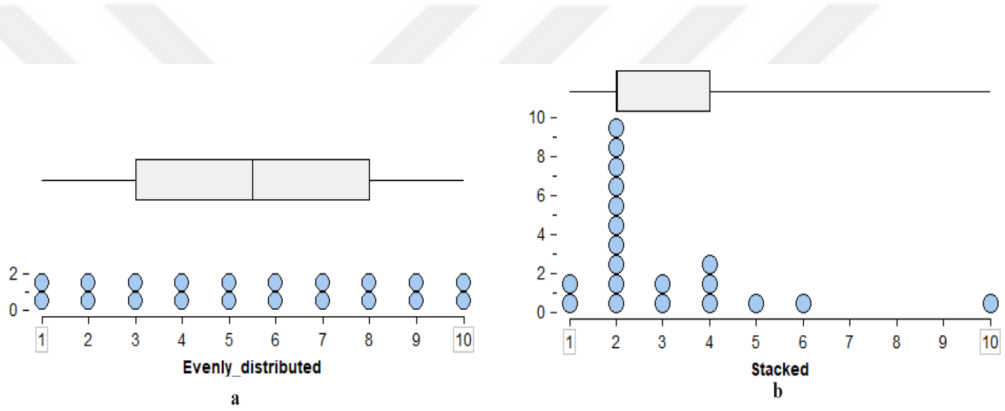


Figure 2.3. The hypothetical data in dot and box plots

In terms of density, it is seen that the quartiles in the box plot are evenly distributed in Figure 2.3a. However, in Figure 2.3b, Q1 and Q2 (median) are the same value, and the first part of the box is stuck on two. In the box plot, the narrower quartile is denser because each quartile contains an equal number of data values. On the other hand, the whisker of the box plot in Figure 2.3b is extended to the right due to a data value (10). In other words, it is skewed because the data value shows a trend opposite to the majority's position. It would appear as skewed to the right in the probability distribution (See Figure 2.4c). In this regard, the skewness and density of a distribution can be interpreted concerning each other.

2.2.4 Skewness

The last aspect of the distribution is skewness (See Figure 2.2). Skewness is a measure of the asymmetry in distribution. When a distribution is symmetric or bell-shaped, the skewness of the distribution is not mentioned. Beyond school statistics, various techniques for measuring skewness have been explored in advanced statistics. Since these advanced statistical techniques are not within the scope of this study, the calculation of the skewness value is not discussed here. However, the interpretation of skewness is important as it gives insight into how the data is distributed. More specifically, it reveals the shape of the distribution.

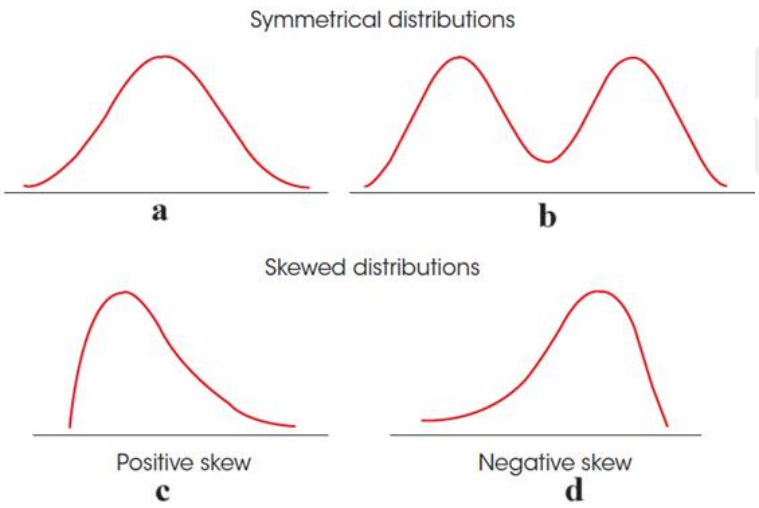


Figure 2.4. Different shapes of distributions (Gravetter & Wallnau, 2015, p. 48)

The shape of the distribution can be symmetrical (Figure 2.4a and 2.4b) or skewed (Figure 2.4c and 2.4d). When a line is drawn through the center, the images on the right and left of the line are symmetrical in a symmetrical distribution. Symmetrical distribution is also informally called bell-shaped because it is more often encountered with the distribution in Figure 2.4a. In a symmetrical distribution, there is an equal taper off gradually towards the ends and clustering in the middle. On the other hand, in a skewed distribution, there is a one-sided narrowing, with Figure 2.4c

narrowing to the right and Figure 2.4d narrowing to the left. Therefore, they are also characterized as right-skewed (e.g., positive) and left-skewed (e.g., negative) distributions, respectively. Alternatively, the majority of the data is opposite to the tails. In other words, the data tends to pile up towards the left in a positively skewed distribution, with the tail on the right side, while conversely, the data of a negatively skewed distribution piles towards the right side. In this regard, skewness can be interpreted according to the position of the majority of the data (Bakker & Gravemeijer, 2004). Assuming that the majority of the data corresponds to density, the shape of the distribution depends on density and skewness. Since the shape of the distribution is the basis for reasoning about the distribution (Bakker & Gravemeijer, 2004; Cobb, 1999; Cobb & Gravemeijer, 2006), density and skewness are indispensable aspects of making sense of distribution.

Although studies that focus only on the concept of density or skewness are not encountered in exploratory data analysis, it was observed that these aspects were utilized in data analysis based on the shape of the distribution. For example, Konold et al. (2015) specify four perspectives when looking at data: *pointers, case values, classifiers, and an aggregate*. Data as pointers indicate the larger event from which the data came. In this perspective, the event and the data are not distinguished. The data as case values perspective focuses on individual cases (e.g., *David's favorite color is red*). Data as classifiers group the data consistent with a particular attribute and point out its frequency, such as "two people love red." In data as an aggregate perspective, the data is perceived as a whole based on the context. The shape of the data is also interpreted in addition to the center and variability. For example, a student's "*half like red*" expression in the study (Konold et al., 2015, p. 309) indicates that the student has considered all the data as the student classifies the data and compares these classes. In this comparison process, relative frequency is used as the classes are associated with the whole data rather than each other. Therefore, it can be said that the aggregate perspective requires proportional reasoning, which is described in the next section.

Based on these perspectives, Ben-Zvi and Arcavi (2001) identify two ways of understanding data. The first one is *local understanding*, in which the focus is on individual data value in the data set or representation. For example, in the data as case values perspective, the single data value of David, who likes red, indicates that the data is perceived as local. In the other understanding, *global understanding*, the general pattern/trend in the data is discovered, identified, or explained by various statistical measures or techniques. In other words, one with a global understanding contemplates how data is distributed and how they change in the whole formed by these values, not the individual values (Ben-Zvi & Arcavi, 2001; Cobb, 1999). In the same manner, one considers the whole data when performing statistical measures and techniques. In this understanding, the aspects of distribution (e.g., center, spread, density, and skewness) are conceived as representative of a group of values.

2.2.5 The Studies on Reasoning About Distribution

Many studies have reported that students could not see the values in distribution as an aggregate (Aridor & Ben-Zvi, 2018; Bakker & Hoffman, 2005; Konold et al., 2015; Wild, 2006), namely local understanding of the distribution rather than global understanding (Ben-Zvi & Arcavi, 2001). When interpreting quantities varying in different ranges, they focus only on specific values, such as maximum or minimum values, instead of the distribution of the data. Furthermore, they calculate the measures, such as the mean and standard deviation, without developing a conceptual understanding (Ben-Zvi & Garfield, 2004), but they have problems defining and interpreting them as representatives of the whole data based on a context (delMas & Liu, 2005; Dubreil-Frémont et al., 2014; Pfannkuch, 2007; Turegun, 2011; Watson, 2006; Watson & Callingham, 2003). Similarly, teachers and PMTs also view the concept of distribution as a set of individual values rather than aggregate. For example, Madden (2008) concluded that when comparing the distributions, high school mathematics teachers, like middle school mathematics teachers (Gökçe, 2019), were weak in statistical justification and reasoning on the characteristics of

distributions. Similarly, the PMTs performed poorly in interpreting statistical findings in the context (Frischemeier & Biehler, 2018; Karatoprak, 2011; Kurt, 2016; Mercimek, 2013; Sorto, 2004; Yılmaz, 2019).

The existing literature (Cobb, 1999; Gravemeijer & Cobb, 2006) declares that the problem stems from the fact that statistics are taught as a collection of separate subjects (e.g., mean, median, mode, and data displays). That is to say, in schools, the content of the statistics is taught separately. The middle school mathematics curriculum also supports the declaration. For example, the objectives in the 6th-grade data processing learning area are as follows (MoNE, 2018a, p. 64):

> M.6.4.1.1. Formulating research questions that require comparing two sets of data and obtaining inappropriate data.
>
> M.6.4.1.2. Showing the data of two data sets in a two-way frequency table and a bar graph.
>
> M.6.4.2.1. Calculating and interpreting the range of a data set.
>
> M.6.4.2.2. Calculating and interpreting the mean of a data set.
>
> M.6.4.2.3. Using the mean and range in comparing and interpreting the two data sets.

When it is considered that the objectives are taught to students respectively, data representations, mean, and range are taken separately until the last objective. While the first two objectives deal with creating and representing two data sets, the following objectives focus on the mean and range of a single data set. In addition, the final objective is also restricted to the mean and range. Contrary to such separate topics and concepts, Gravemeijer and Cobb (2006) argue that the concept of distribution covering big statistical ideas, like center, spread, skewness, and density (Cobb, 1999; Bakker & Gravemeijer, 2004; Pfannkuch & Reading, 2006) is the central idea in statistics education. Accordingly, reasoning about the distribution concept makes seeing and interpreting data as aggregate possible. Therefore, this study aimed to investigate the development of PMTs' statistical reasoning about distribution. In this regard, the research question of the study addressed is how

PMTs' reasoning about distribution develops. Subsequently, the following section encompasses the studies on the development of reasoning about distribution.

Although developmental studies have rarely been conducted with teachers and PMTs (Ben-Zvi & Garfield, 2004), those that exist focus on assessing the participants' statistical knowledge and reasoning while comparing distributions rather than on how they developed their reasoning on distribution. For example, Madden (2008) examined how high school teachers evolved to compare distributions using their statistical knowledge as part of a professional development program. The scope of her study included didactic components such as assessment of student work in an activity because the participants were teachers practicing to teach statistics at the high school level. In the other study, Frischemeier and Biehler (2018) presented how PMTs compared distributions on a task from a course designed to improve their statistical reasoning with TinkerPlots. They focused on developing a framework to rate learners' performance while comparing distributions. The framework includes the assessment dimensions instead of the development of the distribution concept.

In particular, the majority of the developmental studies focus on improving students' reasoning about distribution. Cobb (1999), Bakker (2004), and Bakker and Gravemeijer (2004) conducted a teaching experiment in 7th-grade classes to develop reasoning about distribution. In each study, instructional sequences consisting of various activities were developed. The most comprehensive instructional sequence is described in detail in the method section. In general, the instructional sequences progress from exploring qualitative characteristics of collections of data points to exploring qualitative characteristics of distribution (Cobb, 1999). In these studies, the students are asked to explain and justify their solutions by implementing different learning activities in the classroom aimed at learning various aspects of the distribution. The activities progress by first making sense of the data and then reasoning about the shape of the distribution (Bakker, 2004; Bakker & Gravemeijer, 2004; Cobb, 1999). This process, which Bakker and Gravemeijer (2004) completed in less time, consists of three stages in summary. In the first stage, data values are made sense in representations such as bar graphs as a result of measuring the

attributes of a situation. In the second step, density is interpreted in different data groupings, especially in dot plots, by increasing the sample size. Finally, the students argue about the shape of the distribution depending on where the hill is in the distribution.

In the teaching experiment process, the researchers acknowledged the conjectures about potential students' thinking and learning. The conjectures that guide the principles of this dissertation are expressed in the preparation for the experiment section. They can be summarized as follows:

- The shape of the distribution is the basis for the reasoning about distribution (Bakker, 2004; Bakker & Gravemeijer, 2004).
- Students reason better on data that they have a history with (Bakker & Gravemeijer, 2004; Cobb, 1999).

Regarding the first conjecture, it is emphasized that it is necessary to reason about the shape of the distribution to look at the data as a whole instead of individual data values. To interpret the shape of the distribution, Cobb (1999, p. 11) suggests proportional reasoning, "wherein students reason about the parts of a data set as proportions of the whole data set." For this, he situates activities comparing distributions with unequal sample sizes at the end of the instructional sequence. In these situations, students examine the whole data because they consider relative frequencies (Cobb, 1999). Therefore, I present proportional reasoning below.

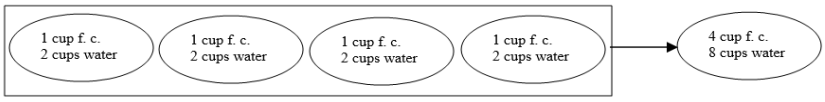
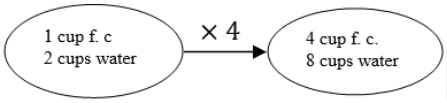
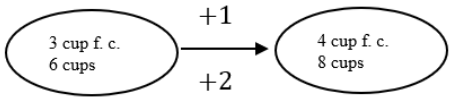
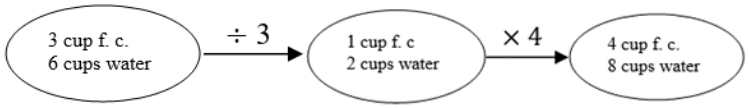
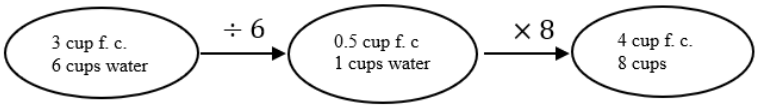
2.2.5.1 Proportional Reasoning

Proportional reasoning is a milestone of mathematics education as it is the center of relative thinking, partitioning, unitizing, quantitative reasoning, rational numbers/fractions, and ratio sense (Lamon, 1999). It is defined as "a sense of covariation and multiple comparisons, and the ability to mentally store and process several pieces of information" (Lesh et al., 1988, p. 93). In other words, it requires understanding a constant relationship between two quantities in a multiplicative

manner. The connection between binary relationships in proportionality leads to interdependence among mathematical concepts within the multiplicative conceptual field (Vergnaud, 1988), including but not limited to multiplication, division, fractions, and ratios (Ezaki et al., 2023; Ölmez, 2018, 2022).

The robust understanding of the invariance relationship between proportional quantities includes reasoning that goes beyond addition, specifically focusing on multiplicative reasoning (Lamon, 2007). Multiplicative reasoning is the coordination of "two composite units in such a way that one of the composite units is distributed over the elements of the other composite unit" (Steffe, 1994, p. 264). This implies that one composite unit is allocated a specific amount of the other composite unit. Unitizing and norming, as described by Lamon (1994), enable the interpretation of proportional situations. In the process of unitizing, a reference unit or a unit whole is established, whereas in norming, the unit is employed to reinterpret the situation. For example, let us consider the context of juice made from 3 cups of fruit concentrate and 6 cups of water. One cup of fruit concentrate per 2 cups of water in 3 cups of juice in the unitizing process is the unit factor and reflects the proportional relationship. In the norming process, it is possible to expand the unit to different situations. For 12 cups of juice, it is necessary to iterate the composite units 1-2, 2-4, 3-6, 4-8. The relationships for the last situation, consisting of 4 cups of fruit concentrate and 8 cups of water, can be established in many strategies summarized in Table 2.1.

Table 2.1 The strategies in proportional reasoning

Strategy	Demonstration of the strategy
Incorrect addition	Fruit concentrate 3 → -3 → 5 → -3 → 8 Water 6 → 8
Building- up	
Abbreviated building- up	
Ratio unit/ building-up	1 cup fruit concentrate for 2 cups of water 
Per-one (Unit factor)	 or 
Within (scalar) strategy	$\frac{3 \text{ cups f. c.}}{6 \text{ cups water}} = \frac{4 \text{ cups f. c.}}{8 \text{ cups water}}$
Between (functional) strategy	$\frac{3 \text{ cups f. c.}}{4 \text{ cups f. c.}} = \frac{6 \text{ cups water}}{8 \text{ cups water}}$

In the incorrect addition strategy, the focus is on the difference despite the multiplicative relationship, and the difference is generalized to other quantities. In the example above, the result is 13 cups of juice, which does not taste the same as the original juice. The concept of ratio is not developed. The building-up strategy corresponds to repeated addition, which is a precursor to grasping the concept of ratio (Heinz, 2000; Kaput & West, 1994). In particular, the reference unit (1 to 2) is iterated until the desired amount is reached. This process, which is additive reasoning, is abbreviated building-up when the reference unit is multiplied by a factor instead of adding one by one. Although the abbreviated building-up strategy represents a multiplicative structure because of the multiplication of each quantity by the same factor (Karagöz Akar, 2007), it is classified as an additive (Bahar, 2016; Lamon, 2007; Ölmez, 2016; Turan, 2017). In the example above, the amount of fruit concentrate and the amount of water for the target quantity are multiplied by the same factor, 4. The reason is that a unit of juice is 3 cups (1 cup of fruit concentrate and 2 cups of water), and 12 cups of juice consist of 4 units. The relationship between the targeted quantity (12) and the reference unit (3) refers to the scalar ratio (Karagöz Akar, 2007). In the ratio-unit/building-up strategy, the focus is on the "x elements for y elements" relationship. For the intended amount in the example, since 1 cup of fruit concentrate corresponds to 2 cups of water, 2 cups of water are added to 6 cups of water, and 1 cup of water is added to the other. In this association, the homogeneity of the situations is more obvious with the phrase "for every" or "for each" (Karagöz Akar, 2007; Ölmez, 2016). However, the relationship is additive because of additive replications of the identified units.

It is the per-one strategy when any paired extensive quantities are relatively reduced to a unit factor (1 to 2 or 0.5 to 1) and then scaled to the desired quantity by the unit. In the example in the table, the first quantity is first transformed into a unit factor (per one), and then the intended quantity is obtained by multiplying it with the scalar factor (4) as in the abbreviated building-up strategy. Although there is some debate about whether this strategy is multiplicative or additive (Heinz, 2000; Kaput & West, 1994), this strategy includes the invariance aspect of proportionality. Karagöz Akar

(2007) emphasized that employing the per-one strategy, also known as the unit factor, can support understanding within-ratio. Conversely, utilizing a building-up strategy could contribute to developing the conception of between-ratio. As evident from the example in the table, the within ratio (scalar strategy) involves using the ratio of different measure spaces, while the between ratio (functional strategy) involves using the ratio of the same measure spaces. These last two strategies in the table are multiplicative (Heinz, 2000) as they involve invariance and covariance between quantities. In addition, the proportion equation strategy or cross-products strategy (Heinz, 2000; Van de Walle et al., 2019) can also establish the relations between and within ratios; however, it does not offer substantiation for proportional reasoning (Kaput & West, 1994).

Although the strategies mentioned above seem to be separated into additive and multiplicative reasoning, the one-per strategy has multiplicative and additive aspects. Therefore, it can be said to be a transitional stage for multiplicative reasoning. Copur-Gencturk et al. (2022), in their study on the proportional reasoning of teachers, added relative reasoning as an intermediate category between additive and multiplicative reasoning levels. According to the results of the study, there are four categories: incorrect, additive, relative, and proportional reasoning. Since they investigated the squareness in terms of proportional reasoning, the examples for the categories are adapted to the example above. Incorrect reasoning is the inability to identify quantities correctly or to reason incorrectly. For the example above, an individual at this reasoning level may think of the amounts of fruit concentrate and water separately. Differently, they might reach more cups of juice in total (13>12) as in incorrect addition. In additive reasoning, the focus is the constant difference between the quantities. One with additive reasoning thinks that the difference between fruit concentrate and water is always equal. Alternatively, analyzing the ratio table in Figure 2.5 horizontally, one can notice that the difference in fruit concentrate quantities is always 1, and in water quantities, it is 2. In relative reasoning, the difference between the quantities themselves and between each other is taken into account. The figure also shows that in addition to additive reasoning,

the differences between the quantities of fruit concentrate and water in columns of the table are also taken into account sequentially. Proportional reasoning involves comprehending the ratio between two quantities. In other words, it is noticed that the amount of water is twice that of the fruit concentrate, or the amount of the fruit concentrate is half of the water.

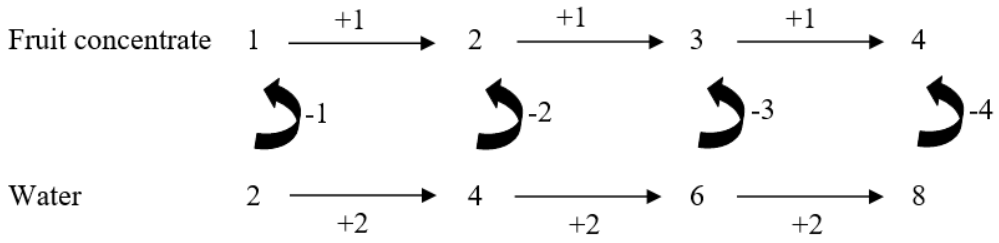
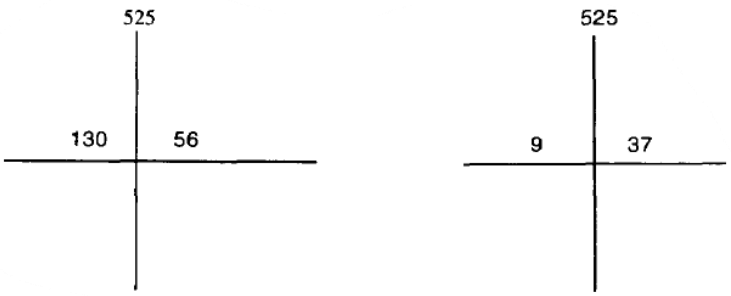


Figure 2.5. Ratio table for the example

In this study, dealing with distributions of unequal sample sizes raises the question of whether to employ additive or multiplicative reasoning. Since the sample sizes are different, comparing the absolute frequencies of the distributions in the specified ranges may lead to misinterpretations. For example, in the study (Cobb, 1999), 7th-grade students worked on the T-cell count data of 186 patients who participated in the old program and 46 patients who participated in the new program to treat AIDS. Without providing numerical evidence, the students could determine which program was more effective based on “the hill” or “the majority” in the distributions. However, they needed proportional reasoning while reasoning based on numbers under the teacher’s guidance. Taking 525 T-cells in the classroom as a reference, they partitioned the distributions into two, as shown in Figure 2.6.



Graphs developed from the third analysis of the AIDS protocol data.

Teacher 2: Could you just argue that this shows really convincingly that the old treatment was better, right, because there were 56 scores above 525, 56 people with T cell counts above 525, and here [points to the graph on the right] there's only 37 above, so the old one just had to be better, there's more people, I mean there's 19 more people in there, so that's the better one. surely.

Figure 2.6. The graphs developed from AIDS data and the teacher's argument (Cobb, 1999, p. 26)

The teacher argued to the students that the old program was more effective in treating AIDS because it increased the number of T-cells (56>37). On the contrary, a comparison below 525 would also reveal that the old program is worse since 130 is greater than 9. Accordingly, the students rejected the argument based on the absolute frequencies and took into account the proportion of data within different value ranges, such as the percentage of T-cell numbers above 525. Drawing conclusions based on relative frequencies, such as percentages, ratios, proportions, or scales, was more valid and accurate due to the differences in sample sizes. It showed that proportional reasoning is crucial in statistical reasoning about distribution, particularly when comparing distributions with differing sample sizes (Bakker, 2004; Cobb, 1999).

Figure 2.6 presents the comparison of the relative frequencies of data partitioned into two at a specific reference point. In studies (Bakker, 2004; Bakker & Gravemeijer, 2004; Cobb, 1999; Gravemeijer & Cobb, 2006) where the *Minitool* was used, the comparison based on proportional reasoning was also characterized by different

partitions of the distributions, such as partition into equal width intervals, partition into equal groups, partition into half, and so on. Accordingly, in addition to proportional reasoning, the use of technological tools is suggested as the tools for reasoning about the shape of the distribution, concerning the first conjecture (Bakker & Gravemeijer, 2004; Cobb, 1999). Therefore, we decided to enrich the classroom teaching experiment in this study with dynamic statistical software. In this context, the following section elaborates on the use of technology to develop reasoning about distribution.

2.2.5.2 The Use of Technology for the Development of Reasoning About Distribution

The use of technological tools has become widespread in schools and statistical education research when analyzing real and huge datasets. It has a great impact on achieving the goals of statistics teaching, as it provides the time to focus on understanding statistical concepts rather than applying procedures and formulas when analyzing data (Chance et al., 2008; Zieffler et al., 2018). Technological tools have various features that facilitate the statistical learning process. For example, the dragging feature of technological tools enables fast, direct, and adaptable changes to visualized data (Biehler et al., 2013; Lee et al., 2012). It also allows students to control themselves with quick feedback in graphical representations. Moreover, the technological tools arrange simulations that model different samples of the real-life (Konold & Kazak, 2008; Lehrer et al., 2007).

Different technological tools are employed to make sense of statistical concepts and skills. For example, Fathom (Finzer, 2001; Hoffmann et al., 2014; Lee et al., 2014), Excel (Oikarinen et al., 2014), Sampling Distribution (Chance et al., 1999), Cabri-Géomètre II Plus (Martins et al., 2014), and CODAP (Common Online Data Analysis Platform, https://codap.concord.org/; Mojica et al., 2019) were used to develop students' concept of data exploration. Two computer-based minitools were developed specifically to improve reasoning about distribution (Bakker &

Gravemeijer, 2004; Cobb, 1999). They allow several ways of structuring data. The first minitool was designed for ordering, partitioning, and organizing up to 40 data values in a horizontal bar graph (Cobb, 1999). In the bar graph, where each bar represents a single data value, the data can be sorted by size and color. In addition, the number of data in any horizontal interval can be determined with the range tool. In the second minitool, bars are replaced with dots. The tool allows users to analyze one or two distributions of up to 400 data values in dot plots. It allows the users to group the data as desired and find the number of data in the groups. In addition, it has four more options for structuring data. They are "partitioning the data into groups of a specified size (e.g., 10 data points in each group), partitioning the data into groups with a specified interval width, partitioning the data into two equal groups, and partitioning the data into four equal groups" (Cobb, 1999, p. 18). The minitools were extended in later years to explore statistical covariation (Cobb et al., 2003). These minitools developed for literature-based instruction were not accessible.

Another technological tool designed for data exploration of students ranging from middle school to university levels is TinkerPlots (Konold & Miller, 2011). The tool has valuable features to describe the characteristics of data. For example, categorical data is represented in different colors, but quantitative data is in the color scale. In TinkerPlots, different graphical representations (dot plots, bar graphs, histograms, box plots, and scatter plots) are created and converted into different formats by dragging dots or changing the scales. The tool also allows the calculation of the mean, median, mode, and range, as well as MAD and SAD. In addition, it has the aforementioned way of data partitioning in the second minitool: creating your groups, partitioning the data into groups of a specified size, partitioning the data into groups with a specified interval width, partitioning the data into two equal groups, and partitioning the data into four equal groups. TinkerPlots, based on NCTM standards, appears to be a combination of the three minitools above. With the reference line, frequencies and percentages of data values above and below a cut-point can be easily observed. Moreover, its sampler feature provides the creation of different samples.

Based on the advantages of TinkerPlots in improving reasoning about distribution (Fielding-Wells, 2018; Frischemeier & Biehler, 2018; Hammerman & Terc, 2004; Koparan, 2019; Makar et al., 2011; Meletiou-Mavrotheris et al., 2009; Watson et al., 2020), TinkerPlots was used as data visualization and modeling tool in this study. However, Lee et al. (2014) concluded that taking advantage of technology for advanced data representations is not always evidence of reasoning development. In addition to the use of technology as a media that may facilitate instruction, how the instruction is administered is also important for development (Biehler et al., 2013). In this context, the study is an example of how instruction works using technology.

As aforementioned, proportional reasoning and the use of technology have been essential for reasoning about distribution in the context of the first conjecture related to reasoning about the shape of the distribution. The other conjecture is related to the student's familiarity with the data. The researchers (Bakker & Gravemeijer, 2004; Cobb, 1999) acknowledge that students reason better on data that they have a history with. The conjecture is related to learners' understanding or experiencing the data collection process (Cobb, 1999). In other words, the focus is on the development of individuals in statistical reasoning through their own experiences. This is the reason why this study is based on RME. Furthermore, since the developmental process takes place in the classroom discourse orchestrated by the instructor, the emergent perspective focusing on the development of the classroom community rather than individual students also guides the study. In this respect, the following section presents the theoretical framework on which the current study is based.

2.3 Theoretical Framework

This study is grounded in two theoretical frameworks. On the one hand, it draws on Realistic Mathematics Education (RME) for designing, planning, and implementing classroom instructional activities. On the other hand, it uses the emergent perspective to interpret the participants' reasoning, which is socially constructed within the learning environment. These theories are described in detail below.

2.3.1 Realistic Mathematics Education

RME is a theory specific to mathematics education, developed by the Dutch mathematics educator Hans Freudenthal. Although statistics and mathematics are separate disciplines (Cobb & Moore, 1997), this theory is also used in the teaching and learning of statistics as statistics is taught within the mathematics curriculum and is closely related to context (Bakker, 2004). It is based on understanding and dealing with problem situations that are experientially real to students. In other words, students solve statistical problems based on realistic situations in the classroom, which allows them to engage in statistics as a human activity (Freudenthal, 1991). According to Treffers (1987), RME has five interrelated tenets.

1. Phenomenological exploration: A rich and meaningful context is needed to construct the concepts. This tenet is closely related to statistics learning because statistics deals with data based on context. All the activities implemented in this study include data from situations that can be experienced in real life or research.

2. Using models and symbols for progressive mathematization: Progressive mathematization, vertically or horizontally, covers a developmental process from intuitive to formal mathematical concepts. Many instruments, such as models, schemes, diagrams, and symbols supporting meaningful learning, can be used in this process, and students reach generalizations and abstractions. In this study, the inscriptions that the participants generated and the things they used in their works or said are signs of their mental constructions.

3. Using students' own constructions and productions: Students must generate their strategies and ideas in a way that makes sense to them in the learning process. Essentially, they take ownership of their learning journey, guided by a specific objective. In the study, participants are expected to reason with their structured data representations or solution methods for the instructional activities in small group work and class discussions.

4. Intertwinement: This tenet relates its relevance to other domains. The distribution that is the main objective of the instructional activities is an overarching concept that includes the center, variability, shape, density, skewness, relative frequency, probability, proportionality, and causality in statistics education and is dependent on them (Bakker & Gravemeijer, 2004; Pfannkuch & Reading, 2006).

5. Interactivity: Collaborative interaction among students in a classroom discourse facilitates learning for both themselves and their peers. This principle highlights the impact of shared experiences in social settings on learning.

According to RME, students engage in phenomenological explorations of intertwined concepts or ideas through their constructions and products in activities with rich and meaningful contexts within socially constructed learning environments. As part of the progressive mathematization process, students explore mathematics with the support of the teacher and instructional design. In this context, three principles of RME are described (Gravemeijer, 1994; Gravemeijer et al., 2000). The first principle is *guided reinvention*, where students experience learning mathematics as a process akin to the discovery of mathematics itself. In this study, the instructor guides the participants in the re-invention of concepts during the implementation of instructional activities. In this regard, having a history of the data to be studied, understanding the context, the methods employed for data collection, the underlying purpose of the research on the collected data, etc., are the points considered when implementing the activities. In this way, the students resemble statisticians as they analyze data.

Another principle of the RME is *didactical phenomenology*. Freudenthal (1983, p. ix) states that a didactical phenomenology can show "the teacher the places where the learner might step into the learning process of mankind." In other words, it describes the potential experiences and educational contexts through which a student engages in the learning process. Consequently, the principle is closely related to the

first tenet, which emphasizes rich and meaningful context to support progressive mathematization. It can be challenging to find activities that support both individual activity and class discussions (Gravemeijer et al., 2000). However, due to the variability in statistics, statistical problems generally lend themselves to discussion and interpretation in different ways. Furthermore, when the established classroom norms encourage students to explore different solution strategies, the students can collaboratively experience and progressively develop sophisticated problem-solving approaches (Gravemeijer et al., 2000). These norms will be described in the emergent perspective in the next section.

The last principle of the RME is *emergent models*. In the second tenet framework, students use a variety of constructions and products, ranging from informal to formal, in the progressive mathematization process. At first, students' informal ways of reasoning, which are situated, serve as "models *of*" (Gravemeijer, 1994). Over time, these "models *of*" evolve into "models *for*" that are more mathematical, encompassing a variety of symbols. For example, a horizontal bar graph initially serves as a model *of* data, while "four-equal groups display as a graphical representation of the distribution that started to function as a model *of* a model for reasoning about distributions" (Gravemeijer, 2002, p. 5).

Considering that instructional activities take place in the classroom, it is logical to assess classroom practices and events through a socio-constructivist perspective. This is because there is engagement between the teacher and students, as well as among students themselves and with the activities. This study also addresses the development of statistical reasoning, especially about distribution, as collective learning of the classroom community in the social environment of the classroom. Cobb (2000) and his colleagues suggest the emergent perspective as an interpretive framework for understanding classroom learning by analyzing what is happening in the classroom. The emergent perspective is described in the next section.

2.3.2 Emergent Perspective

The emergent perspectives arise as a remedy for the endeavor to comprehend the learning of mathematics within the social dynamics of the classroom (Cobb & Yackel, 1996; Gravemeijer & Cobb, 2006; Yackel & Cobb, 1996). In this framework, students' development is interpreted by analyzing both the individual (psychological) and social dimensions of the classroom microculture. Table 2.2 outlines the framework.

Table 2.2 The interpretative framework for analyzing individual and collective activity at the classroom level (Cobb & Yackel, 1996, p. 6)

Social Perspective	Psychological Perspective
Classroom social norms	Beliefs about own role, others' roles, and the general nature of mathematical activity in school
Sociomathematical norms	Mathematical beliefs and values
Classroom mathematical practices	Mathematical conceptions

Classroom social norms, sociomathematical norms, and classroom mathematical practices collectively shape the social culture of the classroom. According to Cobb and Yackel (1996), classroom social norms are identifiers of the classroom participation structure. In other words, they are the expected behaviors of individuals within the classroom. In traditional classroom settings, students are often required to deduce the teacher's intentions rather than express their interpretations. Conversely, in learning environments characterized by classroom discussions, students' obligations vary. For example, the students produce, explain, and justify the solutions, as well as attempting to comprehend explanations offered by their peers, express their own ideas about others' arguments, and question alternatives in conflicting situations. From a psychological perspective, students' beliefs about their

own roles and the roles of others, along with their understanding of the nature of mathematics, are linked to classroom social norms (Cobb &Yackel, 1996). As the norms are established, students tend to adjust and restructure their beliefs accordingly.

The social norms mentioned above can be applied to classrooms where all disciplines, such as science and literature, are taught. When these norms are enacted through actions and interactions specific to mathematics in classroom discussions, they become sociomathematical norms (Cobb & Yackel, 1996). Given that this research focuses specifically on statistics and recognizes mathematics and statistics as separate disciplines (Cobb & Moore, 1997), the term employed in this study is *sociostatistical norm* instead of sociomathematical norm. Sociostatistical norms include different, sophisticated, effective, and acceptable solutions and explanations for statistical problems. Sociostatistical norms in the study are established as students solve a statistical problem in the classroom using various approaches, resulting in statistically meaningful and acceptable solutions along with justifications (Stephan, 2020). That is, students generate ideas and provide evidence to validate those ideas to persuade everyone in the socially dynamic classroom culture if necessary. In such a student-centered classroom, all students' participation is supported in rich classroom discourse (Erath et al., 2021; Hershkowitz & Schwarz, 1999) by ensuring their intellectual autonomy (Yackel & Cobb, 1996). When students engage in making statistical decisions and judgments within a classroom setting characterized by predefined sociostatistical norms, they actively construct their individual statistical beliefs and values as contributors to the classroom statistical community (Cobb & Yackel, 1996). In this context, there exists a reciprocal interaction between psychological aspects involving students' statistical beliefs and values and social dynamics represented by the established sociostatistical norms. In essence, sociostatistical norms and specific statistical beliefs and values reflexively influence and reinforce one another (Cobb & Yackel, 1996).

The final social dimension of the framework is the classroom mathematical practices. A mathematical practice refers to the established patterns of behavior,

communication, and symbolization of mathematical concepts at a particular point in time (Gravemeijer & Cobb, 2006). The classroom community accepts mathematical practice as a mathematical truth without requiring any justification (Cobb & Yackel, 1996). Consequently, classroom statistical practices highlight "the taken-as-shared ways of reasoning, arguing, and symbolizing established while discussing particular *statistical* ideas" (Cobb et al., 2001, p. 126). Individual students' interpretations of the ideas actively contribute to the emergence of these taken-as-shared ways. Therefore, the connection between the statistical conceptions of individual students and classroom practices is reflexive (Cobb & Yackel, 1996).

The emergent perspective deals not only with the development of individual minds but also with the specific social environments in which these minds are engaged. (Cobb & Yackel, 1996). In this regard, learning is the "collective learning of the classroom community" that participates in classroom activities (Cobb et al., 2001, p. 113). Accordingly, classroom practices serve as indicators of community learning. Since this study addresses the development of statistical reasoning about distribution as collective learning of the classroom community in the social dynamics of the classroom, it investigates classroom statistical practices that the PMTs performed during the classroom teaching experiment. Revealing the classroom statistical practices that show reasoning about the distribution of the PMTs corresponds to the answer to the second sub-research question of the study.

As a result, the previous research has demonstrated that the PMTs' reasoning about distribution can be developed through the use of technology in classroom teaching experiments (Bakker, 2004; Bakker & Gravemeijer, 2004; Cobb, 1999; Gravemeijer & Cobb, 2006). Building upon these studies, the current study aims to investigate the development of PMTs' reasoning about distribution as they participated in the classroom teaching experiment enriched by dynamic statistical software. This study aligns with instructional approaches from previous studies grounded in RME, the emergent perspective, proportional reasoning, and the integration of technology, all discussed in this chapter. The next chapter elaborates on further details regarding the methodology employed in this study.

CHAPTER 3

METHODOLOGY

This chapter introduces the methodology employed for this study. It contextualizes the study by providing information on how the study was conducted and the ethical issues in this study. In this regard, the chapter comprises the following sections: design of the study, participants of the study, data collection process, data analysis, trustworthiness, researcher role, and ethical considerations.

3.1 Design of the Study

The main concern in the study is to investigate the development of the preservice mathematics teachers' (PMTs) reasoning about distribution. To achieve this goal, the study employs an instructional intervention to improve the PMTs' reasoning about distribution. By this consideration, the methodological framework chosen for this study is design research, which involves a systematic study of designing, developing, and assessing the influence of an educational intervention or innovation (Bannan, 2010; Plomp, 2010).

Researchers engage in various terms to describe design research, including design experiment (Brown, 1992; Cobb et al., 2003; Collins, 1992), developmental research (Gravemeijer, 1994), and design study (Confrey, 2006). Despite these different labels, they agree on the "engineering" of the intervention (e.g., programs, teaching/learning strategies, materials, products, or systems) based on theoretical orientation in authentic contexts such as the classrooms rather than the controlled laboratory environments. Design research stands apart from experimental studies,

which involve manipulating variables, and action research, which primarily addresses educational challenges (Bakker, 2019).

Design research has five characteristics, although not all may appear clearly in a study (Bakker, 2019; Cobb et al., 2003). The first characteristic is the aim of design research. It is conducted to develop an instructional theory about the learning process and the means to support that learning (Gravemeijer & Cobb, 2006). The researchers design an intervention based on theories in the real world and also contribute to theorizing. In other words, "actionable knowledge and theories of action" are produced (Bakker, 2019, p. 46). Another characteristic is that the theory being developed must work pragmatically. That is, the design based on a theory of learning must be applicable or repeatable by practitioners in contexts other than the study. Bakker (2019) denotes the transferability of the theoretical conclusions, models, or concepts reached from the findings of the specific cases to other contexts as theoretical or analytical generalizations.

The most distinctive characteristic of design research is its interventionist nature. To reciprocally understand and change something in educational settings, a design is deliberately enacted, and new forms of learning are investigated. The educational intervention occurs in a sequence of activities that lead to new learning and practices (Mintrop, 2016). The learning process is implemented and tested simultaneously. Therefore, design research is a transformative and conjecture-driven design (Confrey & Lachance, 2000). This bilateral structure is the prospective and reflective characteristic of design research. While conjectures are being tested, the conjectures can be refuted, or revisions can be made to the design by creating alternatives during the intervention. The recurring adjustments made to improve learning in an iterative process reflect the cyclical nature of the design research, which is its final characteristic.

This study addresses the development of PMTs' reasoning about distribution by working on the activities assigned to them in the classroom. It focuses on understanding how PMTs reason about distribution, extending beyond the individual learning experiences of a single student or a small group of students (Steffe & Thompson, 2000). Since the study occurs in the social learning ecology of the classroom, the study's methodology is a classroom teaching experiment, a special form of design research introduced by Cobb (2000). In a classroom teaching experiment, the research team conducts a series of teaching sessions in the classroom where students participate in learning, in which one of the researchers in the team has the role of the teacher leading the instruction in the study.

Classroom teaching experiments are based on theory and contribute to the theory (Cobb, 2000). Namely, the activities enacted in the classroom inevitably live on the theory. The symbiotic relationship between the theory and the learning activities provides insight into implementing learning activities in the classroom. Cobb (2000) characterizes this process in the mathematics teaching cycle he adapted from Simon (1995). Figure 3.1 shows the teaching cycle in the classroom teaching experiment.

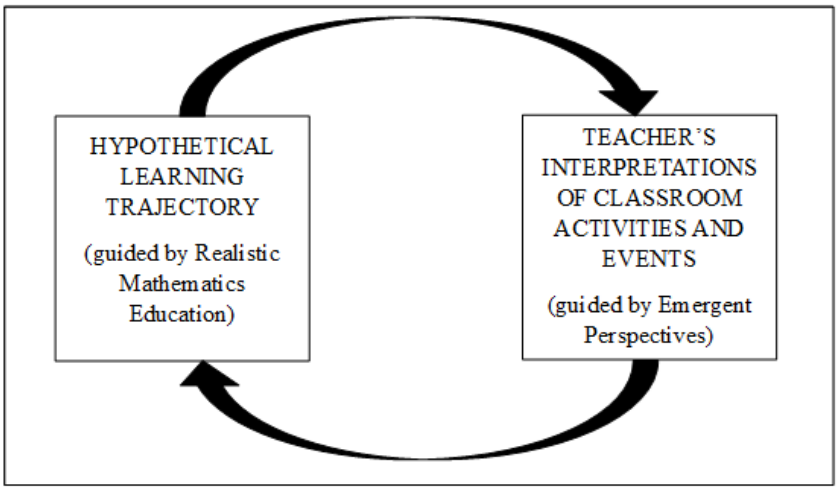


Figure 3.1. Teaching cycle in the classroom teaching experiment

The first component of the figure on the left is the hypothetical learning trajectory (HLT). Simon (2020) defines the learning trajectory as the path of learning that might progress anticipated by the teacher. It includes the learning goals that students are expected to learn, the learning activities they might engage in, and conjectures of the learning process – how students' thinking and understanding about what to be learned can evolve in this process. In this study, as a research team, my supervisor and I formulated an HLT to improve PMTs' reasoning about distribution. The HLT is explained in detail in the section on the data collection process, including instructional sequence design in the phases of the design research.

Considering the theory-based characteristic of the teaching experiment (Cobb, 2000; Cobb et al., 2003; Confrey & Lachance, 2000), the HLT is expected to have roots in the domain-specific theory. In this study, the theoretical framework guiding the design and organization of instructional activities within the HLT is Realistic Mathematics Education (RME). More explicitly, this study envisions that PMTs in the teaching experiment engage with the learning activities that are realistic to them for the progressive mathematization of intertwined statistical concepts (Gravemeijer, 1994; Treffers, 1987). Moreover, the teacher's interpretation of classroom activities and events in the other component of the figure on the right is also situated within the RME framework since the teacher seeks to understand classroom practices for students' learning in an innovative learning ecology created under the guidance of HLT. On the other hand, considering that the practices are activated in the classroom, it is reasonable to evaluate the activities and events in the classroom from a socio-constructivist perspective because in the classroom, there is interaction between teacher-students/student-students. In this context, Cobb (2000) and his colleagues suggest the emergent perspective as an interpretive framework to make sense of what is happening in the classroom. The emergent perspective focuses on classroom and individual learning (Cobb, 2000; Cobb & Yackel, 1996). As this perspective, a specific empirical methodology, stresses analyzing learning activities situated in the

social context of the classroom (Cobb, 2000), there is a mutually reinforcing relationship between HLT and the teacher's interpretation. The two theories that guided the study were explained in detail in Chapter 2. The practical implications of them are presented in the data collection process section.

Each enacting of the HLT in the classroom means repeating the teaching cycle in many sessions throughout the classroom teaching experiment. The microcycles serve to construct a local instruction theory (Gravemeijer & Cobb, 2006). This instruction theory provides insight into the instructional sequence that makes learning easy for students. Accordingly, this study uncovers the HLT that can be used for the development of PMTs' reasoning about distribution. The following section elucidates the backgrounds of the PMTs who are subjected to the HLT.

3.2 Participants of the Study

The current study was conducted with preservice middle school mathematics teachers registering in the Elementary Mathematics Teacher Education program at a public university in Ankara. This program under the Department of Mathematics and Science Education is a four-year undergraduate program in which the education language is English. To train qualified middle school mathematics teachers at national and international standards, the program provides field courses such as Calculus I-II, Basic Physics I-II, Introduction to Probability and Statistics I-II, mathematics education courses such as Methods of Teaching Mathematics in Middle Schools I-II, general pedagogy courses such as Introduction to Education, Educational Psychology, and standard courses such as English for Academic Purposes I-II, Computer Applications in Education. They are must courses. The list of the must-courses in the Elementary Mathematics Education program is presented in Appendix A. In addition, to graduate, the PMTs must take elective courses from the department or other departments according to their interests and abilities.

According to the standards determined by the Council of Higher Education (CoHE, 2019) for teacher education programs, these elective courses are usually the mathematics education courses offered by the program. This study was conducted within the scope of such an elective course, MSE482- Projects in Mathematics and Science Education, offered in the spring semester of 2020-2021. Detailed information about the course content is presented in the next section.

After obtaining the necessary ethical approval (Appendix B) from the relevant institution, my supervisor, who was also the course instructor, and I, as the teaching assistant of the course, announced the study on statistical reasoning to all PMTs. Then, the PMTs who were willing to learn the content enrolled in the course. In other words, the participation in the study was voluntary. However, we had a limited number of one-year licenses of TinkerPlots, which was used as statistical software in the study. Due to the limited availability license of TinkerPlots, the course capacity was restricted.

The participants of the study were 14 preservice middle school mathematics teachers (11 females and three males). While two of them were final-year students of the program, the others were third-year students. All PMTs had taken the course of STAT201- Introduction to Probability and Statistics I in previous semesters. This course covers the statistical content relating to descriptive statistics, probability, combinatorial methods, conditional probability, random variables, univariate and bivariate distributions, expectation, variance, covariance, correlation, central limit theorem, and estimation. Therefore, they were familiar with the statistical measures (the mean, median, mode, range, interquartile range, standard deviation, and variance), graphical representations, hypothesis testing, and correlation. Moreover, two 4th-grade students took the course MSE342- Methods of Teaching Mathematics in Middle School, which is about teaching the mean, median mode, range, and

graphics at the middle school level. In line with this information, the participants of the study can be said to know basic statistical concepts.

3.3 Data Collection Process

The data of the study was collected based on the phases of the design research. The design researchers have described the phases in different ways. For example, Bannan (2010), in the integrative learning design framework, categorizes the phases of design research as the informed exploration, design/development/enactment, and evaluation of local and broad impact. On the other hand, Nieveen and Folmer (2013) reduce them to the preliminary research, development or prototyping, and summative evaluation phases. Gravemeijer and Cobb (2006) called these phases to prepare for the experiment, experiment in the classroom, and conduct retrospective analyses. These three phases (i.e., preparing for the experiment, experimenting in the classroom, and conducting retrospective analyses) guide the design, development, and evaluation of the classroom teaching experiment in this study. Below, I describe these three phases in detail.

3.3.1 Preparing for the Experiment

The intended purpose of design research is to provide an instructional theory that is empirically grounded (Gravemeijer & Cobb, 2006). Following this purpose, the first phase of design research involves preparing for developing local instructional theory. In this phase, a conjectured local instruction theory, including a tentative learning process based on the learner's thinking and understanding, is generated by considering learning goals or instructional endpoints (Gravemeijer & Cobb, 2006). The starting point of the preparation for a classroom teaching experiment is the determination of the learning goals (Gravemeijer & Cobb, 2006). Since establishing

the learning goals requires history, tradition, and assessment, researchers can start with problem situations. As explained in detail in the previous sections, the distribution characteristics are not understood, and there is a need to encourage students to think beyond individual data based on the existing literature. To achieve this, firstly, teachers who are important factors in teaching should grasp the data set as aggregate. If there are problems with this, they must be addressed before teachers start practicing. Therefore, as a starting point for the classroom teaching experiment, a pretest containing five questions was administered to the PMTs to measure their existing statistical knowledge about the distribution. Then, semi-structured interviews were held to detail their responses to the questions in the pretest. The data obtained from the test results and the interviews before the experiment revealed both the theoretical intention of the study and the milestones of the teaching experiment arrangement. The PMTs' low reasoning about distribution confirmed the problem above and that teaching the notion of distribution was the ultimate intent.

At this phase, the aim was to generate a local instruction theory of the development of the PMTs' reasoning about distribution that can undergo elaboration and refinement during the classroom teaching experiment. The development of the local instructional theory involved reviewing the literature, mathematics curricula, and textbooks related to distribution. In this process, first of all, the conjectures that allow to reconceptualise reasoning about the distribution were ascertained as the theory includes conjectures about the possible learning process and the means that support the learning (Gravemeijer & Cobb, 2006). The conjectures involving having a history of data and reasoning about the shape of the distribution that requires proportional reasoning and the use of technology were mentioned in Chapter 2. They guided the principles of the teaching experiment to support the development of the PMTs' reasoning about distribution. The principles of the present study that the researcher adopted according to suggestions in the related literature (Cobb & McClain, 2004) were as follows:

- The focus is distribution, which is an overarching statistical notion.
- The instructional activities combine statistical ideas such as center, spread, density, and skewness to promote the reasoning of the PMTs.
- Elaborating the data generation process is necessary to understand the context. The classroom activity structure is based on a social learning environment, including a whole class discussion of the data generation process, individual or small group work to analyze the data, and a whole class discussion of the data analyses.
- The computer-based tool the PMTs used is TinkerPlots.
- The classroom discourse orchestrated by the instructor is grounded on sociostatistical norms, consisting of different, sophisticated, effective, and acceptable solutions and explanations of the PMTs.

According to these principles, the HLT followed in this study was formulated, with the flexibility to be adjusted and refined as the teaching experiment proceeded. The HLT, consisting of learning activities, learning goals, and a hypothesized learning process for the design of instruction (Simon, 2020), was based on previous research focusing on improving 7th-grade students' reasoning about distribution (Cobb, 1999). Professor Cobb provided the initial instructional sequence including the learning activities used in his project's research, upon contact of the researcher via e-mail (See Appendix C for permission to use the activities). The learning activities in the initial instructional sequence were designed for 7th-grade students with different cognitive levels in different socio-cultural environments. Since this study focussed on the development of PMTs' reasoning about distribution, the researcher and instructor, as the research team, decided that the initial sequence could not be used directly depending on the participants and the context of this study. For this, the research team met weekly to design the HLT of the study, which was obtained by adapting the initial instructional sequence. Below, I describe the integration process of the initial instructional sequence into the HLT of this study.

3.3.1.1 Integration of the Initial Instructional Sequence into the HLT of the Study

With the consideration of improving the PMTs' reasoning about distribution, the activities in the initial instructional sequence were adapted by revising their content and sequence. The HLT, including these activities, was generated based on the PMTs' reasoning throughout the study. Table 3.1 compares the order of the labeled activities in the initial instructional sequence and the HLT of this study.

Table 3.1 The order of the activities in the initial instructional sequence and the HLT of this study

The activities in the initial instructional sequence	**The activities in the HLT of this study**
1. Sneakers	1. Printer
2. Driving to school	2. Tracking a herd
3. Driving to school with Minitools 1	3. Missing word
4. Tracking a herd	4. Soccer game
5. Migration	5. Batteries
6. Hot air balloons	6. Hot air balloons
7. Reasoning about the mean	7. Watermelons
8. Understanding the mean	8. Resting heart rate
9. Batteries	9. Backpacks 1
10. Watermelons	10. Backpacks 2
11. Braking distance	11. Anorexia
12. Helmets	12. Ambulance
13. Resting heart rate	
14. Backpacks 1	
15. Backpacks 2	
16. AIDS	
17. Anorexia	
18. Migraine	
19. Ambulance	

The first activity of the sequence, "sneakers," lets students think about dealing with data. The "sneakers" activity required reasoning on aggregating different ranked data and the relationship between totaling and averaging the data ranks. The content of the statistics in the curriculum at the middle school level is limited to the statistical measures- the mean, median, mode, and range- and representations-frequency table, object and picture graphs, tree diagrams, bar graphs, pie charts, and line graphs- (MoNE, 2018a). The statistical content expands with standard deviation and histogram in high school (MoNE, 2018b). More complex statistical content, such as interquartile range, variance, and box plot, varies according to the curriculum of the programs in higher education (CoHE, 2019). Although data characteristics are formally emphasized in high school objectives (MoNE, 2018b), students deal with data at all levels of statistics education. As the activity was compatible with the curricula, it was decided to start the teaching experiment with such an activity related to data and data analysis. However, the context was changed to buying a printer because it was more familiar to PMTs to generate data. To assist a project coordinator while purchasing a printer, the PMTs' task in this activity was to determine five factors, instead of 8-10 criteria in the "sneakers" activity, rank the factors from the most important to the least important in their project teams, and then create a class list. In purchasing a printer context, the activity was named "aggregating different ranked data-printer." The PMTs were anticipated to complete the task using frequency, total, or mean.

The second and third activities of the instructional sequence related to data representations in the technology were not included in the HLT. This study primarily focused on comprehending the basic statistical concepts describing the distribution, such as central tendency and variability. Therefore, different graphical representations in TinkerPlots were decided to be more appropriate for subsequent activities. Consequently, the "tracking a herd" activity, originally the fourth in the instructional sequence, occurred in the HLT following aggregating different ranked

data. as its learning goal was to understand the average. The “estimating an unknown whole through the average-tracking a herd” activity in the HLT included developing a method for finding the number of elephants in the aerial photograph of an elephant herd in Figure 3.2 without counting them one by one. Some possible strategies that could be used for this activity were: 1) making groups, guessing how many there are in each group, and adding all numbers, 2) making a group with a fixed number and estimating how many groups fit into the whole, 3) counting the number of elephants lengthwise and widthwise and multiplying these, and 4) making a grid, choosing a typical box, and multiplying the number of elephants in this box by the number of boxes in the grid. The next “migration” activity in the instructional sequence was not included in the HLT because it was an alternative to the “estimating an unknown whole through the average-tracking a herd” activity in terms of the learning goal and the activity structure.

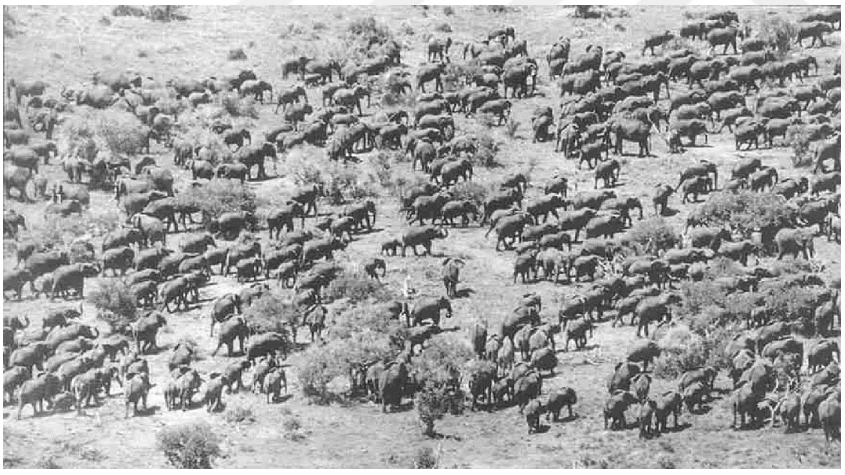
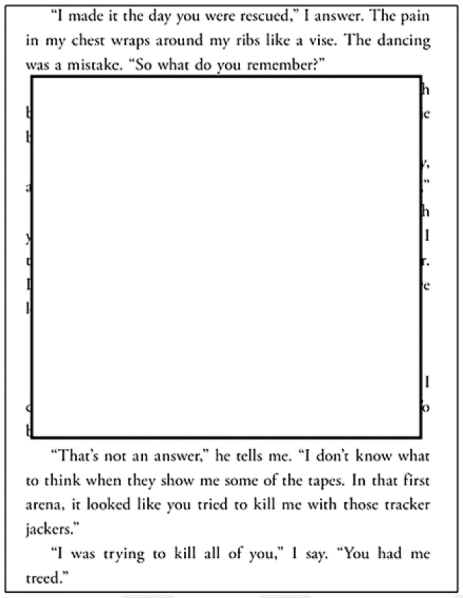


Figure 3.2. The aerial photograph of the elephant herd

When analyzing the class sessions in the weekly meetings, we noticed that the participants tended to use the mean formula while aggregating different ranked data and estimating an unknown whole through the average. After estimating the total

elephant number in the picture, the PMTs were asked to find the average numbers in three four-by-three tables in which they thought of the numbers in the cells as the number of elephants per cell. In the original sequence, the extension corresponded to the seventh activity, "reasoning about the mean," after the "hot air balloons" activity in the sixth order. The PMTs effortlessly calculated the mean, but the mean did not reflect the average numbers in the tables. At this particular moment, it was seen that there was a need for a conceptual understanding of the mean beyond its calculation. The understanding of the structure of the mean equation was anticipated in the "understanding the mean" activity in the eighth order of the instructional sequence. The horizontal bars on the Minitool used in the activity were manipulated as desired. The tool could equalize the different units by adding the more to the less in horizontal bars. Namely, the mean could be explored by compensating the different lengths between bars in the tool. This modeling, which corresponds to the fair sharing or leveling interpretation of the mean, cannot be realized in the TinkerPlots used in this study. The "understanding the mean" activity in the sequence was removed in the HLT as the meaning of the mean could be understood through contextual activities rather than using technology. That is to say, the activities, from "tracking a herd" to "understanding the mean" in the instructional sequence, were combined except for the "generating different samples in weight context-hot air balloons" activity in the HLT (See Table 3.1).

To shift the PMTs' focus from calculating the mean to understanding the meaning behind it, it was concluded that there was a need to draw PMTs' attention to variability in the retrospective analysis of classroom events at weekly meetings. Therefore, the "estimating an unknown part through the average-missing words" activity that requires finding the number of missing words on the book page in Figure 3.3 with a clear explanation of how to figure it out (Wilburne & Kulbacki, 2014) was included in the HLT to encourage participants to consider variability.


"I made it the day you were rescued," I answer. The pain in my chest wraps around my ribs like a vise. The dancing was a mistake. "So what do you remember?"

"That's not an answer," he tells me. "I don't know what to think when they show me some of the tapes. In that first arena, it looked like you tried to kill me with those tracker jackers."

"I was trying to kill all of you," I say. "You had me treed."

Figure 3.3. The book page whose middle is empty in the third activity of the HLT

The activity required estimating the number of invisible words on the book page, whose middle is empty in Figure 3.3, and providing a clear explanation of how to figure it out. As can be seen from the figure, there are traces of some letters to the left and right of the empty. The paragraphs above and below the empty consist of lines of different lengths. One paragraph above has three lines, while the two below have four and two lines, respectively. The last lines of the below paragraphs involve only one word. Accordingly, the distribution of the visible words on the page necessitates addressing variability in fulfilling the task. In this context, the activity aimed to understand the fair sharing/leveling interpretation of the mean due to the different characteristic lines on the page. While finding the number of words, the PMTs could use the solutions as follows: 1) measuring the height of a visible paragraph, determining the number of words in the paragraph, and rating the number of words according to the height of the empty, 2) determining the number of words in a visible region, dividing it by the number of the visible line, determining how many lines are in the space, and multiplying the number of words obtained by the

number of lines in the empty, and 3) categorizing the lines, such as tabbed, complete, or incomplete, finding the average number of words in each category, determining how many lines in each category are in the space, multiplying these numbers by the average number of words obtained for each category, and adding the multiplication results.

In weekly meetings, we decided that it was essential to understand the balance point in addition to the fair sharing/leveling interpretations of the mean. It was assumed that interpreting the mean as a balance point could provide insight to the PMTs about how the data is distributed. In other words, we hypothesized that analyzing both the center and variability of the data is essential for reasoning about distribution. Accordingly, the "measuring variability in distributions with small sample sizes-soccer game" activity, where the mean could be interpreted as a balance point, was added to HLT. It was the first activity within the HLT to use numerical data, following a series of activities that predominantly featured qualitative data. The activity asked which of the two soccer teams was better, having scored the same average points from the soccer games they played. The points scored by the teams in the activity were given in dot plots, as illustrated in Figure 3.4.

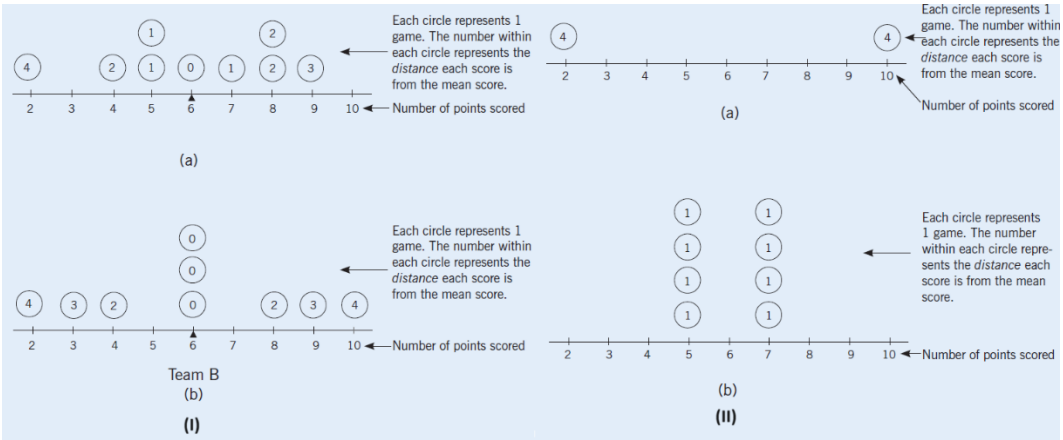


Figure 3.4. The points scored by the two teams in each of nine games (I) and two games (II) (Kader & Mamer, 2008, pp. 39-40)

In the plots, the dots were above the points, and the distances from the mean were written inside of the dots. The task of the PMTs was to choose the better team by examining the two situations. In the first situation (I), both teams played nine matches, while in the second (II), Team A played two matches, the other eight matches. What was common in both situations was that the teams had equal average scores. In other words, distributions in the situations had equal and unequal sample sizes but the same mean and median. Therefore, the PMTs could determine the better team based on variability. Variability in the distributions of the scores in soccer games could be measured depending on the range of the scores and the distance of the scores to the mean. More specifically, the PMTs could use the solutions as follows: 1) reading the highest and lowest scores from graphs and comparing one or both of them, 2) determining the differences between the highest and lowest scores and deciding according to the range, 3) calculating the distance from the center of all scores and adding the deviations of each score, and 4) adding the distance of each score from the mean score and dividing the sum of absolute deviations by the sample size. In the last two solutions, the mean could be interpreted as a balance point depending on the absolute deviations from the mean- the sum of absolute deviations

(SAD) and the mean of absolute deviations (MAD). Accordingly, the soccer game activity could provide the PMTs to establish a relationship between center and variability by considering the balance point interpretation of the mean.

Given the assumption that the PMTs could make sense of the center and variability of data using statistical measures in previous activities within the HLT, the subsequent activity chosen was the "building arguments based on statistical measures that are appropriate for the context-batteries." Regarding center and variability, the comparisons based on range, consistency, total, and the mean of the data sets constituted valid arguments to decide which battery brand was better. More specifically, the battery brand with more variability would be inconsistent, or the battery with a higher lifespan would be better. In the activity, the PMTs were expected to justify their arguments by structuring or organizing the data differently. The activity in the instructional sequence presented the lifespans of 10 batteries from two different battery brands using horizontal bar graphs in Minitool. The tool enabled the sorting and partitioning of data in various alternative ways. Similarly, TinkerPlots, which was used in this study, also possessed these functionalities. However, the PMTs still needed to become familiar with TinkePlots. For this reason, the data of the activity in the HLT was represented in a table instead of a graph. Thus, data input to TinkerPlots could be introduced on this data. In the HLT, similar to sequence, it could provide the ability to see that the bars represented individual data values. In deciding which battery brand was better, the PMTs were envisioned to use the following methods: 1) dividing the batteries into the top ten and identifying one with more frequency, 2) reading extreme values, searching for the brand with larger maximum values, and avoiding small minimum values, 3) calculating the total or mean and determining the higher one, and 4) looking at the slight difference between data values based on a reference point for consistency.

In the "building arguments based on statistical measures that are appropriate for the context-Batteries" activity, the reasoning could depend on range, consistency, and total by using different statistical concepts, such as minimum and maximum values, SAD, MAD, and the mean. The PMTs compared the distributions based on total and mean for longevity, and range and SAD for consistency. Regarding the context of the statistical problem, they put forward two valid arguments focusing on the total or consistency. Since both arguments concluded that the same brand of battery was better, in the next stage, it was decided to give a context in which they could experience a challenge due to the different results reached with these arguments. The next "watermelons" activity in the sequence met the condition. In the activity, the data consisted of the weight in pounds of 10 watermelons randomly selected from two fruit distributors. Different watermelon distributors were preferred according to total and consistency. The differences in the arguments showed how important context was for solving the statistical problem and that there was no single correct answer. To reinforce the PMTs' ability to perform in different contexts, the "building arguments based on statistical measures that are appropriate for the context-watermelons" activity took place in the HLT. The few data values could be structured in the bar graph of TinkerPlots, which also supported seeing the bars representing individual data values as in the previous activity. The PMTs could solve the problem regarding watermelons as follows: 1) calculating the total or mean and determining the higher one, 2) reading extreme values, calculating the difference between the minimum and maximum values, and searching for one with less spread, and 3) dividing the watermelons into the top ten or different number based on a cut-point and identifying one with more frequency.

The random selection of the watermelons was also crucial in the activity. The weights of watermelons from the distributors could be different from the data presented. In this regard, the random sampling enabled the PMTs to infer that the watermelons bought later from the distributors would possess similar weights to the

data in the activity. Based on the reasoning about sampling, to pay attention to the relationship between the sample and population, following the "building arguments based on statistical measures that are appropriate for the context-Watermelons" activity, the HLT involved "generating different samples in weight context-Hot air balloons" activity in the fifth order of the instructional sequence. This activity was different from the others in that it had no specific data other than the piece of information that a hot air balloon could carry eight adults, excluding the balloon driver. The PMTs' task was to determine how many 7th-grade students could safely ride this balloon, taking into account only the weight. It was envisaged that the PMTs would proportion the means of adults and students or divide the estimated total weight that the balloon could carry by the average weight of the students. Different samples of a population riding the balloon could have been, provided that the average and the number of people remained constant. Based on a representative sample of the population, we extended the content of the activity in the weekly meetings. Besides solving the problem, it was decided that the PMTs were asked to invent data according to the results they found and inscribe it in TinkerPlots. Thanks to the simulation feature of the tool, it was planned to reach the population by increasing the number of samples generated or combining actual weight data as in the study of Bakker (2004, pp. 219-222). The revisions to the activity aimed for the PMTs to interpret how the data was distributed, taking into account variability beyond the mean, which would lead to the reasoning about the shape of the distribution.

All activities after the "braking distance" activity in the instructional sequence were aimed at interpreting the shape. The "braking distance" activity aimed to construct and justify arguments according to the problem context. The aim was also included in the previous activities regarding batteries and watermelons. On the opposite of these activities, the "braking distance" activity focused on the lowest values for the fastest stopping of cars. The activity was excluded from the HLT for several reasons. The first reason was that it did not address a new learning goal. The learning goal to

justify arguments in the problem context was achieved in previous activities. The other reason was the data of the activity. The data was braking distances of cars in two brands measured ten times under the same conditions. The ten braking distances could be represented in a bar graph in TinkerPlots. The number of data needed to be higher to think of the shape. The PMTs had already dealt with larger data in the "generating different samples in weight context-hot air balloons" activity. Therefore, the HLT followed the subsequent activities with a larger sample size in the instructional sequence.

More extensive data were presented with dot plots in the following activities of the instructional sequence. To make reasoning about the shape, it was first aimed to understand the shift in hill, clump, or cluster of data with equal sample size. The "helmets" and "resting heart rate" activities in the instructional sequence served this purpose. In the "helmets" activity, the effect of using the heat liner on reducing body temperature was investigated, with 60 football players with lining in their helmets and 60 without lining. The distribution of the data of both groups looked right-skewed so that a decision could be made according to the majority. In the "resting heart rate" activity, it was aimed to decide whether smoking affects resting heart rate. Data from heart rate measurements of 84 smokers and 84 non-smokers were represented in Figure 3.5. Non-smokers were skewed to the right, while smokers appeared more evenly spread. This allowed data to be structured differently in TinkerPlots, such as equal grouping and partitioning into equal-width intervals beyond the majority. For example, the PMTs would compare the numbers below or above a cut-point and identify one with more frequency. The "understanding the shifts in hill, clump, or cluster of data with equal sample size-Resting heart rate" activity was included in the HLT because of the opportunity to use different solution strategies and the larger sample size. The activity related to helmets skipped in the teaching experiment took place in the pretest (Appendix E).

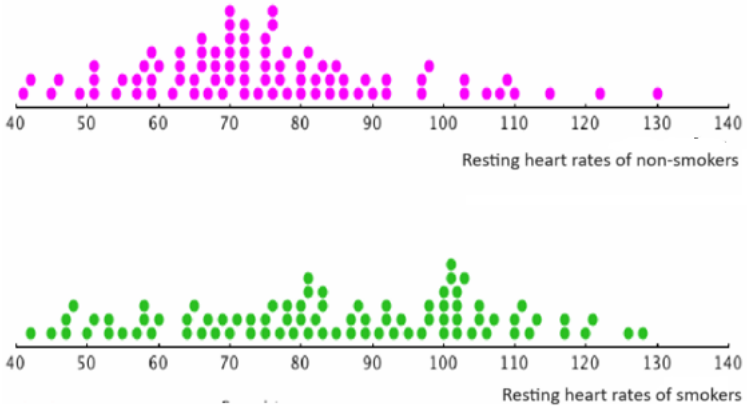


Figure 3.5. The data of the activity related to resting heart rate

When comparing distributions with equal sample sizes, using absolute frequencies was appropriate. Relative frequency could also be used to compare the distributions. However, the absolute frequencies could lead to misinterpretations in distributions with unequal sample sizes due to different proportions. Using relative frequency would yield more valid results for the distributions with unequal sample sizes. To recognize the differences between the use of absolute and relative frequencies, it was necessary to examine situations where the number of samples was both the same and different in the same context. The goal was achieved by backpack activities in the instructional sequence. Accordingly, it was decided to include these activities, which were important in the transition from absolute to relative frequency, in the HLT.

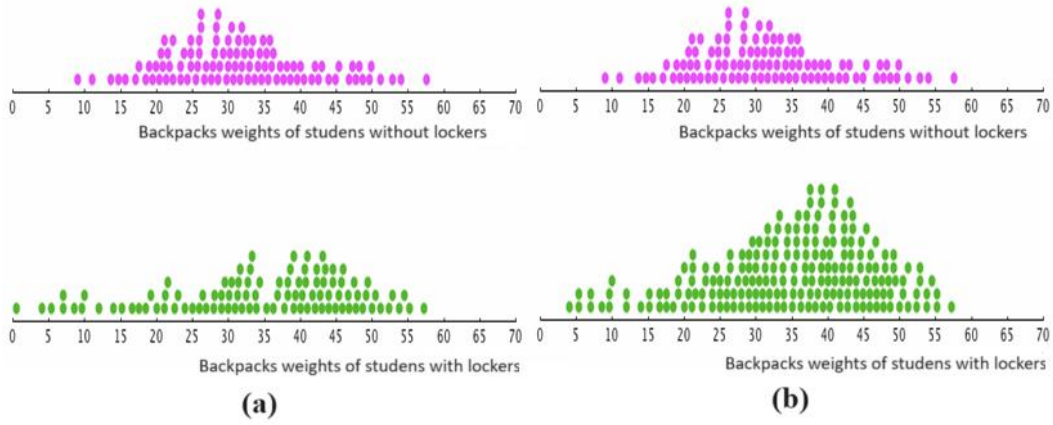


Figure 3.6. The data of the activities related to backpack weights of the same (a) and different (b) numbers of students

The activities required investigating whether the school lockers reduced the weight in students' backpacks. In the first activity with data in Figure 3.6a, the data consisted of the weights of the backpacks of 100 8th-grade students with and 100 without lockers. In the other activity with data in Figure 3.6b, the number of students with lockers increased. There were 199 students with lockers and 100 students without lockers. It means that students with lockers were about twice as many as those without. That is to say, the ratio had to be considered when making comparisons. In both situations, similar solution strategies could have been used, thinking of using relative frequency versus absolute frequency in the latter. In this regard, the "comparing distributions with unequal sample sizes-backpacks 2" activity in the HLT required the use of proportional reasoning due to different sample sizes. While developing arguments for the backpack context, the PMTs could reason about the shape depending on the location of the majority, as in the previous activity. The distribution of students with lockers was skewed to the left, while the distribution of students without lockers appeared more evenly spread in both situations. TinkerPlots' different data partition features could have been used to justify the

arguments. The PMTs could perform the methods as follows: 1) partitioning the data into two with a reasonable cut-point for the context and comparing absolute/relative frequency below or above the point, 2) partitioning the data into fixed-width intervals and describing the hill or comparing absolute/relative frequency in a reasonable interval, and 3) partitioning the data into equal groups and interpret being packed up or spread out in the bins in terms of density.

When comparing distributions, the PMTs could concentrate on context-specific parts of the data they structured or organized in TinkerPlots. For example, in the activities related to backpacks, it was enough to focus on low values in the left parts of the distributions to decide whether or not to install lockers for schools. However, the study was intended to reason about the data as a whole. That is to say, beyond concluding, the focus was "recognizing the pattern in the data and supporting the pattern with different graphical representations." According to Cobb (1999) and his colleagues, every part of the distribution was considered due to proportional reasoning when distributions were not evenly sampled. Therefore, in the HLT, care was taken to ensure that subsequent activities included unevenly sampled distributions.

The instructional sequence had four activities that required comparing distributions with unequal sample sizes after the "backpacks 2" activity. They were "AIDS", "anorexia", "migraine", and "ambulance". All of them were related to health. In "AIDS" activity, two treatments were administered to increase the T-cell count of AIDS patients to treat the disease. Similarly, in the "anorexia" activity, two treatments aimed to increase the weight of anorexia patients. Conversely, in the "migraine" activity, the drugs aimed to reduce the patients' relief time after taking one of the two drugs. In the "ambulance" activity, reducing the response time for ambulances to move faster to the scene was essential. In other words, two out of these four activities necessitated reasoning in favor of high values, while the

remaining two called for reasoning in favor of low values. The time in the teaching experiment process was insufficient to implement all four activities. Therefore, it was decided to implement two activities among them, considering the criteria upon which the activity contexts were oriented. More specifically, it was agreed that one would focus on decreasing and the other on increasing. The "anorexia" and "ambulance" activities were selected for the HLT in weekly meetings. In the HLT, the activities were called recognizing the pattern in the data and supporting the pattern with different graphical representations. Since the shapes of the distributions of the "AIDS" and "migraine" activities indicated the answer, they were not included in the HLT. For example, in the "AIDS" activity, one distribution was right-skewed while the other was left-skewed. Since they were distributed in almost equal ranges, it could be concluded that the left-skewed distribution concentrated at higher values. The distributions of the activities related to anorexia and ambulance were more challenging as they overlapped. In this regard, the PMTs might further structure graphical representations in TinkerPlots to support their arguments. Namely, they could develop more evidence-based solution strategies.

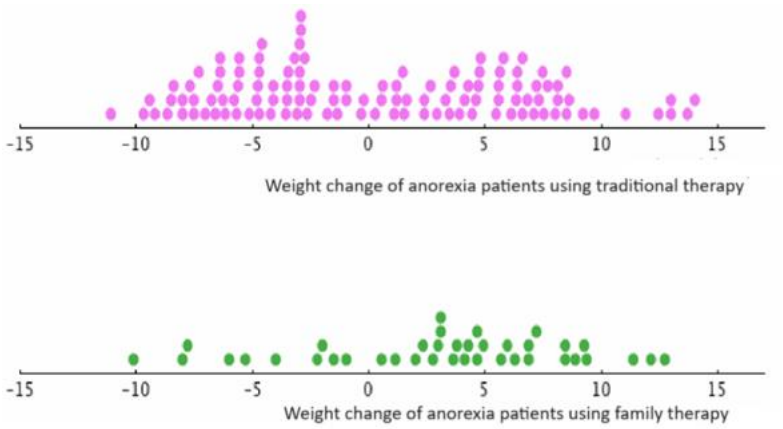


Figure 3.7. The data of anorexia patients using traditional therapy or family therapy

In the “recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia” activity, the weight changes of 117 patients using traditional therapy and 39 patients using family-based therapy in the two months after the start of the therapy were recorded. The data in Figure 3.7 consisted of negative values for weight loss and positive values for weight gain. The data distribution in both therapies was not skewed and was evenly distributed. Since it was difficult for the PMTs to see the pattern in the data, they needed to structure the data to interpret it according to the shape. In TinkerPlots, they could justify their arguments by partitioning the data in different ways. The PMTs would partition the data into two with a reasonable cut-point like 0 and compare the percentage or proportion in each distribution above the cut-point, indicating a weight gain. Below this cut-point was weight loss because of negativity. Another way could be that the PMTs partition the data in half and note the difference in the median’s location in the distributions. The PMTs could also interpret quartiles in terms of density by partitioning the data into four equal groups instead of two groups.

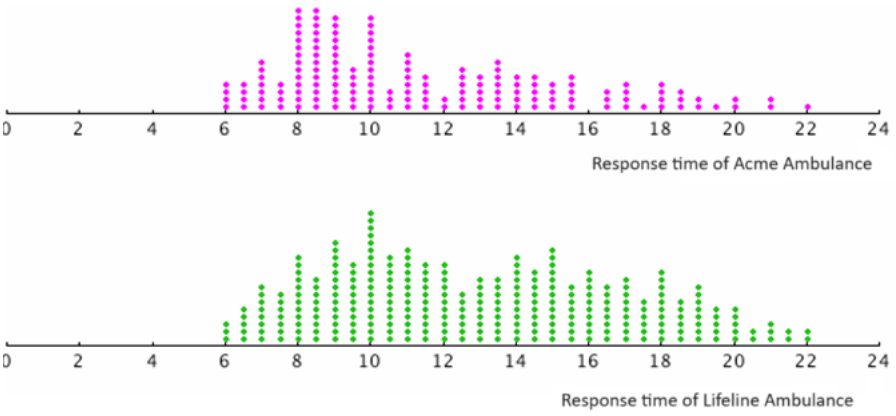


Figure 3.8. The data on response times of ambulances

The final activity in the HLT was the “recognizing the pattern in the data and supporting the pattern with different graphical representations-ambulance” activity.

The response time of two ambulance companies to a school's calls for one year constituted the activity data. There were 157 emergency calls for Acme ambulance and 258 for Lifeline ambulance (Figure 3.8). These numbers showed that the sample size increased compared to previous activities. That is, the PMTs would be interested in more extensive data. The Acme ambulance with the small sample size appeared to be right-skewed, while the Lifeline ambulance seemed evenly distributed. They were distributed within the same range. The PMTs might have thought that the right-skewed distribution was more advantageous. Given that the school's criterion for selecting one of the two companies was based on achieving a faster response time, it would have been reasonable to prefer a right-skewed distribution with a concentration of low values. However, there was no one-to-one correspondence between the distributions because the sample sizes differed. Lifeline Ambulance might have looked more spread out because it had more data. The difference between the number of samples required both using relative frequency instead of absolute frequency and structuring the data differently. As in the previous activity, the PMTs could have compared relative frequencies according to a reasonable cut-point and partitioned the data into equal groups or intervals with equal widths in TinkerPlots. In the background, they would obtain a box plot if they divided the data into four equal groups and a histogram if they divided the data into equally spaced intervals.

As a result, it was decided to implement 12 activities in the classroom teaching experiment under the guidance of the HLT. Table 3.2 presents the last version of the HLT implemented in the classroom teaching experiment to improve PMTs' reasoning about distribution.

Table 3.2 The last version of the HLT implemented in the classroom teaching experiment

Taken-as-shared interests/Activity	**Topics of statistical discourse**	**Imagery and possible PMT actions**
1. Aggregating different ranked data-printer *Assignment 1:* What did we learn?	Identification of the data occurring most frequently (mode), Explaining the relationship between totaling and averaging the data ranks	Converting categorical data to numerical data Re-ranking the data based on the mode, total or mean
2. Estimating an unknown whole through the average- tracking a herd *Assignment 2*: Teaching the mean meaningfully (Uccellini, 1996)	What is the average? Explaining the relationship between the total and the mean	Making groups/grids/columns/rows in the picture and iterating one of them to the whole in a reasonable way
3. Estimating an unknown part through the average-missing words	Connecting center and variability Interpreting the mean- fair sharing/leveling	Measuring the height of different parts of the whole and rating the quantities Iterating categorized parts to the whole in a reasonable way
4. Measuring variability in distributions with small sample sizes-soccer game	Connecting center and variability Interpreting the mean- balance point Understanding the properties of measures of variability	Measuring variability with the range, interquartile range, SAD, MAD, standard deviation, or variance
5. Building arguments based on statistical measures that are appropriate for the context-batteries *Assignment 3:* Data invention	Justifying the statistical arguments about the context based on the range, consistency, total, mean, etc.	Calculating statistical measures in TinkerPlots Ordering the data in the bar graphs using TinkerPlots Counting data above or below a reference line in TinkerPlots
6. Generating different samples in weight context-hot air balloons	Thinking about average within variability Understanding representativeness-the relationship between samples and populations Interpreting the shifts in data	Proportioning between different measurements of the samples Generating data in different graphs in TinkerPlots Increasing sample size of different distributions by sampler in TinkerPlots

Table 3.2 (continued)

7. Building arguments based on statistical measures that are appropriate for the context-watermelons	Justifying the statistical arguments about the context based on the range, consistency, total, mean, etc.	Calculating statistical measures in TinkerPlots Ordering the data in the bar graphs using TinkerPlots Counting data above or below a reference line in TinkerPlots
8. Understanding the shifts in hill, clump, or cluster of data with equal sample size-resting heart rate	Interpreting the frequencies in different parts of data representations Capturing the bin on more density	Creating dot plots by partitioning data with equal width/count dividers or dragging and dropping the data points in TinkerPlots Counting data above or below a reference line in TinkerPlots
9. Comparing distribution with equal sample sizes-backpack 1	Interpreting the frequencies in different parts of data representations Capturing the bin on more density	Creating dot plots by partitioning data with equal width/count dividers or dragging and dropping the data points in TinkerPlots Counting data above or below a reference line in TinkerPlots
10. Comparing distribution with unequal sample sizes-backpack 2 *Assignment 4:* Understanding Center and Spread (Kader & Mamer, 2008)	Reasoning about shape Interpreting the frequencies in different parts of data representations Justifying the use of proportional reasoning	Creating dot plots by partitioning data with equal width/count dividers or dragging and dropping the data points in TinkerPlots Calculating ratio or percentage in the structured data representations in TinkerPlots
11. Recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia	Reasoning about shape Interpreting the frequencies in different parts of data representations Justifying the use of proportional reasoning Talking about box plots and histograms Thinking of the density in narrow or wide intervals	Creating dot plots by partitioning data with equal width/count dividers or dragging and dropping the data points in TinkerPlots Calculating ratios or percentages in the structured data representations in TinkerPlots Converting structured dot plots to box plots or histograms in TinkerPlots
12. Recognizing the pattern in the data and supporting the pattern with different graphical representations-ambulance	Reasoning about shape Interpreting the frequencies in different parts of data representation Justifying the use of proportional reasoning Talking about box plots and histograms Thinking of the density in narrow or wide intervals	Creating dot plots by partitioning data with equal width/count dividers or dragging and dropping the data points in TinkerPlots Calculating ratios or percentages in the structured data representations in TinkerPlots Converting structured dot plots to box plots or histograms in TinkerPlots

3.3.2 The Design Experiment

The second phase comprises conducting the design experiment. My supervisor and I enacted the classroom teaching experiment within the MSE482- Projects in Mathematics and Science Education elective course in the 2020-2021 spring semester. The main objective of this course was to help the PMTs study theoretical and practical needs regarding mathematics education. On the one hand, the focus of this course is on carrying out the current research study and, on the other hand, on improving the statistical reasoning of the PMTs. While doing this, the statistical content in Table 3.3 in the context of learning activities was dealt with. The table was the final version of the syllabus schedule followed in the course.

Table 3.3 The schedule was followed in the course

Week 1	Overview and organization of the course- Assessment *(Pretest)*
Week 2	What is statistics?
Week 3	Data, Data Collection and Organizing
Week 4	Average- Measures of center
Week 5	Average- Measures of center and variability
Week 6	Variability- Measures of variability
Week 7	Comparing distributions- Shape
Week 8	Comparing distributions- Density
Week 9	Comparing distributions with unequal sample sizes
Week 10	Comparing distributions with unequal sample sizes
Week 11	Hypothesis testing
Week 12	Bivariate Distribution/ Correlation
Week 13	Project Presentations
Week 14	Assessment *(Posttest)*

As can be seen from the table, assessments were made instead of instruction in the first and last weeks of the course, which lasted 14 weeks in the semester. In Week 13, projects prepared by groups comprising two PMTs were presented. It was one of the requirements of the course. The course about teaching and learning in statistics education was quite comprehensive. It ranged from aspects of doing statistics to inferential statistics in higher education. The course content included statistics education beyond the middle school curriculum because of hypothesis testing and bivariate distribution/correlation (MoNE, 2018a). In addition, the course had a pedagogical dimension as well as statistical content in the table. Since the study focused on the PMTs' reasoning about the distribution, the period between Week 3 and Week 10, the bold part in the table, was taken within the scope of the dissertation by ignoring the pedagogical issues. The content of these weeks, including data, data collection and organization, the center, variability, density, and skewness aspects of the distribution, comparing distributions with different characteristics, was explained in detail in the previous section and summarized in Table 3.2.

The classroom teaching experiment was conducted over eight weeks, three hours a week. Each class session of the experiment was 50 minutes long, with a total of 24 class sessions. There was a 10-minute break time between the class sessions. Every week, we acted in an anticipatory thought experiment. The thought experiments were conducted about a week before the implementation. They were on the instructional activities to be carried out in the classroom, how they would be realized, and what the PMTs would learn from them. Since these were weekly, three class sessions in a week constituted a microcycle of the design experiment. In thought experiments between each microcycle, we retrospectively evaluated the previous microcycle and planned the next microcycle. The cyclic process included testing the conjectured local instruction theory formulated in the preparing for the experiment phase and (re)designing the instructional activities, classroom norms, and specific aspects of the design to understand how the theory worked in the actual process (Gravemeijer & Cobb, 2006).

The local instructional theory of the study was the HLT in Table 3.2. It included "taken-as-shared interests/activities", "topics of statistical discourse", and "imagery and possible PMT actions" when dealing with the activities (Simon, 2020). The HLT emerged throughout this ongoing process. The first version of it in the first thought experiment before the first microcycle was the initial instructional sequence that Professor Cobb provided. In each weekly thought experiment, this sequence was revised and became the HLT. The reasons for the revisions in the content of the HLT were explained in the previous phase. There were also revisions in the implementation of the activities.

At the beginning of the teaching experiment, more time was devoted to implementing activities to establish classroom norms. In the early days, one activity could be practiced per week. More specifically, in the first five weeks, the PMTs worked on "aggregating different ranked data- printer," "estimating an unknown whole through the average-tracking a herd", "estimating an unknown part through the average- missing words", "measuring variability in distributions with small sample sizes-soccer game", and "building arguments based on statistical measures that are appropriate for the context-batteries" activities, respectively. In the fourth thought experiment, it was decided to implement the "building arguments based on statistical measures that are appropriate for the context-watermelons" activity after the activity related to batteries in the next week. However, the activity related to batteries was maintained for the entire fifth week. Because the PMTs encountered TinkerPlots for the first time this week and practiced with the data of the activity. In addition, these activities were similar to the learning goals and the structure of the data. In both cases, the PMTs would justify the arguments they developed for the problem context. Since justifying arguments was also necessary for all subsequent activities, implementing the activity related to watermelons was abandoned. Therefore, the sixth week continued with the "generating different samples in weight context-hot air balloons" activity. It was envisioned that this activity would last a week due to adding data generation and different samples to its content. However,

the PMTs completed the activity in two class sessions. In the last class session of the sixth week, the "building arguments based on statistical measures that are appropriate for the context-watermelons" activity was applied as it was ready from the previous week.

In other weeks, the PMTs worked on at least two activities in the HLT. They could complete three learning activities in the seventh week. These were "understanding the shifts in hill, clump, or cluster of data with equal sample size-resting heart rate", "comparing distributions with equal sample sizes-backpacks 1", and "comparing distributions with unequal sample sizes-backpacks 2" activities. The PMTs studied the "recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia" and "recognizing the pattern in the data and supporting the pattern with different graphical representations-ambulance" activities in the last week of the teaching experiment. They might have completed the activities faster than at the beginning for two reasons. The first one might be the technological tool used. Dot plots represented the data of the last five activities. In TinkerPlot, the plot button was indexed to dots for a variable. A dot plot was formed when the PMTs dragged the variable to the horizontal or vertical axis. The PMTs quickly adapted to using the different features of TinkerPlot as they also practiced the dynamic statistical software bought for them for a year outside of class. The second reason may be that they are used to the classroom norms. Because the classroom norms described below were established at the beginning of the teaching experiment, they became stereotyped towards the end. The PMTs could foresee the instructor's expectations, how to work on the activity, and what to do next.

The instructor was my supervisor. She implemented the learning activities in the HLT in one classroom. Unlike the study conducted with 7th graders (Cobb, 1999), this process took place on an online learning platform due to the COVID-19 pandemic. The platform was the Zoom application, which had breakout rooms for working in small groups. Additionally, as a participatory observer, the researcher, who was also the teaching assistant of the course, also intervened in implementing

learning activities. In addition to the weekly thought experiments, the researcher took an active role in presenting the activities and data files to the participants, the questions and their order in the presentations in the background, and especially the applications of TinkerPlots for the activities. Activities, visuals, or graphical representations in the activities were introduced by sharing the PowerPoint files on the screen. Data files of the activities were sent to the participants from the chat box of the Zoom application or via WhatsApp. While the data was in spreadsheets at the beginning of the teaching experiment, later on, according to the progress of the PMTs' use of TinkerPlots, the data was directly in the TinkerPlots file.

The instructional learning activities in the HLT were implemented based on the emergent perspective, the theoretical framework of the study. That is, the learning environment was organized with the assumption that learning occurs in the social context of the classroom (Cobb & Yackel, 1996; Gravemeijer & Cobb, 2006; Yackel & Cobb, 1996). Concerning this framework, which interprets classroom discourse and communication, the following path was pursued while implementing the activities. Firstly, the whole class discussion was held to understand the context of the instructional activity. The discussions were on the variables and the critical criterion for the context of instructional activities and the data generation process. For example, in the first activity, the PMTs discussed issues such as the use of a printer, what features a printer can have, and whether they have a printer or not in printer activity. Based on the conversations, they identified the critical variable for context. Although they experienced generating the data in the first activity, they discussed how data in the visuals or graphs would be collected for the critical variable identified in other activities. For example, in the "recognizing the pattern in the data and supporting the pattern with different graphical representations-ambulance" activity, they agreed on the response time as the measured time between the call being made and the ambulance arriving at the scene.

Secondly, individual or small group work was put in which participants analyzed the data in the activities. Participants performed data analysis in groups of 4-5 people

except for the "measuring variability in distributions with small sample sizes-soccer game" and "building arguments based on statistical measures that are appropriate for the context-watermelons" activities, which were held individually. The instructor expected the PMTs to solve statistical problems in multiple ways with acceptable explanations and justifications. She established the classroom *sociostatistical* norms by challenging them to generate more than one solution method and to justify their solutions by providing evidence. The different, sophisticated, effective, and acceptable solutions and explanations enabled the PMTs to learn with intellectual autonomy (Cobb, 1999; Cobb & Yackel, 1996; Gravemeijer & Cobb, 2006). During this ongoing process, the instructor and participatory observer switched between the groups to monitor the participants' work. In groups, the participants put all their solutions in writing on the whiteboard application of the online learning platform. They also saved the inscriptions in TinkerPlots that they worked on as files. In this process, the researcher also quickly took notes of the different, complex, and valuable arguments for the learning purpose to determine the sharing order of the solutions produced in the groups in the following whole class discussion.

After solving the statistical problems in groups or individually, the breakout rooms were closed and everyone gathered in the main room. In a synchronous environment, the instructor asked the participants to present their arguments. Based on the researcher's notes, the whole discussion started with the most primitive solution. The group presenting this solution also explained their alternate solutions both verbally and in writing through screen sharing. Other groups similarly performed this process without repeating the same solution methods already produced by others. In the class discussion, the PMTs made sense of the solutions of others and compared and contrasted them with their own. They discussed which was more accurate or effective for solving the statistical problem in the activities. While this approach of the instructor set up sociostatistical norms, on the one hand, it also led to the emergence of classroom statistical practices (CSPs) because they evolved

explanations beyond justification for particular statistical ideas or concepts (Cobb & Yackel, 1996; Gravemeijer & Cobb, 2006; Yackel & Cobb, 1996).

The PMTs developed the statistical ideas or concepts in a classroom culture designed according to RME. As the learning activities in the HLT were real for the PMTs, they reinvented the ideas and concepts in the learning goals of the activities. In other words, they reasoned and acted about the statistical problem in the activities in a personally meaningful manner (Gravemeijer & Cobb, 2006). They realized the phenomenological exploration by mathematizing their statistical knowledge from the reality of learning activities or the PMTs' prior statistical experiences. In the classroom environment, the compelling learning activities generated the need for the PMTs to reason about the distribution concept. To analyze the data given in the activities in a reasonable and meaningful way, they had to consider how the data was distributed. While analyzing the data, after the "measuring variability in distributions with small sample sizes-soccer game" activity, they organized the numerical data in different graphical representations in TinkePlots. Thereby, they were able to scrutinize each part of the distribution to interpret the data reasonably. In the activities where TinkerPlots was used, different organizations of the data in the graphs allowed to reason about the distribution. The procedures used for statistical measures at the beginning of the teaching experiment also served as submodels to interpret the distribution. Namely, the *models of* activities transformed the *models for* reasoning about the distribution (Gravemeijer, 1994).

The PMTs could describe how the data was distributed in the activities with the analysis buttons of TinkerPlots (hats, counts, average, and meter). For example, mean, median, mode, and midrange could be calculated with the average and meter buttons, and variability measures such as SAD and MAD could be calculated with the hats button. The counts button showed the frequencies or percentages of the data in a particular space. These descriptions might be supported by structuring the data in different ways in TinkerPlots. The PMTs could structure the data by ordering, stacking, and dividing it into intervals or groups. In particular, different

representations and inscriptions were models for thinking about the distribution of PMTs. It was worth noting that when the PMTs were working on the activities, the instructor did not give clear guidance on how these models were used. Depending on the context of the activity or the characteristics of the data, the PMTs structured the data themselves. That is to say, there was no intervention in their thinking process. The participatory observer only provided technical information in case of difficulties according to the action the PMTs wanted.

While implementing the instructional activities in the HLT, the PMTs were also assessed. For this, four assignments outlined in Table 3.2, aligned with the learning goals, were incorporated into the HLT. They not only provided a formative assessment of what the PMTs had learned during the teaching experiment but also shaped the design of the subsequent learning sequence. At the end of the process, a posttest was administered to reveal how much the PMTs improved their reasoning about distribution. Semi-structured interviews conducted immediately after the test administration were also used to substantiate this development. Details about the data collection tools are provided in the following section.

3.3.2.1 Data Collection Tools

This study aims to investigate the development of PMTs' reasoning about distribution as they participated in the classroom teaching experiment based on RME and enriched by dynamic statistics software. Concerning the purpose, it clarifies the PMTs' statistical reasoning before and after the classroom teaching experiment and documents their collective development in reasoning about the distribution by analyzing CSPs emerging throughout the teaching experiment. For this, the study's data comprised the pretest and posttest, semi-structured interviews, video recordings of classroom and debriefing sessions, the PMTs' written artifacts, and field notes. The data collection tools are explained in detail below.

3.3.2.1.1 Pretest and posttest

Pretest and posttest were implemented to assess the PMTs' reasoning before and after the teaching experiment. They comprised five equivalent questions in terms of statistical content on the distribution. According to the distribution aspects, the questions were based on center, spread, density, and skewness (Bakker & Gravemeijer, 2004). While the first three questions required reasoning about aspects of a distribution, other questions were about comparing the two distributions using these aspects. Although the test questions are detailed below, the pretest and posttest are given in Appendix E and Appendix F, respectively.

The first question is the same in the pretest and posttest. It asks to determine the better class regarding the number of books read in four classes during a semester. The number of books for the classes is in Figure 3.9.

Below is the total number of books read by 4th-grade students in different classes in a school during a semester. The principal wants to go on a trip with the better class in reading. With which class do you think the principal should go on a trip? Explain your answer with its reasons.

Class	The total number of books students read during a semester
4A	11, 13, 9, 8, 9
4B	7, 6, 12, 9, 9, 11
4C	10, 9, 9, 10, 7, 12
4D	11, 13, 10, 2, 12

Figure 3.9. Question 1 in the pretest and posttest

As can be seen from the figure, the classes have data from different numbers of students. While six students in classes 4A and 4D reported the number of books they read during the semester, data for classes 4B and 4C was collected from five students. The means of the data sets are very close to each other. They are 10, 9, 9.5, and 9.6, respectively. There are notable differences in their standard deviations. The standard

deviation values are 2, 2.28, 1.64, and 4.39. Similarly, the ranges also differ. The range values are 5, 6, 5, and 11, respectively. In addition to these values, the mode and median are also essential to deciding the better class regarding reading books. The mode values are 9 for classes 4A and 4B, 9 and 10 for class 4C, and no for class 4D. For the median, the values are 9 for class 4A, 9.5 for classes 4B and 4C, and 11 for class 4D. In this question, the PMTs are expected to choose the class with high average and low variability for the better class. In this sense, the question requires reasoning about statistical measures. Class 4C may be the best because of both consistency and higher average. However, the criterion of high average and low variability can also be achieved by excluding the outlier, 2, in class 4D.

Question 2 in the pretest and posttest (See Figure 4.2) are similar in structure. It is a true-false item. There is a left-skewed distribution in the pretest instead of a right-skewed distribution in the posttest. The questions were asked to evaluate the shape of the distribution according to the positions of the measures of center (mean, median, and mode) and the location of half of the sample for the appropriate average. That is, they are related to reasoning about the shape of the distribution and the interpretation of statistical measures. Garfield (2002, p. 5) categorizes "understanding why measures of center, spread, and position tell different things about a data set *and* knowing which are best to use under different conditions" in reasoning about statistical measures. The questions also include reasoning about representations of data as they require "understanding the way in which a plot is meant to represent a sample, … being able to see beyond random artifacts in a distribution to recognize general characteristics such as shape, center and spread" Garfield (2002, p. 5). While reasoning about the shape and measures, the posttest requires an explanation of the reasons for choosing each item; the pretest does not have this explanation part. The disclosure for the pretest was made in semi-structured interviews before the teaching experiment. For Question 2, the PMTs are expected to choose false, true, true, and false in the pretest and true, false, true, and false in the posttest, respectively. Since the median is greater than the mean in the left-

skewed distribution in the pretest, exactly half of the students are above the average if the median is average. Similarly, since the mean is greater than the median in the right-skewed distribution in the posttest, exactly half of the students are below the average when the median is the average. Accordingly, the PMTs will be able to name the shape of the distribution, infer the position of the measures of the center, and interpret the average for these positions.

Professor Cobb sent some assessment items in the appendix with the instructional activities. The remaining questions of the tests were retrieved from these assessment items. Question 3 in the tests (See Figure 4.3) is one of them. It reports the data in the horizontal bar graph showing the money earned by 36 graduates working as graphic designers in their first year. It is related to reasoning about representations of data. It consists of two parts. In the first part of the question, there are three statements in the pretest and four in the posttest. They give information about the minimum and maximum values of the distribution, the average (mean), and the range where the majority of the data falls. The extra statement in the posttest specifies the quartiles. The first part asks for the choice of a useful statement representing the data. In the second part of the question, it is expected to explain the reasons for the choice. While reading the data in the salary context, judging the extreme values, interpreting the majority, and drawing conclusions based on the average are the focus. In other words, the question requires reasoning about representations of data to recognize the general characteristics of the distribution, such as shape, center, and spread.

Question 4 requires comparing two distributions with equal sample sizes beyond interpreting aspects of a single distribution while reasoning about shape and shifts in data. The fourth questions in the pretest and posttest (see Figure 4.4) are similar in structure but in different contexts. The pretest includes the “helmet” activity of the initial instructional sequence mentioned in preparation for the experiment section. In this question of the pretest, the effect of lining suggested to reduce heat exhaustion of football players is asked to be interpreted based on the temperature data inside the helmets of 60 players with lining and 60 without lining. Question 4 in the posttest

includes distributions showing the amounts of zinc per liter taken from 100 different river sites before and after a new treatment method. Its context is investigating the effect of the treatment developed to reduce the amount of harmful zinc in a river.

For Question 4 in tests, it is expected to choose the distribution with low values depending on the context. Since the spreads of the data are the same in the distributions, except for a few extreme values, it is possible to reason qualitatively over the shapes of the distributions. The distributions in the pretest are similar in terms of shape. However, in the posttest, the left-skewed distribution before the treatment has become skewed to the right after treatment. It is clear that the right-skewed distribution with the same range after the treatment has mostly lower data values in the posttest. The majority, between 55 and 65 before treatment, decreased to 30 to 40 after treatment. In Question 4 in the pretest, the temperature of the players using the lining clumped between 90 and 143, while the clump is between 108 and 162 for the players without the lining. The majority of the distributions with equal sample sizes and similar spread are also a predictor for the average. To present quantitative evidence, when the average of the distributions is compared in the pretest, the average of the distribution with lining (the mean-21.45 and the median-122.5) is around 122, and the other distribution without lining is approximately 142 (the mean-138.4 and the median-142.5). In the posttest, there is an average of 57 milligrams of zinc before (the mean-54.8 and the median-56.9) and 35 milligrams after the treatment (the mean-35.4 and the median-35) in the river. There is a difference of about 20 units in their average for the distributions in both tests. In summary, using lining reduced the temperature in the players' helmets, and the applied treatment decreased the amount of zinc in the river.

Question 5 in the tests is related to comparing distributions with unequal sample sizes. It requires more than one competence beyond that in Question 4. In addition to comparing distributions, relative frequencies should be compared rather than absolute frequencies in the distributions. In other words, proportional reasoning is needed. Accordingly, the last questions in the tests involve both comparing

distributions with unequal sample sizes and using proportional reasoning. Although the last question in the tests focuses on the same competencies, they are different in terms of context and structure (See Appendix E and Appendix F).

The last question in the pretest is about a company's plans to provide free flu shots to its employees. To determine whether the flu vaccine would reduce the number of days of absenteeism due to the flu, the data was collected on the number of days missed for work from 76 employees who got a flu shot by paying themselves and 162 employees who did not get a flu shot. The data is presented in two dot plots representing each employee's missed days in Figure 3.10. By analyzing the data with the support of different data representations, four company executives concluded the effects of the flu shot. The first two conclusions are presented with two-way tables in Part A and Part B. Other conclusions are illustrated by the range in Part C and the box plot in Part D. The question asks whether the company's CEO provides free flu shots to all company employees.

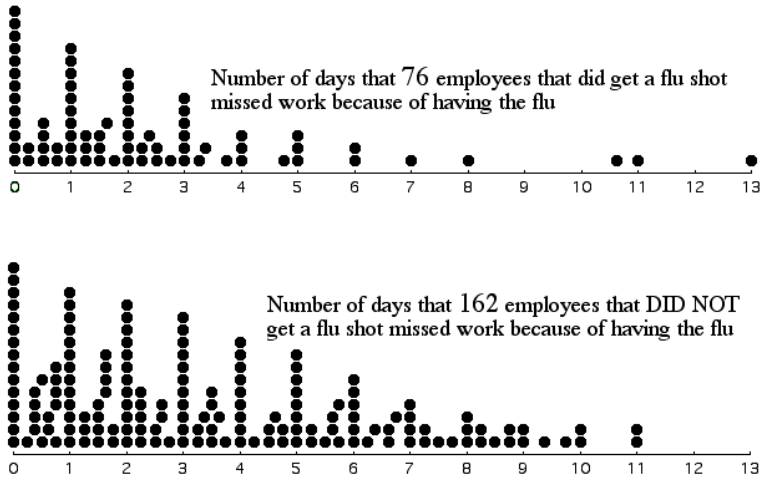


Figure 3.10. The data of the employees who got a flu shot and did not get a flu shot in Question 5 in the pretest

In Part A, absolute frequencies were compared by dividing the data into absenteeism and non-absenteeism. In Part B, taking one day as a reference, the relative frequencies of the data on percentages are interpreted for employees missing work one day or less and employees missing work one more day. Specifically, while Part A requires justification for using proportional reasoning, Part B indicates the way proportional reasoning is used. Since both parts focus on relative frequency due to the unequal number of samples, they measure the use of proportional reasoning. Other parts of the question are related to seeing the whole data with different representations. In Part C, the explanation emphasizes the magnitude of the range. It focuses on one aspect of the distribution, which is spread. Based on the context, a smaller range is reasonable but not enough. Because it needs to give more information about how the data is distributed between the extreme values. The executive in Part D compared proportionally how many people were absent for more than three days in the box plot. The three days are the median of the distribution of people who did not get a flu shot, which is also the third quartile -75 % of the people who got a flu shot. Since the median separates the data into half (50%-50%), the relative frequencies are compared with fractions above three days, a cut point. For the reasoning about distribution, the last two parts of the question expect to explain why the range is insufficient and what the three mean beyond the executives' conclusions. That is to say, they require reasoning about shape and shifts in data to compare distributions structured in the representations, e.g., range and box plots.

The last question of the posttest also requires comparing distributions with unequal sample sizes. It is about medicines developed by a drug company to relieve headaches. The company, with a formula for headaches, produced a new medication. To test the new medication's effectiveness, it was measured how many minutes 150 patients using the old formula and 100 patients using the new formula under the same conditions relieved. The obtained data is shown in two dot plots representing minutes for each patient. It is asked to decide whether to start marketing the new formula by analyzing the data. In doing so, the question consists of three parts. The first part

requires writing a summary of the effectiveness of the new medication based on dot plots in Figure 3.11. In the second part, this summary needs to be justified quantitatively using graphical representations. Since raw data was provided numerically in the posttest and TinkerPlots was used throughout the teaching experiment, it is allowed to use technology for drawing graphical representations. Part 2 is an extension to the original question that Professor Cobb provided. In the first two parts of the question, it is asked to write a summary with reason. In the third part, like the last question of the pretest, it is requested that the statements be evaluated and explained.

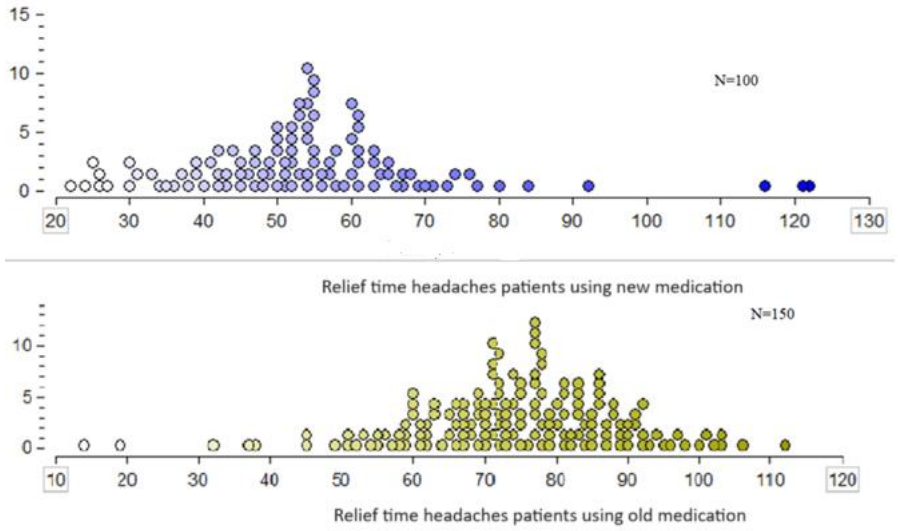


Figure 3.11. The data of the headache patients using new or old medication in Question 5 in the posttest

When reading between the data and considering the aspects of the distributions, the number of samples is not equal. Therefore, the comparison of frequencies is based on proportional reasoning. In terms of the shape of the distributions, unlike the old medication distribution, which is left-skewed, the new medication distribution seems right-skewed. Their dispersions are different. The new medication ranges from 22 to 122, while the relief time of old medication extends from 14 to 112. The outliers in

the distributions reverse the conclusion in favor of new medication. The relief time of the new medication clusters between 40 and 65 minutes, and the majority of the old medication is between 60 and 90. Considering that the average is consistent with the tendency of the distribution, the new medication provides faster relief than the old medication. This conclusion declines Statement A and Statement B based on the extreme values in Part 3 but confirms Statement C, interpreting the average. The distributions with unequal sample sizes in the question can be compared, as the averaging or juxtaposing relative frequencies in certain intervals eliminates the different number of patients in the two groups. It refuses Statement D in Part 3. In other words, using proportional reasoning in distributions with unequal sample sizes gives reasonable results. As seen from the data analysis above, the question requires reasoning about shape and shifts in data and using proportional reasoning when comparing the distributions with unequal sample sizes.

In summary, the questions in the tests are based on recognizing general characteristics of distribution, such as spread, center, and shape, and comparing the distributions with equal and unequal sample sizes for the context. They measure four skills that serve for reasoning about distribution. The skills are reasoning about statistical measures, reasoning about representations of data, reasoning about shape and shifts in data, and proportional reasoning. Table 3.4 presents these skills and corresponding questions in the tests.

Table 3.4 The skills that the questions measure in the tests

	Question 1	Question 2	Question 3	Question 4	Question 5
Reasoning about statistical measures	X	X			
Reasoning about representations of data		X	X		
Reasoning about shape and shifts in data				X	X
Proportional reasoning					X

3.3.2.1.1 Semi-structured Interviews

The semi-structured interviews were conducted with all PMTs individually after administering the pretest and posttest. However, the researcher could not interview with PMT6 after the posttest due to her health conditions. The interviewer, who was also the participatory observer in the class sessions, conducted the interviews on the online learning platform for about 30 minutes. The online interviews were arranged according to the interviewees' available times. As the interview questions were based on the PMTs' responses to the questions in the tests, the interviewer shared the interviewee's answers on the screen during the interview. The entire interview process was recorded with verbal permission from the interviewee.

Although there was no specific interview form, the primary purpose of the interviews was to elaborate on the PMTs' responses. After completing the tests, the interviewer immediately scanned the responses and identified ones that needed clarification. The researcher wrote reminder questions next to the answers she printed out before the interview. The second question of the pretest includes only true-false items. All interviewees explained the reasons for their true or false choices in the interviews. In other questions, although explanations were expected, some answers were

insufficient to understand how the PMTs thought and included unclear expressions. In the semi-structured interviews, they were asked to think aloud about the questions they solved. For example, in the first question in the pretest, PMT7 wrote, "The 4A class students can go on a trip because according to the class averages, the highest average belongs to this class. Also, the differences between numbers are the lowest in this class." In this response, it was unclear which statistical measure PMT7 meant by average. She also needed to state the values to compare the averages and the differences between the numbers of the classes. Therefore, in the interview, questions were asked about what averages were, how they were determined, and why she was looking for the highest average and the smallest difference between numbers, and so on for the first question. In addition to what the PMTs wrote, extra questions were asked about what they said during the interview if they stated. Therefore, the interviews were shaped according to the interviewee's responses.

3.3.2.1.2 Video Recordings of Classroom and Debrief Sessions

The online learning platform where the classroom teaching experiment was conducted had a recording feature. All class sessions were recorded on this platform. Three breakout rooms were created weekly when the PMTs worked in small groups for data analysis. The host of the learning platform, the researcher, could only record one small group work. Also, when she moved between groups, the group work started to be recorded was kept from being stored. The first three weeks of the teaching experiment consisted only of videos recorded by the host. Three PMTs who changed weekly were randomly assigned to record their group work in the following weeks. After the session was over, they forwarded the recordings to the researcher. Thus, the small group works were also recorded. After completing the teaching experiment, the researcher transcribed all the recordings word by word for each week. The transcripts included screenshots from the recordings and remarkable notes about the PMTs' thinking and learning process in the learning environment.

The researcher and instructor met weekly before and after each class session to design the classroom teaching experiment. The meetings taking place on the online learning platform were recorded. Anticipatory thought experiments were employed for the instruction in the meetings before the class sessions. They were related to the HLT to be implemented and the materials to be used. For example, the content and sequence of the questions, visuals, data, etc., in the files and presentations to be used in the classroom sessions were designed in these meetings. Moreover, the researcher and instructor evaluated the prior session by meeting again each week after the class sessions. The meetings formed a basis for meetings before the next class session reciprocally. The researcher watched the recorded meetings and noted the conjectures for the subsequent class sessions. These conjectures were reviewed in the meeting in the same week, and the following class sessions were designed. In other words, two debriefing sessions between weekly class sessions proceeded cyclically.

3.3.2.1.3 The PMTs' Written Artifacts

Another source of data for the study was the PMTs' artifacts. During the teaching experiment, the PMTs wrote reflection papers for four assignments (See Table 3.2). These reflection papers provided insight into the progress of the PMTs in this process. They completed the first reflection paper after the first week. To analyze what the PMTs learned about the printer activity, the first assignment asked them to provide feedback on their thoughts or feelings about the lecture. For the second reflection paper, the paper of Uccellini (1996) about teaching the mean meaningfully was assigned in the second week. The PMTs were expected to describe and interpret the mean in their own words. In the fifth week of the classroom teaching experiment, the PMTs worked on "building arguments based on statistical measures that are appropriate for the context-batteries" activity. Related to the context, in assignment 3, the mission of the PMTs was to invent hypothetical data for three battery brands

according to specific criteria. The criteria were as follows: Brand A is good and reliable; Brand B is good but unreliable; and Brand C's spread is like of A, but the brand is less good. To practice with the technological tool, the PMTs were asked to represent the data in TinkerPlots and explain why it met the criteria. They could report their representations with screenshots in the TinkerPlots file. Alternatively, they could annotate with text directly in the TinkerPlots file. Therefore, the TinkerPlots files with the data were also among the artifacts. Regarding the final assignment, the study of Kader and Mamer (2008) was given in the seventh week. The study provides information about measures of center and spread and includes students' interpretations of dot and box plots for a class activity on memorizing words. Therefore, in the last reflection paper, the PMTs were asked to define variability, discuss the advantages and disadvantages of dot and box plots, and evaluate students' statistical reasoning.

3.3.2.1.4 Field Notes

The researcher was a participant observer during the classroom teaching experiment. She noted the specific classroom events she observed, experienced, and thought as important for reasoning about distribution. She usually focused on the formal and informal thinking of the PMTs, the progress of the class sessions, suggestions and reminders for the next class sessions, etc.

3.3.3 The Retrospective Analysis

The last phase of design research is the retrospective analysis. Although an analysis is carried out after all the data sets are collected, retrospective analysis is enacted during the experiment. The main purpose of retrospective analyses is to contribute to the development of local instructional theory (Gravemeijer & Cobb, 2006). In this study, the complete data set was analyzed in an iterative way to develop the HLT for

reasoning about the distribution. Most of the data analyzed were video recordings of classroom sessions and weekly meetings to debrief. Tests and semi-structured interviews, participants' artifacts, and field notes also formed the data set that was analyzed simultaneously. For example, the analysis of data derived from the pretest and semistructured interview before the teaching experiment provided insight into the starting and end points of the teaching experiment. Other data sets provided formative evaluation in the ongoing process of the teaching experiment. In this regard, especially the weekly meetings where thought experiments were conducted before and after each class session were important. In these meetings, the researcher and instructor refined or refuted conjectures, claims, or assertions about the PMTs' learning processes based on their systematic back-and-forth data analysis. In this process, the learning activities that did not function as expected were eliminated, and those that were effective in the PMTs' learning were reconstructed. The adaptation of the instructional sequence for this study was described in the preparation for the experiment section. As a result of analyses, "a revised, potentially optimal instructional sequence" was obtained, which was the pivotal aim of the retrospective analysis (Gravemeijer & Cobb, 2006, p. 42).

In this study, a theory was developed at "the instructional sequence (local instruction theories) level" (Gravemeijer & Cobb, 2006, p. 46). It also contributed to the development of the domain-specific instruction theory, RME. By applying this sequence in the classroom, the instructor built a bridge between the scientific knowledge produced as a result of the research and the teaching practice (Gravemeijer & Cobb, 2006). The HLT, tentative or proposed during the teaching experiment, was shaped according to the collective reasoning of the PMTs in statistical practices. In other words, these practices guided the reconstruction of the final HLT in Table 3.2. How these practices were documented is described in the data analysis section below.

3.4 Data Analysis

This study had a large amount of data from different techniques. One set of data was obtained from the tests and semi-structured interviews. To describe the data set, qualitative content analysis was performed. In this type of analysis, some degree of interpretation is made to systematically reveal what the data says, dividing the qualitative material into coding units (Schreier, 2012). As aforementioned in the data collection tool section, the test questions aimed to measure four skills corresponding to specific test questions. They were reasoning about statistical measures, reasoning about representations of data, reasoning about shape and shifts in data, and proportional reasoning. Under these skills, the PMTs' reasoning levels were determined. For the reasoning levels, first, the model of statistical reasoning of Garfield (2002) was tried out as a coding frame. Garfield's (2002) model includes five reasoning levels, which are idiosyncratic, verbal, transitional, procedural, and integrated process reasoning. The integrated process reasoning level, which is the most advanced, requires a complete understanding of the statistical process and the explanation of the process with confidence (Garfield, 2002). Similarly, the reasoning about distribution involves relating statistical ideas when analyzing data and interpreting the results obtained according to the context shaped by the data collection process That is, the integrated process reasoning level reflects the reasoning about distribution, which is the focus of the study. Therefore, the PMTs' reasoning levels were classified according to the model of statistical reasoning of Garfield (2002).

Although the model is appropriate for the study due to the integrated process reasoning level, the definitions of the levels do not coincide with proportional reasoning. Therefore, the coding framework was modified. For this, the model of Jones et al. (2004), which includes level definitions for proportional reasoning, was used. However, in this model, there are four levels (idiosyncratic, transitional, quantitative, and analytical). To ensure consistency between the levels, the level of

verbal reasoning was added to Jones et al.'s (2004) model between the idiosyncratic and transitional levels. Furthermore, the levels were named according to Garfield's model (2002). In essence, the adapted version of the model of statistical reasoning in Table 3.5 from Garfield (2002) and Jones et al. (2004) was used to code the PMTs' reasoning in the tests and semi-structured interview.

Table 3.5 The adapted version of the model of statistical reasoning (Garfield, 2002; Jones et al., 2004) to code the PMT's reasoning

	Idiosyncratic Reasoning	**Verbal Reasoning**	**Transitional Reasoning**	**Procedural Reasoning**	**Integrated Reasoning**
Reasoning about statistical measures (Q1, Q2)	Using statistical measures incorrectly without fully understanding them or scrambling them with unrelated information	Inability to apply statistical measures but know them verbally	Identifying one statistical measure	Identifying multiple statistical measures but failing to fully integrate them or understand the process	A complete understanding of the statistical measures, coordination of the rules and features of statistical measures, and explaining the statistical measures with confidence
Reasoning about representations of data (Q2, Q3)	Using some statistical words and symbols for shape, center, and spread of the distribution incorrectly or scrambling them with unrelated information	Expressing shape, center, or spread of the distribution without their numerical values	Identifying only one character of the distribution	Identifying multiple characteristics of the distribution without fully integrating them or understanding the process	A complete understanding of the distribution, coordination of the rules and behavior to describe the distribution, and explaining the process with confidence
Reasoning about shape and shifts in data (Q4, Q5)	Using some statistical words and symbols to compare distributions incorrectly, or scrambling them with unrelated information	Verbal understanding of some concepts about distribution but the inability to apply them to actual behavior	Comparing only one aspect (center, spread, density, or skewness) of the distributions	Comparing multiple aspects (center, spread, density, and skewness) of the distributions without fully integrating them or understanding the process	A complete understanding of the distribution, coordination of the rules and behavior to compare the distributions, and explaining the process with confidence
Proportional reasoning (Q5)	Not using relative thinking	Saying to use relative thinking but not applying it	Using relative thinking qualitatively	Using relative and proportional reasoning in an incomplete and invalid manner	Using relative and proportional reasoning

For example, the first question in tests measured the PMTs' reasoning skills about statistical measures. It was considered at the idiosyncratic reasoning level when the PMT's response to this question included erroneous information about statistical measures or his/her personal experiences and beliefs that were not based on statistical knowledge. The response in which a statistical measure was named without applying in the context was at the verbal level. For example, in the statement "4C is the highest class according to the mean", there needed to be more information about what the mean and its numerical value were, or how it was calculated. Given this information about a statistical measure, it was at the transitional level. However, the response was at the procedural level if information was given about others besides one statistical measure. For example, information about the standard deviation might be provided beside the mean. At the procedural level, the measures mentioned were not integrated. In this question, a center and a variability measure can be interpreted together to determine which class is better. More specifically, a class with a high center measure and a low variability measure can be seen as one that reads many books consistently. The responses where at least two statistical measures such as this were related were addressed at the level of integrated process reasoning.

The researcher coded all PMTs' responses to the questions in the tests and semi-structured interviews in terms of their reasoning levels according to the framework in Table 3.5. In addition, another mathematics educator performed the same process independently of the researcher on the answers of two randomly selected PMTs. When they came together to recode the disagreed codes, they discussed the reasons for the differences and agreed on them. After that, the researcher again recoded all test and interview data sets.

In the findings, the reasoning levels of the PMTs were presented based on the skills measured by the tests and semi-structured interviews. Rather than detailing the reasoning levels for each question, overall levels were determined for the questions covered by the skills (See Table 3.4). A detailed description of the levels was provided by citing examples from the PMTs' responses. The percentages and

frequencies of the PMTs' reasoning levels in each skill were compared before and after the teaching experiment. The comparison illustrated by the tables and graphs in the findings revealed to what extent the reasoning levels of the PMTs improved. That is to say, this analysis addressed the first research question of the study concerning how the PMTs' reasoning about distribution developed throughout the classroom teaching experiment based on RME and enriched by dynamic statistics software.

Apart from tests and interviews, the comprehensive data of the study consisted of video recordings of the classroom and debriefing sessions. Since the teaching experiment was performed in a socially structured learning environment, the data set was analyzed continuously as a part of the ongoing process. Based on the emergent perspective, one of the study's theoretical frameworks, in the retrospective analysis of the data set, the taken-as-shared reasoning ways of the classroom community were investigated. To reveal the classroom's collective reasoning, CSPs were identified. This analysis aimed to answer the second research question of the study: What CSPs emerged throughout the teaching experiment based on RME and enriched by dynamic statistics software?

The three-phase approach developed by Rasmussen and Stephan (2008), based on Toulmin's (2003) model of argumentation, was used to document CSPs. In the first phase of the analysis, the data of the sessions was sorted chronologically and transcribed episode by episode (day by day). The transcripts included small group work and class discussions. The researcher coded class sessions according to Toulmin's argumentation model by watching the videos.

Toulmin's (2003) argumentation model is based on three basic components. These components are claim (C), data (D), and warrant (W). According to the model, undoubted statements in a classroom form the *data* of the argument. A *claim* made based on the *data* is an assertion or conclusion. A *warrant* serves as an inference of argumentation that explains and justifies the claim being made. That is, it establishes a connection between data and claims in the argument. Figure 3.12 presents the skeleton proposed by Toulmin (2003) for a systematic analysis of the arguments.

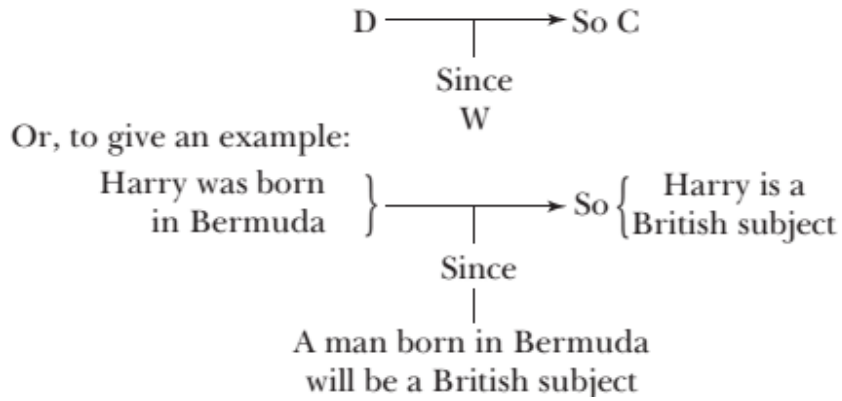


Figure 3.12. The skeleton proposed by Toulmin (2003, p. 92)

The skeleton includes a general example of the components of the argument. Based on the example, the argument can be expressed verbally as follows. Harry was born in Bermuda. Since a man born in Bermuda (i.e., Harry) will be a British subject, Harry is a British subject. As can be seen from the example, the warrant enables the association of other components of the argument, which are visualized with an arrow representation going from data to claim in the skeleton. Based on the degree of this association, Toulmin (2003) built a comprehensive argumentation scheme by adding different components, e.g. qualifier (Q), rebuttal (R), and backing (B). While the *qualifier* describes the degree to which the warrant justifies the claim, the *rebuttal* shows the situations in which it cannot consult the claim. A *backing* provides support for the acceptability and legitimacy of the warrant. Krummheuer (1995), who focuses on data, claim, warrant, and backing components of the argumentation, adapted Toulmin's argumentation model to mathematics. He draws attention to collective argumentation by emphasizing that several individuals contribute to the development of the argument. Since the aim of analyzing classroom sessions in this study is to reveal the communal reasoning ways underlying classroom discussion, the collective arguments of the classroom were investigated in the data analysis.

In the data analysis, the researcher formed argumentation schemes for each claim that appeared in the classroom. That is, I determined the claim, data, warrants, and backing (if it appeared) in the class sessions and generated skeletons for the arguments, as shown in Figure 3.14. At the same time, another mathematics

instructor, who is an expert in argumentation, also formulated the schemes of argumentation independently in the three learning activities in the HLT. The researcher and mathematics educator compared each other's schemes by confirming similarities and discussing differences. In cases of disagreement, they arrived at a compromise for their different argumentation schemes about the same argument by persuading each other. When necessary, they also refuted some argument schemes. Then, the researcher reconstructed the argumentation scheme of all the video recordings of the class sessions and obtained the argumentation logs.

Below is an example of a part of the data analysis. The following excerpt was from a small group study on the "recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia" activity and the group presented the solution in the classroom discussion. To gain weight in the activity, it was asked to justify the treatment the group members recommended to the hospital management.

> PMT3. Let's say plus for the top and minus for the bottom. If we take three times the number of people in all intervals of the bottom one, I think the bottom one will come out more. Family-based therapy is more than 0. But, traditional therapy will come out at values less than zero. So, family-based therapy would be bigger.
>
> PMT12. Exactly, the other side (above 0) is more important for us anyway.
>
> PMT5. Can't we do it directly from percentages?
>
> PMT3. We can also do percentages. The frequencies are 3, 15, 12, 30, 33, 18, and 6, respectively.
>
> PMT1. +5, +7, +12, -12, -8, -9, -1 (See Figure 3.13).
>
> PMT3. Above 0, there are already more minuses.
>
> PMT5. Where traditional therapy is high, it is usually below 0. The other side is more in family therapy.
>
> PMT1. What is the question?
>
> PMT12. Which one would you prefer?
>
> PMT3 and PMT5. Family therapy is more effective.

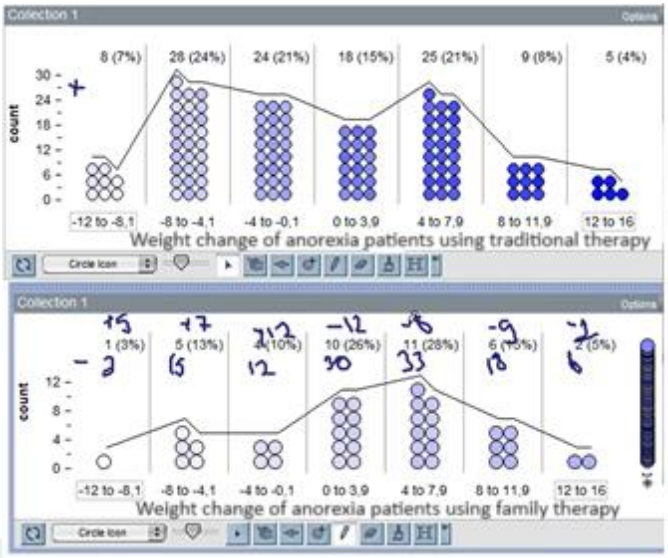


Figure 3.13. The graphs structured by the group in TinkerPlots for the activity related to anorexia

PMT3, a group member, partitioned the data into fixed-width intervals using TinkerPlots. After assigning the signs to two groups, she calculated three times the frequencies in each interval for the family therapy group and subtracted them from the frequencies in the traditional therapy group. Then, she recorded the resulting numbers between the two graphs, as in Figure 3.13. Although she stated that this process can also be performed with percentages, she did not. In comparing relative frequencies due to unequal sample sizes, the group claimed that family-based therapy was more effective because there were more minus signs at higher values. Figure 3.14 illustrates the argumentation scheme for the excerpt. Data 1 and Data 2 were the statements that occurred in the classroom. Data 2 was indicated with a dashed line because it is not explicitly performed. There were three warrants for the claim that family-based therapy was more effective with data comparing frequencies in the fixed-width intervals. The values above and below 0 were interpreted in all warrants.

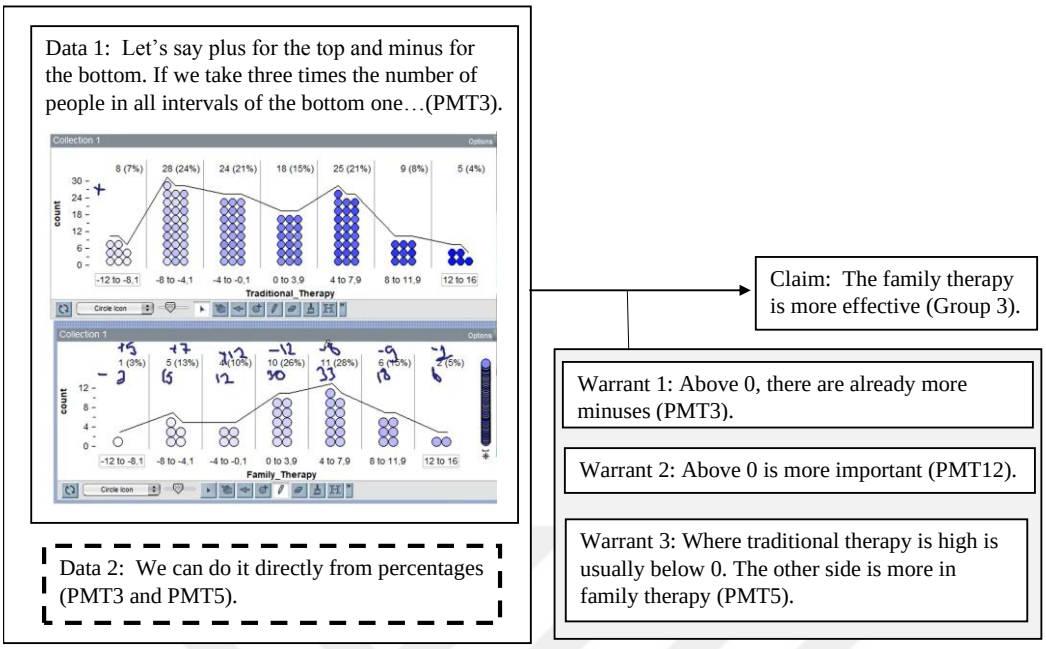


Figure 3.14. The argumentation scheme of the excerpt

In the second phase of the analysis, taken-as-shared (TAS) ideas were determined based on argumentation logs. The arguments containing communal statistical ideas reflected by the argumentation component in a bold color in each argumentation log were examined. During the design experiment, as instructional activities were implemented weekly for specific learning goals, argumentation logs were organized based on each class session. In doing so, the argumentation logs for each class session were constantly compared with the argumentation logs in the other sessions. In this cyclical process, new ideas might emerge, or existing ones might change. This comparison continued until a new idea did not emerge from all argumentation logs. For this, the idea charts were formed to follow back and forth systematically across all class sessions via MAXQADA. The charts included columns for ideas to function as if shared in the classroom, ideas to keep an eye on, and additional comments, including practical and theoretical issues. To decide whether the idea was TAS, it was considered satisfactory to meet any of the three criteria, the first two of which belong to Rasmussen and Stephan (2008, p. 200) and the last added by Cole et al. (2012, p. 200):

1. When the backings and/or warrants for an argumentation no longer appear in students' explanations (i.e., they become implied rather than stated or called for explicitly, no member of the community challenges the argumentation, and/or if the argumentation is contested and the student's challenge is rejected), we consider that the mathematical idea expressed in the core of the argument stands as self-evident.

2. When any of the four parts of an argument (the data, warrant, claim, or backing) shifts position (i.e., function) within subsequent arguments and is unchallenged (or, if contested, challenges are rejected), the mathematical idea functions as if it were shared.

3. When a particular idea is repeatedly used as either data or warrant for different claims across multiple days.

For example, the above argumentation did not include any clear comment about the mean. However, the values of the data points were associated with the average in previous class sessions. Specifically, in the "building arguments based on statistical measures that are appropriate for the context-batteries" activity in the fifth week, it was claimed that the data set with a high mean had a longer lifespan. Therefore, the PMTs turned to the brand with high values. As a different way of reasoning, they took the differences between each reciprocal data value in two data sets by giving signs to the battery brands as above. They deduced that the sign of the sum of the values obtained shows the longer lifespan of the battery. They based this reason on the mean. That is to say, they noticed that the magnitude of the values in the data set estimates the mean size. In all subsequent activities, they identified the data set with low/high mean depending on the context. Before the example excerpt above, it was emphasized that the one with a higher mean is family-based therapy. The idea of "the mean of the data set with high data points is high," which emerged in the previous weeks, was also used in the "recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia" activity. It met all of the criteria, each of which made it a TAS idea.

In the analysis process described above, the PMTS' reasoning about the distribution before and after the teaching experiment and during the teaching experiment was examined independently. However, in this study, it was aimed to investigate the development of PMTs' reasoning about distribution as they participated in the

classroom teaching experiment based on RME and enriched by dynamic statistics software. There was a need to combine the results of these two analyses to reveal how the PMTS' reasoning about distribution developed from the beginning to the end of the teaching experiment. Therefore, the reasoning levels in the TAS ideas that emerged in the classroom were also examined. The reasoning levels included in the framework used in the first data analysis were extended for the concept of distribution. Based on the definition of reasoning levels by Garfield (2002), the TAS ideas were evaluated according to the following criteria.

> Idiosyncratic reasoning level: Knowing some statistical words and symbols about distribution, using them without fully understanding them, often incorrectly, or scrambling them with unrelated information.
>
> Verbal reasoning level: Verbal understanding of some concepts about distribution, but the inability to apply them to actual behavior.
>
> Transitional reasoning level: Identifying correctly one dimension of distribution (center, spread, density, or skewness) and comparing them.
>
> Procedural reasoning level: Identifying correctly the dimensions of distribution (center, spread, density, and skewness) but not integrating them or understanding the process.
>
> Integrated process reasoning level: A complete understanding of the distribution, coordination of the rules and behavior in the statistical process, and explaining the process with confidence.

For example, the PMTs were at the procedural reasoning level while using the TAS idea that "the mean of the data set with high data points is high." They used the idea for the first time in the "building arguments based on statistical measures that are appropriate for the context-batteries" activity. Unlike frequencies in the above solution, they summed the differences between the data values. In the activity related to batteries, they focused only on the mean based on the total differences in the values. Therefore, they were at the transitional reasoning level concerning this idea. However, as seen from the argumentation scheme in Figure 3.14, they also considered the spread of the data over 0 in the activity related to anorexia. For this reason, they were at the procedural reasoning level. Since they did not show a complete understanding of the distribution, they did not move the integrated process

reasoning level for this idea. This idea was addressed at the procedural level instead of the transitional or integrated process reasoning level based on how the PMTs used the idea in the activity.

After determining the TAS ideas and the PMTs' reasoning level in the ideas according to these criteria, the PMTs' collective reasoning about distribution was identified by organizing the TAS ideas in the last phase of the analysis. The collective ways of reasoning, symbolizing, and arguing were classroom statistical practices (CSPs). To explain further, the TAS ideas served as a unit of analysis to obtain the CSP. In the thematic analysis, TAS ideas were themes that served the emergence of CSPs. In this context, the related literature and the events in the classroom guided the ascertainment and interpretation of the practices. For instance, in the above example, the PMTs compared the frequencies in the interval the data set was partitioned instead of data points. They used the proportionality constant to equate the sample size in the distributions. However, they could not think of the distribution as an entity. They interpreted the data set as a collection of data points by comparing the parts partitioned. Based on the relationship between part and whole of the data, Cobb (1999, p. 17) called this practice "exploring qualitative characteristics of collections of data points." Therefore, the practice in the literature was addressed as one practice of this study.

The other data set for this study constituted field notes and the PMTs' written artifacts. The field notes generally contained the reasoning ways of the PMTs that the researcher noticed instantly in the class sessions. In the notes, the researcher stored the solution method used and examples showing it in each class session. I named them depending on the characteristics of the method and the way it was applied. For example, the above method of partitioning into equal intervals and performing arithmetic operations on the frequencies in the intervals was categorized as a plus-minus marking method. In the analysis of the class sessions, the solutions of the PMTs were correlated to the categorized methods. During the analysis, comparisons were made with the field notes while watching the videos of the class sessions. Since the videos presented what compactly happened in the class, the

videos of the class sessions were taken into account. That is, the field notes served as a reminder to support the analysis of the class sessions.

Similarly, the written artifacts of the PMTs were used to support the data analysis of the class sessions. Artifacts comprised individual statements of the PMTs in the form of documents. In these documents, expressions reflecting the ideas and thoughts of the PMTs were coded. For example, in the second assignment, which was asked to describe the mean and explain its interpretations in the second week of the teaching experiment, PMT11 wrote the following:

> The second interpretation is balancing points and it is not a common way to teach mean but it is a very good way because it shows the relationship between numbers very clearly. In balancing points, students can find the distance from the mean by placing their data on a number line. For example, if we look at the distance of the numbers 2, 3, 4, 6, 7, and 8 from 5, the distances of 2 and 8, 3 and 7, and 4 and 6 are equal, which shows us that the number 5 is the average.

In this statement, PMT11 emphasized the distance of data values from the mean in the interpretation of the balance point. The example given by PMT11 was evidence of this statement, which was coded as *equality of distances from the mean*. Categories were created based on the codes of the written artifacts of the PMTs. These categories reflected the ideas of the PMTs. The artifacts were associated with the ideas that emerged in the classroom, as they were an extension of classroom sessions. In this respect, the categories were compared and contrasted in reflection papers and classroom sessions. As a result of the analysis, there were common categories in reflection papers and class sessions. For example, the PMT11's statement in her paper corresponded to the TAS idea that *the sum of the distances of scores above and below the mean is equal*. The findings included quotations, such as PMT11's statement, from the PMTs' artifacts that contain common categories to support the emerging TAS ideas.

Table 3.6 Research questions, corresponding data collection tools, and data analysis

Research Questions	Data Collection Tools	Data Analysis	
1. How did the PMTs' reasoning about distribution develop throughout the classroom teaching experiment based on RME and enriched by dynamic statistics software?	Pretest – Posttest - Reasoning about statistical measures (Q1, Q2) - Reasoning about representations of data (Q2, Q3) - Reasoning about shape and shifts in data (Q4, Q5) - Proportional reasoning (Q5) Semi-structured interviews	The adapted version of the model of statistical reasoning (Garfield, 2002; Jones et al., 2004) Content analysis	The adapted version of the model of statistical reasoning (Garfield, 2002; Jones et al., 2004) Content analysis
2. What CSPs emerged throughout the teaching experiment based on RME and enriched by dynamic statistics software?	Video recordings of classroom and debriefing sessions Written artifacts Field notes	The three-phase approach (Rasmussen & Stephan, 2008) Thematic analysis	

In summary, as can be seen from Table 3.6, in this study, data analysis was performed in two parts. In the analysis of the data set consisting of pretest, posttest, and semi-structured interviews, the reasoning levels of the PMTs were determined based on the skills needed for reasoning about the distribution. Comparing the reasoning levels before and after the teaching experiment, it was revealed how the PMTs' reasoning about the distribution developed. Thus, the first research question was answered. On the other hand, for the second research question, the data set consisting of class sessions, field notes, and the PMTS' written artifacts was analyzed. In this analysis, CSPs that emerged in the classroom throughout the teaching experiment were revealed. CSPs served as indicators of the PMTs' collective reasoning in the social environment of the classroom. In light of the emergent perspectives based on the socio-constructivist perspective, the three-phase approach developed by Rasmussen and Stephan (2008), based on Toulmin's (2003) model of argumentation, was used to document CSPs. In the analysis, firstly, argumentation logs containing argumentation schemes were obtained. Based on the argumentation logs, the TAS ideas of the classroom community were identified in the second phase of the analysis.

In this process, the reasoning levels of the PMTs were also determined. Organizing the TAS ideas, which was a unit of analysis, the CSPs were documented in the last phase. Thus, the PMTs' collective reasoning was examined at the beginning, during, and end of the teaching experiment. In other words, the reasoning development of the PMTs throughout the classroom teaching experiment was addressed, which was the main concern of the research. This data analysis process is outlined in Figure 3.15.

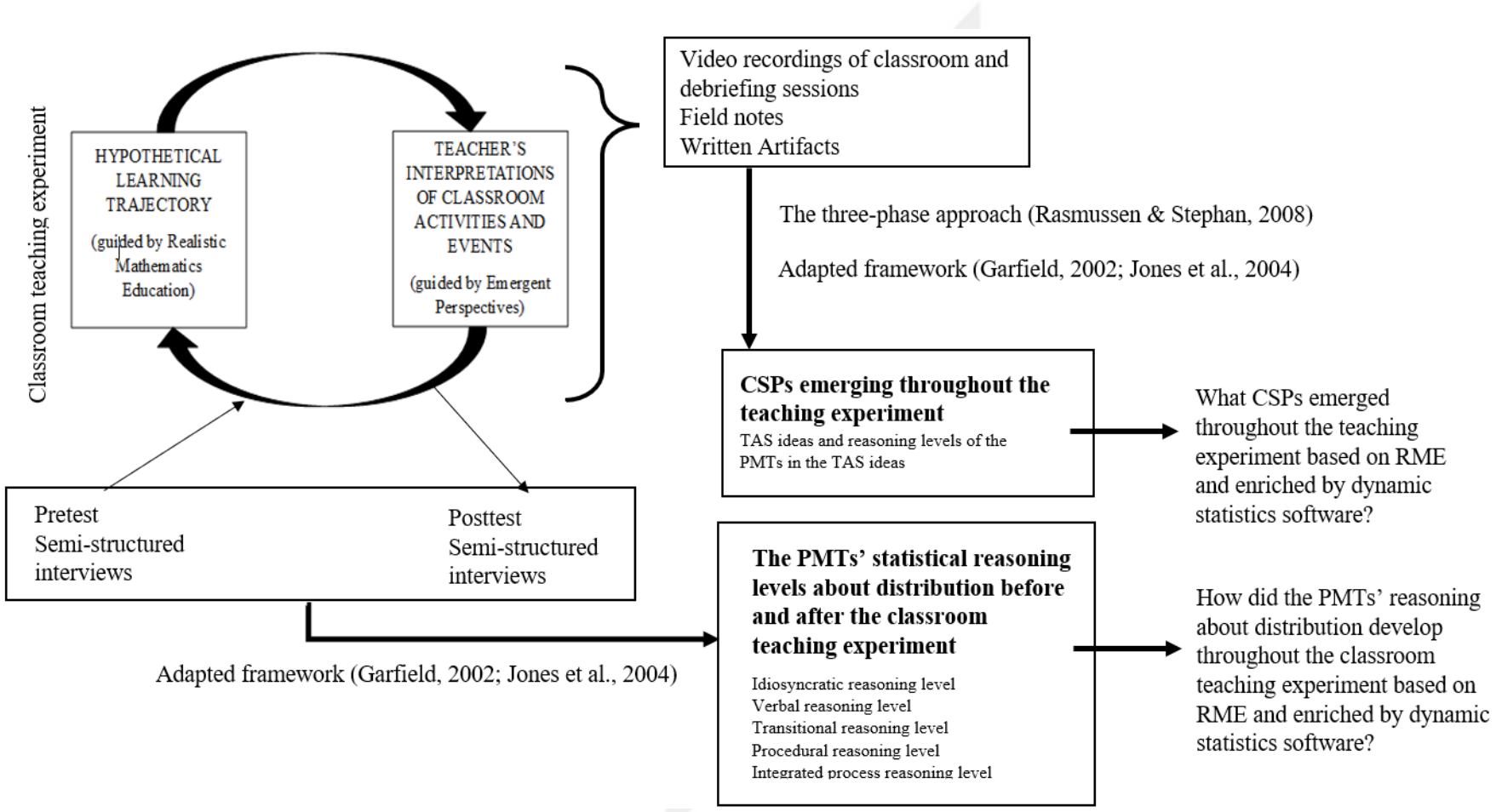


Figure 3.15. The outline of the data analysis process

3.5 Trustworthiness

Many terms are used to describe the quality of a study, such as trustworthiness, authenticity, goodness, relevance, and so on (Savin-Baden & Major, 2013). In this study, trustworthiness is used to describe the quality of the study. Because it aims to accomplish objectivity with the idea that the positions of the researchers can shape the study. The traditional approaches used to ensure the quality of the study are validity and reliability. In the simplest sense, validity refers to whether what is intended to be measured is measured, while reliability is an indicator of the independence of the results from the researcher (Bakker, 2019). Lincoln and Guba (1985) use credibility, transferability, dependability, and confirmability to ensure reliability and validity in qualitative studies. Below, I explained how these terms referring to the quality of the study were addressed in this study.

Credibility reflects the reality of the results (Lincoln & Guba, 1985). One way to ensure credibility is to show the participants' reality (Savin-Baden & Major, 2013). For this purpose, the participants were described in detail. The instructor and the researcher knew the participants because they had long-term interactions with them in previous semesters. In addition, during the design experiment process, they constantly observed the participants, which provided insight into them. For example, PMT9 emphasized using variability measures in all data analysis since the beginning of the teaching experiment. Another way to increase the credibility is triangulation (Savin-Baden & Major, 2013). In this study, data triangulation was performed. The data of the study was collected from multiple sources, e.g., pretest and posttest, semi-structured interviews, video recordings of classroom and debriefing sessions, the PMTs' written artifacts, and field notes. Accordingly, the same conjectures or results were revealed by triangulation of different data types (Bakker, 2019; Patton, 2002; Savin-Baden & Major, 2013). In this sense, video records of the sessions, quotations from the participants' reflection papers, data inscriptions in TinkerPlots,

and field notes were used to verify an observation or interpretation of the research findings.

Transferability is related to external validity, referring to the replicability of the study in different contexts (Gravemeijer & Cobb, 2006; Lincoln & Guba, 1985). It means other researchers should be able to follow the study as reported and obtain similar results. Even if the same instructional treatment is applied in educational research, it is impossible to replicate the study exactly. However, ecological validity in design research provides the replicability of the study in other contexts, rather than the generalizability of the results (Gravemeijer & Cobb, 2006). To ensure transferability, a thick description of research settings was elaborated. Thus, other researchers can adapt their work on this experiential basis.

A thick description also provides the *dependability* of the study, which refers to the persistence of research findings (Lincoln & Guba, 1985; Savin-Baden & Major, 2013). To establish dependability in this study, the context of the study, the participants, and what happened during the design experiment were described in detail. The entire research process, from the start of the study to the final report, was kept on record. In particular, weekly debriefing sessions helped to track the progress of the study in each microcycle.

Different interpretations by other researchers may lead to different results. According to Gravemeijer and Cobb (2006), inferences and conjectures from the design experiment should be justified and reasonable to ensure trustworthiness. That is, *confirmability* should be provided to ensure the quality of the study (Lincoln & Guba, 1985; Savin-Baden & Major, 2013). To make sure that the researcher was neutral in data analysis and interpretation, a peer examination/external audit was conducted for the analysis of the tests, interviews, and video recordings of the class sessions. A mathematics educator coded approximately 15% of the data set of tests and interviews. The inter-coder reliability between the researcher and the mathematics educator was 87.5%. According to Miles and Huberman (1994), this rate was sufficient as the initial coding was not expected to be better than 70%. After

re-coding the entire data set of tests and interviews, the researcher found the intra-coder agreement to be 96.4%. Peer examination was conducted in three learning activities for data from class sessions, which comprised 25% of the data set. For the argumentations in the class sessions, the inter-coder reliability was 84.2%. In addition, the reasoning levels of the PMTs in TAS ideas emerging in the classroom were also determined by two independent coders. Inter-coder agreement for the reasoning levels was 81%. Moreover, to establish confirmability, the data analyses were supported by quotations from the PMTs' written works, classroom excerpts, and the PMTs' data inscriptions in TinkerPlots.

3.6 Researcher Role

The researcher influences the research with his/her knowledge, skills, experience, and personal beliefs (Denzin & Lincoln, 1994; Patton, 2002). Since the researcher takes an active role throughout the entire process of design research, it is essential to describe the researcher's position and perspective in the study. The researcher attended all class sessions and participated in debriefing sessions with the instructor to provide a common interpretation of classroom events (Gravemeijer & Cobb, 2006). Accordingly, the researcher had two roles in this study. The first was that the researcher, I, was the participatory observer in the data collection and analysis process. The researcher implemented the tests, conducted the interviews, took video recordings, collected reflection papers, and performed data analysis. Since the teaching experiment was conducted on the online learning platform, the researcher enabled the learning process to continue in cases where the instructor experienced technical problems such as internet, sound, and screen sharing. In addition to recording the entire process, the researcher presented the activities and required data files to the participants. The researcher was also responsible for instructions regarding TinkerPlots without giving instructions to the participant. The other role of the researcher was in the design of the teaching experiment. The instructor and the researcher designed the instructional sequence in weekly debriefing sessions.

After evaluating the previous week, the researcher introduced the next activity in the initial instructional sequence or the activities considered appropriate to add sequences where necessary. After discussing the implementation of the activities with the instructor, the researcher prepared the activity and assignment files before the class sessions. That is, the researcher was an active participant throughout the entire teaching experiment.

3.7 Ethical Considerations

Before the teaching experiment, the necessary permissions were taken from the Middle East Technical University Human Subjects Ethics Committee (Appendix B). Within the scope of the teaching experiment, the PMTs were informed about the content of the elective course offered in the Mathematics Education program at the university. The instructor who opened the course explained the purpose of the study and the procedures to be followed to the PMTs before registering for the course. The PMTs who wanted to participate in the study voluntarily signed the consent form (Appendix D). They knew that the entire process, including class sessions and interviews, would be recorded. It was guaranteed that their identities would not be revealed. Instead of using pseudonyms for confidentiality, they were randomly numbered. For example, the PMT assigned number three was named PMT3. In addition, the participants were not deceived. The study was carried out in its natural environment.

CHAPTER 4

FINDINGS

The study aimed to investigate the development of PMTs' reasoning about distribution as they participated in the classroom teaching experiment based on RME and enriched by dynamic statistics software. For this purpose, the development of PMTs' reasoning about distribution throughout the classroom teaching experiment was investigated around the two research questions. Firstly, the reasoning levels of the PMTs before and after the classroom teaching experiment were analyzed to examine how their reasoning about distribution developed, answering the first research question. For this, the data collected through pretest, posttest, and semi-structured interviews were analyzed according to the adapted version of the model of statistical reasoning (Garfield, 2002; Jones et al., 2004). The data collection tools measured the skills of reasoning about statistical measures, representations of data, shape and shifts in data, and proportionality necessary for reasoning about distribution. In this context, the PMTs' reasoning levels were determined by content analysis based on these skills.

As a result of the classroom teaching experiment conducted in a social classroom environment, there was an improvement in the PMTs' reasoning about the distribution. The classroom statistical practices that emerged during the teaching experiment demonstrated how this improvement was achieved, which addressed the second research question. To reveal the CSPs, the data from video recordings of classroom and debriefing sessions, written artifacts, and field notes were analyzed using the three-phase approach (Rasmussen & Stephan, 2008). In this process of thematic analysis, the TAS ideas that constitute CSPs were identified. Since the CSPs that emerged from TAS ideas that PMTs declared common in the classroom indicated the development of the PMTs' reasoning about distribution during the

teaching experiment, the practices were presented in conjunction with TAS ideas. While elaborating on the TAS ideas, the reasoning levels of the PMTs in the TAS ideas were also addressed as in the first research question. In other words, the reasoning levels of the TAS ideas were determined according to the adapted version of the model of statistical reasoning (Garfield, 2002; Jones et al., 2004).

In summary, how PMTs' reasoning about distribution developed at the beginning, during, and at the end of the teaching experiment was examined based on the two research questions. In this regard, the chapter was organized into two sections: 1) The PMTs' reasoning levels about distribution before and after the classroom teaching experiment and 2) The emerging classroom statistical practices throughout the teaching experiment.

4.1 The PMTs' Statistical Reasoning Levels About Distribution Before and After the Classroom Teaching Experiment

Based on the first research question of the study, how the PMTs' reasoning about distribution developed was revealed by comparing their reasoning levels before and after the teaching experiment. As seen in Table 3.6, the reasoning levels were determined by analyzing the data obtained from pretest, posttest, and semi-structured interviews according to the adapted version of the model of statistical reasoning (Garfield, 2002; Jones et al., 2004). According to the model, five reasoning levels exist—namely, idiosyncratic, verbal, transitional, procedural, and integrated process reasoning (See Table 3.5 in data analysis for details). Additionally, the analysis identified four skills that serve to reason about distributions. As detailed in the data collection tools section of Chapter 3 (Table 3.4), these skills and corresponding questions in the instruments are as follows: reasoning about statistical measures (Q1 and Q2), reasoning about representations of data (Q2 and Q3), reasoning about shape and shifts in data (Q4 and Q5), proportional reasoning (Q5). In this section, the reasoning levels of the PMTs are organized based on these skills. In other words, the reasoning levels were presented under the headings of reasoning about statistical

measures, reasoning about representations of data, reasoning about shape and shifts in data, and proportional reasoning. The stacked bar graphs in Figure 4.1 demonstrate PMTs' reasoning levels based on the skills to exhibit the overall change before and after the teaching experiment.

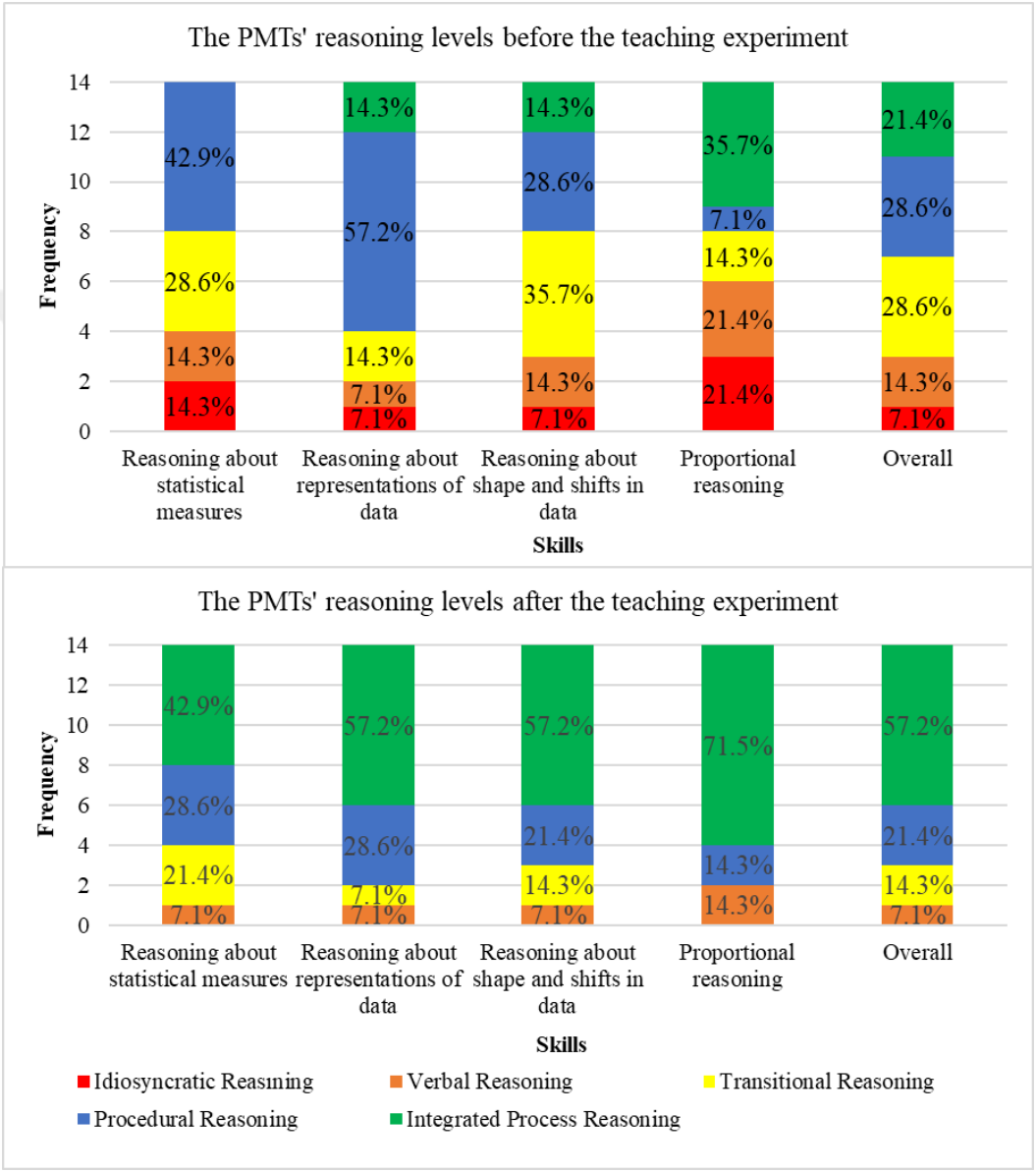


Figure 4.1. The stacked bar graphs of the PMTs' reasoning levels based on skills before and after the teaching experiment

According to Figure 4.1, the PMTs performed at each reasoning level before the teaching experiment, as illustrated in the overall column in the first graph. In other words, the reasoning levels of the 14 PMTs participating in the study were dispersed across five reasoning levels. Specifically, one PMT misused statistical words and symbols about the distribution at the idiosyncratic level or confused it with unrelated information. Two out of 14 PMTs (14.3%) at the verbal reasoning level stated the concepts about distribution but could not apply them practically. At the transitional reasoning level, 28.6% of the PMTs used only one aspect of the distribution. Similarly, four out of 14 PMTs (28.6%) at the procedural reasoning level identified all distribution aspects but faced challenges while integrating them into the statistical process. At the integrated process reasoning level, 21.4% of PMTs demonstrated a comprehensive understanding of distribution, successfully coordinated rules and behaviors within the statistical process, and confidently explained the process.

Following the teaching experiment, no PMT exhibited the misuse of statistical knowledge or provided irrelevant information at the idiosyncratic level. The ratios of the PMTs at other levels were as follows: 7.1% at the verbal reasoning level, 14.3% at the transitional reasoning level, 21.4% at the procedural reasoning level, and 57.2% at the integrated process reasoning level. According to the graphs, there was a decrease in the number of PMTs at the end of the teaching experiment at idiosyncratic, verbal, transitional, and procedural reasoning levels. On the other hand, the ratio of PMTs at the integrated process reasoning level increased from 21.4% to 57.2%. That is to say, the PMTs' reasoning levels shifted to integrated process reasoning. After the teaching experiment, more than half of the PMTs fully understood the distribution at the integrated process reasoning level. This progress also occurred in each skill. The number of PMTs at the integrated process reasoning level at least doubled for each skill. Since the indicators of the skills at the levels are different (See Table 3.6), it would be more meaningful to describe the PMTs' reasoning levels based on skills. Below, the reasoning levels of the PMTs are presented according to their respective skills.

4.1.1 Reasoning about Statistical Measures

The first skill in the tests was reasoning about statistical measures. It includes "understanding why measures of center, spread, and position tell different things about a data set; knowing which are best to use under different conditions, and why they do or do not represent a data set;(…) knowing why a good summary of data includes a measure of center as well as a measure of spread and why summaries of center and spread can be useful for comparing data sets" (Garfield, 2002, p. 5). Accordingly, in the pretest, posttest, and semi-structured interviews, the PMTs were expected to use and reason the appropriate statistical measure for the context, considering center, variability, and shape when interpreting and comparing data sets. The first and second questions in the pretest and posttest address this skill. Question 1 (See Figure 3.9) asked to determine the better class regarding the number of books read in four classes during a semester. Data values were expected to be consistent with each other and have high averages for a better class based on context. Class 4C fulfilled these criteria. In addition, when the outlier was excluded in the data of class 4D, 4D would become a better class.

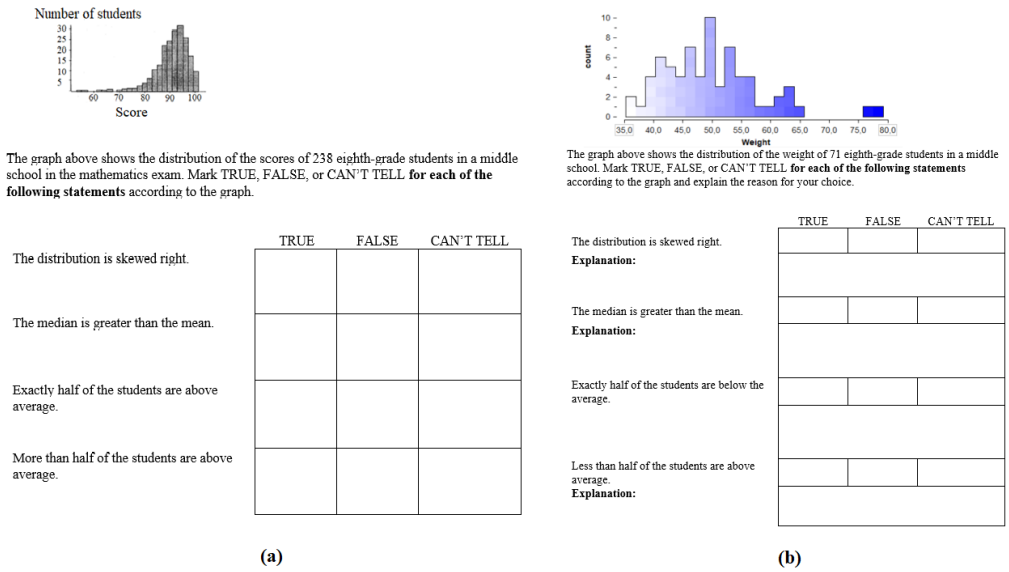


Figure 4.2. Question 2 in the pretest (a) and posttest (b)

Question 2 involved the selection and interpretation of statistical measures depending on the shape of the distribution. It required interpretation using the median for measures of center because of the skewed distribution in the pretest and posttest (Figure 4.2). Since the median is greater than the mean in the left-skewed distribution in the pretest, exactly half of the students are above the average if the median is average. Similarly, since the mean is greater than the median in the right-skewed distribution in the posttest, exactly half of the students are below the average when the median is the average.

While determining levels for reasoning about statistical measures, misuse or confusion of statistical measures with unrelated information or personal experiences was categorized at the idiosyncratic level. At the verbal reasoning level, only the names of statistical measures were mentioned without assigning any quantitative values. Identifying one of these statistical measures was required for the transitional reasoning level. When more than one measure was used without integrating them, it indicated the procedural reasoning level. At the integrated process reasoning level, more than one statistical measure was utilized in context with a comprehensive understanding and proper association with each other. Table 4.1 shows the reasoning levels of PMTs while reasoning about statistical measures before and after the teaching experiment.

Table 4.1 The reasoning levels of PMTs while reasoning about statistical measures before and after the teaching experiment

	Before		**After**	
	f	**(%)**	**f**	**(%)**
Idiosyncratic Reasoning	2	(14.3%)	-	-
Verbal Reasoning	2	(14.3%)	1	(7.1%)
Transitional Reasoning	4	(28.6%)	3	(21.4%)
Procedural Reasoning	6	(42.9%)	4	(28.6%)
Integrated Process Reasoning	-	-	6	(42.9%)
Total	**14**		**14**	

The PMTs' reasoning levels about statistical measures improved considerably after the teaching experiment. Since 14.3% of the PMTs misused statistical measures or confused them with irrelevant information, they were categorized at the idiosyncratic level while making contextual interpretations using statistical measures, before the teaching experiment. They failed to interpret the formal properties of the mean. For example, PMT5 at this level before the teaching experiment completed the missing data due to different sample sizes with the mean. However, she could not think that adding a datum whose value corresponds to the mean of the existing data would not change the mean. Similarly, PMT4 also claimed that the median is larger than the mean in a right-skewed distribution in Question 2 of the posttest by omitting the effect of outliers on the mean. However, no PMT reasoned about statistical measures at the idiosyncratic level after the teaching experiment. There were a few PMTs at the verbal reasoning level who mentioned statistical measures but did not apply them before or after the teaching experiment. At the verbal level, judgments were made based on range, average, or total without giving any numerical value. The below response of PMT6 to Question 1 in the posttest was at the verbal reasoning level.

> Since the number of students is different in different classes, the sum of the number of books would not be enough to find the best class in terms of reading. Therefore, we should use means by dividing the sum by the number of students in each class. The highest mean is for the best class (PMT6).

Although she stated the mean formula, she could not provide numerical evidence. Moreover, she did not decide which class, based on the criterion she determined (e.g., the highest mean). Stating the positions of the mean, median, and mode in a distribution as a rule without interpretation was also an indicator of the verbal reasoning level. At the verbal reasoning level, statistical measures concerning the shape of the distribution could not be interpreted logically. PMT6, at this level after the teaching experiment, compared the measures of center for the right-skewed distribution, "Most of the frequencies are bunched on the left side of the graph. The mode of the data is on the left side. In the right-skewed graph, the median is always greater than the mean."

Regarding the reasoning about statistical measures, 28.6% of the PMTs were at the transitional level before the teaching experiment. This rate decreased to 21.4% after the teaching experiment. At the transitional reasoning level, the PMTs took one of the statistical measures as a criterion, such as the mean, total, sum of absolute deviations (SAD), mean of absolute deviations (MAD), or standard deviation. For example, while PMT4 chose the class with the most books read, PM13 preferred class 4A with the highest mean, claiming that the mean is more appropriate than the median or mode. A striking situation at this level was the overuse of the mean as average. Even in the skewed distribution, the PMTs used the mean as average at the transitional reasoning level. By comparison with this level, at the procedural reasoning level, the PMTs considered more than one statistical measure without integrating them. However, it was necessary to associate the statistical measures with each other at the integrated process level. Before the teaching experiment, most of the PMTs (42.9%) who identified more than one measure were at the procedural reasoning level. In contrast, no one could relate the statistical measures that they mentioned. For example, PMT3, at the procedural reasoning level, calculated the distributions' standard deviation and MAD without associating them. In addition, the mean and standard deviation of the data sets were not interpreted together at the procedural reasoning level. Conversely, after the teaching experiment, most of the PMTs (42.9%) could associate the statistical measures at the level of integrated process reasoning. At this level, they could determine the best statistical measure representing the data for the context of the two questions addressing reasoning about statistical measures. In Question 1, PMT5's reasoning about statistical measures at the integrated process level was noteworthy. She explained how the mean represents the data despite the outlier in the distribution. The quote was her response in the posttest.

> Since the sample size is different, I look at class mean rather than totals. Mean is the average number of books expected from a person in the class. 4A=10, 4B=9, 4C=9.5, 4D=9.6. The mean of 4A is higher than the others; it seems to be more successful. But there is an outlier in the 4D class. Outlier decreases mean. If we do not take two into account, the mean is 11.5. I synchronize the

sample size of all classes. When we take the highest four values of each class, the highest mean belongs to 4D (PMT5).

When comparing the means in the posttest, she noticed the outlier in the 4D class and removed it from the data. She concluded that the highest mean belongs to the 4D class instead of the 4A class. In other words, she was able to associate center with variability. In Question 2, the PMTs at the integrated process reasoning level used the median for the skewed distributions. For example, in the semi-structured interview after the teaching experiment, PMT7 emphasized using the median with a graphical representation. She converted the histogram given in the posttest to a box plot in TinkerPlots and calculated the numerical values for the statistics. She said using the mean was a big mistake and preferred to use the median.

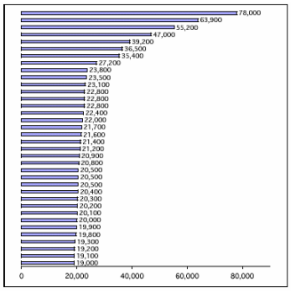
To sum up, the teaching experiment contributed to developing the PMTs' reasoning about statistical measures. Concerning the first skill, the number of PMTs at the transitional and procedural levels before the teaching experiment shifted to the integrated process level after the teaching experiment. In addition, those at the idiosyncratic and verbal levels also moved to the transitional and higher levels. Next, I present the change in the reasoning level of the PMTs for another skill: reasoning about representations of data.

4.1.2 Reasoning about Representations of Data

The second skill in the tests was reasoning about representations of data. It included "understanding how a plot is meant to represent a sample, understanding how graphs may be modified to better represent a data set; being able to see beyond random artifacts in a distribution to recognize general characteristics such as shape, center and spread" (Garfield, 2002, p. 5). For the skill, in the pretest, posttest, and semi-structured interviews, the PMTs were expected to read the data in the representation and interpret it by considering the general characteristics of the distribution, such as center, spread, skewness, and density within the given context.

Questions 2 and 3 in the pretest and posttest addressed reasoning about data representations. Question 2, which included skewed distributions, required an interpretation of the sample relative to the average (Figure 4.2). Exactly half of the sample was above or below the average (e.g., the median) because the distribution was skewed in both pretest and posttest. Question 3 in Figure 4.3 asked to interpret the distribution's extreme values, averages, clusters, and quartiles in the context of salary. While reading the data, it was necessary to judge the extreme values like 78.000 dollars, interpret the majority, and conclude based on the appropriate measure of center (e.g., not the mean).

The graph below shows salaries of 36 individuals who graduated last year from the Hurley School of Graphic Design. These are the salaries they earned the first year after graduating.



The School wants to publish this information in their newsletter to give current students an idea of how much they might expect to earn their first year working. But they don't have room to print the entire graph.

Part 1. Below are three possible summaries they are considering for their newsletter. Which summary do you think will give students the best sense of how much they might expect to each their first year working?

____ a. "The salaries for last year's graduates were quite varied and ranged all the way from a low of $19,000 to a high of $78,000."

____ b. "Last year's graduates earned an average (mean) salary the first year of $27,000. The highest salary was 78,000."

____ c. "The majority of last year's graduates made between $19,000 and $23,800 their first year. A few made considerably more than that, with one earning 78,000."

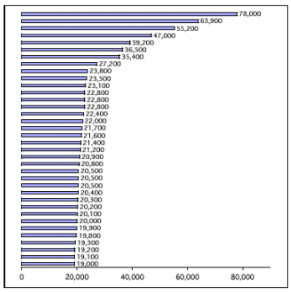
Part 2. Briefly say why you wouldn't choose the other two summaries.

I wouldn't choose summary ___ because ----------

I wouldn't choose summary ___ because ----------

(a)

A college newspaper conducted a survey of college graduates who had finished their studies one year ago. These graduates were asked how much they earned during their first year working. The goal of the study was to collect and publish information that would help college students get a sense of how much money they could expect to earn once they graduated. The newspaper editors considered that this information could be especially useful for freshman and junior students when deciding which major to choose. The graph below shows the information that was collected on 36 graduates who had majored on graphic design. Each bar shows the amount of dollars that each of the 36 graduates earned in their first year working as graphic designers.



Part 1. Below are four possible ways of reporting the data shown in the graph to their readers. Which statement do you think will give student readers useful information?

____ a. "Our data shows that entering salary for a graphic designer ranges between $19,000 and $78,000 a year."

____ b. "Our data shows that the twenty-five percent of the people entering the field of graphic design that earned the least, earned between $19,000 and $20,200 in a year. The fifty percent in the middle earned between $20,300 and $23,500, and the twenty-five percent that earned the most, earned between $23,800 and $78,000."

____ c. "Our data indicates that the average annual salary for a graphic designer entering the field was of $27,000."

____ d. "Our data shows that most people entering the field of graphic design made between $19.000 and $23.800 in their first year working."

Part 2. Briefly say why you wouldn't choose the other three statements.

I wouldn't choose summary ___ because ----------

I wouldn't choose summary ___ because ----------

I wouldn't choose summary ___ because ----------

(b)

Figure 4.3. Question 3 in the pretest (a) and posttest (b)

While determining levels for reasoning about representations of data, misuse or confusion about shape, center, and spread of the distribution or personal experiences was categorized at the idiosyncratic level. At the verbal reasoning level, only the names of the characteristics were mentioned without assigning any quantitative values. Identifying one of the distribution characteristics (shape, center, and spread) was required for the transitional reasoning level. When more than one characteristic was used without integrating them, it indicated the procedural reasoning level. At the integrated process reasoning level, coordinating the characteristics to describe the distribution was performed in context by confidently explaining the statistical process. Table 4.2 shows the reasoning levels of PMTs while reasoning about representations of data before and after the teaching experiment.

Table 4.2 The reasoning levels of PMTs while reasoning about representations of data before and after the teaching experiment

	Before		**After**	
	f	**(%)**	**f**	**(%)**
Idiosyncratic Reasoning	1	(7.1%)	-	-
Verbal Reasoning	1	(7.1%)	1	(7.1%)
Transitional Reasoning	2	(14.3%)	1	(7.1%)
Procedural Reasoning	8	(57.2%)	4	(28.6%)
Integrated Process Reasoning	2	(14.3%)	8	(57.2%)
Total	**14**		**14**	

Regarding the reasoning about representations of data skill, before the teaching experiment, there was a PMT at the idiosyncratic level who used incorrect information about the center, spread, and shape of the distribution. He expressed the left-skewed distribution in Question 2 of the pretest as right-skewed. However, after the teaching experiment, no PMT's reasoning about the skill was at the idiosyncratic level. The rate of the PMTs at the verbal reasoning did not change before and after the teaching experiment. One out of 14 PMTs was at the verbal reasoning level in both periods. The PMT6 correctly specified the shape of the distribution but could

not state anything about the sample. In the semi-structured interview, she expressed, "It would not be said anything as we do not know the number of samples." However, the number of samples was given in both questions. In addition, the sample size was not needed to interpret the position of half of the sample according to the average. At the transitional reasoning level, a few PMTs mentioned where the data was concentrated or how the outlier affected the mean value before (14.3%) and after (7.1%) the teaching experiment. That is, they focused on a single characteristic of the distribution (e.g., shape or spread). They could make some judgments based on the distribution's range, average, density, or extreme values with their numerical values.

The PMTs were at the procedural reasoning level if they identified characteristics of more than one distribution and could not integrate them. Before the teaching experiment, 57.2% of the PMTs were at the procedural level, whereas after the teaching experiment, this ratio decreased to 28.6%. In Question 3, the PMTs at this level either stated that the mean was outlier sensitive or that the mean should represent the majority. In other words, they estimated the average with skewness or density. However, they could not relate the majority and outliers in using the mean for the average. For example, the response of PMT7 at the procedural reasoning level to Question 3 (See Figure 4.3) was as follows:

> Extreme values do not show us the real situation. The average is 27000. 78000 is the outlier, which increases the mean. So, the knowledge in c doesn't give a general inference. Statement b does not include any interpretation of the data. There is just the knowledge according to the graph (PMT7).

Although PMT7 was able to explain the effect of the outlier on the mean, she could not interpret the meaning of the intervals consisting of quartiles in statement b by integrating density. In other words, she stated that the mean did not represent the data because of only the outlier. On the other hand, the PMTs at the integrated process reasoning level could coordinate all general characteristics of distribution according to the given graph. In other words, they could justify the sample's representativeness in the graph according to general characteristics such as shape,

center, and spread. Before the teaching experiment, 14.3% of the PMTs were at this level, and after the teaching experiment, the percentage increased to 57.2%. In Question 2, they interpreted the graph based on the median as the average due to a skewed distribution with an outlier. For example, by referring to the median, PMT7 confirmed that exactly half of the students were below the average. In Question 3, the PMTs at the integrated process reasoning argued that knowing only the mean was not enough about the distribution as extreme cases moved away from the interval where the majority of the distribution was located. For example, PMT9 clarified her reasoning for Question 3 in the posttest as follows.

> Just giving range here is a very inadequate measure. Since the range is the simplest of the variability measures. Although measures of central tendency are inevitable for some data groups, it is necessary to look at the distance of each value from the mean, such as the SAD, MAD, and standard deviation, to determine the variability in all salaries for the data group here. What is the distribution around the mean? The central tendency and variability should be looked at together. The box plot, as in summary b, will provide more comprehensive information. Summary d would be an understatement to say the majority of the distribution. We should also look at the outlier, max/min values, etc. Information given in this way is not even valid as it does not cover all 36 people (PMT9).

Briefly, according to Table 4.3, the PMTs' reasoning about representations of data was generally concentrated at the procedural level. The cluster at the procedural level (57.2%) before the teaching experiment progressed into the integrated process reasoning level at the end. However, there was no considerable change in the other levels due to the small number of PMTs; the PMTs' reasoning about representations of data improved after the teaching experiment. Following that, the reasoning level of the PMTs about shape and shifts in data was investigated below.

4.1.3 Reasoning about Shape and Shifts in Data

Reasoning about the shape and shifts in data was another skill required when comparing distributions. It referred to interpreting the location of the majority, cluster, clump, or hill of the data to find pattern/global shifts in the data (Bakker &

Gravemeijer, 2004; Cobb, 1999). That is to say, the main concern in the scope of the skill was to see the data as a whole by integrating all aspects of the distributions to compare them. When comparing distributions, the PMTs were asked to compare different aspects of the distribution contextually appropriately. More specifically, to compare distributions, they needed to judge the extreme values and average, interpret and conclude on the spread, density, and shape by considering the entire distribution. The PMTs compared both distributions with equal sample sizes and distributions with unequal sample sizes. In the pretest and posttest, Question 4 required comparing distributions with equal sample sizes, while Question 5 involved comparison distributions with different sample sizes. Since these questions required seeing the data as a whole and deciding on the appropriate distribution for the context, they were evaluated for reasoning about shape and shifts in data.

Question 4 in the pretest and posttest had a different context. In the pretest, the impact of helmet lining on reducing heat exhaustion in 60 football players was investigated. Both temperature data from helmets with and without lining showed a right-skewed distribution. However, Question 4 of the posttest included one left and one right-skewed distribution of data from 100 river sites before and after a treatment method, aiming to reduce harmful zinc in the river. In both questions, the expectation was to select the distribution concentrated on the lower values. As seen from the questions in Figure 4.4, the distributions with lining and after the treatment were clustered on the left side.

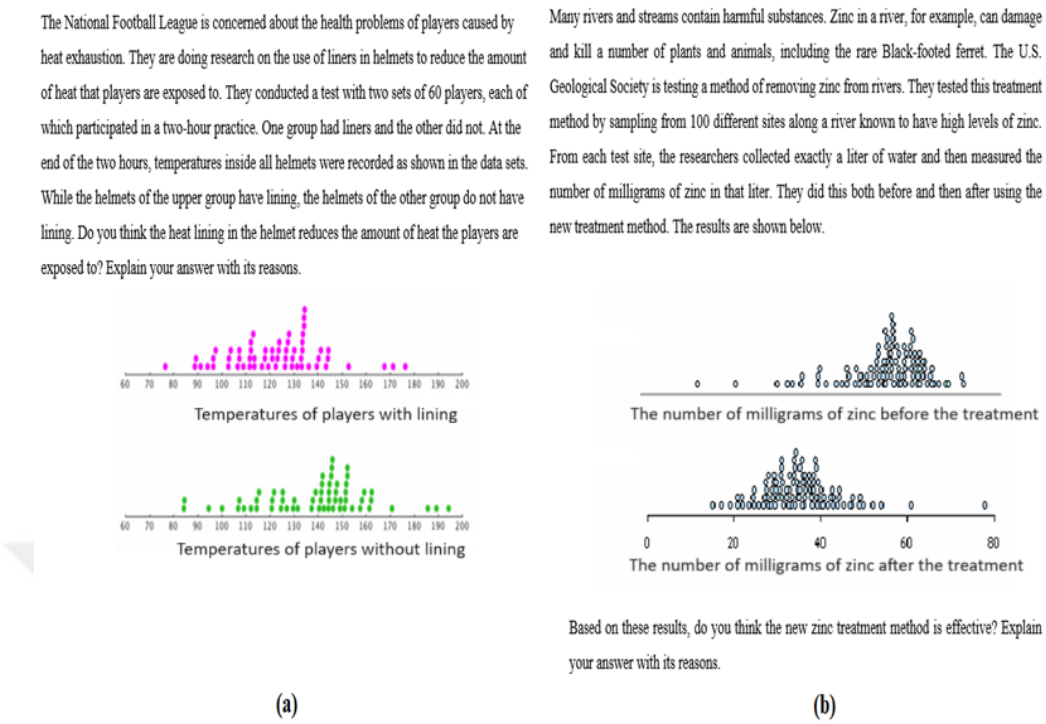


Figure 4.4. Question 4 in the pretest (a) and posttest (b)

The fifth question in the pretest and posttest were long and consisted of structurally different multi-stage items. Question 5 in the pretest asked whether the company's CEO should provide free flu shots to all employees. The PMTs were expected to evaluate the conclusions of four executives utilizing different representations based on the distributions with unequal sample sizes (See Appendix E for further details). Question 5 in the posttest involved comparing distributions with unequal sample sizes for medicines developed to alleviate headaches. In the question, in addition to comparing distributions, the PMTs were asked to structure the data with graphical representations to support their conclusions (See Appendix F for further details). Similar to Question 4, in Question 5, the favor should be towards the distribution clustered on lower values.

When determining levels for reasoning about shape and shifts in data, misusing some statistical words and symbols or scrambling them with unrelated information or personal experiences was categorized at the idiosyncratic level. The fact that the concepts related to distribution remained only in words without practicing was

evaluated as verbal reasoning. The reasoning levels of the PMTs about shape and shifts in data were assumed at the transitional level when the PMTs made comparisons based on only one of the distribution's center, spread, density, and skewness aspects. Multiple distribution aspects were mentioned at the procedural reasoning level but could not be integrated while comparing distributions. When all aspects were used by associating them and comparing distributions was explained confidently, it showed the integrated process reasoning level. Table 4.3 indicates the reasoning levels of PMTs while reasoning about shape and shifts in data before and after the teaching experiment.

Table 4.3 The reasoning levels of PMTs while reasoning about shape and shifts in data before and after the teaching experiment

	Before		**After**	
	f	**(%)**	**f**	**(%)**
Idiosyncratic Reasoning	1	(7.1%)	-	-
Verbal Reasoning	2	(14.3%)	1	(7.1%)
Transitional Reasoning	5	(35.7%)	2	(14.3%)
Procedural Reasoning	4	(28.6%)	3	(21.4%)
Integrated Process Reasoning	2	(14.3%)	8	(57.2%)
Total	**14**		**14**	

According to Table 4.3, the PMTs' levels while reasoning about shape and shifts in data improved substantially after the teaching experiment. Before the teaching experiment, most of them (64.3%) reasoned at the transitional and procedural levels. That is, they compared distributions according to at least one aspect of distribution. After the teaching experiment, their reasoning about shape and shifts in data was generally at the integrated process level (57.2%). That is to say, when comparing distributions, they were not only able to use at least two aspects of distribution but also to relate them. On the reasoning level basis, before the teaching experiment, only one PMT stated her ideas rather than statistical knowledge at the idiosyncratic

level. For Question 5 in the pretest, she noted that there were not enough samples in the research for the flu shot and that the flu was caused by the differences in the bodies of these people, such as their genetic codes, cancer treatment, and immune system. After the experiment, no one was at the idiosyncratic level where statistical terms were misused or irrelevant information was given. At the verbal reasoning level, a few PMTs mentioned how to compare distributions but did not apply them before (14.3%) and after (7.1%) the teaching experiment. They could not express the numerical values clearly, even though they made judgments based on the distribution's range, average, density, or extreme values. For example, PMT8, at the verbal reasoning level, answered Question 4 in the pretest based on density qualitatively, "I will say the lining in the helmets reduces the amount of heat. Because in the without-lining chart, the points have higher values."

At the transitional reasoning level, beyond verbal reasoning, the PMTs compared the distribution quantitatively according to an aspect of the distribution. In addition, they could not justify the reason for using that aspect. Before the teaching experiment, 35.7% of the PMTs compared the distribution according to the value of one of the center, spread, skewness, and density aspects. After the teaching experiment, this proportion at the transitional reasoning level was 14.3%. Below are examples of the PMTs' answers comparing distributions according to the mode, density, range, median, and mean, respectively. The explanations of PMT1 and PMT14 were given for Question 4, while the others pertained to Question 5 in the pre-test.

> When I examine the data distribution, the first is in the 55s, and the next is in the 35s. The mode is down. That is why it is affected (PMT1).
>
> Those with lining helmets are generally concentrated between 110 and 140, while those without lining helmets are concentrated between 140 and 160. This clearly shows that the lining effectively lowers the heat (PMT14).
>
> I think this analysis will also be helpful. As the number of days that have not arrived increases, a situation may arise that may be harmful to the workplace. The less time you do not come to work, the better (PMT3).
>
> As can be understood from the graph, there is more gap between the lower and upper half of the data in no flu shots, 3 (PMT6).

> The new medication mean is 54.18, while the old one is 73.96. As a result, I can say that the new medication reduces the duration of headaches by approximately 20 minutes (PMT4).

Contrary to transitional reasoning, at the procedural reasoning level, 28.6% of the PMTs before the teaching experiment and 21.4% of the PMTs after the teaching experiment mentioned more than one aspect of center, spread, density, and skewness to compare distributions but could not associate them with each other. For example, when comparing distributions with equal sample sizes in Question 4 of the pretest, PMT11 gave information about the center and spread of the distributions without integrating them. She stated, "It reduces the temperature because, with lining, the average is approximately 120-130, and the maximum value is nearly 180. However, the average is approximately 150 without lining, and the maximum value is nearly 200." Similarly, at the procedural reasoning level, the aspects could not be associated with each other when comparing distributions with unequal sample sizes in Question 5 in the tests. In the pretest, Question 5 included conclusions based on range and box plot representations. While interpreting the conclusions, the PMTs could not explain why the range was insufficient and what three means in the box plot by integrating them with other aspects. Unlike in the pretest, for Question 5 of the posttest, the PMTs made their inferences and structured the data in representations. Their reasoning at the procedural level included at least two aspects but was based on only one of them without establishing a relationship. For example, although PMT13 stated the shape of the distribution based on the density, she compared the distributions according to the mean in Question 5 in the posttest. Essentially, when comparing distributions, she could not demonstrate a comprehensive understanding of shapes and shifts in data. Her answer with graphs is below.

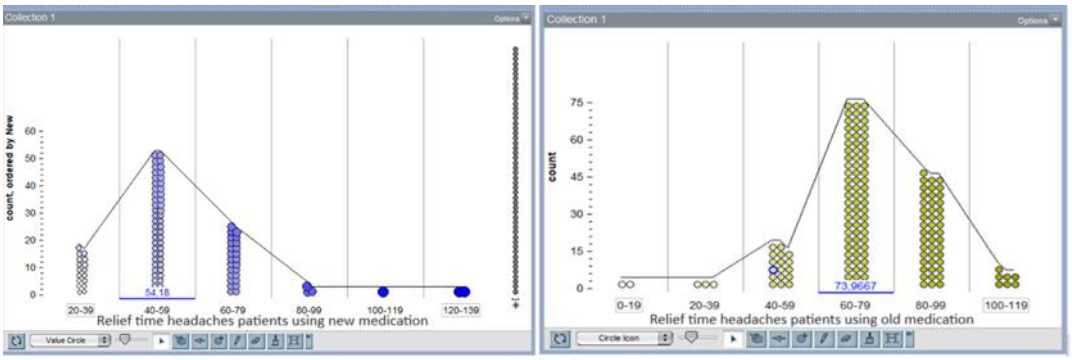


Figure 4.5. The graphs that PMT13 drew in TinkerPlots in Part 2 of the posttest

> I can compare percentages here, but a paradox can arise when it is broken down into pieces. I look at the means to avoid the paradox. The new medicine mean is 54, and its graph shape is right skewed. It shows that the data is concentrated on the left side, and we want to decrease the headache time. However, when we look at old medicine, its mean is 73, and its shape is a normal distribution. It helps to see that new medicine is an effective way. Moreover, their ranges are close, 20-130 and 10-120 (PMT13).

Despite expressing the shape, PMT13, who suggested comparing the percentages for each interval she drew, compared the means. The shape of the old medicine was not normal. Despite the incorrect association of density with shape, she did not justify the center with density. Moreover, she ignored the effect of the extreme values due to the equal range corresponding difference between the values.

The inferences were made by associating distribution aspects at the integrated process reasoning level. For example, when comparing equally sampled distributions in Question 4 in the pretest, PMT2 confidently explained the relationship between spread and center. Before the teaching experiment, his reasoning was as follows.

> The mean is approximately 125 with lining groups. However, the mean is 145 on average when we consider the without-lining groups. Also, the max heat is 176 in the lining group, while the max heat is 195 in the without-lining group. Moreover, considering the minimum heat, it is 78 in the lining group and 85 in the without-lining group. The range also shifted to the right. Therefore, we definitely say that heat lining reduces the amount of heat the players are exposed to (PMT2).

PMT2 at the integrated process level compared minimum, maximum, and mean values. In addition, he could justify that the temperature distribution in helmets

without lining shifted to the right with the help of range by combining three values. In other words, he used the range as a supplement to the mean by integrating variability and the center of the distributions. After the teaching experiment, he could also interpret the shapes of the distributions in terms of the mean and density and compare the interquartile ranges. It showed that he was able to relate all aspects of the distribution.

When comparing distributions with unequal sample sizes at the level of integrated process reasoning, the distributions' center, spread, density, and skewness were fully integrated, and a conclusion was confidently reached. Before the teaching experiment, the inadequacy of range alone in Question 5 in the pretest to compare distributions was explained with density, average, and outliers. In addition, the PMTs at this level interpreted the density according to five values in the box plot. For example, PMT8 commented on the range and box plot in Question 5 of the pretest in the decision-making process as follows.

> We can see the range in this analysis (including the box plots). However, range alone is not a factor that can assist in decision-making. In this analysis, we see five different values for each group. They are max, min, median, 1, and 5.25 values. Each of these values gives us more information about the distribution of data. Although the group with flu shots is spread over a wider area from 0 to 13, the majority of them (75%) are absent, which is less than three days. This shows that spread over a wider area does not result in more absenteeism. I also think the distribution of employees who did not get a flu shot is more reliable. Therefore, I think this analysis will give the CEO more accurate information and help her decide (PMT8).

PMT8 claimed that the box plot already gave information about the range, consisting of five values (minimum and maximum values, first quartile-Q1, Q2-median, and third quartile-Q3). She could deduce the spread of the data as well as the range in the box plot. Accordingly, she explained that a reliable conclusion could not be obtained by looking at only the extensive range instead of density. In the context of the question, an extensive range was a disadvantage. However, when she compared the third quartile-Q3, which is 75% of the data, she demonstrated that the distribution with the more extensive range concentrated in the smaller values. This inference

referred to the fact that evaluating data from a single aspect may lead to misinterpretations, and therefore, how the data is distributed should be considered. Distinctly, she was able to interpret the shift in the data by relating spread and density. Therefore, her reasoning about shape and shifts in data was at the integrated process level.

After the teaching experiment, the PMTs compared distributions with unequal sample sizes in Question 5 of the posttest at the level of integrated process reasoning, considering different aspects of the distributions. They concluded according to the representations in which they structured and organized the data in TinkerPlots. They usually support the density of the data with the center or spread. Thus, they could interpret the shift in the data. For example, PMT8 generated the dot plots in Figure 4.6 and explained them as follows.

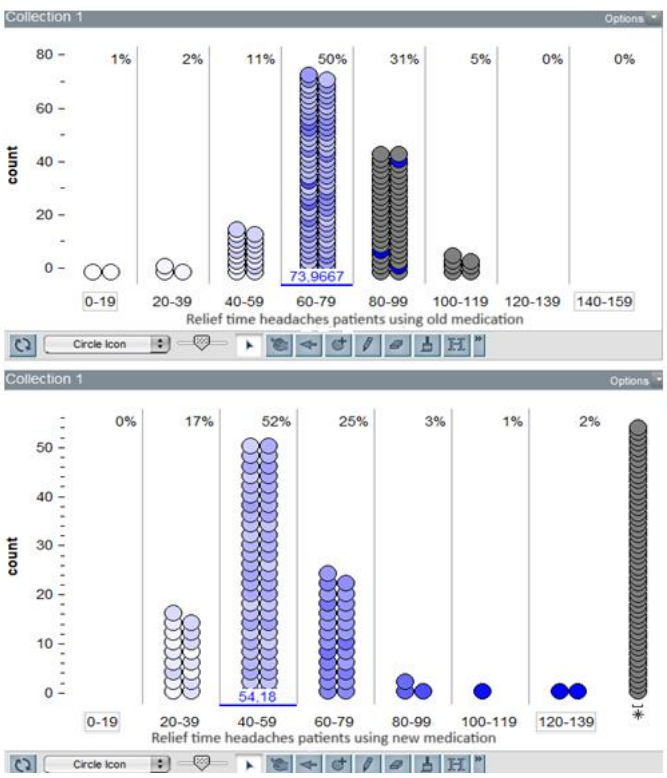


Figure 4.6. The graphs that PMT8 drew in TinkerPlots in the posttest

> While the duration of action of the old medication was clustered around 75, the duration of action of the new medication was clustered around 55. Also, the ranges of both data are almost equal. As can be easily seen from these graphs, the old medication data is left skewed while the new medication data is right skewed. In addition, if we take 70 minutes as a cut point, 66% of patients using old medication waited longer than 70 minutes for the pain to go away. This rate is 12% in new medication. The medications' mean is 73,96 and 54,18, respectively. This means that the average effect time of the new medications is less than that of the old medication. Also, the standard deviations of these two data are 15.64 and 17.78, respectively. This shows that the consistency is close to each other (PMT8).

While comparing the distributions with unequal sample sizes, PMT8 considered the upper part of the distribution based on the idea that individuals' headaches should go away within 70 minutes. In other words, she attended to the low rate of the upper part in the new medication by comparing the eight equal-width intervals of the graphs in Figure 4.6. She realized that both distributions' range and standard deviation were similar regarding variability. In the interview, she explained the higher standard deviation in new medication with more extreme values, such as 10 or 120 minutes. According to her, this was inconsistent data. However, she cared about concentrating the data at low values rather than consistency according to the context. Since she considered the distributions' shape and mean, as well as their variability, she was at the integrated process reasoning level while comparing distributions. That is to say, she could examine the whole distribution through different aspects. In this statistical process, PMT8 confidently interpreted the aspects according to the context while reasoning about the shape and shifts in the data.

In the example above, distributions were compared based on percentages above and below the cut point, which was 70 minutes. In addition to comparing the rates in equal-width intervals concerning a specific reference point, the differences in percentages in each interval were processed by two PMTs. For instance, PMT3 partitioned her graph as in Figure 4.7, and she explained her reasoning as follows:

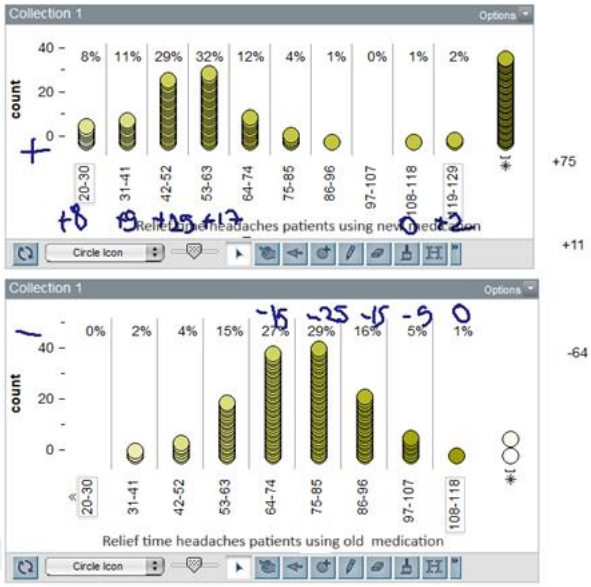


Figure 4.7.The graphs that PMT3 drew in TinkerPlots in the posttest

> First of all, since the sample sizes are not the same, we cannot make an inference by looking at these data directly. Dividing the data into groups and taking percentages for each group gives a better result. I compared the same ranges as a percentage. I got the difference between them in the comparison. If the old drug is bigger, it is (-); if the new drug is bigger, it is (+) when I compared all the percentages like this. Add the values of the upper one +75 and the lower -64. From here, the total value was +11. I saw that the effect of the new drug was intense in shorter periods, which was my expectation. So, the new drug acts faster (PMT3).

PMT3 focused on the lower values based on the context to reduce the duration of the relaxation. She was interested in the dispersion of signs assigned to the distributions to decide which medication was more effective. She favored the medication with a sign that was more at low values. In the semi-structured interview, she explained her reasoning in more detail. She said that the + values were in the majority in the first four intervals of the above graph, and the - values in the upper values were in the majority in the below graph. She also obtained this result by adding the values in each interval. When she summed the differences in her description (+11), there was a trend in favor of the new medication whose sign was +. Although this reasoning may seem like comparing individual data values in the distributions, it can be seen as a transitional stage to view the distribution as a whole, as PMT3 considered all

parts of the distributions by comparing percentages. Essentially, PMT3 was looking for where the shift in those distributions was. In addition, she also compared relative frequencies in the intervals, not individual data values. She took particular care to ensure the comparison was within the same interval.

When reasoning about shape and shifts in data, at the integrated process reasoning level, data analysis was extended beyond the interpretations of the relative frequencies in equal-width intervals. Some PMTs examined the shifts in data based on an equal proportion of data values within the intervals with different widths. That is to say, the PMTs structured the box plot with 25% of the data in each interval with different widths for the last question of the posttest. They compared quartiles with different aspects of distribution when reasoning about shape and shifts in data. For example, PMT12 compared quartiles in box plots in Figure 4.8 and expressed the range each quartile fell as follows:

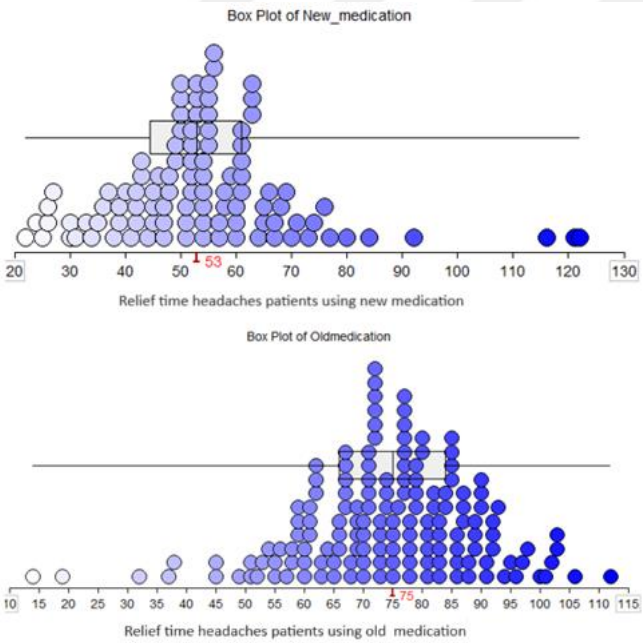


Figure 4.8. The graphs that PMT12 drew in TinkerPlots in the posttest

> In the new medication, %25 of the patients' headaches have gone less than 44.25 minutes. %50 of the patients' headaches have gone less than 53 minutes,

> and %75 of the patients' headaches have gone less than 61 minutes. On the other hand, %25 of the patients' headaches have gone less than 66 minutes in the old medication. 50% of the patients' headaches have gone less than 75 minutes, and %75 of the patients' headaches have gone less than 84.25 minutes (PMT12).

In addition, she wrote that "the average is a typical value; it is representative of the data. In this case, the median should be used as an average since there are outliers" in Statement C in Part 3 of Question 5 in the posttest. In the interview, she stated that she could use only the medians that were 53 and 75 minutes. However, she preferred interpreting the box blot because it provides quartiles for how the data is distributed. PMT12 was at the integrated process reasoning level, as density was the focus of her reasoning when reasoning about shape and shifts in data. She could also relate center with skewness.

As a result, the teaching experiment contributed to developing the PMTs' reasoning about shape and shifts in data. While comparing distributions, the number of PMTs at the transitional and procedural levels before the teaching experiment shifted to the integrated process level after the teaching experiment. Those at the idiosyncratic and verbal levels also moved to the transitional and higher levels. After the teaching experiment, the decrease in the first four levels transformed the increase at the integrated process reasoning level. The ratio at the integrated process reasoning increased from 14.3% to 57.2%.

4.1.4 Proportional Reasoning

Another skill used for reasoning about the distribution was proportional reasoning. The PMTs' use of proportional reasoning was revealed by their answers to the last question in the tests. The above mentioned that in Question 5 of both pretest and posttest, the PMTs were expected to compare distributions with unequal sample sizes. More specifically, in the pretest, the number of data obtained from employees with and without the flu shot was different to see whether the flu shot reduced absenteeism from work. Seventy-six employees got flu shots, and 162 employees

did not. That is, the number of people who did not get vaccinated was about twice as many as those who did. Therefore, relative frequencies rather than absolute ones are required to be compared. However, the absolute frequencies were compared in Part A of the pretest. The comparison of frequencies was as follows. "Of the 28 people that had zero absentees, 15 did not get a flu shot, and 13 did. This indicates that flu shots were ineffective because more people who were not immunized had zero absentees." Although it seems reasonable to tend towards a distribution with a high number of employees with zero absentees based on context, it is expected that there will be approximately 26 employees who have flu shots. When proportioning the sample sizes, it can be inferred that those who have flu vaccination are less absent. The PMTs were expected to recognize and explain this contrast. Part B of the pretest had a two-way table with relative frequencies (percentages) concerning a cut point of one-day absentees. It included the conclusion that "41% of the people who got a flu shot missed work one day or less, versus 26% of the people that did not get a flu shot. The flu shots seemed to have reduced the number of days that employees missed work." Since the context required missing work less, there was supposed to be a bias in favor of the higher frequency in the part with fewer than one-day absentees. In this case, since the comparison was based on comparing percentages rather than absolute frequencies, the PMTs were expected to justify and confirm this conclusion.

In Question 5 of the posttest, a company with a drug formula for relieving headaches had developed a new medication formula. To test the effectiveness of the new formula, the data was collected from patients taking one of the two drugs. The number of patients taking the new and old medication to reduce headaches was also different. There were 100 patients on the new medication and 150 on the old medication. In other words, the number of patients on the old medication was 1.5 times as many as those on the new medication. The statement "I would not conclude anything from these data. Because there are different numbers of patients in the two groups, there is no fair way to compare the two formulas" in Part 3 of the question required disagreeing. Because a reasonable conclusion could have been reached with

the multiplicative relationship. Again, proportional reasoning was needed to make a reasonable comparison. The PMTs had to compare relative frequencies in their own words and inscriptions in Part 1 and Part 2 of the question. For faster headache relief, there was an effect in favor of the formula with more frequency in smaller values in the distributions. In this regard, the new medication was effective due to the higher proportion at low values.

In summary, the responses for Part A in the pretest and Statement D of Part 3 in the posttest showed the PMTs' justification for using proportional reasoning. Part B of the pretest and Part 1 and Part 2 of the posttest addressed how the PMTs used proportional reasoning and their interpretations for using relative frequency. Due to the structure of Question 5, the reasoning levels of the PMTs for proportional reasoning were presented by combining these parts of the questions. As given in Table 3.6 in the data analysis, the reasoning levels of the PMTs were determined according to the following criteria. They were at the idiosyncratic level when they could not think relatively and used additive reasoning. At the verbal reasoning level, they encouraged using relative thinking without implementing it. In other words, although they justified the need for a proportional relationship due to the difference in sample size in Part 1 of the pretest and Part 3 of the posttest, they could not apply this in the other parts of the questions. Answers for qualitative proportional reasoning without using numbers indicated the level of transitional reasoning. At the procedural reasoning level, they employed relative and proportional reasoning in incomplete and invalid ways. They were at the integrated process reasoning level when using relative and proportional reasoning completely. Table 4.4 presents the reasoning levels of the PMTs for using proportional reasoning before and after the teaching experiment.

Table 4.4 The reasoning levels of PMTs for proportional reasoning before and after the teaching experiment

	Before		After	
	f	**(%)**	**f**	**(%)**
Idiosyncratic Reasoning	3	(21.4%)	-	-
Verbal Reasoning	3	(21.4%)	2	(14.3%)
Transitional Reasoning	2	(14.3%)	-	-
Procedural Reasoning	1	(7.1%)	2	(14.3%)
Integrated Process Reasoning	5	(35.7%)	10	(71.5%)
Total	**14**		**14**	

Table 4.5 shows that three out of 14 PMTs were at the idiosyncratic reasoning level before but not after the teaching experiment. They focused on the differences between the numbers given in Part A. In other words, they compared absolute frequencies. For example, PMT6 confirmed the use of absolute frequency for the unequal sample size as follows:

> The number of employees is too low to make a decision. 76 employees got a flu shot, and 162 did not. We should involve another department of the company in the experiment. In this interpretation (Part A), there are only 15-13=2 differences. In addition to that, the research is trying to find the effect of the flu shot on the employees' absentees, not trying to find the employees' zero absentees. This data (Part B) may give more detailed information to decide because it takes less and more than one day. It increased the timeframe. At first, it looked at 0, but here it is, looking at the total. However, the number of employees is still low enough to decide on this (PMT6).

PMT6 thought that it was not representative of the population based on the idea that the numbers were low. In addition, she drew attention to the adequacy of looking at 0 days for non-attendance. According to her, the number of days missed should also be considered, as in Part B. Similarly, the PMTs at the verbal reasoning level claimed the use of percentages to compare all the data. However, they did not implement it with quantities in their action. 21.4% of the PMTs before and 14.3% of the PMTs after the teaching experiment expressed the necessity of using relative frequency without implementation at the verbal reasoning level. They also could not identify

the actual reason for using relative frequencies, which was attributed to unequal sample sizes. In addition, they did not make comparisons between numbers. For example, before the teaching experiment, PMT9 stated that the use of percentages is a more statistical language as follows:

> These 28 randomly selected from those without absenteeism are valid for the whole number. It would be more persuasive if he could conclude all the people without absences here. I think the second graph is appropriate because he grouped the vaccinated and non-vaccinated individuals according to their absenteeism and expressed this with the percentage usage, which is a common statistical language. We can observe the vaccine's effect by comparing these percentages (PMT9).

After the teaching experiment, the PMTs at the verbal reasoning level repeated the expression in Statement D that "there are different numbers of patients." For example, PMT4 agreed that a conclusion could be drawn from the data and stated that there were different numbers of patients as a justification. However, he did not explain how to conclude. On the contrary, the PMTs who advocated using proportional reasoning based on different sample sizes were at the transitional reasoning level. They explained qualitatively the necessity of using relative thinking according to the criteria for verbal reasoning level, but there was no attempt to apply it quantitatively. For example, before the teaching experiment, PMT14 confirmed the use of percentages for different sample sizes, as follows.

> The difference in the total number of vaccinated, 76, and unvaccinated workers, 162, is large. It cannot be a defining quality. There is a need for equality between those who are vaccinated and those who are not. The sample size must be equal. The percentages in the second table ensure this (PMT14).

Before the teaching experiment, although two PMTs at the transitional reasoning level did not provide numerical evidence for proportional reasoning, they were able to highlight the need to compare relative frequencies because of unequal sample sizes. The PMTs who advocated using percentages (14.3%) did not use them. However, after the teaching experiment, the PMTs who mentioned the use of relative frequency implemented this idea. In other words, there were no PMTs at the transitional reasoning level after the experiment.

The PMTs' levels regarding proportional reasoning for the procedural or integrated process level were determined according to the implementation of proportional reasoning. The PMTs who used proportional reasoning incorrectly or incompletely were at the procedural reasoning level. One PMT before and two PMTs after the teaching experiment were at this level. For example, PMT10 interpreted Part A and Part B in the pretest before the teaching experiment as follows:

> The number of people is different. It could be commented like this (Part A) if it was the same. This table shows that the employees who did not get a flu shot missed work more than one day. However, I think there is a misconception because even if the rate decreases when the employees get a flu shot, the number still not coming for more than a day is more as a percentage in both. 59% is more than 41%. Again, 74% is more than 26%. It is confusing (PMT10).

Although PMT10 compared the percentages due to unequal sample sizes, she considered rows of the table. She used the percentages invalidly and compared the vaccine groups among themselves. This unreasonable way of comparing different parts of different wholes made her unable to use proportional reasoning properly. Therefore, she was at the procedural reasoning level while using proportional reasoning. After the teaching experiment, the use of proportional reasoning was observed through the graphs structured in Part 2 of Question 5 in the posttest. For example, PMT9 partitioned the data into three intervals with equal width, as in Figure 4.9. She explained her reasoning in the test as follows:

> Although there are different sample sizes, we can analyze them with some methods regardless of the number of people. For example, I tried to make equal-width intervals in the three groups. We are trying to reduce the time. When I look at the longest durations, there are 53 patients in the old medication's highest values, while the other has four patients. Again, when we look at the lower values, five patients are in the distribution of the old medication, while there are 57 patients in the new medication (PMT9).

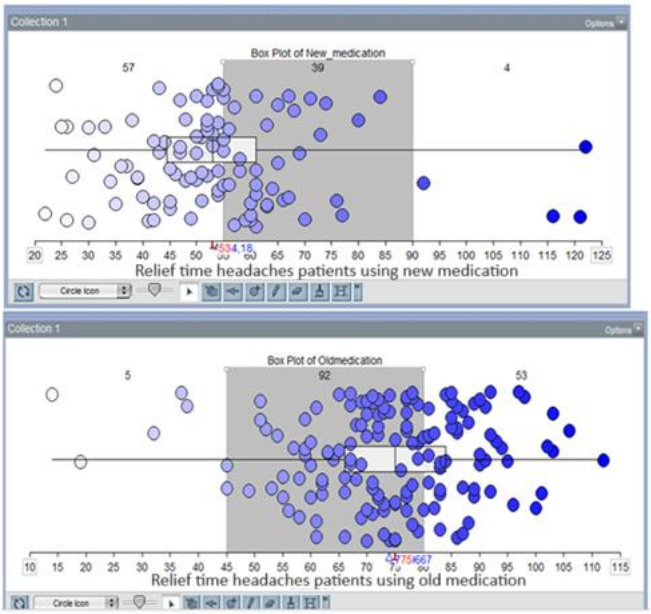


Figure 4.9. The graphs that PMT9 drew in TinkerPlots in Part 2 of the posttest

From her explanation, PMT9 was aware that the number of samples was different even though she was comparing absolute frequencies at the extremes of the graphs. In the semi-structured interview, she stated that she ignored the difference in sample sizes and that looking at the percentage would be more reliable. Although she adapted absolute frequency to percentage, she incompletely used proportional reasoning. Therefore, PMT9 was at the procedural level in using proportional reasoning.

The PMTs at the integrated process level, who explained the reasoning for using the proportions for unequal sample sizes with their numerical values, compared the relative frequencies correctly. They compared the proportions completely. For example, after the teaching experiment, PMT7 compared percentages using a value as a cut-point, 79 minutes. Below are the graphs she drew in Part 2 and her reasoning.

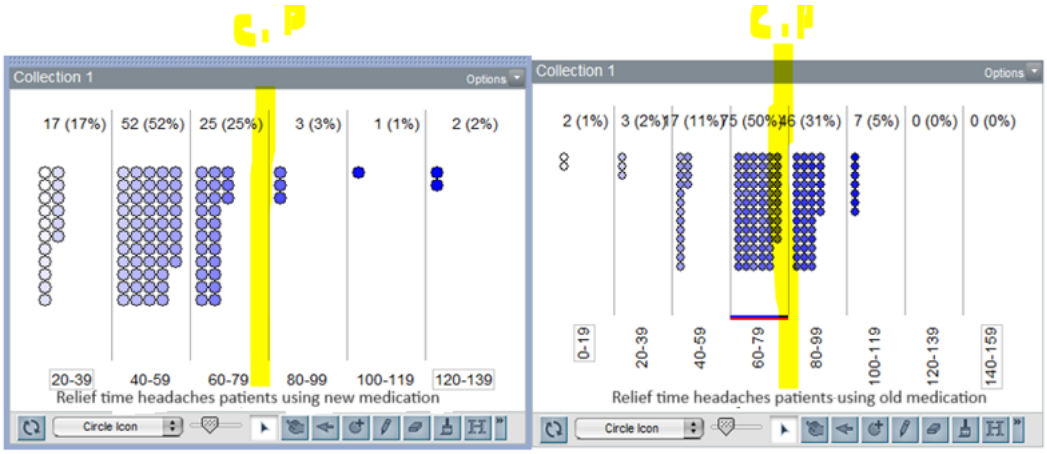


Figure 4.10.The graphs that PMT7 drew in TinkerPlots in Part 2 of the posttest

> When comparing these treatments, we want minimal time for the patient to report that the headache is gone. In other words, it is the situation we want to have in a small number of minutes. When we examine the new treatment method, we see that extreme values exist, but for us, exceptions should not affect the overall situation. In this case, we can conclude by determining an important threshold value for us and comparing percentages below and above this value. I set 79 as a threshold value and made the inference that this treatment does not work if the number of patients who say that the headache has passed over 79 minutes increases. In this sense, 36% of patients using the old treatment method are above this value, and 6% of the patients using the new treatment method are above this cut point. The difference is too much, and we can say that the new treatment method is very effective (PMT7).

PMT7 reasoned that what was desired was a lower relaxation time. When making comparisons, she looked at rates of relaxation times over 79 minutes. Although she focused on over 79 minutes, ignoring exceptions of the new medication in her answer, she was also aware that the rates of the below 79 minutes could be compared. In the interview, she did the opposite of the upper for below 79 minutes, explaining that the higher the ratio of the bottom, the better the medication. In comparing both the upper and bottom ratios, she reached a consistent result, namely that the new treatment was more effective. Similarly, before the teaching experiment, the PMTs at the integrated process reasoning level correctly compared the percentages in the two-way table. For example, PMT5 interpreted the table in Part B of Question 5 in the pretest as follows.

> No, a generalization cannot be made by looking only at employees who do not have absenteeism. Since the number of people not getting a flu shot is much higher, it is natural to have a higher number than those who get a flu shot. The percentages show that the flu shot worked. Those who do not have the flu shot often take more than one day off, 74%. It can also be compared to less than one day or one day. Those who take leave of one day or less have more vaccinations—41% vs 26% (PMT5).

Before the teaching experiment, 35.7% of the PMTs completely used relative and proportional reasoning, while after the teaching experiment, the ratio of the PMTs at the integrated process reasoning level was 71.5%. In other words, the number of PMTs doubled after the teaching experiment. Therefore, it can be said that the PMTs' use of proportional reasoning improved after the experiment.

To summarize, the reasoning about distribution encompassed the skills of reasoning about statistical measures, reasoning about representations of data, reasoning about shape and shifts in data, and proportional reasoning. Following the teaching experiment, the number of PMTs at the idiosyncratic, verbal, and transitional levels for all skills decreased after the teaching experiment. Similarly, there was also a decline in the number of PMTs at the procedural level; an increase from 7.1% to 14.3% occurred only while using proportional reasoning. On the other hand, the number of PMTs at the integrated process reasoning level increased in all skills. The increase rates at this level were 42.9% in each skill of reasoning about statistical measures, data representations of data, and shape and shifts in data, and 35.7% for proportional reasoning. That means the PMTs' reasoning levels shifted considerably to the integrated process reasoning level after the teaching experiment. After the teaching experiment, more than half of PMTs (57.2%) at the integrated process level could associate the aspects of distributions. They had a complete understanding of the distribution. More specifically, they could interpret the distribution's center, spread, density, and skewness logically by integrating them while deciding based on the context. In light of the results, it was concluded that the PMTs' reasoning about distribution improved after the classroom teaching experiment enriched by dynamic statistics software, which answered the first sub-research question of the study.

4.2 The Emerged Classroom Statistical Practices Throughout the Teaching Experiment

This study concerns how the PMTs' reasoning about distribution developed throughout the classroom teaching experiment based on RME and enriched by dynamic statistics software. In the above section, it was shown that the PMTs' reasoning about the distribution developed based on reasoning about statistical measures, reasoning about representations of data, reasoning about shape and shifts in data, and proportional reasoning skills. This development emerged after the classroom teaching experiment based on RME and enriched by dynamic statistics software. Since the teaching experiment was carried out in the social environment of the classroom, to reveal how PMTs' reasoning improved, it was necessary to examine what happened in the classroom during the teaching experiment. Accordingly, an analysis was conducted to identify CSPs that nurtured this reasoning development, aligning with the emergent perspective—the theoretical framework underpinning the study. This examination answered the second research question of the study: What CSPs emerged throughout the teaching experiment based on RME and enriched by dynamic statistics software? CSPs were taken-as-shared reasoning, arguing, and symbolizing ways that evolve when discussing the distribution concept (Cobb et al., 2001). To reveal the CSPs, the TAS ideas that constitute them were identified. These TAS ideas were confirmed through argumentation schemes, as they were deduced from the arguments articulated by the classroom community while engaging in activities within the HLT. While elaborating on the TAS ideas, the reasoning levels of the PMTs in the TAS ideas were also addressed according to the adapted version of the model of statistical reasoning (Garfield, 2002; Jones et al., 2004). It showed how the PMTs' reasoning development progressed while engaging in activities during the teaching experiment. All these examinations revealed how the PMTs' reasoning about the distribution has evolved throughout the classroom teaching experiment.

The CSPs promoted by TAS ideas evolved and intertwined throughout the process. For example, the PMTs were also concerned with the CSP of selecting appropriate statistics for the context while dealing with all other CSPs. In addition, the TAS ideas were used in more than one CSP. The ideas were discussed where they were the primary purpose of the learning activities in the HLT. In addition, the reasoning levels of the PMTs while using TAS ideas were presented when TAS ideas emerged. The ideas were given in italics under the related CSP. In this section of the findings, the TAS ideas and emerging CSPs were reported in the order outlined in Table 4.5. The table demonstrates five CSPs containing TAS ideas related to reasoning on distribution that emerged throughout the classroom teaching experiment.

Table 4.5 The CSPs and TAS ideas that emerged throughout the classroom teaching experiment

CSP1. Reasoning about the average as the signal in a noisy process
If the sample sizes are the same, the mean and the total say the same thing.
Average is the typical value representing the data.
The mean distributes data items equally.
The mean is responsive to outliers.
The sum of the distances of scores above and below the mean is equal.
CSP2. Measuring of variability in data
Normality is not a criterion for variability.
If the number of samples is different, the MAD is used rather than the SAD.
The MAD is a predecessor of the standard deviation.
CSP3. Selecting appropriate statistics for the context
The measures of variability support each other.
Connecting the measures of center with the measures of variability leads to more reliable results.
Unlike the range, the interquartile range reduces the effect of outliers/extreme values.
CSP4. Exploring qualitative characteristics of collections of data points
The sample varies between extreme cases based on the specific value/mean
The mean of the data set with high data points is high.
The dispersed distribution is unreliable.
The data is concentrated where the mean is

Table 4.5 (Continued)

CSP5. Exploring qualitative characteristics of distributions
The hill of the data set is in the bin with the most frequency.
The intervals allocated to quartiles have different widths.
The density is higher in the narrower intervals of the data partitioned into equal sizes.
The location of the measures of center judges the shape of the distribution.

4.2.1 Classroom Statistical Practice 1: Reasoning about the average as the signal in a noisy process

Five ideas about the average and its formal properties became taken-as-shared in the first three weeks of the teaching experiment, in which the PMTs worked on the first three activities in the HLT to make sense of the average considering variability. They were as follows:

If the sample sizes are the same, the mean and the total say the same thing.
Average is the typical value representing the data.
The mean distributes data items equally.
The mean is responsive to outliers.
The sum of the distances of scores above and below the mean is equal.

The HLT followed in the study started with an understanding of data and data analysis. In this respect, at the beginning of the teaching experiment, the PMTs tried to determine the most important factor while purchasing a printer. The activity related to the printer required creating ranked multiple factors that might be important, making operations and transformations on those ranks to combine them, and, finally, generating relationships between and among quantities transformed for decision-making. For this, PMTs first discussed the use of a printer, what features a printer can have, and whether they have a printer or not. The discussion allowed them to understand the relevant characteristics of the problem, where the data came from, and the data collection process. To familiarize the PMTs with the context of the activity, such conversations took place before all the activities in the HLT. Then, the instructor asked what matters when buying a printer for a project, and PMTs determined five factors in small groups by arguing why the factors were necessary.

The factors specified by the groups were different from each other. Some factors corresponding to the same features were even named differently. For example, cartridge productivity and toner consumption stood for toner/cartridge quality. This confusion of variability in the factors referred to the noisy process. The instructor then requested that similar factors be given the same name. In this case, if the factors listed by the groups were similar to the previous ones, they were categorized by a common name with similar characteristics. The factors determined by the PMTs in small groups were price, toner quality, print quality, electricity usage, paper type, consumer comments, connection (WIFI/Bluetooth), service, number of prints, and printer size. To create a list of five factors that the class agreed on, they first considered the factors common to all groups (e.g., toner quality and print quality) and then categorized factors that refer to similar features under a more inclusive factor. For example, the number of prints corresponding to the number of papers printed by one cartridge was included in toner quality. Afterward, connection and price with higher frequency were accepted as the class's factors because the two groups also preferred them. The service was the last factor on the grounds that it was a priority due to guarantee and accessibility conditions. Accordingly, the factors for the class list determined were toner quality, print quality, connection, price, and service. The PMTs then worked in the same small groups to enumerate from most important to least important factors. The groups presented their rankings to the class along with their reasons. These lists were written on the whiteboard of the online platform for everyone to see. Table 4.6 demonstrates different rankings determined by the groups.

Table 4.6 Groups' ranking of the factors from most important to least important

Group 1	**Group 2**	**Group 3**
1. Print Quality	1. Toner Quality	1. Print Quality
2. Connection	2. Print Quality	2. Toner Quality
3. Service	3. Service	3. Connection
4. Price	4. Price	4. Service
5. Toner Quality	5. Connection	5. Price

The end mission of the PMTs in this activity was to generate a single class list by aggregating ranked data in Table 4.6 and to determine the most important factor while purchasing a printer. However, they ranked the factors differently from most important to least important in small groups, and three lists emerged. PMT10 reacted to this mission with the word- "but very different!" which indicated the difficulty of dealing with the variability in the context. The instructor asked the groups to use different strategies and justify their accuracy to aggregate the ranked data. The intervention, which the instructor repeated throughout the teaching experiment, provided the milestone in establishing sociostatistical norms in the classroom from the first class session. The excerpt below illustrates PMTs' reasoning while compiling one class list based on the different rankings of the same factors in the class discussion.

> PMT2. Print quality is first in two groups and second in the other. We can put it first because it is the most preferred. Therefore, the ranking of Group 2 is eliminated directly. Toner quality is first, second, and last place. Since toner quality is first in Group 2, whose selection of print quality I ignored, I can put the toner quality in second place. I chose Group 3's list since the connection and service are similar.
>
> Instructor. You compared the groups within themselves. Any other ideas?
>
> PMT13. I took the mean of the ordinal numbers of the factors. Connection is second place in Group 1, 5th in Group 2, and 3rd in Group 3. When I added these numbers and took their average, it was 3.3. The means are 1.3 for print quality, 3.3 for connection, 3.3 for service, 4.3 for the price, and 2.6 for toner quality. The order from highest to lowest is print quality, toner quality, connection, service, and price. However, their order can change since the means of connection and service are the same. Group 3's ranking is the closest.
>
> Instructor. What is PMT13's reasoning?
>
> PMT10. Since the best is in the first place, she took the lower mean for the more important in the ranking.
>
> PMT12. But there is an outlier. Toner quality is 5th in the first group but ahead in the others. This is effective when calculating the mean.

The PMTs used different strategies to deal with the challenge of the same factors with different rankings. In this process, the instructor asked whether there were different strategies, encouraging them to share their other strategies with the class.

PMT2 used the mode or frequency-based strategy to determine the ordinal data occurring most frequently in the first and second places. He chose one of the existing lists, thinking it was enough to determine the location of the first two factors. There was also a quantification of the factors determined by the PMTs. That is, the ordinal data were converted to numerical data. In this context, the numbers corresponding to the place of the factors were used. Reasoning the factor with a small mean as more important, PMT13 took the mean of the ordinal numbers in front of the factors, which PMT10 clarified. Regardless of PMT12's objection that the mean was affected by the outlier, the class calculated the means of the factors. That is, they added the numbers assigned based on the order and separated this sum equally to each group. The partitive division performed to obtain how much each group has equally did not result in an integer number for the order. However, it worked in the ranking. Whichever factor was smaller in number was considered more important. The class dialog continued as follows.

> PMT14. On the contrary, I summed it up by giving five points for the first, four points for the second, and so on. The point of print quality is 14, connection eight, service eight, price five, and toner quality 10. The order does not change.
>
> Instructor. Why do you think you get the same ranking with different methods?
>
> PMT12. If we look at the total and average scores, we find the same ranking as the number of samples is the same. We divide the sum by three, so there is no need to do one more operation for the mean. If the sample sizes are the same, the mean and the total say the same thing.

Another strategy was introduced by PMT14, which was reversing the numbers for the factors. He reached the same ranking as the previous one in which factors with the same mean had the same total. PMT12 explained the justification of this sameness with the equality of the number of groups for each factor. In this case, the PMTs ranked the factors based on how much more or less of the quantities were from each other.

The different strategies that the PMTs had used so far, namely, the frequency-based, mean, and total strategies, provided the diversity of the data on which the claim was

based in the argumentation of the class. The argumentation scheme that the PMTs presented in the classroom to re-rank the data was summarized in Figure 4.11.

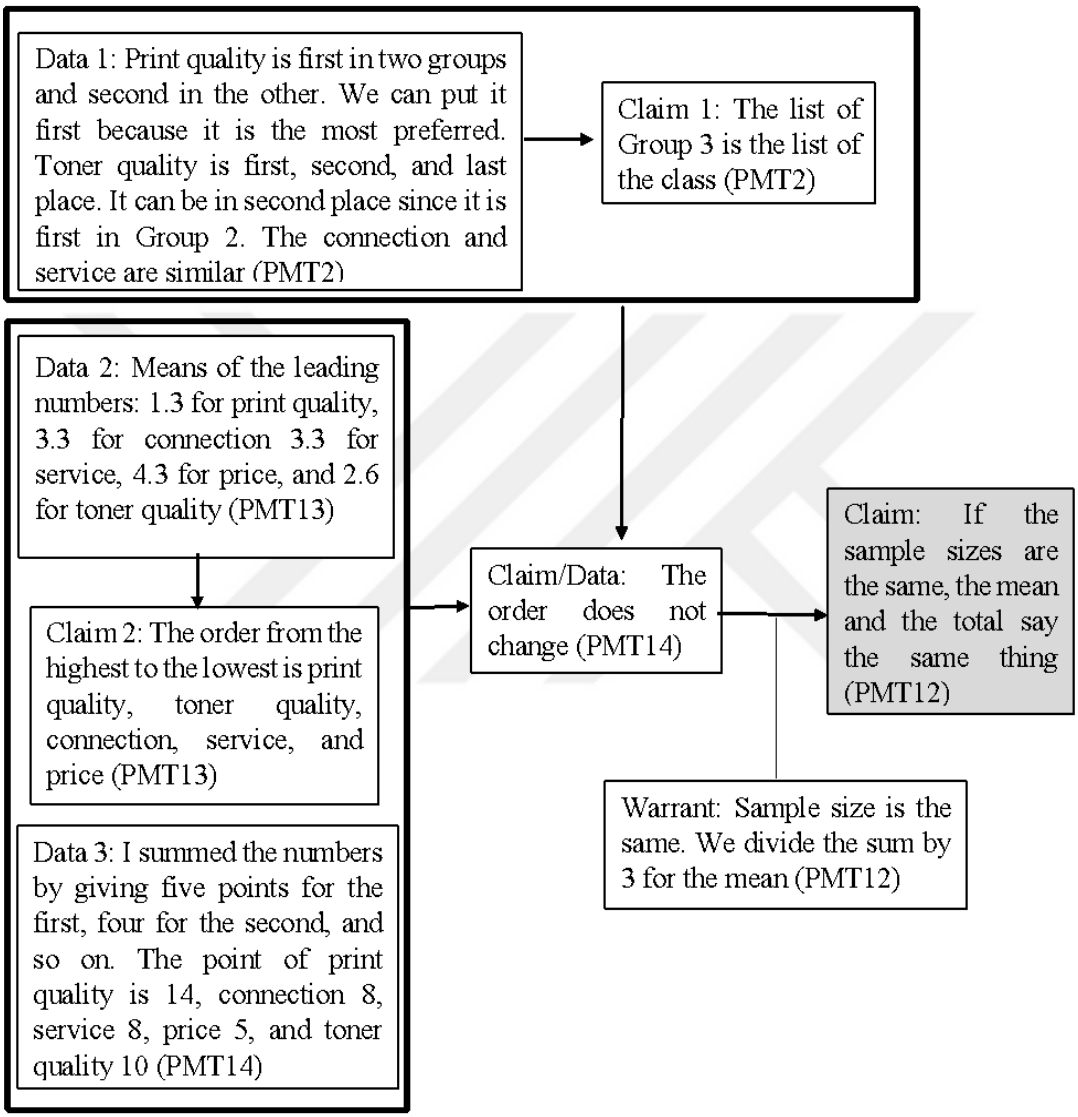


Figure 4.11. The argumentation scheme related to ranking the data

The claim that the ordering in the class list did not change was reached by using three strategies. It was also the claim/data at the center of the argumentation scheme since it was the data of the final claim. The final claim represented in dark color in the scheme, emerging in the continuation of the argumentation, was independent of the

frequency-based strategy because, on the instructor's question, PMT12 justified that claim based on the mean formula, including the sample size and total. The idea of "*if the sample sizes are the same, the mean and the total say the same thing*" was challenged by no PMT and shifted into warrants in all subsequent sessions of the teaching experiment. Thus, this idea became TAS. While aggregating the different ranked data, the PMTs used the mean or total to eliminate the variability. More specifically, they concluded that the total and the mean correspond to the same matter based on the mean formula since the total values representing each factor in each ranked list were divided by the number of ranked lists, 3. In the process of the emergence of this idea, the PMTs' reasoning was at the transitional level as they focused only on the mean.

Regardless of the strategy used throughout the session, it took time to determine the order of connection and service. The mean and total values of connection and service were the same. PMT5 suggested using their frequencies to deal with this dilemma. She stated: "Since the service is in the third order in both Group 1 and Group 2, we think of it as a mode and put the service in the third order and the connection in the fourth order." This order differed from the initially determined class list. Alternatively, PMT9 endeavored to solve the uncertainty using measures of variability and came up with the idea of the range.

> PMT9. I ranked by looking at the variability of these points, that is, range—the print quality, 1, 2, 1. The largest minus the smallest is one. The range of toner quality is four because of five minus one.
>
> PMT12. The ranges of connection and service are different.
>
> Instructor. What do these differences mean?
>
> PMT9. It tells you how preferences have changed. If the range is small, the choice difference is slight, as if everyone agreed. There was much distinction in the toner quality.
>
> PMT12. We can test the correctness of ranking with range. The ranges of print quality, service, and price are one. Everyone agreed, so their order was correct.
>
> PMT9. If the range is one, there is no difference in preferences. For example, there were many differences when choosing toner quality.

PMT10. So, the range alone did not work. It is not reliable.

PMT13. We get more reliable results if we associate the range with the total or the mean. Please do not use the mean by itself; interpret the measures of center by considering the measures of variability.

PMT12. Yes. Connecting the measures leads to more reliable results.

According to the meaning attributed to the range, PMT12 interpreted the one difference in the selection as the correct placement of the factors in the aggregated class list, confirming that everyone agreed on the small range. Based on this interpretation, there were different preferences for the factor with a large range. For example, the range of the toner quality was the largest. That is, toner quality, which was an important factor for Group 2, was least important for Group 1. Although PMT12 expressed this difference with the concept of outlier while the mean strategy was discussed, it needed to be considered by the other classroom community members to become a TAS idea. In the "aggregating different ranked data-printer" activity, the big difference in range for toner quality required considering relative positions of the factors rather than equipartitioning of the total, i.e., the mean. Even if the appropriateness of using the mean due to the outlier in this activity was not examined, it came to the fore when discussing the effect of an outlier on the mean in the following class sessions.

Comparing the ranges according to the place values of the factors was not considered sufficient when determining the order of the factors in the aggregated class list. For example, three factors with the same range (e.g., print quality, service, and price) existed. It was emphasized that using the range with the mean or total is trustworthy. In the class discussion, no one objected to the idea that "*connecting the measures of center with the measures of variability leads to more reliable results.*" Furthermore, all participants stated that they learned this idea in their reflection papers. For example, the self-evaluation of PMT12 was as follows:

I learned that using measures of center and variability together in the decision-making process will give us more accurate information… We could not comment on consensus or disagreement if we just looked at the mean.

> Likewise, we could not comment on the data distribution if we only looked at the range.

However, the participants did not use that idea in the printer activity to re-rank the factors. Therefore, the idea of "*connecting the measures of center with the measures of variability leads to more reliable result"* was kept on an eye to be TAS. Since the PMTs called upon it in later activities of the HLT when choosing the suitable measure, particularly for comparing distributions, it became a TAS idea. However, the idea was presented under the third CSP, selecting appropriate statistics for the context. Therefore, the TAS idea that emerged in the activity related to printers is that "*if the sample sizes are the same, the mean and the total say the same thing*."

Following the understanding of the data and data analysis, the learning goal in the HLT was to make predictions about the population using the average, explicitly connecting the average to the number of samples in the population. This relationship was initially intuited during the first activity in the first week of the teaching experiment. To delve deeper into this connection and enhance the conceptualization of the mean, "estimating an unknown whole through the average-tracking a herd" activity was introduced in the second week. It required estimating the total number of elephants in a region without counting them one by one. For this, the aerial picture showing the elephant herd was first projected. After discussing how scientists could track the population in this region, the PMTs were asked to assume the role of scientists and determine how many elephants were in the picture without counting every single one. Once they were sure that the statistical problem was understood, they tried to estimate the number of elephants in small groups using more than one solution method. In groups, group-specific strategies were developed. For example, Group 1 tried to estimate the number of elephants by proportioning the area of the picture to the one elephant's area in the picture. They could not apply the method with the numbers. However, in the class discussion, PMT5 in Group 1 suggested dividing the area of the picture for each elephant, warranting that there are elephants almost everywhere. Although the class accepted this approach as scientific and accurate in finding the exact number, it was abandoned because the number of

elephants was needed again. Alternatively, Group 1 shared how they made groups with a specific number of elephants. Other small groups also applied the method. Figure 4.12 presents the grouping with ten elephants of Group 3 shared in the class discussion.

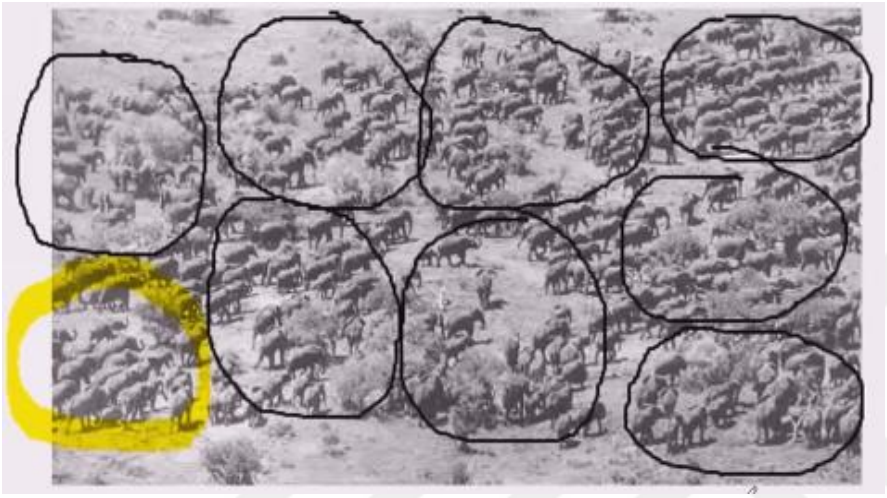


Figure 4.12. Group 3's grouping of the elephants

> PMT2. We have circled ten elephants in the figure.
>
> PMT7. This (yellow circle) was where it seemed the densest. I tried to randomly select a sample with this dense that I could spread to the general public. We thought about how many of these samples there are.
>
> Instructor. Are you creating a group of ten or ten groups?
>
> PMT7. A group of ten. We iterated it as an image.

The instructor questioned whether the focus was on the number of groups or the number of elephants in the group. In this sense, whether the division was partitive or quotitive was interrogated. Since the focus in all small group works and the class discussion was the number of elephants in each group, a partitive division was treated in this method. However, the number of groups was also considered when finding the total number of elephants. The PMTs assigned both multipliers, and the result was obtained by scaling the number in the specified group according to the number of groups. PMT7's description of the group as "*spread to the general public*"

generalizing coincides with the idea that the number in the group is average. In other words, this group also represents other groups in the picture.

Dividing the picture into grids was a common solution method in the class. Each small group divided the picture into different equal parts. For example, in Group 3, the picture was divided into equal 14×8 squares. Figure 4.13, divided into ten equal parts, was used in the class discussion. The excerpt below illustrates PMTs' reasoning while estimating the number of elephants in the class discussion.



Figure 4.13. The picture is divided into ten equal parts

PMT14. The third square in this picture represents the typical elephant herd.

Instructor. What does the typical mean?

PMT1. Average.

Instructor. How did you decide that grid three is typical?

PMT4. Because it is neither the most nor the least intense one.

Instructor. Why not others? For example, why is the grid below three not typical?

PMT4. It could be sparser.

PMT11. The fifth is also very intense.

PMT5. There are almost 30 elephants in the third one. There are also ten grids in total. We can say that there are 300 elephants in this region by multiplying them.

Having noticed that equal parts were not squares, the instructor encouraged them to justify why this grid was typical. The class was convinced of this typical grid based on PMT4's interpretation of the elephant density on the grids. In addition to the least intense grid directed by the instructor for clarification, PMT11's immediate counterexample showed that others were active in the class discussion. Depending on the density of the elephants in the grids, PMT13 in Group 2 shared her strategy, including creating a new typical grid by averaging two extreme values. The conversation continued as follows:

PMT13. I averaged the numbers in the fifth and eighth grids instead of the third. I found the minimum (17) and maximum (42) values and divided their sum in half, 29.5. That is, there are 30 elephants in an average grid.

Instructor. What does the value you found mean?

PMT12. Neither the most intense nor the least intense. The middle of the two.

Instructor. Then, did you multiply by the number of grids or half of it?

PMT5. The number of grids.

PMT2. Half of the number of grids.

PMT10. Grid number! Because 30 tells us the average number of elephants in a grid.

PMT2. Yes! I was confused.

PMT13. When we multiply by the number of grids, we have 300 elephants.

PMTs. Same result.

The midrange of these values was the number of elephants in the typical grid that PMT13 re-produced by taking the mean of the extreme values beyond the existing grids. That means the average is the midpoint or midrange for PMT13. While the determined number was iterated to the other grids, all PMTs were alert to the instructor's distracting question about the number of iterations except PMT2, who thought the mean represented the number of elephants on two grids. Fortunately, he cleared up the confusion thanks to PMT10's explanation regarding the midrange

corresponding to the average number of elephants on a grid. The PMTs reached the same number of elephants in both ways, summarized in the argumentation scheme in Figure 4.14. In this case, by chance, the midrange of the extreme values paralleled the number of elephants on the determined typical grid. However, the midrange could be quite different from the typical value representing the data when there are significantly different values, such as the outliers. Therefore, using a midrange for the average may not always be reliable. No one in the class voiced this idea, and the midrange of extreme values was considered typical.

Instead of both horizontal and vertical lines on the grids, Group 2 divided the picture into equal rows using only horizontal lines. PMT14 in the group presented the solution as follows: "There are 14 rows in the picture, and one has almost 15 elephants. So, there are 15x14, around 210 elephants." Here, 15 elephants in one row also represented the elephants in the other rows. It was stated that this could also be implemented for columns, but it was not applied.

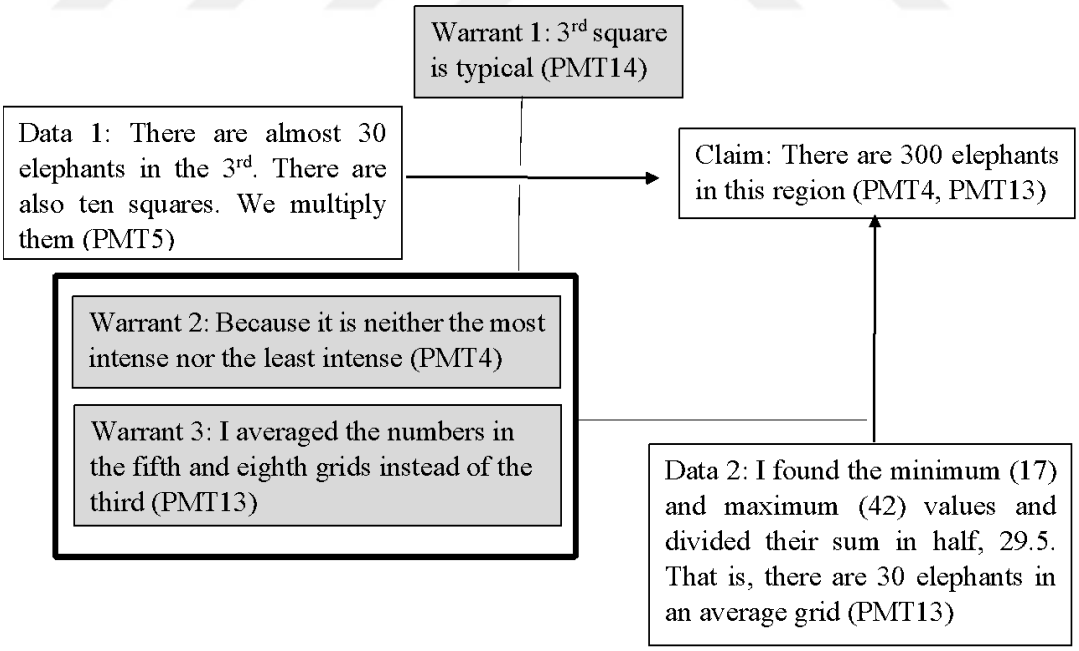


Figure 4.14. The argumentation scheme for the typical value representing the data

The idea of *average is the typical value representing the data* that emerged from the warrants represented in dark colors in the argumentation scheme in Figure 4.14. It was not challenged both in all small group work and in whole-class discussions, and it was used to reach the number of elephants. That is to say, the PMTs interiorized it without backing or warrants. Therefore, it was a TAS idea of the class. The PMTs calculated the number of elephants by iterating the grids, rows, columns, or groups as typical units in class discussions. The generalization of the typical value in units, regardless of the unit, is based on homogeneity, where the same number is in every case. In other words, despite different data values from the mean, only one value represented the whole data set. The mean reduced the data into a single value. The PMTs assumed the same number, or average, in each unit by matching the particular case with other cases. To find the intended value (total number of elephants), they multiplied the typical value in the units by the number of units. Similar to the activity in the previous week, the PMTs utilized the mean formula to reach the total from the mean. During the emergence of the TAS idea that *average is the typical value representing the data*, the PMTs were at the transitional reasoning level as they identified and used one dimension of the distribution.

In the last class session of the second week, three different 4x3 tables with numbers were given instead of elephants, and it was asked which cell of the table is the average of the numbers in the table. Instead of finding the average cell, the PMTs practiced calculating the mean, median, and mode. At the end of the lesson, they were assigned an assignment paper (Uccellini, 1996) containing equal distribution of values in the data and balance point interpretations of the mean to write a reflection paper. They were asked to read this paper and reflect on what they understood about the mean.

The homogeneity between the units used in "estimating an unknown whole through the average-tracking a herd" activity brought up the idea that *the mean equally distributes the values in the data set* since each unit was assumed to have an equal number in all iterations. The unitizing process that the PMTs performed was observed when they constructed a reference unit, e.g., typical value, and then iterated

that unit without changing it to other units. In other words, the reference unit represented all units. In this unitizing process, the PMTs dealt with a single unit spreading equally in all units. The PMTs mentioned this idea of equal distribution of data values during the "estimating an unknown part through the average-missing words" activity implemented in the third week of the teaching experiment. The primary learning goal of this activity was to reason about the average considering variability. In the activity, a picture of a book page with a rectangular space in the middle was sent to the PMTs and shared in the online learning platform. The PMTs' mission was to reasonably estimate how many words were missing in the middle of the page and clearly explain how they figured it out. In other words, the PMTs would estimate the unknown (e.g., number of missing words) through what is known (e.g., words that appear). They had employed the relationship between the total and mean in the previous week. Moreover, they consistently applied the mean formula due to their prior knowledge. However, in this activity, it was unclear how to determine the mean as the words that seemed to be referenced were scattered differently across the page. That is to say, different numbers of visible words on the different lines of the page posed a challenge for them. The PMTs had to cope with the uncertainty that arose because of the variability. They first worked on the activity in small groups and then shared their estimations with reasons in the class discussion.

Each group tried to estimate the accurate number of words by producing more than one solution method based on the sociostatistical norms of the classroom. One of the first strategies that came to mind was assigning an average number of words per line and multiplying it by the number of lines, a method used in the "estimating an unknown whole through the average-tracking a herd" activity. For example, PMT5 in Group 2 explained her claim in the group as follows: "I summed the numbers in the visible lines. It is 85. There are nine lines, so I divided it by nine. The result is like 9.4. If we accept nine words in each line, there are 108 words from 12x9." However, in the other strategies of the group, more words, such as 130, 144, and 153, were reached. Because they had a challenge with the number assigned to the lines as in the other groups, this situation drew attention to the variability in the data.

Another difficulty was that it was unclear how many lines were in the space. Group 3 suggested using an evenly distributed table for spaces to deal with this. Below is the solution and the table (Figure 4.15) presented by the group in the class discussion.

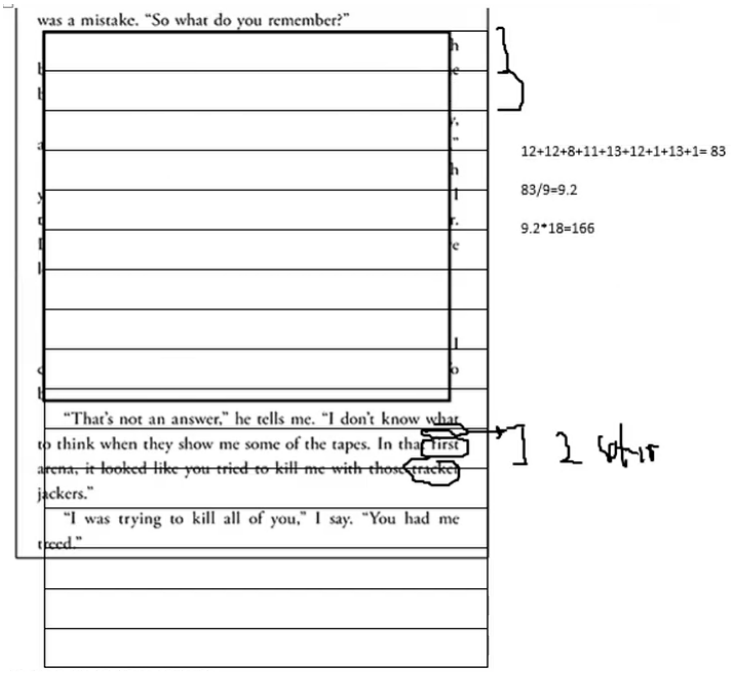


Figure 4.15. The table was drawn by Group 3 on page

> PMT6. When we divided the total number of words (83) into the number of lines (nine), we obtained 9.2. An average of 9.2 words per line. As you can see, we drew a table (Figure 4.15) to be evenly spaced. Since a row of the table corresponds to about two lines, there are 18 lines in the invisible part. When we multiply 9.2 by 18, the space has 166 words.
>
> Instructor. How did you use average here?
>
> PMT6. We assume that we have 9.2 words in each line.
>
> PMT10. Each row in the table does not correspond to two lines.
>
> PMT6. There are some spaces in the table with no lines, but some rows cross, such as the lines where I marked. Therefore, they complement each other.

In the class discussion of the activity, PMT6 based her claim, "the space has 166 words", on the equal distribution of words in each line. In addition, she also

benefitted from this assumption when determining the number of lines in the table because, according to her, there were two lines of words in each row of the table. Figure 4.16 shows the argumentation scheme PMT6 stated in the classroom to find the number of missing words on the page.

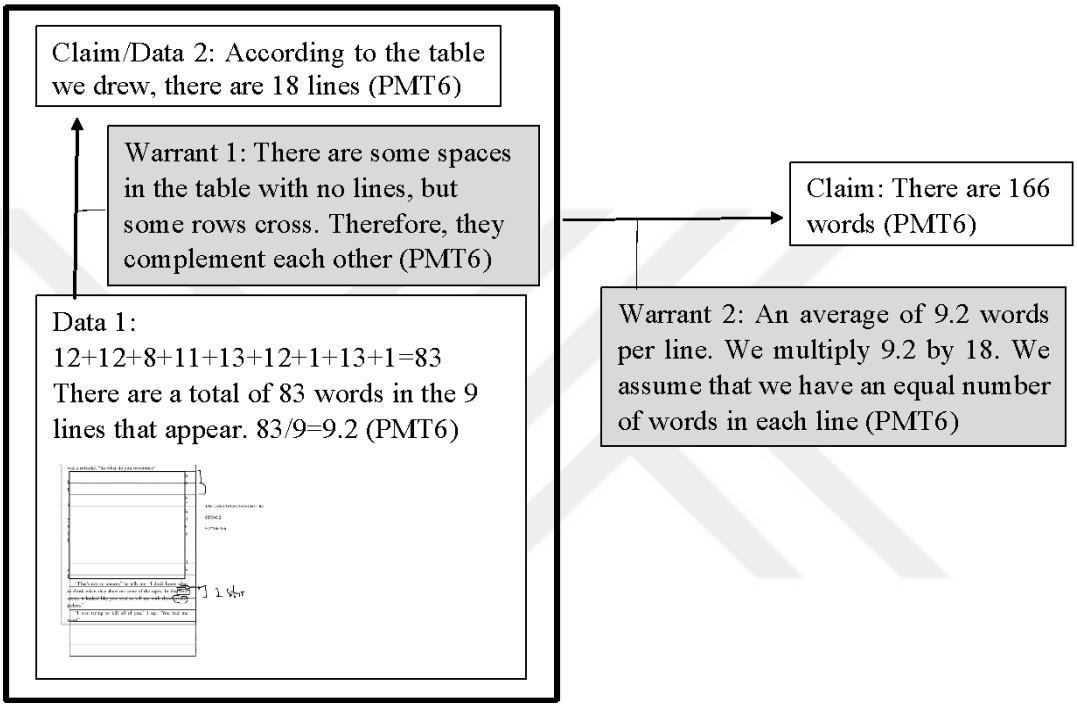


Figure 4.16. The argumentation scheme regarding the equal distribution of words and lines

As can be seen from the warrants represented in dark colors in the argumentation scheme in Figure 4.16, PMT6 distributed equally both the number of words per line and the number of lines per row of the table. This reasoning of per-one provided homogeneity because each or every item was attributed to an equal number of quantities. In other words, the fair sharing/leveling interpretation of the mean was employed in the solution. The resulting value for the number of words, the mean in this argumentation, was generalized to every line. Accordingly, the unit that was the number of words per line was multiplied by a factor- the number of lines.

In the class discussion, like PMT10's objection, there were disagreements about the number of lines in the middle of the page. Another challenge was that the lines were not identical. Therefore, they attributed different numbers to the number of lines in the space and the number of words in the lines. One of them was the claim that there were 12 lines determined in the space instead of 18 and 12 words instead of 9.2 due to the first visible line as a reference. Considering the variability in the number of words in different lines, the lines were categorized in each small group work. The following conversation occurred for classified lines as complete and incomplete in the class discussion.

> PMT13. We tried to consider lines as complete and incomplete lines. If it comes to the end of the page, it is a complete line with 12 words, as in the first line in the visible part. Incomplete lines are the third line with eight words and the last lines with one word in the last paragraphs. There is an average of three words in incomplete ones.
>
> PMT2. You defined two different means, one for the complete and one for the incomplete.
>
> PMT13. The rectangle has ten complete lines and three incomplete lines (Figure 4.17).
>
> Instructor. How did you decide if there was a line there?
>
> PMT13. If you saw a trail of letters at the end, this line is complete, for which there are 120 words from 12x10. You can see a letter at the beginning of the line for three incomplete lines, but not at the end, 3.3x3, ten. This is 130 out of a total of 120+10.
>
> PMT12. There can also be words in the blank. That is why it is more than 130.
>
> PMT2. It can be a title or a question and answer. There are at least five words there.
>
> PMS12. At most, ten words will fit there. Otherwise, it would appear out of the lines.
>
> PMT8. Since there may also be a picture, there might be 130-140 words.

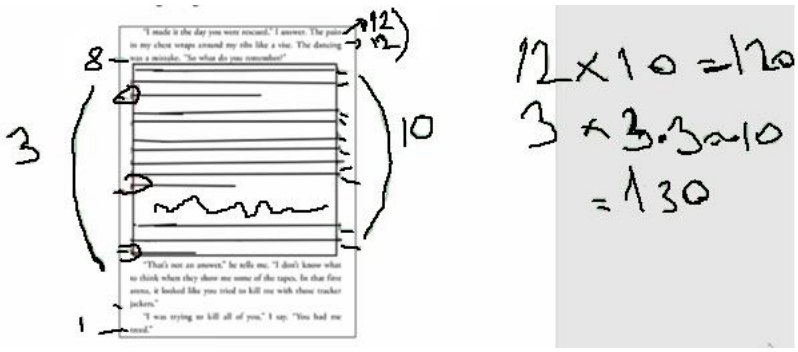


Figure 4.17. The solution for complete and incomplete lines

As PMT2 clearly stated, a different mean for each line category was assigned. Upon the teacher's question, PMT13 presented the letter appearances on the right and left of the space as evidence to determine the number and category of the invisible lines. The number of words above the last invisible paragraph determined in the space in which there is the scribble instead of the line was ambiguous. It was claimed that there might be 0-10 words in the space based on the maximum and minimum values. At the end of the discussion, no one challenged the suggested total of 130-140 words in the space. In this process, PMTs made sense of the weighted mean by considering the variability of the number of words in the categorized lines. Nonetheless, they used an equal distribution of words in each line category in each solution of the statistical problem. The idea that *the mean equally distributes data items* was a TAS idea because it did not require any warrants or justifications. In addition, it was used in subsequent class sessions as in every solution in the "estimating an unknown part through the average-missing words" activity. In addition, in their second reflection papers, all PMTs could explain equal sharing/leveling interpretation of the mean reasonably. For example, PMT14 pointed out two strategies for the way of equal sharing:

> The first point I will touch on is the equitable sharing approach. It is the method most used by children in their daily lives. Children divide the whole, a certain amount, in a way that falls equally to everyone. These may be some types of food such as chocolates, biscuits, and so on. They use different strategies to share them. For example, they can divide the whole by the number of people or share it by giving each one at a time. The basic strategy is always equal sharing (PMT14)

In the solution in Figure 4.15, the PMTs reached the total number of missing words by focusing only on the mean through the number of lines and words distributed evenly. Therefore, they were at the transitional reasoning level in this solution. However, in the solution above, where they categorized the lines, they tried to stabilize the uncertainty by recognizing the differences in the lines. Since they considered variability as well as the mean, they were able to get higher than the transitional reasoning level. Although they identified different categories, they did not explain the density of these categories in the space and the effect of each of them on the total number of missing words. In this context, they reached the procedural level in the TAS idea that *the mean equally distributes data items*.

Equal sharing of items was supported by the reasoning that incomplete lines complement each other and form a complete line. PMT2 described complementation as the balancing of the different lines. For the number of words in the incomplete line, the words were counted in the last lines of the three paragraphs that appeared. There were eight, one, and one word. The mean, 3.3, was obtained when ten words were distributed equally on three lines. In the continuation of the class discussion, PMT6 questioned retrospectively how this value was found, and the effect of one-word lines at the end of the pages was discussed.

> PMT6. Have you included the "So what do you remember?" line in this 3.3 value as an incomplete line?
>
> PMT13. Yes. There would be no trace at the end of it if it were in the space.
>
> PMT6. You found the mean of the three numbers.
>
> PMT2. All three are incomplete lines, but one has eight words, and the other only one.
>
> Instructor. What do you think?
>
> PMT13. Yeah. The one-word lines decreased the mean.

Everyone in the classroom agreed that the one-word lines at the end of the page lowered the mean. Therefore, they were convinced that the number of missing words in the middle of the page could be at least 130.

After the end of the "estimating an unknown part through the average-missing words" activity, they discussed the paper given as an assignment in the previous week. The paper included two interpretations of the mean: fair sharing/leveling and balance point. PMT12 explained the idea of equal sharing interpretation of the mean by using the example of distributing an equal number of candies to everyone. She said, "We level off from the one with more candy to the one with less candy." The instructor kept talking to get the PMTs to think about the difference in the number of candies people have. In this class session related to equal sharing/leveling interpretation of the mean, the primary learning goal was the idea that the mean compensates the data values in the distribution. However, conservation evolved in terms of the shape of distribution.

> Instructor. How does adding or removing candies affect the number of candies people will take?
>
> PMT2. If it is higher than the mean, people take more candies.
>
> Instructor. In the reverse situation?
>
> PMT12. It decreases the mean.
>
> Instructor. Can you imagine the initial distribution of the candies and the subsequent distribution for leveling? (No sound). What is distribution?
>
> PMT6. Normal distribution, skewed distribution.
>
> Instructor. Can you draw them? (PMT5 drew Figure 4.18) What do these graphs say about the mean?
>
> PMT2. The mode is greater than the mean and median, as most people have the highest value in a left-skewed distribution. When ordering, the median is to the left of the mode; it can be right in the middle.
>
> PMT10. Because the area under the curve to the right and left of the median will be equal.
>
> Instructor. Where is the mean in Figure 4.18 (b)?
>
> PMT7. It will be smaller than the median.
>
> PMT10 and PMT3. It will be between the median and the mode.
>
> PMT2. Right, because higher values are more to the right of the median.
>
> PMT13. No. I will explain it with the help of an example. We distributed candies to five people. Assume they are high, 200, 200, 300, and 100.

However, Ali has two candies. Here, the mode is high, but the mean is lower because of Ali's number of candies as an extreme value.

PMT9. The mean is affected more by the outlier when leveling off.

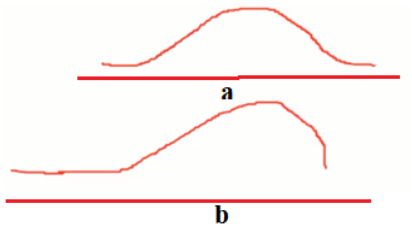


Figure 4.18. Normal (a) and left-skewed (b) distributions plotted by PMT5

The PMTs could make sense of the change in the mean according to the relative size of new data to the mean. To emphasize the changes, with the leveling-off function of the mean, the instructor asked to compare the state of the candies before and after they were distributed evenly. However, the PMTs recalled their prior knowledge and identified the distribution concept with the shapes. The whole class knew the mean, median, and mode as measures of center. While discussing their positions in a left-skewed distribution (Figure 4.18 (b)), the location of the mode was quickly detected based on its definition. PMT10 justified that the median is lower than the mode on the ground that the median divides the area under the curve in the graphical representation into two equal parts. However, the class disagreed about the location of the mean. Contrary to the claim that the mean was between the median and mode, PMT13 proved that the mean was the lowest in the candy example she created hypothetically, supported by the outlier concept. Her argumentation scheme is given in Figure 4.19.

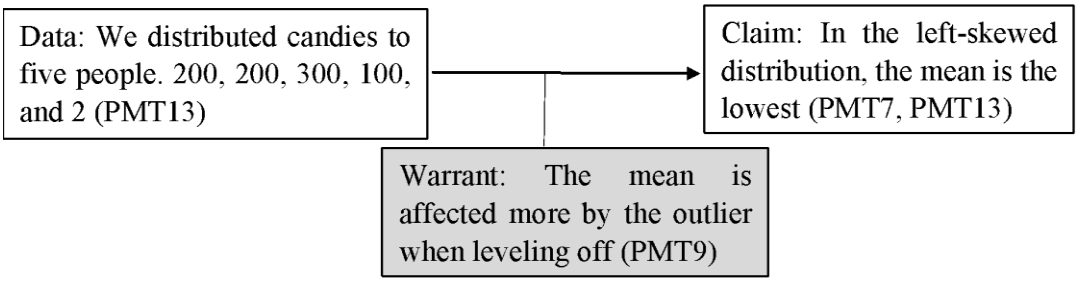


Figure 4.19. The argumentation scheme regarding the effect of the outliers on the mean

One-word lines at the end of the page reduced the mean in the "estimating an unknown part through the average-missing words" activity. In fact, the idea of "*the mean is responsive to outliers*" had been claimed for the first time in the first week of the classroom teaching experiment. This idea also provided a rationale for selecting appropriate statistical measures for the context, which will be explained later. Therefore, the idea became TAS with no challenge or justification. In addition, the PMTs used the idea repeatedly in the subsequent class sessions. With this idea in mind, they could determine the mean and outlier in a data set. Although they expressed the center and skewness of the distribution, they remained at the procedural level because they did not provide an explanation based on the interpretation of equal sharing/leveling of the mean. In other words, they did not interpret the density or variability of the distribution to describe how the data was distributed while reasoning about this idea.

Based on the TAS idea that *the mean is responsive to outliers*, the whole class agreed that using the other measures instead of the mean for the average in distributions with outliers was more appropriate. They conceptualize the average that represents the whole data. As with skewed distributions, the mean would not represent the data when there were outliers. PMT1 explained the situation in her first reflection paper as follows:

> If there is an outlier in the data, I need to develop statistical approaches that will exclude this outlier. I learned that in these approaches, I should use the median or mode, not the mean, to get a result closer to the actual result (PMT1)

One of the learning goals in the HLT was the balance point interpretation of the mean, which was also included in the paper assigned. While explaining the paper by the PMTs, after the equal sharing/leveling interpretation of the mean was associated with the outlier, the interpretation of the balance point was discussed in the last class session of the third week. For the interpretation, PMT7 shared the example in her reflection paper with the class, and the following dialog emerged.

> PMT7. We can consider the number line as the basis for the balancing point. Let me give an example of a set with 1, 3, and 5. When we mark these numbers on the number line we draw, we notice that 1 and 5 are equivalent to 3. Both are two units apart, and they meet in the middle, that is, at the 3. In this case, 3 is the balancing point. It also means the mean.
>
> Instructor. What is the meaning of the balance point?
>
> PMT2. When we look at the sum of the differences from the balance point, the left and right sides must be equal.

PMT7 used the number line to justify the meaning of the balance point on three data values. The other PMTs in the class accepted her argument about the balance point interpretation of the mean. PMT7's argument about the mean as the balance point also included the idea that "*The sum of the distances of scores above and below the mean is equal*," which is represented in dark color in Figure 4.20. It was the collective idea of the class that is a TAS idea, as the idea was reused in the later weeks of the teaching experiment.

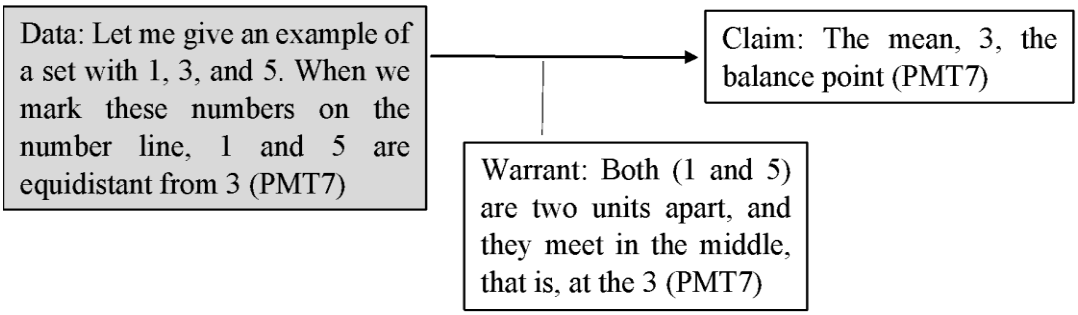


Figure 4.20. The argumentation scheme regarding the balance point interpretation of the mean

Regarding the idea, the PMTs seemed to be at the level of procedural reasoning since they did not integrate the mean and variability based on the distances of the data from the mean. However, in the repeated use of this, they reasoned about how the data was distributed with the spread of the data around the mean as well as the distances. Therefore, the PMTs' reasoning about the TAS idea could reach the level of integrated process reasoning. For example, in a follow-up of the dialog, they compared the data distribution around the mean. The instructor provided data on the points that PMT7 stated above. However, she increased the sample size to six in different distributions. Although there were different distributions, the means were still three.

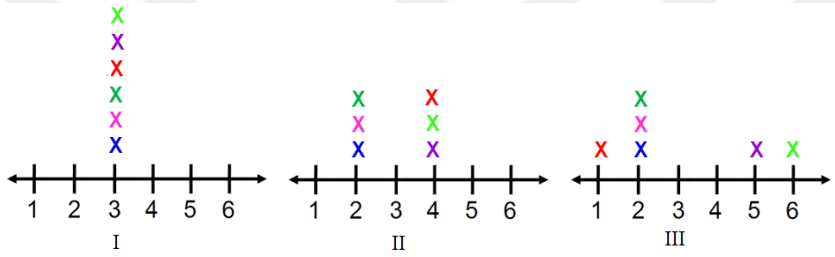


Figure 4.21. The distributions for balance point interpretation of the mean

Instructor. OK. Let us talk about the figures.

PMT13. The center is always three. Later, they are distributed to the sides.

Instructor. What is the difference between those three figures?

PMTs. Spread out.

PMT4. There is no variance in the first data. It is 0. There is little variance in the second data. The data is spread out around the mean in the last figure.

Instructor. Let us distribute the data differently. At first, I put the green from Figure 4.21 (I) at 6. I broke down the balance. Where should I place purple and red to keep it in balance?

PMT2, PMT14, and PMT4. One and two.

Instructor. Why?

PMT12. The total distance should also be three on the left side.

Instructor. What happens if I put both on one?

PMT13. The balance is disturbed. It lies down to the left.

In the above conversation, the PMTs implemented the TAS idea that "*the sum of the distances of scores above and below the mean is equal.*" Using the data in Figure 4.21, they explained the variability in the data with the term spread out. When the instructor suggested distributing the data differently, it was determined correctly by calculating the distances from the mean. They ensured that the total distances to the left and right of the mean were equal. Conceptually, they were aware of the invariance between two distances, i.e., the ratio of the sum of the distances on two sides is constant in each situation. They also wrote in the reflection papers that when they calculated the distances of each data value from the mean, taking into account the direction, the sum was 0. They knew the disequilibrium if the sum of the distances above and below the mean differed. Although PMT14, in his reflection paper, wrote that "the average value will be close to the dominant side," it was not explicitly shared in the class. Regarding the properties of the distances, PMT9 wrote in her paper to emphasize the mean as the center of the data:

> This balance point interpretation is less known and used. The data is represented as an X on the number line, and the difference between each data item's value and the mean's value is calculated. Since the equality in the sum of the distances of the data items above and below the mean, the mean is interpreted as the balancing point of the data. This interpretation clarified why the mean is seen as the center of data and used in central tendency measurements (PMT9).

As can be deduced from the teaching experiment so far, in all the above activities ("aggregating different ranked data-printer", "estimating an unknown whole through the average-tracking a herd," and "estimating an unknown part through the average-missing words"), variability could not be ignored. Each had a prediction based on the given data with some uncertainty. In this process, the average, mostly the mean, was used as a tool. It was a stable value but represented the data, including variability. According to Konold and Pollatsek (2002, p. 262), the central tendency has a stable component summarized with measures of center and a variable component that reflects the variability of data values. They describe the average as "the signal in noisy processes" to make the situations involving uncertainty arising from the variability stable. The five TAS ideas emerging about the average

constituted the first CSP- Reasoning about the average as the signal in a noisy process. They were 1) *If the sample sizes are the same, the mean and the total say the same thing*; 2) *Average is the typical value representing the data*; 3) *The mean distributes data items equally*; 4) *The mean is responsive to outliers*, and 5) *The sum of the distances of scores above and below the mean is equal*. Based on the TAS ideas revealed by the class community, integrating center and variability was the first step in reasoning about distribution. While initially, the PMTs only focused on the relationship between the mean and the total at the transitional reasoning level, in later activities, they also started to consider variability, which corresponded to the procedural reasoning level. Namely, they conceptualized the central tendency reflecting both the center and variability of the data.

When the PMTs dealt with activities in class sessions, they talked about the measures of variability, such as range and variance. For example, in the above conversation about the distributions in Figure 4.21, PMT4 specified the degree of variability with the variance. He compared the distributions according to their variance based on the distances of the points from the mean. By the matter, we agreed that the main focus of the learning activities should be variability without ignoring the center in the weekly meetings.

4.2.2 Classroom Statistical Practice 2: Measuring of variability in data

Three ideas about the measures of variability became taken-as-shared in the fourth week of the teaching experiment, in which the PMTs worked on "measuring variability in distributins with small sample sizes-soccer game" activity to interpret variability in data with the center. The TAS ideas were as follows.

Normality is not a criterion for variability.
If the number of samples is different, the MAD is used rather than the SAD.
The MAD is a predecessor of the standard deviation.

The learning goal of the soccer game activity was to make sense of the center through variability and to measure the variability in the activity data. The activity allowed us

not only to apply the balance point interpretation of the mean admitted in the previous week but also to examine the variability. Data representations such as graphs and tables were not used in HLT until the fourth week. The data was presented with graphs in soccer games for the first time. In Week 4, dot plots representing two teams' points in the study of Kader and Mamer (2008) were presented. The number within each dot displayed the distance of each score from the mean. Figure 4.22 shows the points scored by the teams in each of the nine games. The following excerpt was observed in the class after determining the teams' scores, namely raw data, according to the graphs.

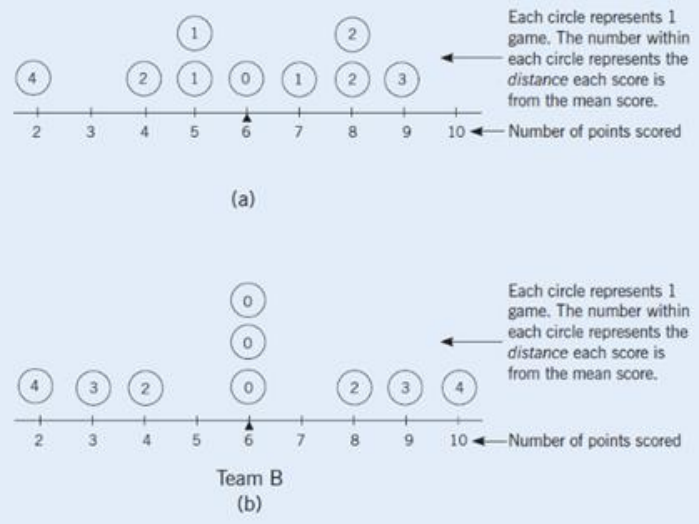


Figure 4.22. The points scored by two teams in each of the nine games (Kader & Mamer, 2008, p. 39)

Instructor. OK. How can we interpret these two plots?

PMT7. The mean of both is six.

PMT6. Variability is low in Team B because there is an equal distance. For example, 4 and 8, 3 and 9, and 2 and 10 have equal distances from the mean. In Team A, there are many differences.

PMT12. It is the opposite. Variability is lower in Team A because Team B is further away than the mean. They are closer together in Team A.

Instructor. How do you decide?

PMT6. We can add the differences. For Team A, it is 16. For Team B, it is 18.

PMTs. Team B has a bigger variability.

Instructor. Why?

PMT6. Because the variability is the spread of the scores around the mean.

While interpreting the plots, the mean was determined instantly by calculating or stating the number on the circle showing 0. It was agreed that the means of both distributions were the same. Regarding the variability in the plots, PMT6 pointed out the different distances of the scores to the mean for high variability. However, PMT12 challenged the idea of equal distances from the mean for low variability. In this respect, PMT6 calculated the sum of the absolute deviations (SAD) for the two distributions by adding all distances. She was convinced that the variability was higher in Team B based on the distance from the mean. Moreover, the PMTs could also have reached the same conclusion by looking at the range in the distributions. The ranges were seven and eight, respectively. Instead of the range, the PMT6's definition for variability gave more information about how the data was distributed, as SAD provided an idea about the size of the data values to the center. In other words, the PMTs started to think of the relative size rather than only the absolute. A possible reason for this might be comparing two distributions rather than a single distribution, even though their means were the same. Comparing the variability of the distributions with the same center could lead to a more reasonable conclusion in determining which team was better. This variability interpretation, which developed based on the symmetrical distribution, continued with the distribution concept under the instructor's guidance.

Instructor. How about the distribution of them?

PMT2. Team B has a normal distribution, but Team A does not.

Instructor. What about Team A?

PMT13 and PMT14. Right-skewed.

PMTs. Left! Left-skewed.

Instructor. It is not exactly left skewed, but it also is not a normal distribution. What do you think of their variability?

> PMT6. Err… Team B's variability is more than Team A's. Then, if it is normal, we cannot say that its variability is low. Normality is not a criterion for variability.

From their prior knowledge, they could decide whether distributions were skewed or symmetrical based on their appearance. They were sure that the data of Team B with a symmetric appearance was a normal distribution. They concluded that Team B's variability was higher based on SAD. Upon this conclusion, PMT6 claimed that *normality was not a criterion for variability*. The PMTs did not provide evidence for the normality of the distributions, such as skewness (-0.43 and 0) and kurtosis (-0.41 and -0.97) values for Team A and Team B, respectively. However, it was understood that what was meant by normality was whether the distribution was symmetrical. There was confusion between normality and symmetry. In this class session, the idea was written with normality as they used normality instead of symmetry. The idea that "*normality is not a criterion for variability"* became the TAS idea as no one challenged it, and even if challenged, it was rejected. In addition, in the following sessions, the distribution shape was used to reason about the shift in the distribution for variability, and the positions of the components changed in Krummheruer's argumentation model. Figure 4.23 displays the argumentation scheme regarding the relation between normality and variability.

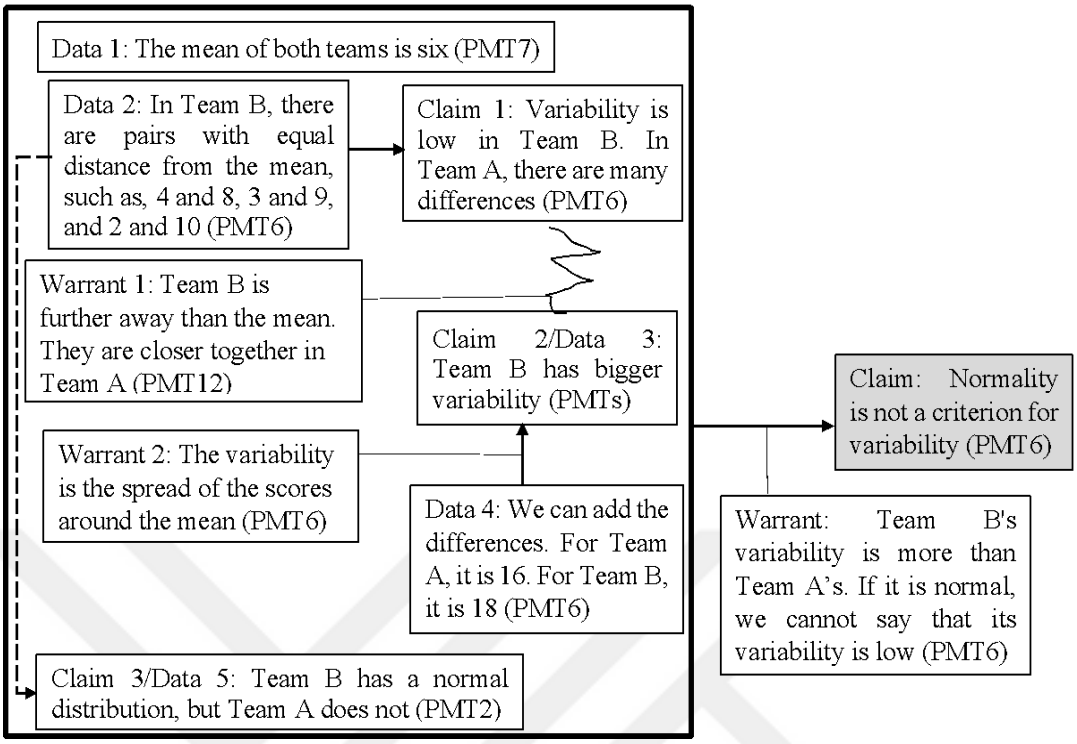


Figure 4.23. The argumentation scheme regarding the relation between normality and variability

According to the figure, the final claim represented in dark color emerged based on the shape of the distributions (Data 5) and variability of the distribution (Data 3) derived from the mean (Data 1) and the sum of distances from the mean (Data 4). The idea that Team B has low variability due to its symmetric distribution (Claim 1) was rejected based on Data 5, which includes the sum of absolute deviations from the mean. The class agreed that normality is not a criterion for variability based on the relationship shown by the zigzag in the scheme. The TAS idea that *normality is not a criterion for variability* emerged based on the data component of the argumentation. The data components included information about the characteristics of the distribution (e.g., center, spread, and shape). That is, the PMTs could associate the center, spread, density, and skewness aspects of the distribution. They exhibited a complete understanding of how the data was distributed. Therefore, they were at

the level of integrated process reasoning while establishing a relationship between normality and variability.

Another purpose of the teaching experiment session was to reason about the degree of variability. The PMTs could measure the variability of distributions with the range or sum of the absolute deviations (SAD) in the first-class session of Week 4. The focus in the second class session of the week was measures of variability. After introducing SAD, the PMTs studied another case, including two distributions over the same context in Figure 4.24. Team A played two soccer games this time, while Team B played eight. That is, the distributions consisted of two and eight scores. Both the mean and SAD of the distributions were the same.

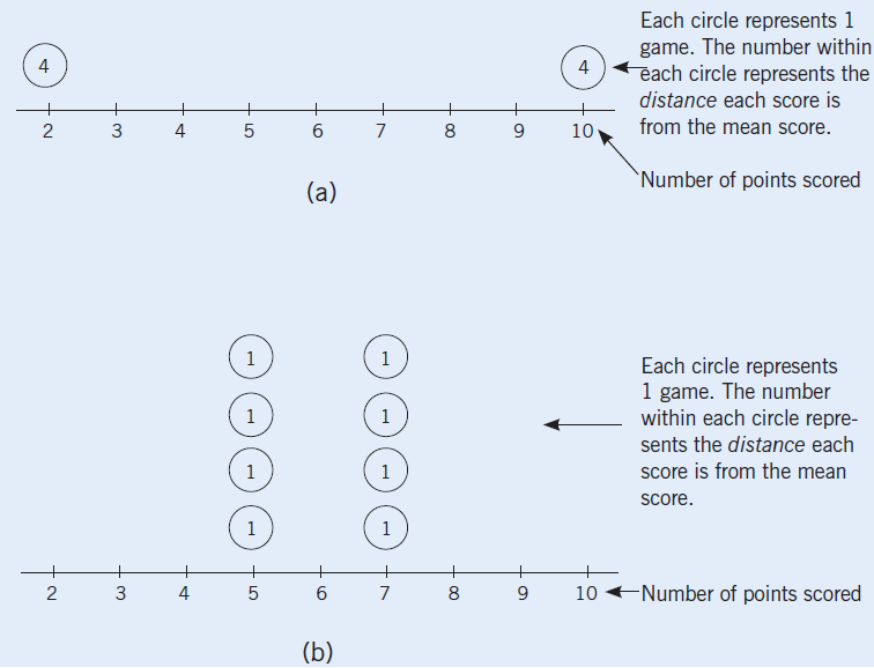


Figure 4.24. The points scored by the two teams (Kader & Mamer, 2008, p. 40)

First, the scores of the teams were determined. However, PMT9 stated there should have been a circle with 0 for the mean, as in the previous case. With immediate feedback, PMT2 explained that there was no data with the mean value. This situation

illustrated that participants were active in the process. She was also convinced by the instructor's question about how many cases were in the distributions.

> PMT10. SAD is eight, and both are equal.
>
> PMTs. The means are the same: six.
>
> Instructor. What is the difference?
>
> PMT2. The number of samples is different.
>
> PMT9. We can think of the above as outliers.
>
> Instructor. How can we interpret it?
>
> PMT13. First, we need to take the average of the distances because, for the distance to be eight, two people are enough in the first while the other has eight people. I mean the average distance from the mean for each score.
>
> Instructor. Do they have the same variability?
>
> PMT3. The number of samples is important. Their variability is different. The above one is farther from the mean, but the other's data is closer to the mean.
>
> PMT13. If the number of samples is different, we look at the distance value from the mean for each individual. It is not enough to look into the distance.

While describing the distributions, they noticed that the mean and SAD were the same, and the sample sizes were different. PMT9 intuited that the variability of the scores was high in Team A by describing the data as outliers. She thought Team A's distribution was more spread because a few data were too far apart. When the instructor asked for a conclusion, PMT13's idea of the mean of absolute deviations (MAD) justified the intuition. Eight data sets in the other gave the value (SAD) given by two data sets in one distribution. According to the justification of PMT13, the first distribution had four times more power for variability. Indeed, if the sample size of both distributions were equalized to be a similar distribution, the four two and four ten values in the first distribution (2, 2, 2, 2, 10, 10, 10, 10) would be a total of 32 units away from the mean. The difference in sample size formed a prerequisite for proportional reasoning. The number of samples would affect the magnitude of the SAD as it would change the number of distances from the mean when calculating the SAD. Although the PMTs could not express this proportional relationship clearly, PMT13 attempted to explain the mean of absolute deviations (MAD)

definition. She defined MAD as "the value of the distance from the mean for each individual." The MAD was another measure of variability used in the class session besides SAD. The difference in sample size was seen as a prerequisite for using MAD while comparing the variability of distributions. PMTs adopted the idea of using MAD instead of SAD if the number of samples is different and even used it in all subsequent class sessions. Therefore, the idea of "*If the number of samples is different, the MAD is used rather than the SAD*," represented in dark color in the scheme, became TAS. Regarding the idea, the PMTs were at the level of procedural reasoning since they did not integrate the variability based on distances of the data from the mean with other aspects of the distribution. Even though they determined the mean and interpreted the measures of variability, they do not explain how the data is distributed by integrating them. Figure 4.25 summarizes PMTs' arguments for the use of MAD and SAD.

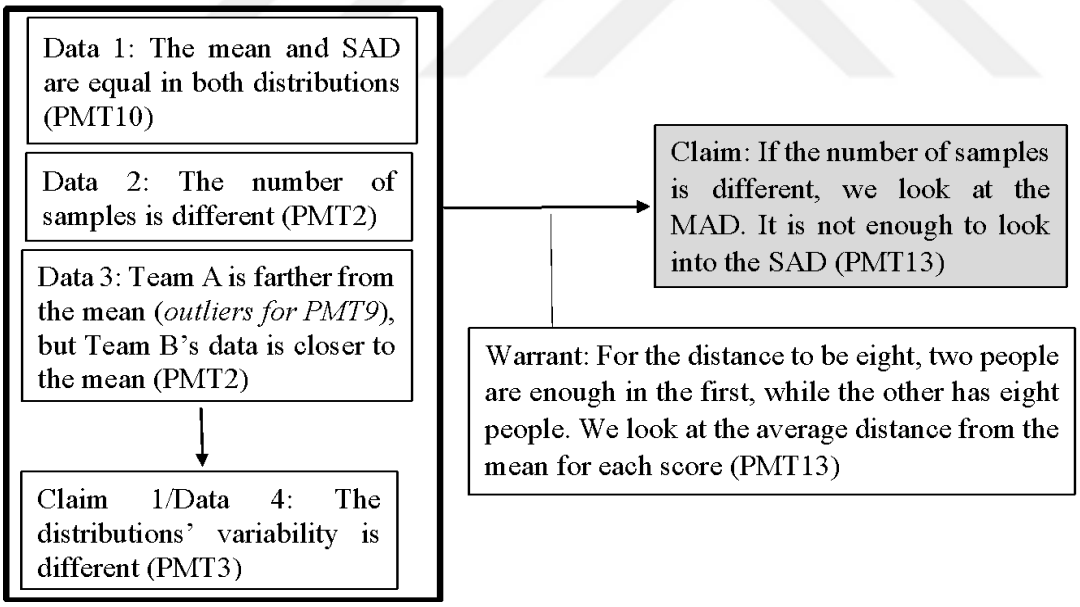


Figure 4.25. The argumentation scheme regarding the use of SAD and MAD

While introducing variability measures in the class session, PMT10 considered MAD as variance since MAD is determined by dividing the SAD by the number of samples.

PMT10. Then, the variance is the mean of the absolute deviations (MAD).

PMT6. We take the absolute values of deviations to eliminate negativity. MAD is not variance; there are squared values in the variance.

PMT5. When we take the absolute deviations from the mean, we already make it positive. Why do you need to take the square? We are already measuring the variability.

PMT6. The goal is not only to make positive but also to increase the value by its own multiple.

PMT7. For the standard deviation, we take the square root of variance. MAD serves the same logic.

PMT10. However, the standard deviation and the MAD are not the same. The MAD is the simpler version of the standard deviation, the predecessor.

From their previous statistics and probability courses, PMTs knew all statistical formulas, including standard deviation and variance. These formulas were used to solve PMT10's confusion. The explanations of PMT6 and PMT7 based on the formulas of variance, standard deviations, and mean of absolute deviations from the mean formed data in Figure 4.26. The explanations persuaded the whole class that the MAD and standard deviation were different because of their mathematical acceptability. MAD is the sum of the distances divided by the number of samples, while standard deviation is the square root of the sum of the squared distances divided by the number of samples. In other words, MAD is the division of SAD by the sample size. On the other hand, in the standard deviation, the sum of the squares of the distances is divided by the sample size. The formulas for standard deviation and MAD are as follows. In the formulas, σ denotes the standard deviation, x_i signifies individual data points, $\bar{x}$ represents the data set's mean, and n corresponds to the total number of data points.

$$\sigma = \sqrt{\left[\sum_{i=1}^{n}(xi - \bar{x})^2\right]}/\sqrt{n}$$

$$MAD = \left[\sum_{i=1}^{n}|(xi - \bar{x})|\right]/n$$

As can be seen from the formulas, the sum of the squares of the distances in the square root is not the same as the sum of the distances. In standard deviation, the square root is taken after summing squares of the distances. However, in MAD, the

distances are summed directly. Based on this difference, in the class discussion, PMT6 and PMT7 explained that standard deviation and MAD were not the same. With the persuasion of PMT10, who was on the opposite view, the idea of "*MAD is a predecessor of the standard deviation*" became a TAS idea. It was unchallenged, and in the subsequent class sessions, the PMTs used MAD and standard deviation as different measures of variability. The PMTs were at the transitional reasoning level because they focused only on variability while interpreting the standard deviation and MAD.

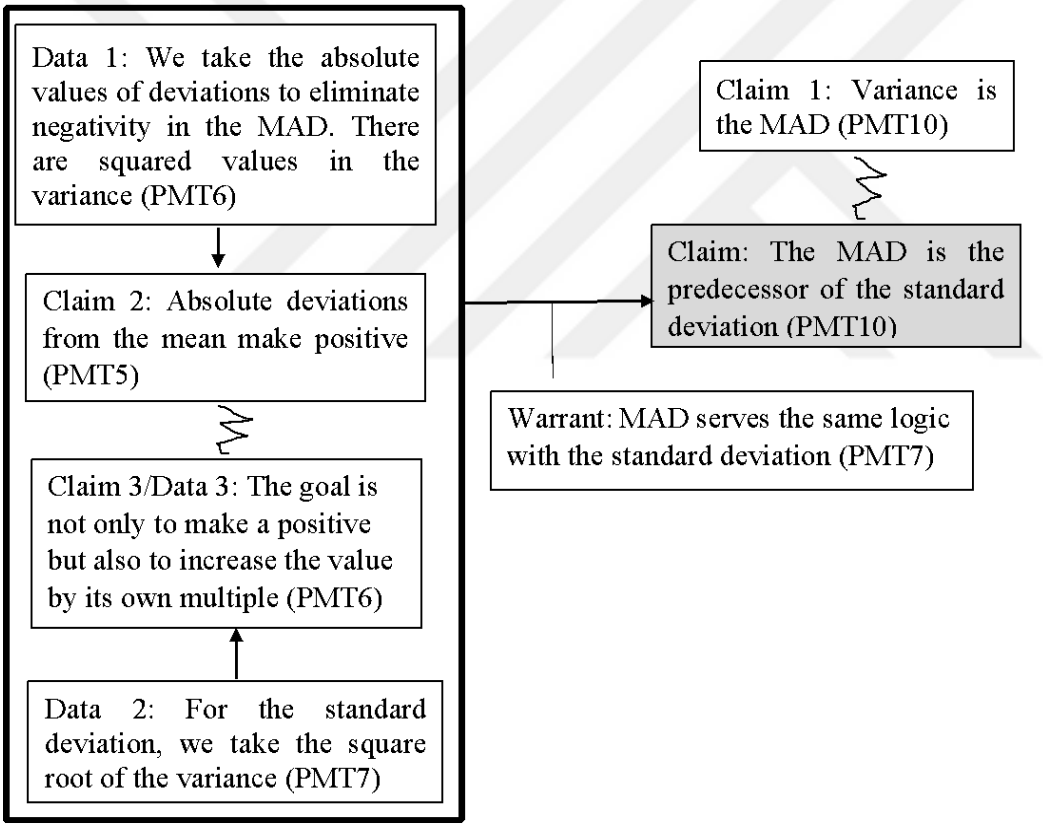


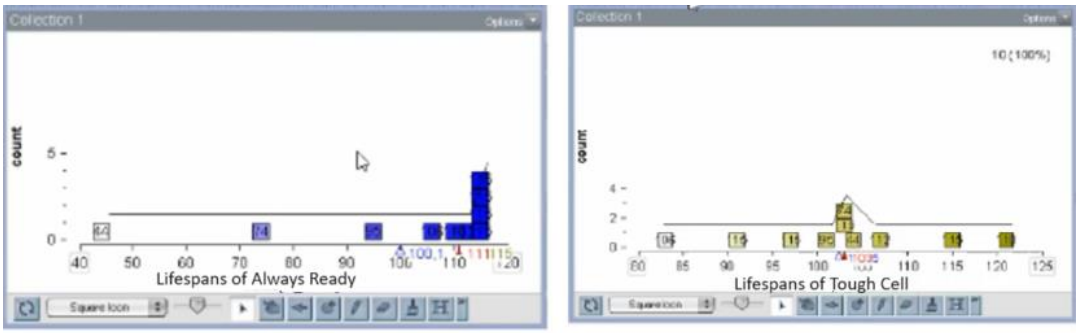
Figure 4.26. The argumentation scheme regarding the relation between MAD and the standard deviation

In the HLT, the main learning goal of the fourth week was variability and variability measures. The PMTs knew what these measures (range, interquartile range, SAD, MAD, standard deviation, and variance) were and how they were calculated. The

range, which was called the simplest variability measure, was expressed over the first situation of the soccer games in Figure 4.22. In the first situation, the range of Team A was seven, and the range of Team B was eight. The range was said to be used because there were no extreme cases in these distributions. However, in the last situation of the soccer games in Figure 4.24, the range was not used because both the number of samples of the distributions was different, and the values in Team A were very extreme. Similarly, since the highest and lowest 25% of the data were eliminated in the interquartile range, it was said that the actual variability was obtained. It was also stated that mean, median, and mode are more likely to fall within this range. These comments for the interquartile range can be a trace of the fact that the PMTs were beginning to see the distribution as a whole. Due to the lack of warrant or backing in the argumentation, the PMTs' interpretation of the interquartile range became a TAS idea that was unlike the range, as *the interquartile range reduces the effect of outliers/extreme values*. Since it was used based on the data in the activities in the subsequent class sessions, the idea related to the range and interquartile range, which were the measures of variability, was presented under later CSP- Selecting appropriate statistics for the context.

In the last class session of the fourth week of the teaching experiment, the data of activity related to batteries to be applied the later week were given. The aim was both to become familiar with the data and to explore the software's features by practicing on TinkerPlots. The PMTs entered ten data values for both battery brands, Always Ready and Tough Cell. Then, two plots were drawn using squares instead of dots. The mean, median, and mode values were calculated on the plots by using the "Averages" button of the software. Regarding the interpretation of the plots, PMT12 claimed that the Always Ready brand was a left-skewed distribution (Figure 4.27a) by ranking the mode, median, and mean from the largest to the smallest. These values were not equal for the Tough Cell brand, but they were very close to each other, 102.2 for the mean and 103 for the median. Although it was not an exactly normal distribution (Figure 4.27b), the expression "close to normal" was used. Based on the exercise made in this class session, it can be deduced that the PMTs determined the

shape of the distributions according to the size of the measures of the center. Indeed, they compared the mean, median, and mode for the distribution shape from the second-class session in the second week to the last week of the teaching experiment. The idea about the location of the measures of center judging the shape of the distribution was presented in the last CSP- Exploring qualitative characteristics of distributions.



(a) (b)

Figure 4.27. The plots drawn by PMTs for the two battery brands in TinkerPlots

To sum up, in the fourth week of the teaching experiment, the PMTs were interested in normality, SAD, MAD, variance, standard deviation, range, and interquartile range. Except for the concept of normality, the others are measures of variability. The normality was also associated with variability while working on the soccer game activity. Therefore, the three TAS ideas that emerged with normality, MAD, SAD, and standard deviation revealed the second CSP- measuring of variability in data. The TAS ideas were 1) *Normality is not a criterion for variability*, 2) *If the number of samples is different, the MAD is used rather than the SAD*, and 3) *The MAD is a predecessor of the standard deviation*. Regarding these TAS ideas, the PMTs described, modeled, and explained the variability in the data. They compared the measures of variability with each other, which allowed them to investigate the features of the measures. Challenging tasks given to the PMTs to compare distributions led them to reason about variation. In this context, it was concluded that the reasoning about variation was necessary to describe how the data was distributed.

Since the centers and even the SAD values were the same in the tasks, it also supported the conjecture of not separating the study of center and variability that emerged in the first CSP.

4.2.3 Classroom Statistical Practice 3: Selecting appropriate statistics for the context

Selecting appropriate statistics for the context was carried out in all activities performed during the teaching experiment process. Therefore, this practice was intertwined with other CMPs and TAS ideas. Since the findings were presented in the chronological order of the teaching experiment, the ideas under this practice were presented retrospectively, connecting them with the ideas that emerged in the previous and subsequent class sessions. In this sense, there was no linear flow in the reporting of ideas. In this section, I present ideas about selecting appropriate statistics for the context of the learning activities in the HLT. Three ideas about this CSP became TAS in the fifth week of the teaching experiment, in which the PMTs worked on the "building arguments based on statistical measures that are appropriate for the context-batteries" activity. The TAS ideas were as follows.

The measures of variability support each other.
Connecting the measures of center with the measures of variability leads to more reliable results.
Unlike the range, the interquartile range reduces the effect of outliers/extreme values.

The learning goal of the "building arguments based on statistical measures that are appropriate for the context-batteries" activity was to justify the valid arguments by applying concepts such as range, consistency, total, and mean within the given problem context. Furthermore, the PMTs were intended to organize the data in a manner that would support these arguments, including actions like partitioning or grouping the data by using TinkerPlots. Before giving the data that they had plotted the previous week, they talked about electronic devices that use batteries, how long it takes to change these batteries, things to make batteries last longer, how long it

takes for their phones to run out of charge, etc. Then, a table given in the last class session of the fifth week of the teaching experiment was again presented to the PMTs. It consisted of ten data values that were the lifespans of the batteries of the two brands. The PMTs were asked to determine which brand was better by providing reasons. They were invited to work in small groups to come up with different solutions. Before moving on to groups, one person was assigned to each group as the recorder of the group work. This responsibility of recording was assigned to different PMTs every week. The small group work for the "building arguments based on statistical measures that are appropriate for the context-batteries" activity took about ten minutes.

If the battery's life span is longer, it can be used for a longer time, and it is more reliable if these times do not differ in different uses. Therefore, in the activity, the PMTs were expected to prefer a distribution with a high average and low variability, depending on the context. Different statistical measures or concepts met these criteria in small groups. The PMTs did not cite the shape of the distributions discussed in the previous week. The first thought was that the battery brand with a high mean was good for longevity. The PMTs predicted that the distribution with higher values would have a higher mean before calculating the mean. They could quickly review all data values due to the small sample size in the distributions. For example, PMT8 stated that the Tough Cell brand was better because it had more than 90 values. Regarding this interpretation, they started to use the idea that "*the mean of the data set with high data points is high.*" In addition, since the number of samples was the same, ten in both brands, they used the TAS idea that "*if the sample sizes are the same, the mean and the total say the same thing*" in the first week. They agreed that taking the mean and collecting the lifespans served the same. However, the two groups voted for the means instead of the total because they would use means when calculating MAD or SAD. In addition, they quickly calculated the mean of the distributions by clicking a button in TinkerPlots. Again, since the number of data is equal, they used SAD instead of MAD. It showed they used the TAS idea that "*if the number of samples is different, the MAD is used rather than the SAD*" in the previous

week. For example, after the totals were said by Group 3, PMT10 explained that "we can look at its variability over SAD, as there is no need for MAD if the samples have the same number of data. We learned that in the last week."

Measures for variability differed, while the mean was always used for the center in the groups. Group 1 made a descriptive analysis by entering the data into Excel. They reasoned using the mean, standard deviation, and range from the obtained inputs. Although group members tried to relate the level of reliability in the inputs to hypothesis testing, they could not conclude because of the reversible relationship. Therefore, they did not present it in the class discussion. The group described the Always Ready brand as unsafe because of the greater variability, with 23.61 versus 10 in standard deviations. In this sense, the PMTs used the TAS idea that "*the dispersed distribution is unreliable*" presented under CSP4-Exploring qualitative characteristics of collections of data points. The range also supported the claim of unreliable distribution with larger variability. The range of the Always Ready brand was 71, while the other was 38. In distributions with equal sample sizes, the degree of variability of the Always Ready brand was more than doubled for both standard deviation and range values. This situation, where multiplicative comparison did not make for variability, was expressed as "there was a big difference" by PMT14. The other group members approved his claim. This result achieved with both standard deviation and range in Group 1 was obtained with SAD and range in Group 3 and only SAD in Group 2. This aroused the TAS idea that "*the measures of variability support each other*." In subsequent sessions, the PMTs continued using different viability measures to support their claims. Since the PMTs focused only on the measures of variability rather distribution of the data while using the idea, they were at the transitional reasoning level. They were interested in the standard deviation, SAD, and range. The following excerpt was from the class discussion on using multiple measures for variability in the second-class session of the week. Although they calculated the values of each measure specified in the group work, they summarized them in the class discussion. The solution of Group 3 in Figure 4.28 was first presented.

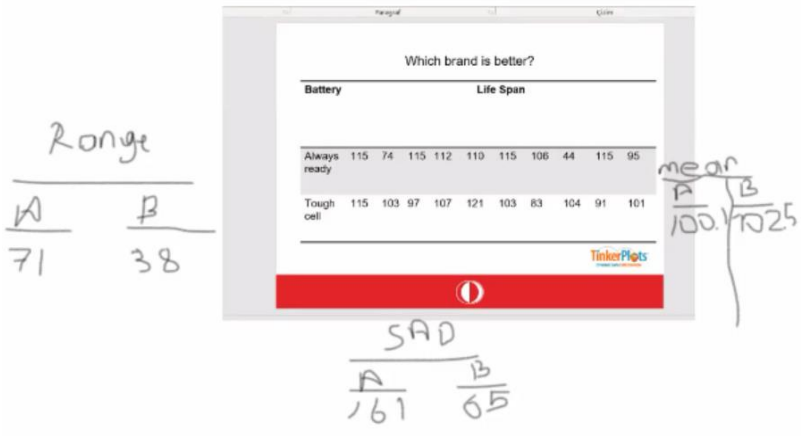


Figure 4.28. The solution of Group 3 for the activity related to batteries

> PMT10. After looking at the mean (100.1 for Always Ready and 102.5 for Tough Cell), it is enough to look at SAD as both have the same number of data. We do not need to look at MAD. When we calculated it, B (Tough Cell) looked more reliable anyway, but we still looked at the range. The range has supported this SAD. That is why we chose B (Tough Cell). Instructor. Because both measures support each other. So, what does the low SAD or low range mean?
>
> PMT2. It means that lifespans are close together. There is not much difference between them.
>
> Instructor. Which brand is better?
>
> PMT11. For example, there are 44 and 74 in the first one. They somewhat lowered credibility. We thought the latter was better. In other words, we found it more reliable because it sounded more consistent.

In addition to the SAD, range, and standard deviation, all PMTs were aware of individual values that differed across the data, such as 44, for variability. They considered these suspicious values to be the source of the inconsistency or higher variability. In other words, the outliers increased the range, that is, variability in the data. In fact, like the mean, the range was also responsive to outliers. It was the basis of the TAS idea that "*unlike the range, the interquartile range reduces the effect of outliers/extreme values*" in this CSP, selecting appropriate statistics for the context. However, it had not yet been articulated clearly as the PMTs focused on the context in the class session. Since Group 3 preferred the better criterion of consistency, they turned to the brand that did not contain data values far from the general data and had low variability.

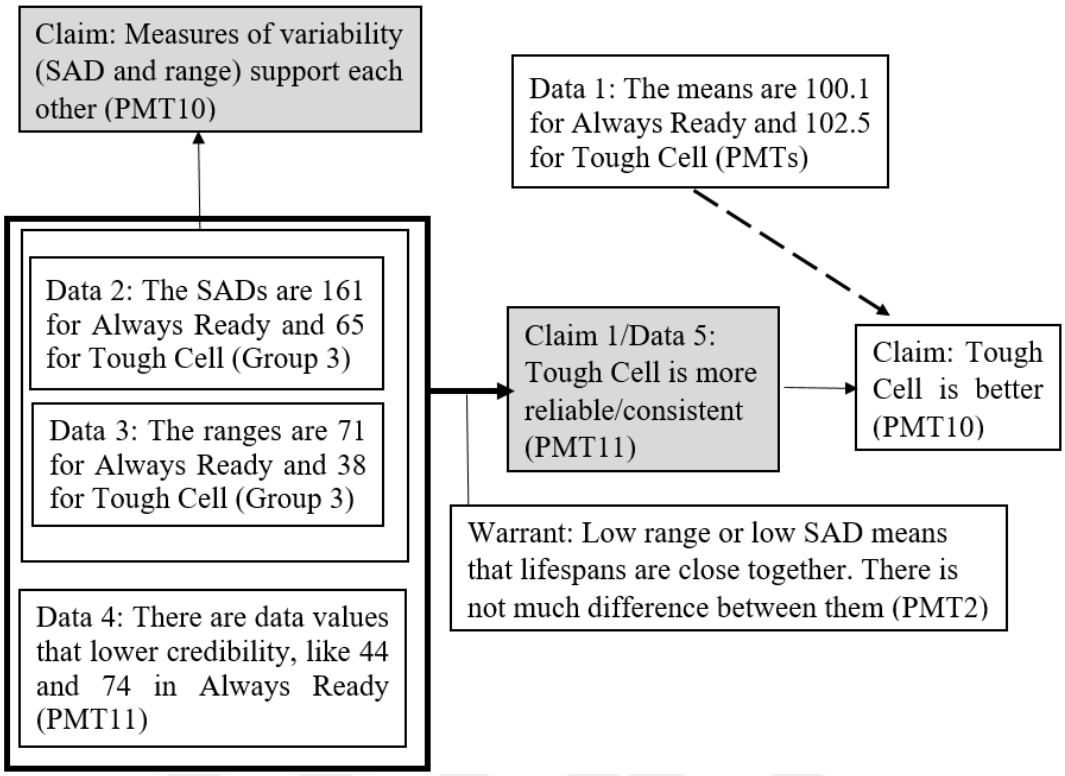


Figure 4.29. The argumentation scheme for the use of measures of variability for the activity related to batteries

Figure 4.29 shows the conclusion that the PMTs reached based on the measures of variability in the class discussion. It contains two TAS ideas represented in dark colors in the scheme. The first is that "*the measures of variability support each other."* The measures are the range and SAD in this argumentation. The other idea is that *the dispersed distribution is unreliable.* This idea is implicit in Claim 1. Therefore, it is reported under the next CSP- Exploring qualitative characteristics of collections of data points where it is clearly stated. In this argumentation in the figure, the result of "Tough Cell is better" obtained based on variability was also implicitly supported by the mean. However, no relationship between the center and variability could be established, as shown by a dashed line in the figure. Continuing the class discussion, Group 1, which concluded with the mean and standard deviation, needed to look at another measure because they judged the mean values to be very close.

PMT7. While comparing which brand is better, we preferred to look at the mean first, and we saw that the Always Ready brand was 100.1 and the Tough Cell brand was 102.5. The means are very close to each other, but the average of Tough Cell is higher. Let us look at the standard deviation; let us be precise.

Instructor. Where did you calculate these?

PMT7. We did it in Excel. When we look at the standard deviation, we found 23.61 for Always Ready and 10 for Tough Cell. Since the standard deviation of Tough Cell is less, we thought that Tough Cell had better longevity. If the deviation is small, it is a good sign.

PMT11. Its lifespan is also too long because it has a higher mean.

As seen from the dialogue above, they were unsure about their choice because the difference between the mean values was small. It was unclear why they rated the 2.4-minute difference between the means as large or small. In this argument, they needed help explaining relative size. They thought this value did not affect the result and used the standard deviation from other measures to decide. They concluded that the one with the smaller variability had a longer lifespan, which does not coincide with the idea above that the distribution with low variability is more consistent. In other words, they interpreted low variability as higher values. They tended to corroborate their results by using different measures. However, they had trouble choosing appropriate equivalent statistics (measures of variability for longevity and measures of variability for consistency) for the context. Regardless of which statistical measure was selected, everyone in the class practiced using multiple measures in contextual decision-making. That is to say, when they encountered the statistical problem, they tended to calculate all possible measures such as the mean, range, MAD, and SAD. However, they considered that one of them should be a measure of center, generally, the mean, and the other should be a measure of variability. Therefore, they used at least two measures in all learning activities during the teaching experiment.

In the following excerpt from the class discussion, the PMTs in Group 2 impressed individual data values in the distributions for the activity related to batteries

PMT3. We thought the mean and SAD like other groups. Then, PMT13 said plus for Always Ready and minus for Tough Cell. Then we looked at the difference between them—for example, 115-115 in the first one. The

> difference between them was 0. We wrote it in the middle. In the other, the difference between 74 and 103 was 29. However, because we called the bottom minus, it became -29. So, it went down. Then 97-115. The difference between them is +18. This one went to the above. That is how we did it all, one by one. Then, we gathered all the differences. Then, we saw that it was negative (-24). That is why we said this (Tough Cell) would last longer.
>
> PMT12. Since it was negative after that, we came to the following conclusion: the ones below, namely those with Tough cells, have a longer battery life.
>
> Instructor. What did you do? What does it mean to do one-to-one correspondence and get the difference between the data?
>
> PMT13. Currently, the number of data here is the same. In fact, our experiment would have yielded almost the same result as the mean. However, I do not think it is a good method. Because Always Ready has extreme value, like 44. However, Tough Cell has values that are intertwined with each other. That is why our experiment with the mean or a method close to this mean did not give us a reliable result.
>
> Instructor. What do you think?
>
> PMT10. However, this already confirms the longevity of Tough Cells—extreme values in the other. We said Always Ready is inconsistent before. It is both durable and consistent.

This excerpt clarified the mean part of the argumentation in Figure 4.29. They added the center to their previous arguments based on variability in the activity. Here, they embodied the idea that "*connecting the measures of center with the measures of variability leads to more reliable res*ults." They seemed to be at the level of integrated process reasoning in terms of reasoning about statistical measures since they suggested relating center and spread aspects of the distribution. However, although they integrated the measures, they were at the procedural level when reasoning with this TAS idea since they did not deal with the distribution by thinking about other aspects, such as density and skewness.

The idea first appeared in the first week of the teaching experiment, which proved it to be a TAS idea. In the activity related to batteries, they concluded that the Tough Cell brand was better regarding variability and center. In the center component of the argumentation scheme in Figure 4.30, the sum of the difference between values by matching one-to-one indicated which distribution had higher values. While

sequentially subtracting the values might seem like comparing their relative magnitudes, this approach was equivalent to comparing the sums of all values in the distributions since the final operation was to combine them. When dealing with totals of the values in the distributions, they compared their means, as the PMT13 noticed. This way of reasoning also verified the TAS idea that "*if the sample sizes are the same, the mean and the total say the same thing.*"

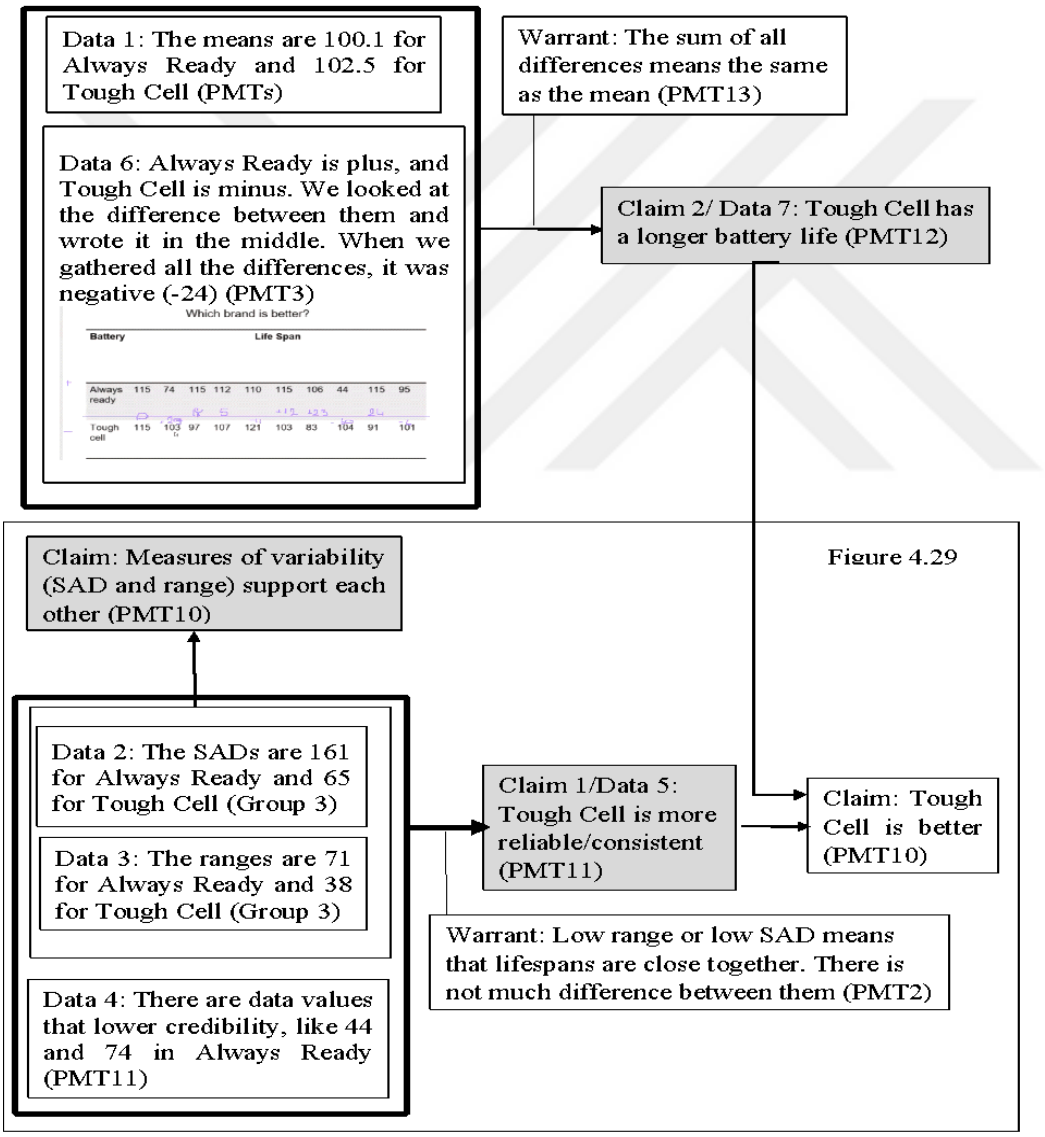


Figure 4.30. The argumentation scheme where measures of center and variability were connected for the activity related to batteries

Figure 4.30 contains three TAS ideas represented with dark colors. They were "*the measures of variability support each other*," "*the dispersed distribution is unreliable*" in Figure 4.29, and "*the mean of the data set with high data points is high.*" These are discussed in the next CSP. According to the scheme in Figure 4.30, the PMTs supported the result of "Tough Cell is better" by longevity based on the higher mean or total in the above part of the figure and by reliability based on low variability in the below part of the figure that corresponded to Figure 4.29. That is to say, they used both measures of center and variability to support each other in determining the result. The PMTs explained the reasoning based on the magnitude of the data values in the distributions with the seesaw metaphor by inspiring the balance point interpretation of the mean. In other words, they thought that whichever distribution has the larger values, there is a disposition toward that direction. In the above example, since the values assigned to minus are more in total, it was selected in favor of Tough Cell. If it were marked with a plus, the sum of the differences would be reversely positive. In this approach, the opposite result could be reached if the reasoning was made according to the number of signs in the differences instead of summing the differences because the differences in the solution made by the group had five positive values (18, 5, 12, 23, and 24), four negative values (-29, -11, -60, and -6) and one neutral value (0). Accordingly, Always Ready could have been preferred since the number of positives was higher. On the contrary, the PMTs were interested in the whole data rather than comparing the individual values, thanks to the sum of the differences in the data values.

After the decision in favor of Tough Cell, at the end of the last class session in the fifth week, Assignment 3 was given to the PMTs to evaluate and reinforce what they learned about consistency and lifespan in the context of batteries. There were no restrictions, such as group or individual, for the PMTs to do the assignment. While six PMTs completed it in pairs, the others did it individually. The assignment required the invention of three battery distributions consisting of ten data values meeting the following criteria and plotting them.

Brand A is good and reliable.

Brand B is good but unreliable.
Brand C's spread is like A, but the brand is less good.

All PMTs, except PMT14, who generated the data in Excel, visualized the data they invented in TinkerPlots and explained why the data met the criteria specified in their reflection papers. While PMT6 interpreted her data on dot plots in TinkerPlots, the other twelve PMTs drew horizontal bar graphs for the data using the software's ruler feature. All PMTs met the required criteria in their data. That is to say, they could make sense of the need for a high average to qualify the battery as good and low variability to be reliable. Everyone compared the means of the data for high quality. For reliability, MAD or SAD was calculated in TinkerPlots. PMT13 and PMT7 stated that the standard deviation value can be estimated with the MAD values. It showed they used the idea that "*the MAD is a predecessor of the standard deviation*." In addition, PMT6 and PMT14 calculated the standard deviations outside of TinkerPlots. The PMTs stated that no extreme values should differ from the general data for reliability in their assignments. For example, the graphs drawn by the PMT2-PMT4 pair in TinkerPlots (Figure 4.31) and their explanations were as follows:

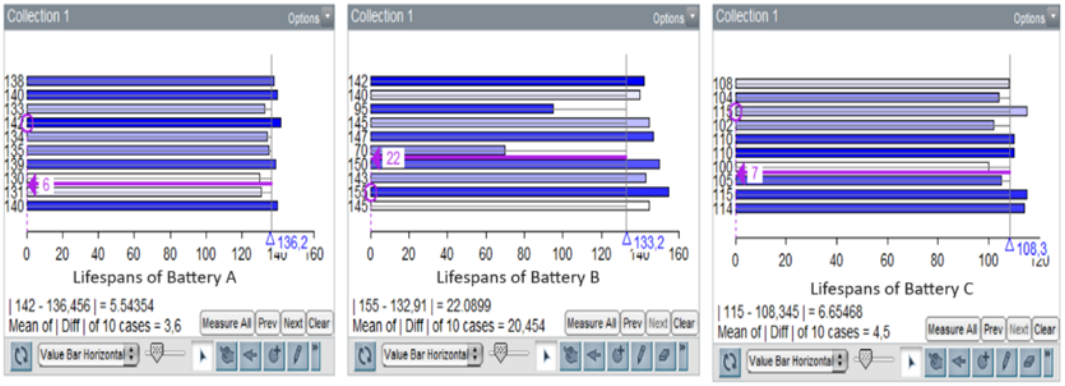


Figure 4.31. The horizontal bar graphs of the three battery brands drawn by the PMT2-PMT4 pair in TinkerPlots

A is a reliable and good brand because the mean and samples are very similar to each other, and there is no extreme value. In addition to that, MAD is 3,6, so variability is low. Therefore, it is reliable.

> B is a good but not reliable brand because the mean is high, and most of the samples have high values. However, it is unreliable since there are two extremely small numbers, 70 and 95. Also, the MAD of the B brand is 20, implying high variability. Therefore, it is not reliable.
>
> C is less good than others because the mean is lower than the other two brands. However, Brand C is reliable because the samples are very close to each other, and there is no extreme value in Brand C. Moreover, the MAD of brand C is 4,5, which indicates that the variable is low, like Brand A. Therefore, Brand C is reliable, but it is less good.

As can be seen from their explanations, they had kept the mean values high for a good brand and closed the data to each other for reliability. They found the low mean value from the values which were low and close to each other for the C brand, which treated bad and reliable brand. They selected two outliers in the non-reliable brand B, lowering both the mean and the reliability relative to brand A. The outliers increased the MAD value, thus reducing the consistency. One or two outliers in the data were generated for brand B in all assignments. In this sense, everyone in the class attended that the outliers extend the variability in the data in terms of both range and deviations from the mean. Therefore, they focused on where the majority was to interpret the distribution. Consequently, they tended to use the interquartile range rather than the range. While discussing variability in the second class session of the fourth week, PMT6 clearly said the tendency: "If there are extreme values, these affect the estimation to be made because of so many differences. IQR reduces the effect of extreme values." No one in the class objected, and in the following activities, they compared the interquartile ranges on the box plots, especially when working on activities with a large number of data. Therefore, the idea that "*unlike the range, the interquartile range reduces the effect of outliers/extreme values*" became TAS. This idea was included under CSP3, selecting appropriate statistics for the context, as it was about which statistical measure would be more appropriate to use for variability. Although the PMTs compared range and interquartile as measures of variability in this idea, their explanations based on the majority of the data extreme values in their assignments showed that they generated the data by relating aspects of the distribution (e.g., center, spread, density, and skewness). Therefore, they were at the integrated process reasoning level.

As a result, in the fifth week of the teaching experiment, the PMTs were interested in building arguments based on statistical measures. To make a decision, they utilized statistical measures appropriate to the context of batteries. For this, they needed to comment on both the center and variability measures. They also evaluated the measures among themselves. In this respect, the three TAS ideas regarding statistical measures emerged while selecting appropriate statistics for the context-CSP3. They were 1) *The measures of variability support each other*, 2) *Connecting the measures of center with the measures of variability leads to more reliable results*, and 3) *Unlike the range, the interquartile range reduces the effect of outliers/extreme values*. Based on the TAS ideas revealed by the class community, it was concluded that considering multiple statistical measures was crucial to drawing more accurate and reliable decisions for a statistical context. Even if the PMTs were interested in one criterion for the context, they could provide more than one piece of evidence through different measures for statistical results. When considering more than one measure, they also noted which was more meaningful for describing the distribution. For example, although the range and interquartile range were both measures of variability used for consistency, the interquartile range was preferred to describe the overall distribution because it did not include extreme values.

4.2.4 Classroom Statistical Practice 4: Exploring qualitative characteristics of collections of data points

Four ideas about the qualitative characteristics of the collection of data points became TAS in the sixth and seventh weeks of the teaching experiment, in which the PMTs worked on three learning activities ("generating different samples in weight context-hot air balloons," "building arguments based on statistical measures that are appropriate for the context-watermelons", and "understanding the shifts in hill, clump, or cluster of data with equal sample size-resting heart rate"). The TAS ideas were as follows.

The sample varies between extreme cases based on the specific value/mean

The mean of the data set with high data points is high
The dispersed distribution is unreliable
The data is concentrated where the mean is

The first three ideas emerged in the sixth week during the "generating different samples in weight context-hot air balloons" and "building arguments based on statistical measures that are appropriate for the context-watermelons" activities, while the last one was encountered for the first time during the "understanding the shifts in hill, clump, or cluster of data wit equal sample size-resting heart rate" activity in the seventh week. Unquestionably, the ideas from the previous weeks continued to be used. Regarding this, these four TAS ideas were presented concerning the ideas of previous weeks.

Until the sixth week of the teaching experiment, the focus of the HLT was developing an understanding of center and spread, which are the distribution aspects. This served as the ground for developing reasoning about distribution, which constituted the study's primary goal. The next step was to explore the concept of the distribution. A challenging situation was needed to encourage the PMTs to consider the distribution. For this, the sixth week started with the "generating different samples in weight context-hot air balloons" activity without any data. The learning goal of the activity was to think about the average within variability and understand the representativeness relationship between samples and populations. The context of the activity implemented was as follows: There was a balloon that could safely carry eight adults, not including the driver. It was asked how many seventh graders this balloon could safely carry. The PMTs would determine the number of students who would ride in this balloon in the role of the teacher of these students. They were familiar with the context from their travels in Cappadocia or from the experience, if not of their own, of the people around them. Upon the instructor's questions, they talked about their balloon riding experiences, the working system of the balloon, the maximum number of people who can ride the balloon, etc. After understanding the problem, they studied the activity in small group work.

In the small groups, there were many suggestions for the average weight of a person. Some of these were finding the average weight of the seven people in the group,

determining the weights by gender, or taking the 75 kilos accepted in the elevator as a reference. An idea was suggested in Group 1 but needed to be discussed to be a communal view of the class. PMT7 in the group said, "There are eight people. Let us divide them into four women and four men. Then let us take those four as two thin and two overweight." In this case, the distribution between extremes was considered rather than the sample size. This situation raised the TAS idea that "*the sample varies between extreme cases based on the specific value/mean.*" Since the idea referred to data distribution according to specific values, the PMTs seemed to be at the level of integrated process reasoning. However, PMT2 opposed it by explaining that instead of separating men and women, it should be a standard thing that covers all overweight and thin. This approach narrowed the reasoning levels of the PMTs in the group to the transitional level because they focused only on the center. By setting the means for both students and adults, they solved the statistical problem in the "generating different samples in weight context-hot air balloons" activity, where no data was given. That is to say, they focused on a single value, representing the sample instead of a distribution.

The PMTs in Group 1 used different approaches when determining the weight of a standard adult and a student. For the two samples, they judged the average weight by maximum and minimum weight or gender. In terms of the magnitude of the weight, they randomly guessed the maximum (100 kg) and minimum (40 kg) values. They negotiated that more students could ride if the adult had the maximum weight, but the student had the minimum weight. They talked in the group that if they accepted both the student and the adult in maximum or minimum weight, the average value would correspond to the same number of students. Therefore, they decided to handle the average weight of adults and students according to gender. They used the table they accessed from the internet. The table included information about average weight according to gender and age. In the table, the average weight of male and female students was almost the same (50 kg). In adults, they took an average of 75.8 kg for males and 66.9 kg for females. They calculated the number of students by considering whether the adults were all females or all males. In another calculation,

the average weight of adults was the average of males and females (71.35 kg). That is to say, they had eliminated the gender variable. In different approaches, the number of students varied between 10-13. Afterward, they decided that the number of students was 12, thinking the average was 75 kg in the elevators. While determining the number of students, their reasoning was the ratio of the total weight the balloon could safely carry to the means set for both samples, which was observed in all small group works. The group's solution shared in the whole class discussion is given below.

> PMT5. The table says 49.8 kg for 13-aged students and 70-75 kg for adults.
>
> PMT9. Let us assume 50 kg for the students.
>
> PMT2. I have always seen 75 kg in elevators. Then, if we take 75, the total is 600 kg (75x8). It is divided by 50. The balloon can carry 12 students.

The PMTs considered extreme values when executing estimated values for means. For example, PMT3 in Group 2 stated, "It can be very overweight; it can be very extremely underweight." As in this description, extreme values or outliers for weight were considered in all groups. It showed that they were aware that *the sample varies between extreme cases based on the specific value/mean.* Therefore, they started to think at the level of integrated process reasoning by considering other aspects of the distribution besides the center. Group 2 found this range through an internet search. Since 7th graders were approximately 13 years old, these ranges were between 34-63.3 kg for female students and 30-67.5 kg for male students. For the balloon context, they took into account the mean of the maximum weights of male and female students (65 kg), considering that a safe ride should not exceed the maximum total weight. Assuming a maximum adult weight of 90 kg, they concluded that about 11 students could be safely ridden using the same solution method (90x8=720 kg; 720÷65 ≅11).

As another solution method in Group 2 and Group 3, they proportioned the means they set without needing the total weight. In the class discussion, PMT6 stated: "We said that one adult (weight) corresponds to almost 1.5 students (weight). The weight of 8 adults is equivalent to that of 12 students". When the instructor asked how they

concluded that, she said they took the ratio of the average weights (75/48). Everyone in the classroom accepted the approach based on proportional reasoning without justification. In this solution method, they were interested in the quantities coming from the same weight measurement space (e.g., between-ratio).

The PMTs generally used two solution methods in the "generating different samples in weight context-hot air balloons" activity. Both solution methods resulted in the same conclusion. For example, the arguments of the PMTs for 12 students that the balloon could safely be carried in the class discussion were as in Figure 4.32.

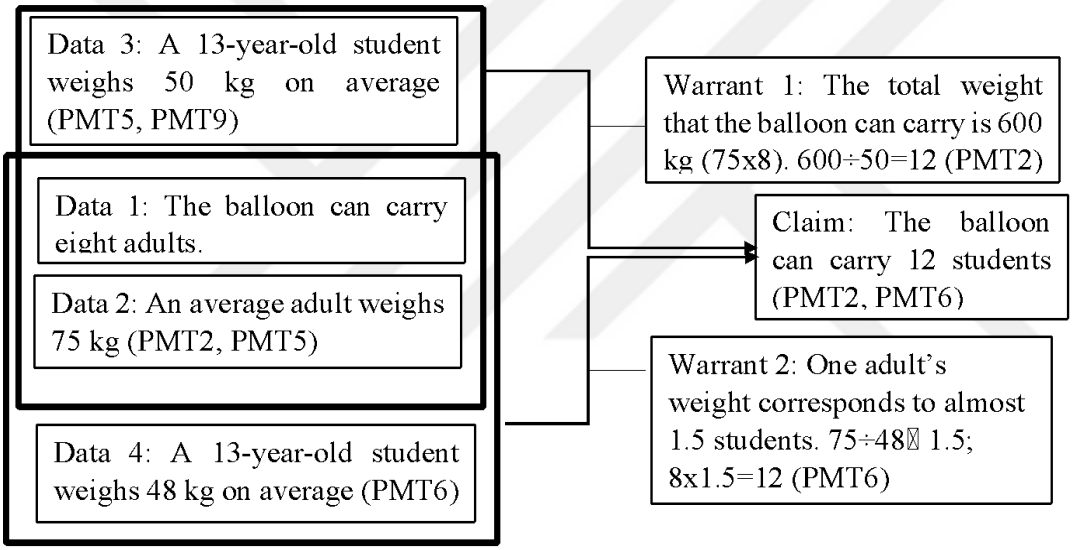


Figure 4.32. The argumentation scheme regarding the balloon can safely carry 12 students

In the argumentation scheme, although the means set in the groups differed by 50 (Data 3) and 48 kg (Data 4) for the average student weight, the same conclusion was reached that the balloon could carry 12 students. The PMTs calculated the total weight by multiplying the means they set by the number of people (Warrant 1), or they compared the mean set (Warrant 2). In the first solution method, multiplication was performed first, followed by division. The other solution method was the opposite. Apart from unitizing in the second solution method, the same reasoning

underlined the two solution methods. The result of dividing the average weights in the second method functioned as a unit factor. In another solution based on Warrant 2 in the figure, the weight of one adult was matched with the weight of two students. That is, the unit factor was two. Thus, it was concluded that the balloon could safely carry up to 16 students. Regarding this conclusion, the instructor asked what its graph looked like in the class. The primary purpose of the instructor was to draw attention to the distribution. PMT2 plotted the graph in Figure 4.33 according to the number of people at the end of the class session.

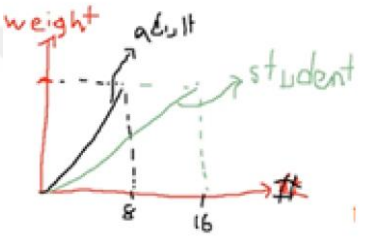


Figure 4.33. The graph is drawn by PMT2 for the weights of students and adults

PMT2 preferred to show students and adults in the same graph, emphasizing the ratio of two between them. He drew attention to the fact that the total weight for the two samples was the same while interpreting the graph for the balloon context. Unfortunately, the number of people on the horizontal axis was continuous. While interpreting the graph, PMT10 objected: "We took the mean as a weight value. However, I do not understand that we are trying to draw it as a distribution." The means set in both solution methods are based on the multiplicative relationship of the average weights, which represent the entire weight distribution. Namely, the distribution was treated as consisting of a single value. The average reduced a set of numbers into one value in which the variability in the distribution was ignored. Therefore, in the next class session, the instructor asked PMTs to invent twelve data values for 7th graders and create a graph for the data in TinkerPlots. The instructor's intervention to generate data in TinkerPlots allowed them to think about the distribution of the weights.

In the second class session of the sixth week, in small groups, the PMTs drew the graphs for the twelve student weights that the balloon could carry in TinkerPlots. The data were determined based on 50 kg, which was the average weight of the students. The following dialog was an excerpt from the data invention process of Group 1 presenting its graph in the classroom.

> PMT2. Let us consider level off or balance. If the mean is 50, we get 40 and 60.
>
> PMT4. We can get partial, too. 35 kg may be the lowest, logically. 35-65, 47-53, 46-54.
>
> PMT9. We set 50 as the middle value; let us set an upper and a lower value. Let us make it between 35 and 65. (PMT4 wrote data values as 35,40, 45, 47, 48, 50,50, 52, 53,55, 60, and 65).
>
> PMT9. For too many values in the 42-48 range (Figure 4.34a), move them more to the right, which becomes a normal distribution.
>
> PMT2. Let us make 47 43 and 53 57 (Figure 4.34b).
>
> PMT9. Weird. The distribution changes completely. As dragging, the shape seems skewed. More at the bottom makes sense, as they are usually around 40.

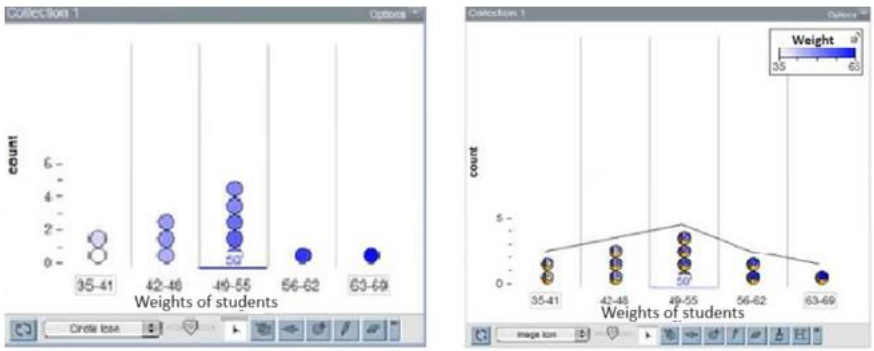


Figure 4.34. The graph drawn by Group 1 for the weights of 12 students in TinkerPlots

They referenced the normal distribution to generate the data. Based on the balance point interpretation of the mean, they created the data consisting of symmetric pairs to 50. TinkerPlots enabled the creation of graphical representations and the quick calculation of the mean. Although the generated data values were bell-shaped, the

appearance seemed different due to scaling or the set intervals in the graphs. As the data values were manipulated in TinkerPlots, the shape of the distribution could be interpreted immediately because the graph in the statistical software changed dynamically. In addition, as they practiced the tool, they noticed more informative descriptors about the characteristics of the data. For example, they used the key button, which indicated that the data was quantitative by the color scale in the upper right corner, and the label button demonstrated the data values above the dots (Figure 4.34b).

Like this group, the other groups first identified the extreme values. Group 2 specified the range between 38 and 55, Group 3 between 38 and 62. They generated the rest of the data according to the mean value. This data generation process showed that the PMTs at the integrated process reasoning level explicitly used the TAS idea that *the sample varies between extreme cases based on the specific value/mean*. While inventing data, they relied on a symmetrical distribution. However, Group 2 presented the left-skewed graph in class discussion. In their small group work, the group focused on the total weight, 600 kg, from 12×50. PMT14, one of the group members, drew attention to the variability. He said, "The data set has just random numbers." In the whole-class discussion, they argued that the left-skewed distribution was more realistic when they made the mean about 48 by changing a few values. On the contrary, Group 3 wanted to create a symmetrical distribution by assigning data values equidistant from the mean. Their data consisted of pairs and triplets: 38-62, 42-58, 45-55, 49-50-51, and 45-47-58. The last triple caused more data below the mean. The symmetrical, if not precisely, graphical appearance formed.

Although all groups estimated that the 7th-grade students' weight values concentrated around 40-50 kg like right-skewed distribution, the distributions of the groups were symmetrical and left-skewed, respectively. When the instructor asked if anyone had a different idea, PMT10 suggested looking at the variability of the distribution of Group 1 through SAD or MAD, which was studied in Group 3, whose member was PMT10. The MAD value of the distribution, 6.5, was calculated on the

horizontal bar graph in TinkerPlots. Inspired by that inscription, after calculating the MAD, the other groups compared the variability of their distribution based on these values, which were 4.16 and 5.6, respectively. Regarding the shape of the distribution, it was seen that the variability of the symmetrical distribution was higher than that of the distribution that was not perfectly symmetrical or skewed. This situation supported the TAS idea that *normality is not a criterion for variability*. PMT10 also compared the variability of distributions belonging to Group 1 and Group 2 through their range. Group 1's data was distributed from 35 to 65, and Group 2's was spread from 38 to 55. The range of the distribution of Group 2 was again low, as was the MAD value. It validated the TAS idea that *the measures of variability support each other*. The class discussion continued with the following dialogue.

> Instructor. What would happen if one of the students who wanted to ride weighed 90 kg?
>
> PMT4. The mean increases.
>
> PMT2. Since ours is a normal distribution, the graph reverts to right-skewed.
>
> PMT5. The variance increases.
>
> Instructor. Why?
>
> PMT5. The distance to the mean is increasing. More diversity occurs as a more extreme value comes, but the number of students who will ride the balloon decreases.
>
> PMT13. There is no need to reduce the number of people necessarily. With lower-weight students, 12 students can ride again.
>
> PMT5. Ahaa, yes. We need to reduce one or two of them.
>
> Instructor. Why?
>
> PMT5. The mean and the sum should not change.
>
> Instructor. What has changed?
>
> PMT13. The graph and variance change.
>
> PMT12. If so, it will be right-skewed.

The participants reasoned that when a higher outlier far from the mean was added, the data values had to be concentrated on smaller values to be the same sample size, provided the mean and the total remained the same. Therefore, they were convinced

the weight distribution was right-skewed for the last situation. On the one hand, individual data values (e.g., the outlier) were considered in terms of variability. On the other hand, the participants focused on the shape of the distribution, which allowed the distribution to be viewed as a whole. Considering the distributions they invented, they had exemplified symmetrical and skewed distributions. The PMTs determined the shape of the distribution according to where the data was concentrated. For example, PMT11 called the distribution of Group 2 left-skewed because of the concentration of large values. Conversely, if the data was concentrated at low values, it was called right-skewed. The instructor continued her guided questions to make the PMTs think about the distribution shape of different samples.

> Instructor. What if there are 8th graders instead of 7th graders?
>
> PMT5. Since the 8th graders have more weight, the number of students getting on the balloon decreases. The mean increases.
>
> Instructor. Does everyone agree?
>
> PMTs. Yeah.
>
> Instructor. How is the distribution?
>
> PMT11. There is no clear data. We do not know the distribution.
>
> PMT3. I am not sure if we can compare the two (samples).

There was no consensus on how data would be distributed due to lack of data. When they kept the range small, they could invent distributions of all shapes. Therefore, data from the literature (Bakker, 2004) were presented. First, the shape of three distributions consisting of ten data values (Figure 4.35a) was interpreted. After that, the sample size was increased step by step (Figure 4.35b).

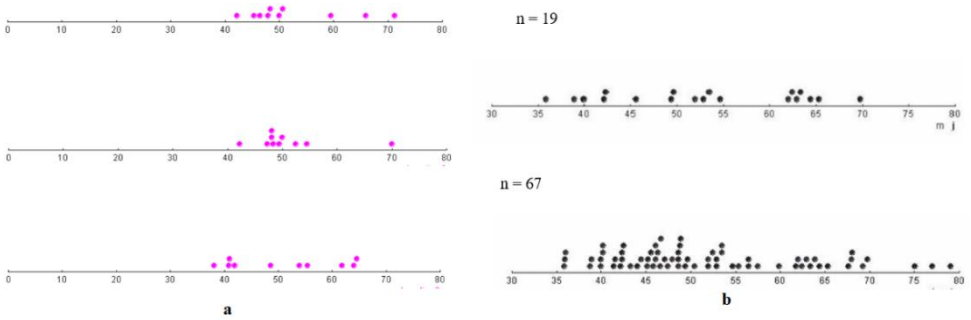


Figure 4.35. Different distributions for weight context (Bakker, 2004, pp. 219-222)

Instructor. Can you describe the graphs (Figure 4.35a)?

PMT5. The first one is right-skewed. Because the majority is on the left, the second one is normal.

PMT11. In the second, if we consider the 70 as an outlier, there is a tail there.

PMT5. Yes, it is right-skewed.

PMT4. The third graph is like a normal distribution. There is data at the beginning and the end.

PMT13-14. It is a normal distribution but with a large standard deviation.

Instructor. What will the shape look like when the sample size increases? Let us compare the weight distributions of 19 and 67 students (Figure 4.35b).

PMT11. In the first, the concentration is to the right of the mean, while in the 67-person distribution, it is to the left of the mean.

PMT4. In the second, the range is too much.

Instructor. What kind of graphic does it become as the sample size increases? For example, consider all the 8th-graders in the district.

PMTs. Right-skewed.

The PMTs could quickly determine the shapes of the existing three distributions of ten data values. In Figure 4.35a, the distributions seemed symmetrical and right-skewed. It had been observed that the data values were distributed in the range of 35-70. They interpreted the shape of distributions according to the position of the majority relative to the mean and whether there was an outlier or not. For example, when comparing the two distributions in Figure 4.35b, the majority to the left of the mean was justified for the right-skewed distribution. However, while interpreting the shape of the distribution, instead of calculating the mean value, its location according

to the majority was estimated. The idea that *the mean is responsive to the outliers* was used. That is, the outliers in the higher values increased the mean value. It formed a basis for the idea that *the location of the measures of center judges the shape of the distribution*, which was discussed in the last CSP, exploring qualitative characteristics of distributions. Although the weights of the students were between 35 and 70 kg, it was thought that the students might be overweight more than 70 kg as the number of samples increased when they imagined all students. For example, the last graph had three weight values greater than 70 kg. Everyone in the class agreed that the weight graph was right-skewed, as the majority was on the left, which was narrowing to the right of the graph. Then, right-skewed samples with larger samples, such as 500, 1000, and 5000, were generated by simulating in TinkerPlots. As the number of samples increased, smooth curve shapes were obtained. In this context, they started to interpret the shape of distributions with large sample sizes while reasoning about the specific values individually in small sample size distributions. Consequently, increasing the number of samples through simulation features of TinkerPlots helped see the whole distribution. Therefore, it was decided in the weekly meetings that the following activities in the HLT should have a large sample.

In the weekly meetings, the PMTs had been envisioned to deal with only the "generating different samples in weight context-hot air balloons" activity in the sixth week of the teaching experiment. They were likely to struggle even more because they were not given ready data. However, the PMTs were able to generate their data quickly. Therefore, the last class session of the week was free. The activities for the next week were planned in the weekly meetings so that the "understanding the shifts in hill, clump, or cluster of data with equal sample size-resting heart rate" activity was ready to be implemented. However, there needed to be more time to complement the activity in a class session. During this period, the PMTs could have recognized the context and done small group work without sharing their work in the class. For this reason, the "building arguments based on statistical measures that are appropriate for the context-watermelons" activity with ten data was implemented in

the last class session. As a matter of fact, we had abandoned implementing the activity related to watermelons in the meeting as it was similar to the activity related to batteries. The extra time was enough to study the activity related to watermelons without small group work. Accordingly, with an immediate change, the "building arguments based on statistical measures that are appropriate for the context-watermelons" activity was studied after the activity related to balloons in Week 6.

Before the "building arguments based on statistical measures that are appropriate for the context-watermelons" activity was introduced, the PMTs were asked to familiarize themselves with the context by talking about whether they had ever ordered juice in a restaurant, how restaurants bought fruits for juice, according to which criteria the fruits were ordered from the distributors, how juice was made in restaurants, etc. For watermelon juice, they mentioned the criteria of sweetness, flavor, price, size, quality of watermelons, and production rate of distributors. When introducing the activity, the instructor mentioned that the watermelons with the same sweetness and quality were bought from two distributors selling watermelons at $2.99 each. The PMTs made sense in that all criteria were constant except for weight. In the activity, the mission of the PMTs as restaurant managers was to decide which distributor they would buy watermelon from. Ten watermelons were randomly selected from Tropical Days Inc. and Fresh Fruits Inc., and their weight was measured in pounds. When the instructor asked why the weight was important, PMT4 said the heavier watermelon had more juice. PMT2 reasoned this by associating it with the juice glass to be owned. In other words, the heavier watermelon would have more juice glasses, and the restaurant would earn more money. PMT5 immediately claimed to choose the one with the higher mean to buy the heavier watermelon even before the data was given. PMT9 was in favor of watermelons being close in weight.

The data was presented in a horizontal bar graph representing the two distributors in different colors and a table. Two different approaches were produced for the activity related to watermelons, as PMT5 and PMT9 suggested before data was given. The first one was to find a distributor selling watermelons that weighed more. Preferring

to use Excel, PMT14 conducted a descriptive analysis of the data. From this analysis, PMT12 tended to Tropical Days Inc. with a larger total instead of the mean of the distribution. Since the number of samples was the same,10, in both distributions, the total or mean would lead to the same result: more juice is obtained from the watermelons of Tropical Days Inc. This result was based on the TAS idea that *if the sample sizes are the same, the mean and the total say the same thing*. The following excerpt is from the discussion on descriptive analysis.

> PMT12. The first has a higher total weight. It is less in the second. A higher total weight means more juice.
>
> PMT10. It may not always give ten such.
>
> PMT14. So, it has no reliability. When the second 10 is bought, values such as 6, 6, 7, 7 can be obtained. Fresh Fruit Inc. is usually around 10.
>
> PMT10. Just looking at this, the first one has 126 kg. The other has 110 kg. However, the standard deviation of Tropical Days Inc. is much higher.
>
> PMT3. Exactly, if it comes, it will be very good, but if it does not come, it will not be good.
>
> PMT10. Very different, very dispersed.
>
> PMT3. The second is more reliable.

The total weight of the watermelons sold by Tropical Days Inc. was more. However, some watermelons were both low and high in weight. PMT10 proved it with a large standard deviation. The standard deviation was 4.13 for Tropical Days Inc. and 0.78 for Fresh Fruits Inc. Based on the sampling idea, PMT14 predicted the subsequent watermelon purchases, reasoning that less juice is likely due to lower weight. Since the second distribution was always around 10 kg, the probability of unexpected underweight watermelons would be less. This discourse of PMT14 gave rise to the idea that the total data set with high data points would be high. The Tropical Days Inc. data had higher values, and both the total and the mean were higher. This showed that the idea of "*the mean of the data set with high data points is high"* had become TAS. They already used it in the "batteries" activity in the previous week. Usually, the PMTs seemed to focus only on the mean. Therefore, they could be at the transitional reasoning level. However, in the "building arguments based on statistical

measures that are appropriate for the context-batteries" activity, the PMTs were interested in the magnitude of the data values based on the range. For example, they said that all values were greater than 80. That is to say, they took into account variability as well as center. However, they were at the procedural reasoning level because they did not explain how the data was distributed based on other aspects like density and skewness. The idea was presented here because it was about the characteristics of collections of data points, and the PMTs examined the data values individually rather than collectively.

As a second approach, they favored low variability rather than making the total or mean high. The MAD and range values supported the variability of the distributions. Put differently, they used the TAS idea that *the measures of variability support each other*. The MAD values were 3.59 and 0.66, respectively, calculated by PMT5 on horizontal bar graphs. The range of distributions was 11.5 for Tropical Days Inc. and 2.4 for Fresh Fruits Inc.

The PMTs had a dilemma in choosing a consistent distribution with low variability or an inconsistent distribution due to high variability but with a high mean. However, they based both approaches on valid reasons. The majority preferred Fresh Fruit Inc. to avoid risks, as it would be unreliable and highly variable. They all agreed that *the dispersed distribution is unreliable* (Claim 2 in Figure 4.36). This idea became TAS because they also used it in the "building arguments based on statistical measures that are appropriate for the context-batteries" activity in the fifth week. Regarding this idea, the PMTs were at the integrated process reasoning level as they evaluated data distribution by relating density and variability. In addition, they considered the center of data. Based on the context, choosing the distributor with a higher mean or total was also considered reasonable to obtain more watermelon juice. It can be seen in the dark colors (Warrant 1 and Claim 1) in the first part of Figure 4.36. PMT4 concluded this dichotomy by selecting the more reliable one as the difference of one kilogram between the means (12.61 kg and 11.08 kg) was small. The two approaches produced by the PMTs were summarized in the argumentation scheme in Figure 4.36.

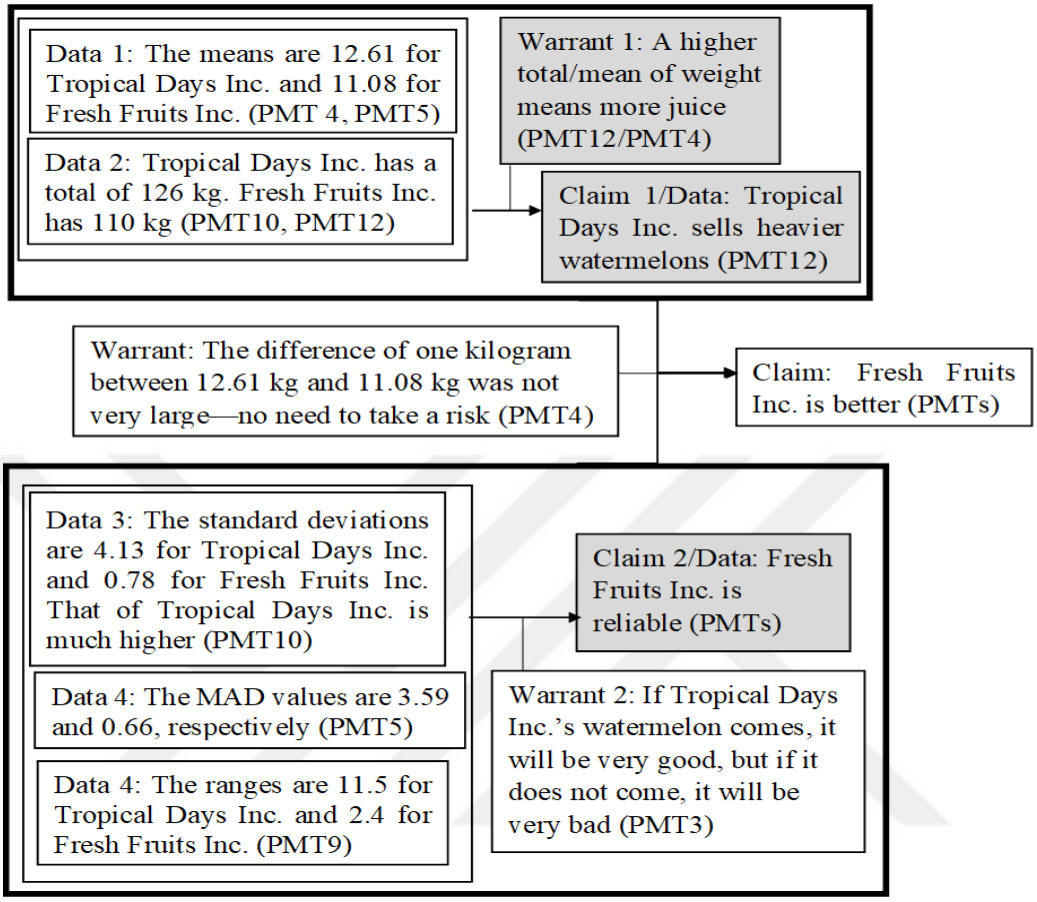


Figure 4.36. The argumentation scheme regarding the watermelon activity

The PMTs tended to evaluate central tendency and variability measures together because they agreed that *connecting the measures of center with the measures of variability leads to more reliable results*. However, they could only focus on one of them due to the data on the "building arguments based on statistical measures that are appropriate for the context-watermelons" activity. Since heavy watermelon was sought in this activity, the sum of the data values or the distribution with a high mean could be selected. Despite that, the PMTs insisted that it was more important that the data be close to each other, considering that there may be different samples. If data values of Fresh Fruits Inc. with low variability were concentrated in low values, they could change their preferences. Based on the data in the activity, they turned to the

distributor with low variability, which was reliable. It showed they used the TAS idea that *the dispersed distribution is unreliable*.

Week 7 started with the "understanding the shift in hill, clump, or cluster of data with equal sample size-resting heart rate" activity. The learning goal of the activity in the HLT was to describe the hill using different data organization and explain the shape and shifts in the hill, clump, or cluster of the data. In the first class session of the week, to familiarize the PMTs with the context, issues such as heart rate, how it is measured, factors affecting heart rate, heart rate during sleep, etc., were discussed. PMT10 described how the heart rate was measured on the left wrist. Accordingly, the PMTs measured their heart rate by counting the number of pulse beats in one minute. The instructor questioned the range of heart rates of normal healthy individuals before proceeding to interpret the data because PMT10 had interpreted that the heart rate should be inversely related to age. In other words, the older the person, the lower the heart rate. Since this range was investigated during the preparation phase of the teaching experiment, the participatory observer, I, stated that the normal heart rate was 60-100. Interestingly, PMT2's heart rate was 104. It was called an outlier because it moved away from the stated range and majority. The data collected from the class was represented in the dot plot by PMT7 in TinkerPlots.

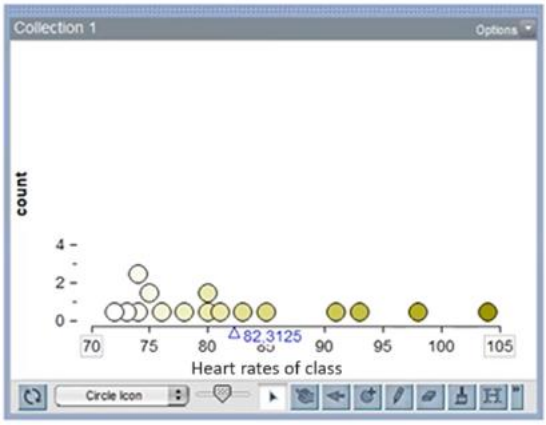


Figure 4.37. The heart rate distribution of the class represented by PMT7 in TinkerPlots

The following class dialog appeared when the instructor asked to interpret the graph.

> PMT10. The data ranges from 72 to 104.
>
> PMT4. The average is around 80.
>
> Instructor. How about the variability?
>
> PMT4. Variability seems too high.
>
> PMT2. Right-skewed.
>
> Instructor. Why?
>
> PMT2. Because the majority is always like this, between 70 and 85. The data will accumulate there, but since I wrote 104, it was pulled from the right.

When interpreting the data, the range and the mean were first mentioned. When the instructor emphasized variability, PMT2's focus shifted to the shape of the distribution. Since his data was separated from the majority, he could explain how the data was distributed. His interest in his data made it easier to reason about the shape of the distribution. He identified the distribution as right-skewed because the majority was on the left, and there was a shrinking to the right. The distribution represented the case in the heart rate of the class. Also, PMT2 had a prediction for the other cases. His comment about where the data cumulated could also be considered for larger sample sizes. Afterward, the instructor introduced the actual case to be investigated.

Two health organizations wanted to research the effect of smoking on the heart rate of young adults. They were selected randomly from 168 people who were 25 years old and had similar weights, health conditions, and eating and exercise habits. Half of them had been smoking since they were 16 years old. The other 84 people never smoke. When asked to comment on two dot plots containing the resting heart rate data of the people, the PMTs first tended to the shape of the distribution. The heart rate distribution of non-smokers was interpreted as right-skewed, whereas that of smokers was close to normal distribution. PMT9 immediately inferred that smokers had a higher heart rate because their pulses were more distributed to the right. Since they had the same range and number of data, this conclusion could be considered

statistically valid. She decided based on the shape of the distributions with the same range from 40 to 130. Following, when the instructor asked them to interpret the similarities and differences of the graphs, PMT6 stated that the clump was between 70-80 for non-smokers and between 100-110 for smokers. The PMTs were invited to work in small groups to draw a quantitative concrete conclusion on these interpretations made by looking at the overall appearance of the graph.

Comparing the means of the distributions was a common solution method for all small groups. They concluded that smoking increases resting heart rate because smokers have a higher mean heart rate that was calculated in TinkerPlots. The mean of the non-smokers' heart rate was 75.22 and 84.58 for smokers. To provide further evidence for this conclusion, the PMTs used different approaches. They partitioned the data into different parts with the divider features of TinkerPlots. Groups 2 and 3 partitioned the data into equal-width intervals with "equal width dividers" in TinkerPlots, as in Figure 4.38. The PMTs compared the normal heart rate range of 60-100, which was told at the beginning of the class session. The following excerpt was from the whole class discussion.

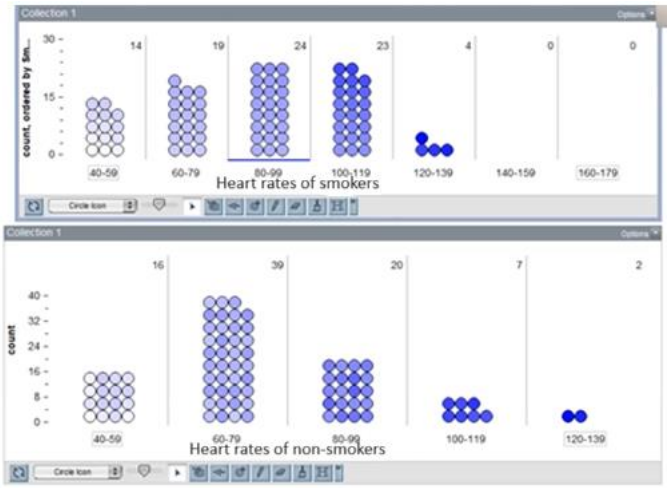


Figure 4.38. The heart rate graphs partitioned into fixed-width intervals in TinkerPlots by Group 2

PMT12. You know, between 60 and 100 is normal. Non-smokers have fewer values greater than 100, nine. Smokers have more values greater than 100; I counted 23 (27 in the graph). That is why it affects.

PMT9. For example, at 130, non-smokers have one value at the extremes while the other has almost two values. Likewise, at 40, there is one non-smoker, while those who smoke are piled up higher.

PMT12. We saw that non-smokers were concentrated between 60 and 79. The smokers were spread over a wider range.

PMT11. Like Group 2, our group (Group 3) also compared values higher than 100. Differently, we noticed that they were both distributed around the mean.

The groups that deliberately partitioned the data into equal-width intervals compared the frequencies of heart rates greater than 100. Smokers had a higher number of people with a heart rate of more than 100 heart rates (27>9), so the inference was against smoking. As a matter of fact, they compared the number of data points falling above a cut-point based on the context of the heart rate. That is to say, they performed additive reasoning with the absolute frequencies at the top of the distribution. They could have reached a different conclusion in other parts of the distribution. Fortunately, every part of the distribution supported their conclusion. For example, comparing the frequencies at the extreme values, PMT9 also showed that there were more non-smokers at the lower values. The intervals with the highest frequencies in the distributions were also compared. As PMT12 explained, the highest frequency in smokers was concentrated in two intervals, 80-99 and 100-119, while in non-smokers, the data was clustered between 60-79. Since the mean values were also calculated in the graphs, PMT11 claimed that the mean was expected within these ranges. This claim, which was introduced in this case, became a TAS idea because the idea that "*the data is concentrated where the mean is"* was used repeatedly in subsequent learning activities. While reasoning about the location of the mean in terms of the density, the PMTs were interested in how the data was distributed. Therefore, they were at the integrated process reasoning level.

Especially when the data related to the activities were introduced, the mean was estimated according to the majority's position by looking at the shape of the distribution without any calculation or analysis. If the data were clustered at low

values, the mean of the distribution was considered low or vice versa. The validity of this idea was later proven by calculating the mean of the data in TinkerPlots. Alternatively, the majority of the distribution, instead of the average, represented the distribution. Because the average, especially the mean, was used to represent data from the beginning of the teaching experiment. Regarding this, the PMTs could not only focus on a single value representing the data but were also interested in intervals in which the majority was located. It showed they went beyond reasoning based on a single value when making decisions about these distributions. In the "understanding the shifts in hill, clump, or cluster of data with equal sample size-resting heart rate" activity, comparing the interquartile ranges that Group 1 made, although not formally, also set a precedent for the transition from single value comparison to interval comparison. The group's dialog in the class discussion was as follows:

> PMT10. We wanted to support the means and calculated MAD values. They are 13.68 for non-smokers and 17.79 for smokers.
>
> Instructor. What does it mean?
>
> PMT5. So, smokers do not seem to have a certain order; they can vary a lot.
>
> PMT13. Both have a maximum value of 130. There is something strange here. Smokers have not shifted to the right.
>
> PMT10. However, what you say is also valid for the value of 40.
>
> PMT2. It does not affect the people at the bottom and the top, but it affects the people in between. It increased the mean by ten.
>
> PMT3. So, it is only the in-between that changes. The bad ones are already bad; the good ones are still good.
>
> PMT13. We were looking at the middle from 25% to 75%.
>
> PMT10. Well, we were taking out the extreme cases. The middle quartiles remain. This middle part increased in the smokers.

Although the MAD value for variability inspired the group, they supported the difference between means, ten units, with the interquartile range. The middle 50% were compared qualitatively, assuming that smoking would not affect individuals at the extremes. It showed that they used the idea that, *unlike the range, the*

interquartile range reduces the effect of outliers/extreme values. The position shift of this part to the right was interpreted as smoking affecting the heart rate. Although they could not create the box plot for the data in TinkerPlots, they were able to come up with the underlying idea without being able to implement it explicitly. Their solutions were summarized in the argumentation scheme in Figure 4.39.

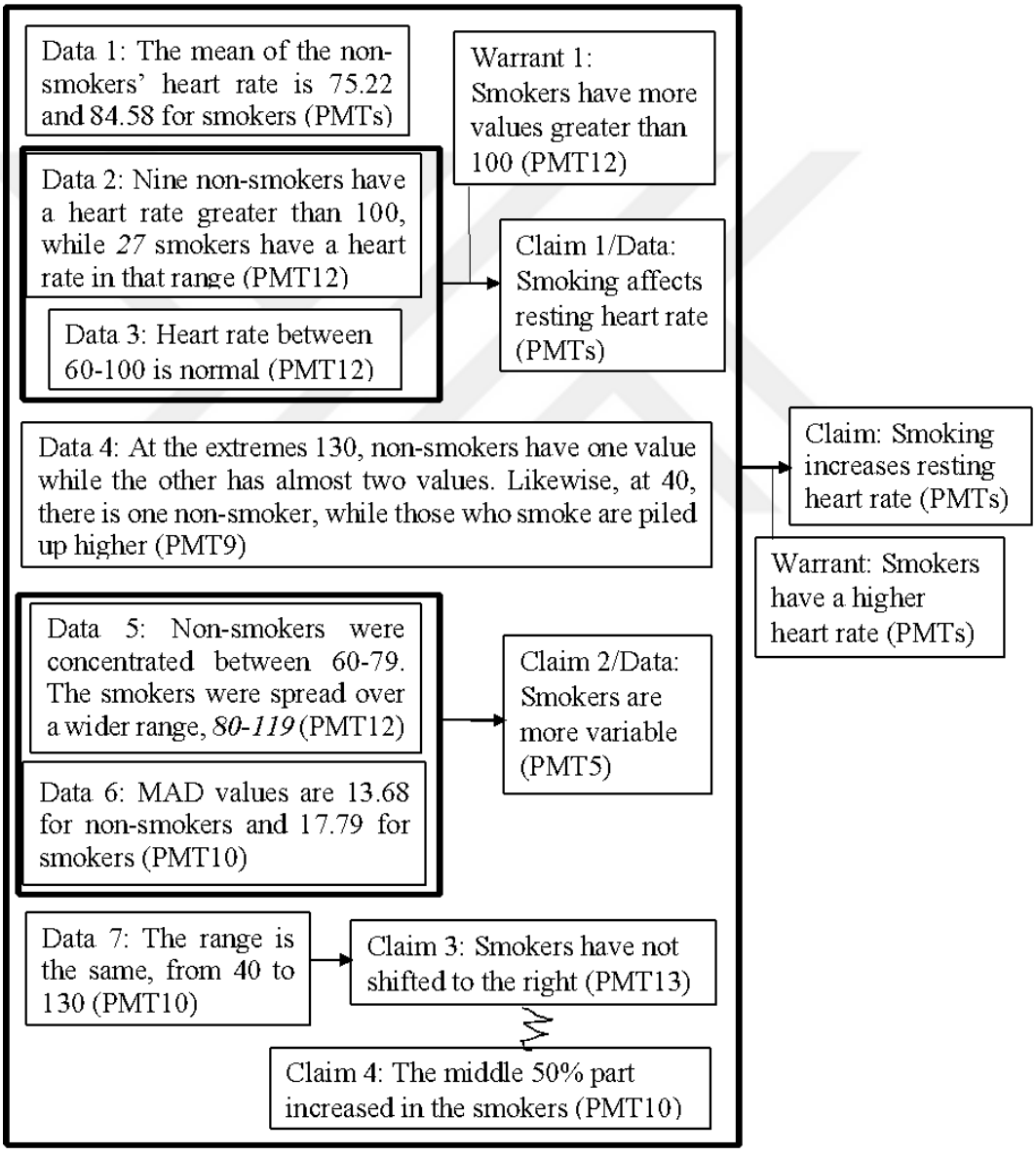


Figure 4.39. The argumentation scheme regarding the effect of smoking on resting heart rate

As can be seen from Figure 4.39, in summary, the PMTs compared distributions according to a single value or frequency in a specific interval. Although comparing frequencies at extreme values was unfavorable to smoking, the PMTs could not justify it concretely. When they partitioned the data into equal intervals or two on the cut-point, they were only interested in one part of the distribution when comparing the frequencies in the relevant intervals. They could have reached the opposite conclusion on the other side of the distribution. In addition, they could have made misinterpretations with additive reasoning if they compared absolute frequencies in the intervals of distributions with unequal sample sizes. At the beginning of the class session, they reached a conclusion based on the shapes of the distributions. Thus, there was no contradiction. Since the number of samples was the same in this activity, not using relative frequencies was not a problem. In this context, no matter which part of the distribution they were interested in, they were still interested in the collection of data points, not the distribution, since they were interpreting how many data values were in that interval. While they were interested in the collection of data values, these TAS ideas laid behind their reasoning: 1) *The sample varies between extreme cases based on the specific value/mean*, 2) *The mean of the data set with high data points is high*, 3) *The dispersed distribution is unreliable*, and 4) *The data is concentrated where the mean is*. In the first two ideas, the location of the individual data values was interpreted based on the spread and center aspects of the distribution. In the other two ideas, the PMTs focused on the spread and center aspects of the distribution as well as the density of the data. In the "generating different samples in weight context-hot air balloons" and "understanding the shifts in hill, clump, or cluster of data with equal sample size-resting heart rate" activities, the PMTs tended to reason about the density in the data when describing distributions, depending on data generation and increasing the number of samples. Consequently, data generation and increasing the number of samples helped to realize the density of the data. The density was associated with either center or spread aspects. The PMTs focused on particular data values or a particular part of the distribution rather than the whole data. For example, in the "understanding the shifts

in hill, clump, or cluster of data with equal sample size-resting heart rate" activity, the PMTs unconsciously studied the histogram in the background while comparing frequencies at fixed-width intervals. Depending on the larger sample sizes, identifying the density served as a transitional step in exploring qualitative characteristics of distribution rather than collection of data.

4.2.5 Classroom Statistical Practice 5: Exploring qualitative characteristics of distributions

Four ideas about the qualitative characteristics of distribution became TAS in the last two weeks of the teaching experiment, in which the PMTs worked on four learning activities (comparing distributions with equal and unequal sample sizes-backpacks1 and backpacks 2, recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia and ambulance). The TAS ideas were as follows.

The hill of the data set is in the bin with the most frequency
The intervals allocated to quartiles have different widths
The density is higher in the narrower intervals of the data partitioned into equal sizes
The location of the measures of center judges the shape of the distribution

The PMTs started to practice exploring qualitative characteristics of distributions partially in the "understanding the shifts in hill, clump, or cluster of data with equal sample size-resting heart rate" activity in which they reasoned the shapes of the distributions represented by the dot plots. However, they might have seen it as a collection of data because they focused on a part of the distribution or reached the same result in every part they structured. They needed to confirm that they viewed the distribution as a whole. Therefore, in continuing the teaching experiment, it was decided that the PMTs would study cases where they could reach different results in different parts of the distributions. The activities related to backpacks investigated in the later class sessions of the seventh week fit this situation.

In the second class session of the week, the "comparing distributions with equal unequal sample sizes-backpacks 1" activity was implemented. Before starting the activity, as in the other activities, to familiarize the PMTs with the context, they talked about the weight of their backpacks, the maximum weight they could carry without damage, the weight of student backpacks in schools, the items carried in the backpacks, the use of lockers in schools, etc. with the instructor's guiding questions. Based on the information that the body can only carry 15% of its weight on the back without damage, as the American Orthopedic Association reported, the PMTs calculated their maximum weight. Finally, the instructor introduced a district's research on whether or not to build lockers in schools.

There were two schools in the same district, one with lockers and one without. For the data, the weights of backpacks carried by 8th-grade students in the schools were measured in pounds. The data were collected randomly from 100 8th graders in each school. Assessing the situation, PMT5 stated that if students' backpacks were lighter in schools with lockers, the district would install lockers in other schools. The data was presented in dot plots. When the instructor asked PMTs what they could say without any statistics, measures, or calculations, they immediately sensed something contrary to the evaluation made by PMT5. They focused on the shape of the distributions when there were no calculations.

The distribution of students with lockers was interpreted as left-skewed, whereas that without lockers was close to normal distribution. PMT10 claimed that students with lockers carry heavier backpacks because the cluster shifted to the right. According to this reasoning, PMT5 declared that it increased the mean of the distribution of students with lockers. At lower values, there were also more students with lockers. One of them even did not carry a backpack at all. PMT4 explained that students with lockers carried backpacks of more varied weights because they were of every weight from 0 to 57 pounds. The other ranged from 9 to 57 pounds. Based on the ranges, it was claimed that the distributions had the same maximum value. The PMTs were invited to work in small groups to support their arguments with inscriptions.

Again, the first approach in all groups was to compare the means of the distributions. In TinkerPlots, they calculated the mean of students without lockers as 31.61 and the mean of students with lockers as 34.45. Therefore, they concluded there was no need to install lockers in schools to reduce the weight students carry in their backpacks. They partitioned the data with "equal width dividers" in TinkerPlots to support the conclusion. Different approaches were used to compare frequencies in the equal-width intervals. One of them was to interpret the difference in frequencies in these intervals, which was seen in Group 1 and Group 2. Below is an excerpt from the class discussion regarding Figure 4.40, including these differences.

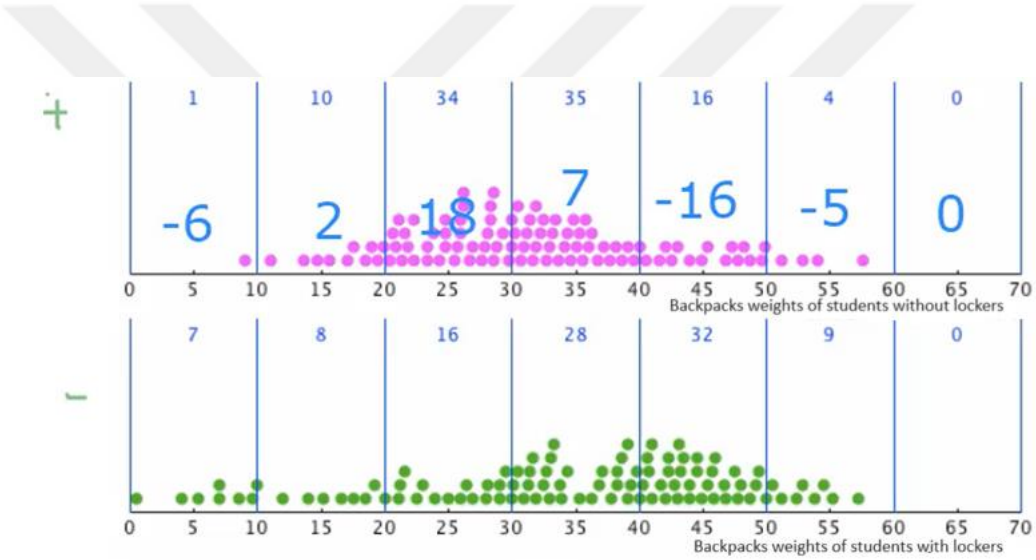


Figure 4.40. The partitioned backpack distributions into equal-width intervals

PMT3. When we assign plus and minus to the distributions, the logic is the same as the mean. The students with lockers carry heavier backpacks.

PMT1. +27 and -27 are reached by comparing the signs. The result is 0. I was undecided as the result went for any sign.

Instructor. What can you say about the intervals?

PMT13. Students with a locker carry less weighted backpacks, between 0 and 10 pounds.

Instructor. So?

PMT2. Since it was -6 in the first interval, more students had lockers.

Instructor. What about between 10 and 20 pounds?

PMT10. The students without lockers are more.

Instructor. The third interval?

PMT10. Again, the students were without lockers.

Instructor. The fourth interval?

PMT3. Those without lockers.

Instructor. In which interval/intervals do students carry heavier backpacks?

PMT10. Fifth and sixth intervals.

Instructor. Which students are more in these intervals?

PMT10. Students with lockers. We assigned minus for the students with lockers. The result is also negative, so there are more students with lockers.

PMT11. Looking at the first interval, you can see the positive effect of using a locker. However, in the last interval, it has a negative effect. Students carry too much weight even though they have a locker. Alternatively, those who carry less weight are the ones who have a locker.

As in the "building arguments based on statistical measures that are appropriate for the context-batteries" activity in Week 5, the PMTs tried to identify which distribution concentrated in high values according to the sign assigned. However, when the written values were added, zero was obtained. This meant that both distributions were equal in terms of the mean or total for the PMTs. They did not conclude according to the mean or total. In this approach, the data values were compared, not the frequencies in the intervals. In other words, it was not the case that the totals or means of the data values in the distributions were equal to each other. Data values in the two distributions were not matched, but the intervals overlapped. More specifically, when interpreting the histogram, it was necessary to interpret how many data values fell in that interval, not the magnitude of individual data points in equal-width intervals. This interpretation showed that PMT1 could not see the distribution as a whole; she was still interested in the collection of data points.

The instructor asked them to compare the intervals one by one to emphasize that the frequencies in the intervals, not data values, were being compared. The PMTs decided which distribution to choose according to the sign of the frequency

differences in the intervals. Because students with lockers outnumbered those without lockers at both the beginning and the end of the distribution, PMT11 experienced a conflict. Group 2 also noticed that the same distribution had more values at both ends. They could relate this to variability since they calculated the standard deviation, 10.11 for students without lockers and 13.16 for students with lockers. PMT12 explained, "It means that students with lockers carry underweight or overweight." To avoid indecision due to low values, they compared the frequencies in the intervals covering higher weight values based on context. PMT14 took a cut-point of 40 in TinkerPlots and concluded that more students had lockers above 40 through the "counts" button of TinkerPlots. There were 41 weight values in students with lockers and 20 in students without lockers. In this range, there were more than twice as many students carrying backpacks as those without lockers.

In the class discussion, the clusters in the distribution partitioned into equal-width intervals in TinkerPlots were compared. The first clusters compared were the intervals with the hill/peak, i.e., the intervals with the highest frequency. It led to the idea that *the hill of the data set is in the bin with the most frequency*. In the partitioning of Group 1, these intervals were 27.0-35.9 for students without lockers and 36.0-44.9 for students with lockers (Figure 4.41a). In Group 2, they were 30.0-39.9 and 40.0-49.9, respectively (Figure 4.41b). The frequencies in these intervals were 36 and 31 for the first partitioning and 35 and 32 for the second. Even though the width of the interval was doubled, there was not much difference in the frequencies on the hill. PMTs did not discuss this relationship between the ratio of the interval widths and frequency. They only compared the frequencies as more or less.

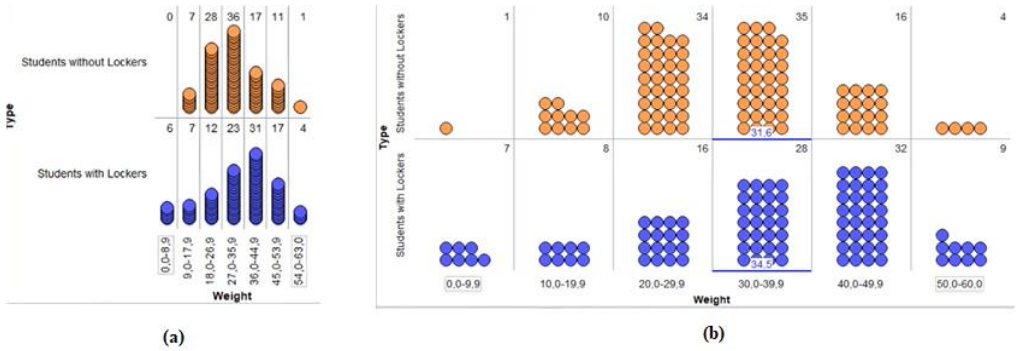


Figure 4.41. The partitioned backpack 1 distribution of Group 1 (a) and Group 2 (b) in TinkerPlots

The other clusters were compared to the intervals adjacent to the interval containing the mean due to partitioning. In the partitioning of Group 3, the mean was in the same intervals, 27.0-35.9 (Figure 4.42), but not in the bin with the most frequency. They only compared the frequencies in the intervals to the left and the right in the class discussion. Although they did not share it in the class, they started to comment on the qualitative characteristics of the distribution in the group work, as seen from the excerpt below.

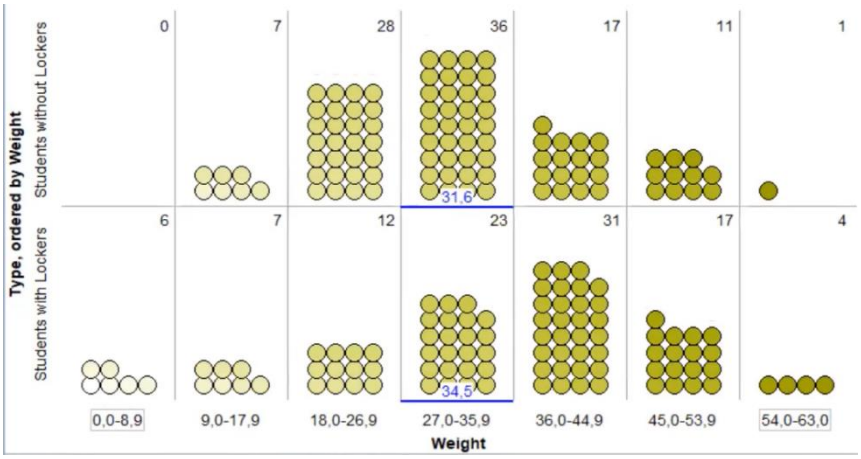


Figure 4.42. The partitioned backpack 1 distribution of Group 3 in TinkerPlots

PMT6. Very interesting. Students with lockers are more in the 36.0-44.9.

PMT3. This means that students should not use the lockers.

PMT8. The means are very close to each other; there are three differences. However, when we look at the distribution, students with lockers carry more weight.

PMT11. It originates from the three intervals above the mean.

PMT6. We can look at the before and after the 27.0-35.9 interval. If we keep this interval constant, the two graphs between 18.0 and 44.9 seem symmetrical. 18.0-26.9 is more at the top, and 36.0-44.9 is more at the bottom. 36.0-44.9 at the top and 18.0-26.9 at the bottom is less.

PMT11. For students without lockers, it is concentrated between 18.0 and 26.9. It is more dispersed for students with lockers divided into three.

PMT7. Again, the students with lockers are more at lower weights. At the same time, 35 students without lockers carry below 27.0-35.9, and 25 with lockers.

PMT11. It is just the opposite.

When comparing the bins with the most frequency in this partitioning, those of students with lockers are in the higher interval, 36.0-44.9. Comparing these common bins in all small groups turned the idea that *the hill of the data set is in the bin with the most frequency* into TAS because no one in the class objected to it, and it was used repeatedly in the following class sessions. Regarding this idea, the PMTs were at the integrated process reasoning level, as when capturing the hill in more density, they considered frequencies in other intervals. In this context, they looked at the whole by investigating the spread and density.

Based on the idea that *the mean of the data set with high data points is high*, PMT11 could relate the difference in their means, around three, to the intervals to the right of the graphs. They concluded that the students with lockers did not use the lockers. However, they reached different conclusions when comparing the graphs' upper and lower three intervals separately. In the class discussion above, it was mentioned that PMT11 experienced this conflict. When we disregarded the mid-interval (27.0-35.9) as a constant data value and compared both distributions from the top and bottom, students with lockers were more concentrated above this value. However, the number of students with lockers who carried weight below this value is not relatively small. It was because students with lockers were also more distributed at the bottom of the graph, as captured by PMT11.

A consistent result was obtained when a similar comparison was made in the middle three intervals, from 18.0 to 44.9, instead of the whole distribution. As PMT6 stated, students with lockers tended to carry heavier weights because they were concentrated in 36.0-44.9, which was the bin, including the hill of the distribution. When they took out the intervals at the extremes and compared them to the middle, they found that the result matched what they had interpreted from the shape of the distribution at the beginning of the activity. That is, students with lockers shifted to the right, where the weights were higher. PMT3, who shared the solutions of Group 1 in the class discussion, justified this result with the box plot they created by trial and error with the "Hats" button of TinkerPlots.

The box plot was part of the learning goal of the activities involving unequal sample size distributions in the HLT. However, Group 1 discovered the box plot earlier in TinkerPlots. Thereupon, it was aimed to closely look over the box plot, which was encountered for the first time during the teaching experiment process. In the last class session of the seventh week, the distributions partitioned into four groups with equal sizes in Figure 4.43 were discussed. The following excerpt was from the class discussion.

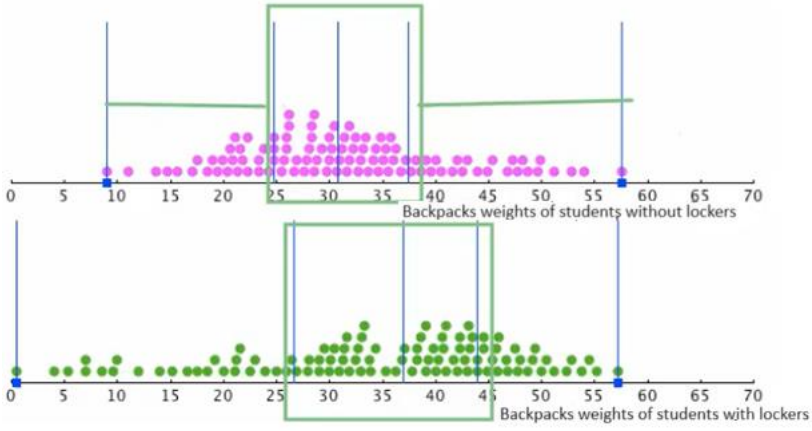


Figure 4.43. The graph is partitioned into four groups with equal sizes

Instructor. What is the difference here?

PMT10. We divided the intervals equally before. It is not equal in this graph.

Instructor. What can you say about this graph?

PMT5. The number of students in the intervals is equal.

Instructor. Can you explain it in detail?

PMT3. There were 100 students. They were divided 25-25. So, there are 25 students in each range. For example, in the first one, there are 25 students from 10 to 25. There are 25 students from 25 to 31 and so on.

Instructor. What is the difference?

PMT3. The number of students is the same, but the intervals are different.

PMT2. For example, 50 students over 37 are in the graph with lockers. On the other, it is over 31. Since there is a 6-pound increase in 50 students, we can say that students with lockers carry heavier backpacks.

Instructor. What else?

PMT12. We can see that those who had lockers carried various backpacks of less weight. The graph also has a wider range, so the number of students is the same as the one above. There is an inverse relationship.

PMT13. In the above graph, the line corresponding to 25 is Q1, Q3 is between 35 and 40, and Q2 is in the middle. Extreme values at the edges.

PMT2. When we remove extreme values, the difference between Q1 and Q3 remains. Half of the students carry between 25 and 37 weights in schools without lockers. However, half of the students with lockers carry backpacks between 27 and 43. Even when we remove the extreme values, the remaining half of the students with lockers carry more. That is why the lockers should not be installed.

Although there was a box plot among the solutions of the PMTs of Group 1, it was just shown without interpretation. The discussion in the last class session of the week started with their solutions to make sense of the idea underlying the box plot. The PMTs structured the data by partitioning it into equal-width intervals in the previous class session. However, in the above visualization, the data were partitioned into different-width intervals with equal sizes. The data was divided evenly to get equal data values in each interval. Since there were four equal groups, the intervals can also be called quartiles. PMT13 used “Qs” for quartiles with formal language. The difference between the two data configurations was quickly recognized due to the different widths of the intervals in the partitioning. While the intervals in the

partitioning of the PMTs were of equal width, and the frequency in the interval was different in the previous session, the frequencies in the intervals in the above partitioning were equal, and the widths of the intervals were different. When PMT3 said, "The number of students is the same, but the intervals are different," she talked about the equal number of students in intervals of different widths. That means *the intervals allocated to quartiles have different widths*. Since nobody disputed the idea that the intervals allocated to quartiles have different widths, it became TAS. Since the PMTs were aware of the varied spreadoutness in intervals with different widths, they were at the integrated process reasoning level based on a complete understanding of the distribution. For example, PMT12 interpreted the quartiles in terms of the width. She explained that the weights in the first quartile were more varied than in the second quartile because the first quartile had a wider range of lower weights related to the backpack weights carried by students with lockers. She was aware that both quartiles had an equal number of data values but different widths. In other words, *the density was higher in the narrower intervals of the data partitioned into equal sizes*. Although PMT12 expressed the inverse relationship, it had yet to be explicitly discussed, while PMTs compared different parts of the distributions with equal sample sizes.

Since the sample size of the distributions was 100, interpreting the quartiles was equivalent to simply adding a percentage notion to the number. For example, as in the explanation of PMT3, 25 students carried backpacks of 10 to 25 pounds in the students without lockers, and 50 students had backpacks of up to 31 pounds. The number of students in these intervals corresponded to 25% and 50% of the data, respectively. In his interpretations, PMT2 first compared the last two quartiles and then the middle two. In both cases, half of the students were compared. While the values he stated in the first interpretation were the medians of the data (37 and 31), in the second one, he compared the interquartile ranges, 25 to 37 and 27 to 43.

The relationship between interval width and density for data partitioned into equal sizes emerged explicitly when comparing distributions with different sample sizes. The sample sizes were differentiated in the other activity related to backpacks, which

was in the same context as the previous activity. There were 100 students without lockers and 199 students with lockers. Namely, the number of students with lockers was doubled in the follow-up activity. As there was no time left in the last class session of the week, PMTs shared their ideas without doing small group work. The instructor's questions were critical for understanding the effect of the sample sizes of the distributions on making a reasonable decision. Firstly, the instructor asked whether they could get the result by partitioning the data into equal intervals. Although the class objected due to the different number of samples, PMT12 suggested proportioning the frequencies in each interval. The class discussion continued as follows.

> Instructor. Why can't we compare frequencies in equal intervals?
>
> PMT10. Because there are always more students at the bottom.
>
> PMT12. We can look at the ratio by dividing each interval by the number of samples.
>
> Instructor. Let us compare the frequencies. There is a partition into equal-width intervals (Figure 4.44).
>
> PMT12. 1, 10, 34, and so on. We need to divide the above by 100 and the below by 200. That way, we can make a comparison.
>
> PMT3. It is almost double. 100 to 199.

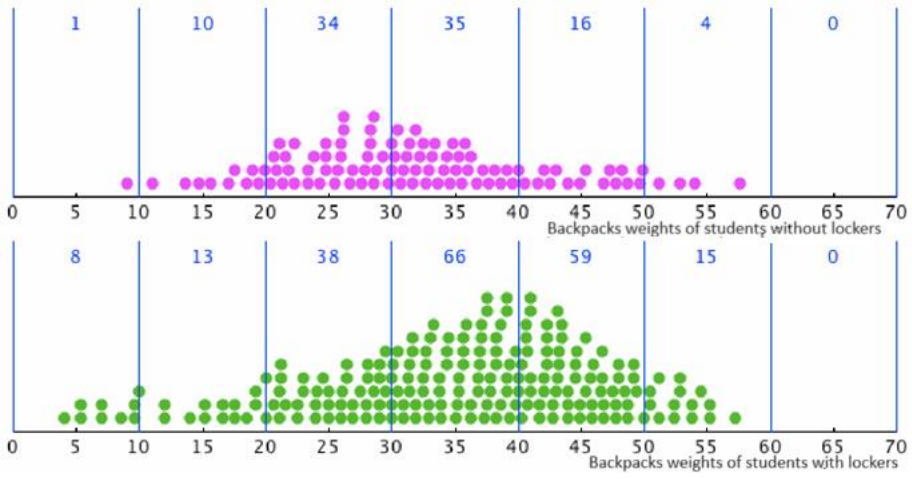


Figure 4.44. The Backpack 2's data partitioned into equal-width intervals

This approach suggested comparing relative frequencies by proportioning the number of samples instead of comparing absolute frequencies. PMT12 proposed to divide absolute frequencies in each interval by the sample size. After this idea to equalize the sample size, PMT3 replied by multiplying the absolute frequencies of the students without lockers by two since the ratio between the sample sizes of the distributions was doubled. They could compare the frequencies obtained in specific intervals or part of the distribution as the sample sizes were equal. However, the ideas put forward were not implemented in equal-width intervals. Instead, the relative frequencies were compared according to 30 in Figure 4.45.

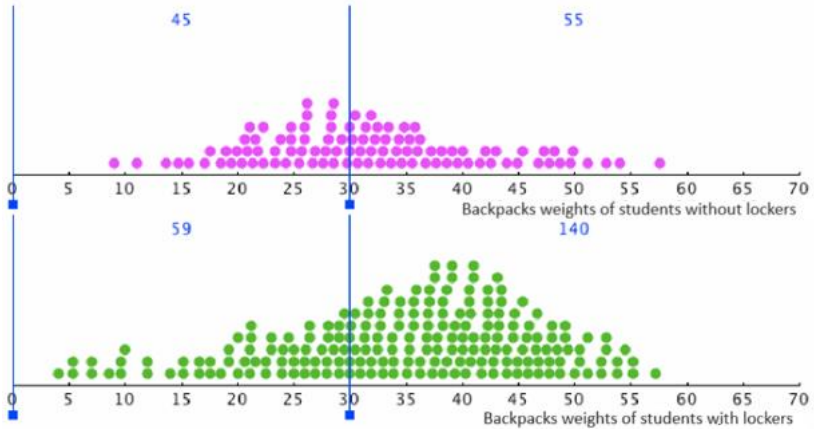


Figure 4.45. The inscription with 30 pounds of the cut-point

> PMT2. There are more students, both below and above 30, in schools with lockers. 140 to 55 above 30, 59 to 45 below 30. There is a conflict.
>
> Instructor. There is a dilemma. Why?
>
> PMT5. However, the sample size is different.
>
> PMT3. There is no conflict if we multiply the 45 and 55 above by two. It is more at low values at the top and more at high values at the bottom.
>
> PMT13. Alternatively, we could take their percentages.

In the dialogue above, when the absolute frequencies in both parts of the distributions were compared, a dilemma was obtained. Since there were more data values in the distribution of students without lockers, there were more frequencies in the low and

high values according to the specific cut point. Regarding the conflict, PMT2 was immediately challenged with a counterexample of values above 30. This showed that PMT2 could evaluate the distribution as a whole because he dealt with all parts of the distribution. In this case, the difference in sample size required the PMTs to reason proportionally. The challenge was eliminated when the frequencies were proportioned to eliminate the difference in sample sizes. When the frequencies of the above distribution were multiplied by two with the suggestion of PMT3, there were more students without lockers at low values (90>59), while there were more students with lockers at higher values (140>110). This comparison confirmed that students with lockers carry heavier backpacks. Although calculating the percentage was also proposed to equalize the sample sizes of the distributions, the calculation was not completed. Instead of the percentages in the intervals, the frequencies were also used in the distributions partitioned into four equal sizes in TinkerPlots.

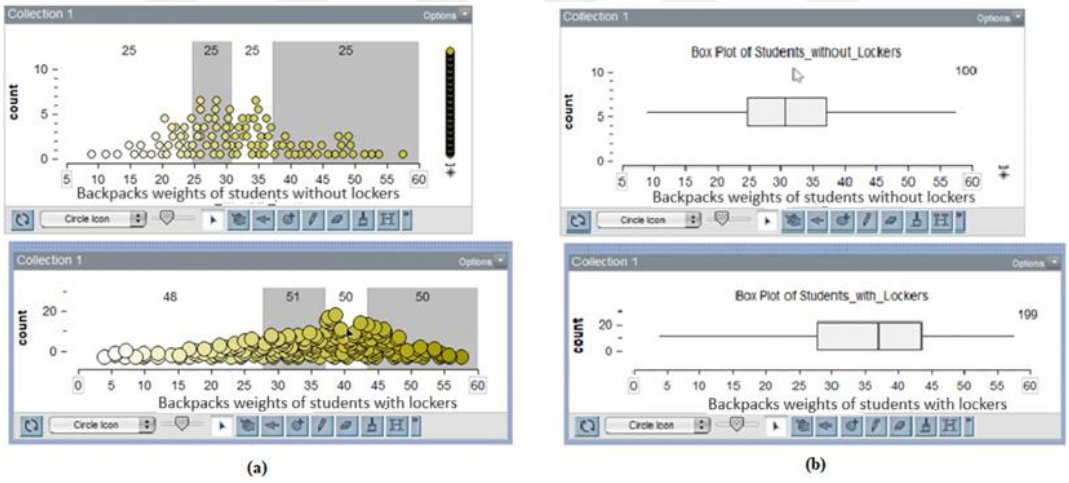


Figure 4.46. The data partitioned into four equal sizes (a) and the box plots (b) for the backpacks 2 in TinkerPlots

As can be seen from the figure, the partition in Figure 4.46a corresponded to the box plot in Figure 4.46b. That is, the data was divided into four equal groups by using "equal count dividers" of TinkerPlots. The boundary values of the intervals were the

five values of the box plot. Regarding the partitioning, the meanings of the box and whiskers on the box plot were discussed, regardless of the context.

> Instructor. Can you tell me the Q values in the box plot above?
>
> PMT10. Initial line.
>
> PMT3. The place with the 25.
>
> PMT12. Q2, median, is in the middle.
>
> Instructor. What else?
>
> PMT11-12. The values at the ends of the stems are also the minimum and maximum values (the expressions of Qs, min., and max. were written on box plots).
>
> Instructor. What about the box?
>
> PMT2. Q3-Q1
>
> PMT10. Aaa! Interquartile range.
>
> PMT1. The quartiles correspond to 25%, 50%, and 75% of the data.
>
> PMT10. In the graph below, Q2 is closer to Q3. In the graph above, it is more or less in the middle.
>
> Instructor. What kind of interpretation can you make?
>
> PMT10. The bottom is left-skewed.
>
> PMT13. 25% of the data fit in the narrower box (3rd interval of the bottom). Most of it piled up on the right and pulled from the left.

When interpreting the box plot, the PMTs knew from the previous activity that the quartiles (Q1, Q2, and Q3) correspond to 25%, 50%, and 75% of the data, respectively. PMT1 also repeated this partition. Here, different reasoning was used about the interquartile range and the distribution shape according to the median's position. All PMTs knew that the formula of interquartile range was Q3-Q1. When these values were labeled on the box plot, it was realized that the box stood for the interquartile range. The interval width was narrower once the median was closer to one of the other quartiles. Correspondingly, the PMTs used the TAS idea that *the intervals allocated to quartiles have different widths* without question.

In the case of students with lockers, Q2 was closer to Q3. That is, the third interval was narrower than the others. However, as PMT13 pointed out, it had the same amount of data, 25% of the data, as in all the other intervals. Even though she did not explicitly say density, the expression “pilling up” showed that she was using the idea that *the density is higher in the narrower intervals of the data partitioned into equal sizes*. She reasoned at the integrated process level because she understood how the data was distributed depending on the density. She also provided a warrant for the distribution's shape based on the median’s location. It was evidence of the TAS idea that *the location of the measures of center judges the shape of the distribution*. Similarly, this idea depends on the distribution shape and integration of center, spread, density, and skewness aspects. It provided a way to make sense of how the data was distributed. Therefore, reasoning based on this idea was at the integrated process level. The reasoning about the shape of the distribution based on the comparison of the mean, median, and mode in Week 2 and Week 6 was performed here only with the median. The distribution was called left-skewed because it was a stacked distribution on the right and a more spread distribution on the left. Indeed, when PMTs interpreted the distributions of the first activity related to backpacks without doing any numerical operations at the beginning of the week, they labeled the students with lockers as left-skewed. In the case of the second activity of the backpacks, the distribution remained almost the same, and the sample size was approximately doubled.

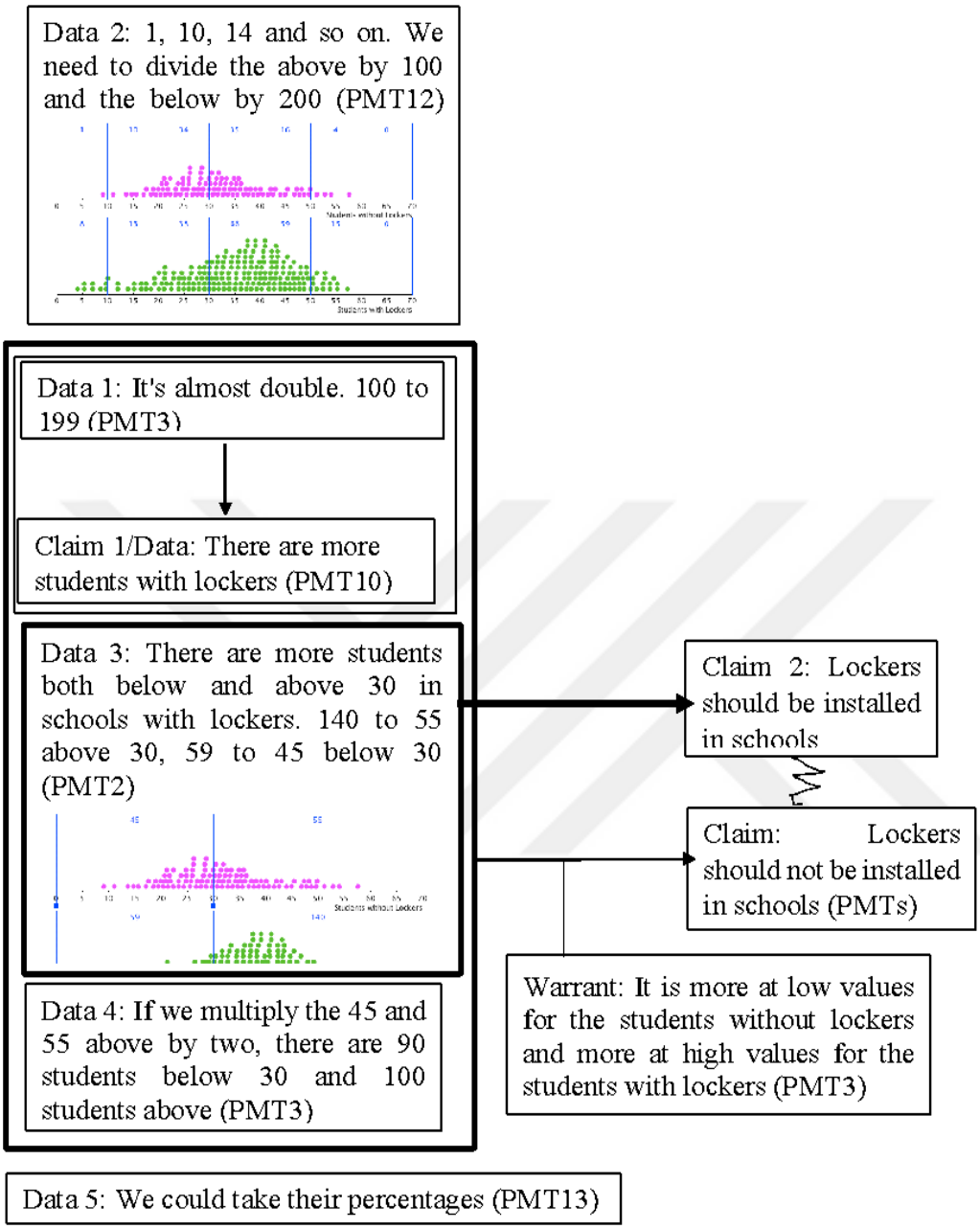


Figure 4.47. The argumentation scheme regarding the effect of lockers on the weights carried in the backpacks

Although the PMTs described the data distribution when interpreting the box plots, they did not make a contextual interpretation. Their arguments (Figure 4.47) for the "comparing distributions with unequal sample sizes-backpacks 2" activity were limited to partitioning the data into equal-width intervals and cut-point inscriptions.

The use of proportional reasoning was envisaged due to different sample sizes for both partitions. In this context, the result was achieved by comparing the relative frequencies concerning the reference point, 30 pounds, in the argumentation. The opposite result was obtained when additive reasoning was used, called a dilemma or conflict. In summary, PMTs needed to use proportional reasoning when the distributions' sample sizes differed to be compared. In the following activities of the teaching experiment, proportional reasoning was at the forefront as the distributions with unequal sample sizes were compared. For example, the use of percentages was an indication of proportional reasoning. In the box plot where the data was divided into 25th percentiles, proportional reasoning was based on percentages. However, despite the instructor's question about the quartiles in the dot plot and box plots representing backpack weights, the interpretation was not accomplished for the "comparing distributions with unequal sample sizes-backpack-2" activity.

How the PMTs interpreted the box and dot plots in a context was revealed by examining their writing on the reflections. For the reflection, the study of Kader and Mamer (2008) was assigned to the PMTs at the end of the seventh week. The assignment paper was related to the distribution concept, focusing on understanding the center and the spread. The paper consists of two parts. The first part discusses the meaning of mean and variability, the measures of variability, and their implementation on the soccer game data. The second part of the paper mentions the activity about remembering words from two lists given by a teacher while doing statistics. List A consists of 20 three-character words, and List B has nonwords. The activity seeks to answer the question of which lists people remember more easily. The paper contains student responses to the teacher's questions on the dot and box plots, as in Figure 4.48. Student responses on dot plots are related to comparing ranges, comparing frequencies or percentages above or below a cut-point, and comparing dispersions. When the teacher asked the same questions for the box plot as the dot plot, students compared quartiles, medians, and interquartile ranges with percentages or identified overlaps and differences in the box plots. For the box plot, the responses compared the data proportionally over the whole or part of the

distribution, whereas for the dot plot, frequencies and individual values were compared. To examine whether the PMTs could see the distribution as a whole, the assignment asked the PMTs to discuss the advantages and disadvantages of the dot and box plots and evaluate the student's responses based on the context of memorizing words. For example, Figure 4.48 shows the graphs and the sample students' responses to the teacher's first question. Based on the graphs, the question asks, "Did all students who studied list A score higher than all students who studied list B? Explain" (Kader & Mamer, 2008, pp. 41-42).

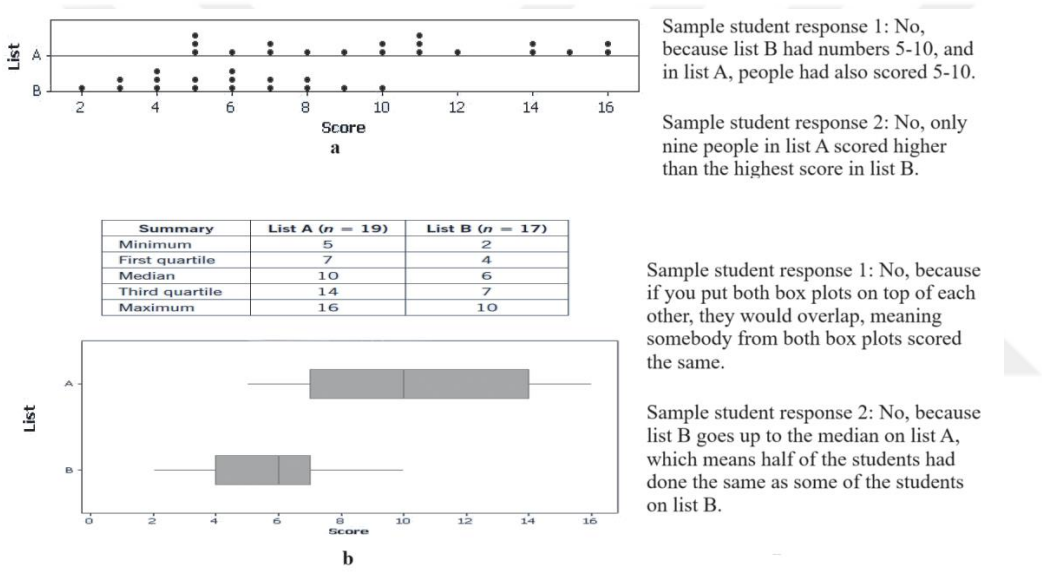


Figure 4.48. Dot (a) and box (b) plots and sample student responses for Question 1 regarding memorizing words in the last assignment

Only PMT5 analyzed students' responses one by one. Her analysis of student responses to the teacher's first question was as follows.

> Q.1: A student said no because List B had numbers 5–10; in List A, people had also scored 5–10. This is a good reasoning as a counterexample.
>
> The other said that no, only nine people in List A scored higher than the highest score in List B. This is a good reason. Only nine people scored higher than the highest score in List B, indicating that the remainder scored lower than the highest score in B (PMT5).

She explained the two students' reasoning for the question with counterexamples. Indeed, the first student falsified the conjecture in the question by stating that some values in both dot plots overlap in the range of 5-10. The second student reached a similar result with nine data values in the range of 11-16, which included only students studying in list A. PMT5 again explained it with the overlapping parts in the dot plots. She recognized that the distribution of students studying list B was below the nine values in the other dot plot. In other words, she explained the students' responses by considering all distributions. Although individual data values were expressed in student responses, PMT5 could evaluate them regarding the entire distribution. In the version of the same question for the box plot (Figure 4.48b), she used the terms in the box plot, such as range, quartile, and median, for the overlapping parts.

> Q.1: The student said no because if you put both box plots on top of each other, they would overlap, meaning somebody from both box plots scored the same. Here, the student discovered that the width of the box plot graph is the same as the range, and when s/he superimposes it, s/he finds that some data is the same.
>
> The other said that no because list B goes up to the median on list A, which means half of the students had done the same as some of the students on list B. Here, the student spoke over the quartiles represented by the box plot and commented before and after the median. This is good reasoning (PMT5).

Other PMTs generally mentioned that the students used statistical language, gave valid answers to the teacher's questions, and had advanced statistical reasoning. For example, PMT12 evaluated the responses as follows:

> Students attempt to interpret the data by looking at the frequencies, distribution of the data, and max-min values instead of directly calculating mean/median in dot plots. They identify and show the box plots' min/max values and quartiles. It shows that they have advanced literacy and statistical reasoning skills. Moreover, students give consistent answers to the questions asked by the teacher using statistical terminology (PMT12).

From the students' responses, they emphasized the difficulty of the interpretation of the box plot. PMT9 wrote, "Although the students could answer the questions meaningfully through the dot plot, it was much more complicated for them to make

this meaning through the box plot and to associate the summary number values with their meanings." She explained the possible reason for this as a five-number summary (minimum value, first quartile, median, third quartile, and maximum value) instead of the data itself. This difficulty was also expressed when comparing dot plots and box plots. All PMTs mentioned that all data values could be seen in the dot plot, but comments were made with percentages in the box plot. More specifically, PMT7 and PMT8 emphasized the necessity of proportional reasoning when comparing the percentages in the box plot. In addition, all PMTs noted that the box plot was more suitable for large sample sizes than the dot plot. They agreed that the box plot gave more information about how the data was spread. Three of the PMTs also associated the way the data was spread in the box plots with the shape of the distribution. For example, PMT10 wrote about the shape of the distribution as follows.

> The advantage of the box plot is that we can easily see the median of the distribution because the Q2 value is the median. The second advantage is that we can interpret graphics, whether they are expected, left-skewed, or right-skewed distributions. It is normal if Q2 is in the middle between Q1 and Q3. If Q2 is near Q3, it is left-skewed; if Q2 is near Q1, it is right-skewed (PMT10).

The PMTs, who reasoned the shape of the distribution in the box plot, interpreted the density according to the widths of the intervals in the data partitioned into four equal groups. That is, they used the idea that "*The density is higher in the narrower intervals of the data partitioned into equal sizes."* Furthermore, PMT10 and PMT12 interpreted the distribution's shape according to the median's location in the box plot. In this regard, they used the idea that "*The location of the measures of center judges the shape of the distribution*," which emerged in the class sessions. In particular, when reasoning about the shape of the distribution, they also considered other parts of the distribution. For example, the PMT10's interpretation of the distance of the median to other quartiles pointed to the rest of the distribution. In other words, the explanations about the box plots were oriented toward the whole distribution. However, in the dot plot, the emphasis was more on interpreting individual data values. The large sample size, the use of proportional reasoning, and

the comparison of percentages for the box plot in the explanations were all reasons to think beyond the data values to how the data was distributed. In addition to what was mentioned previously, PMT2 offered the idea of different sample sizes for the box plots. He wrote in his paper, "Students can easily compare two or more data even if the samples are not equal because the box plot enables students to compare with the percentages." He could judge that even with different sample sizes, it was possible to compare distributions by comparing percentages.

Only PMT2 mentioned comparing distributions with different sample sizes in the reflection papers. Comparing distributions with unequal sample sizes was one of the ultimate learning goals of the HLT. There was also a greater need to interpret the box plot in different contexts. To ensure communal learning in the class, the subsequent learning activities aimed to help the PMTs study distributions with large and unequal sample sizes. Therefore, two activities with unequal sample sizes were implemented in the last week of the teaching experiment. They were "recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia" and "recognizing the pattern in the data and supporting the pattern with different graphical representations-ambulance" activities. The anorexia data had a smaller sample size than the ambulance data.

The "recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia" activity required deciding which of the two methods of treating anorexia was better. To better understand the context before starting the activity, in the first-class session of the last week, eating disorders, types of diseases such as obesity, bulimia, and anorexia, their effects on the body, the PMTs' weight situations, and eating habits were discussed. After eating disorders were classified as excessive weight gain or inability to gain weight, anorexia was emphasized. The instructor defined anorexia as a lack or loss of appetite for food. The PMTs could deduce that weight gain was the primary goal of anorexia treatments. In the "recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia" activity, there were two treatments, traditional and family-based therapy, that served this goal. When the instructor

introduced the activity, the PMTs talked about the characteristics of the treatments. While traditional therapy aimed for the patient to check into a health institution and receive education or treatment, in family-based therapy, family members were also involved in the patient's treatment. A study had already been conducted in which patients were weighed at the end of a three-month treatment. The weight gain or loss of 117 patients in traditional therapy and 39 patients in family therapy was recorded. When the instructor asked what they understood from the situation, PMT8 immediately said that the sample sizes differed. PMT10 supported her by adding that the number of patients in the family-based therapy was less than the number of people who got traditional therapy. PMT2 added that one was three times the other. The PMTs used multiplicative reasoning while generating a solution to the statistical problem in small groups or class discussions. They focused on the reason for the differences between the numbers. They thought that the lower preference for family-based therapy might be due to the greater responsibility of family members.

The data in the "recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia" activity differed from those in the previous activities because the data values had direction (minus or plus) on the horizontal axis of the dot plots. The PMTs were asked what these graphs or dots meant. In addition to PMT10's interpretation of weight change, PMT5 interpreted the values. She said that if the value was positive, the person gained as much weight as that value at the end of three months, and if it was negative, the person lost as much weight as that value. PMT8 stressed that there was no change in those with 0. When asked to comment on two dot plots containing the weight change of the people without doing any calculations or comparisons with numbers, the following dialog was observed.

> PMT10. In traditional therapy, the tail stretches on the right side. There is a pile-up on the left side. So, it is a right-skewed. In family-based therapy, it looks close to normal, but there is no exact shape.
>
> PMT8. I think the below is a bit more left-skewed.
>
> Instructor. What kind of deduction can you make?

PMT8. There are fewer people losing weight than gaining weight in family-based therapy.

The class agreed that family-based therapy was more effective in weight gain. They hypothetically concluded without quantitative evidence. According to their reasoning, people gain or lose weight based on where they think the accumulation is. They worked in small groups to justify their conclusion with numerical data. Common solution methods in all group works were partitioning the data into equal-width intervals and using a cut-point of 0, interpreting the box plot, and mean/median comparison. The class discussion started with the solutions of Group 1. They first divided the data into four equal-width intervals in TinkerPlots. Since there was not much difference between the ranges in the distributions, they could make a one-to-one comparison. They preferred to compare percentages above 0. Group 1 presented two different representations in Figure 4.49, which they interpreted similarly.

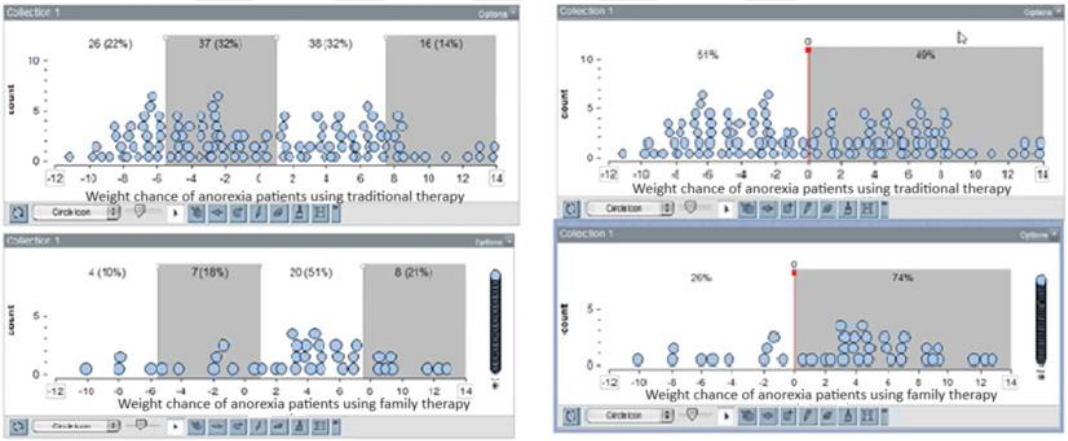


Figure 4.49. The anorexia distributions partitioned by Group 1 in TinkerPlots

PMT13. We need to make people gain weight, right?

PMT2. So, if I look at the ones above 0, the one below is over 72%. Above, 32+14 is more than 46% (Figure 4.49a).

PMT13. The cut-point method compared the numbers below and above the cut-point. Since their numbers are not equal, we can multiply the data of the graph below by three since the above is a multiple of three of the below and compare them.

PMT2. We can also look at the percentages to the left and to the right of it.

PMT9. It is 74% in family therapy and 49% in the other (Figure 4.49b).

As the main goal in the context was to gain weight, everyone in the class focused on positive values. Therefore, they engaged in values greater than 0. For example, in the dialog above from the work of Group 1, when they partitioned the data into four equal-width intervals, the boundaries of the intervals were not 0 (See Figure 4.49a). PMT2 summed the percentages in the last two intervals for this data structuring. They could interpret the exact values in the cut-point method. As said before the group work, all PMTs knew that the numbers of the samples were not equal in the distributions. Accordingly, Group 1 compared the percentages, even though they suggested using a factor of three in the distribution of family-based therapy. Group 3 performed this suggestion by rating the numbers in the intervals in Figure 4.50.

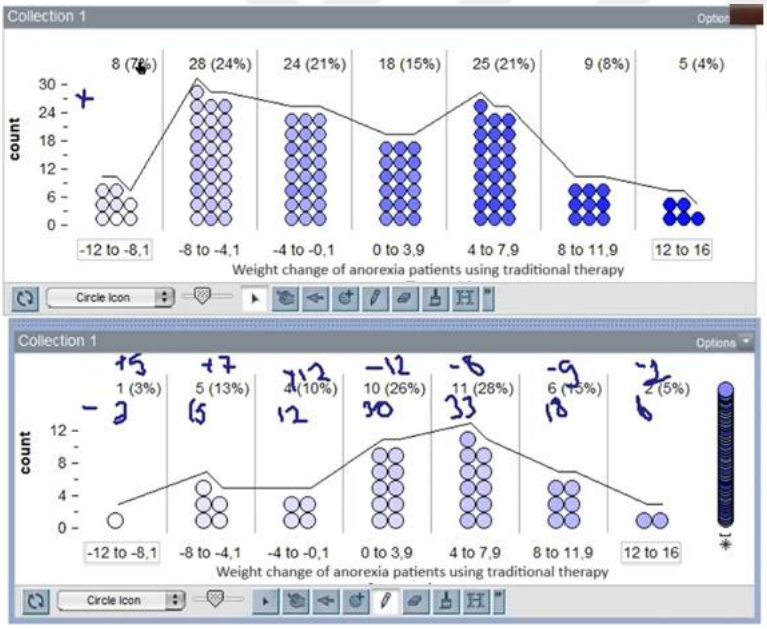


Figure 4.50. The anorexia distributions partitioned by Group 3 in TinkerPlots

In their group work, they compared the percentage values above and below 0 on the same partition in Figure 4.50 structured with equal width dividers in TinkerPlots. PMT5 summarized the solution, "In traditional therapy, 52% of patients lost weight, and 48% gained weight. However, 74% gained weight in family-based therapy, and

26% lost weight. Family-based therapy works better." As a different solution method, they used the ratio between the numbers (117/39) to equalize the distributions in terms of sample size. They assumed that the distribution would not change when they increased the sample size by proportional reasoning. In the family-based distribution with the small sample size, they multiplied the frequency in each interval by three and took the difference between the frequencies. In the group work, PMT3, who proposed this solution, had observed that values above 0 were always negative and values below 0 were always positive. This showed that she had considered every part of the distribution and interpreted the sign of values based on the context. However, in the class discussion, she explained that they had added the differences and yielded a negative result. The accurate result was 0. She had miscalculated in the second interval. Since the sum of the frequencies was equal to the number of samples, it was obvious that the result would be 0 since three times the number of samples of family-based therapy was equal to the number of samples of traditional therapy. In this approach, it was more important to know how much data rather than what the data was. According to the result of PMT3, they concluded in favor of family-based therapy as they assigned the minus. This method used in previous "building arguments based on statistical measures that are appropriate for the context-batteries" and "comparing distributions with equal sample sizes-backpacks-1" activities was similar to comparing the means. The mean of family-based therapy (3.13) was higher than traditional therapy (0.36). However, in the data partition, frequencies were compared rather than data. It allowed the PMTs to see the distribution as a whole.

Regarding a different partition of the data, Group 2 created a box plot in TinkerPlots by dividing the data into four equal counts. Groups 1 and 3 directly created the box plot in TinkerPlots and compared the quartiles. The simple one was to compare the medians in the box plots. In the class discussion, Group 2 justified that the median of family-based therapy (4) was higher than that of traditional therapy (0). In the group work, they also noticed the position of the median according to the other quartiles. For example, PMT10 in Group 2 said, "Q2 is close to Q3 at the below but

closer to Q1 at the above.” However, they could not mention the shape of the distribution. Except for Group 2, the others made inferences about the shape of the distribution according to the position of Q2 in the box plot. In the class discussion, the following dialog emerged based on the box blots in Figure 4.51 drawn by Group 1.

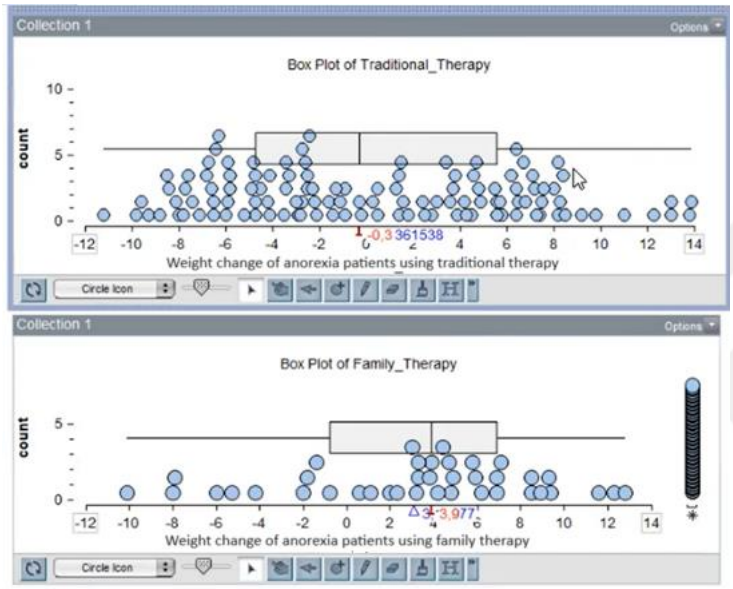


Figure 4.51. The box plots drawn by Group 1 in TinkerPlots for the activity related to anorexia

> PMT13. The measure of center we look at changes depending on whether the distribution is skewed or not. For example, if it is skewed, we use the median.
>
> Instructor. How do you decide whether these graphs are skewed or not?
>
> PMT13. According to the location of the median and the mean. For example, in traditional therapy, the median is smaller than the mean, so it is more concentrated on the left. On the contrary, the one below is left skewed because the median is larger.
>
> PMT10. We can also look at the width of the boxes. If Q2 is closer to Q1, there are more low values. There is density on the left. That is why extreme values remain on the right.

PMT2. The whiskers also give ideas. The one above is skewed to the right. Since they are all 25%, that is, there are a few values on the right. So, it is right-skewed. The one below is more spread to the left, so it is left-skewed.

PMT13. What if the median is exactly in the center of the box inside and the wishers are the same length?

PMT10. It is normal.

Instructor. Which therapy did you decide was effective in treating anorexia?

PMT12. In our group (Group 3), we decided that the above is right-skewed and the below one is left-skewed. In the right-skewed, the mean is larger than the median. However, in the one below, the median is bigger than the mean. This means that in family-based therapy, more than half of the people are above the mean. The mean is already positive, which means it is good because it is in the direction of weight gain.

In skewed distributions, the mean may not represent the actual data as it is affected by outliers. Based on the TAS idea that "*the mean is responsive to outliers*," PMT13 compared the median in graphs defined as skewed. In Week 3, the whole class first expressed that using the other measures instead of the mean in distributions with outliers was more appropriate. Group 2 had already compared medians. The instructor tried to orient the discussion towards the shape of the distribution. For the shape of the distributions, the PMTs used three different approaches, including the proposed solutions that favor family-based therapy. They were represented by dark colors in Figure 4.52 since they reflected the TAS ideas. The first one was to compare the magnitudes of the measures of center (Data 7). In descending order of magnitude, the distribution was skewed toward the mean. Underlying this was the fact that *the mean is responsive to outliers*. Regarding the measures of center, the mean was smaller than the median in family-based therapy, so it was a left-skewed distribution. In traditional therapy, the values were very close to each other. When the values were compared, the mean was slightly larger than the median, interpreted as right-skewed. In the second approach, the location of the median in the box was evaluated (Data 6). The median inclined to Q3 in the box produced a left-skewed distribution or vice versa. In the first two approaches, the PMTs used the TAS idea that "*the location of the measures of center judges the shape of the distribution.*" They reasoned at the integrated process level because they understood how the data was distributed.

Towards the end of the discussion, based on the location of the mean and the median, PMT12 in Group 3 inferred that more than half of the data was greater than the mean in family-based therapy. Since the median divides the data in half and the box plot divides the data into four equal groups of 25%, she could make sense of half of that data from the middle line, the median, in the box plot. Concerning the shape, the left-skewed distribution with higher values was better in the context. Since the mean was also positive, she contextually justified that family-based therapy provided more weight gain. Unlike measures of center, PMT2 interpreted the lengths of the whiskers, which was the last approach for the distribution shape (Data 8). Since the box plot has 25% of the data in each interval, longer whiskers are more dispersed. Conversely, the shorter ones are also denser. The whiskers in the box plots embody the TAS ideas; *the intervals allocated to quartiles have different widths*, and *the density is higher in the narrower intervals of the data partitioned into equal sizes*. Rather than individual data values, considering where the data was concentrated according to frequencies in certain intervals encouraged the PMTs to explore the qualitative characteristics of distributions. Expressly, reasoning about the shape provided to think about the distribution rather than the collections of the data while comparing the distributions.

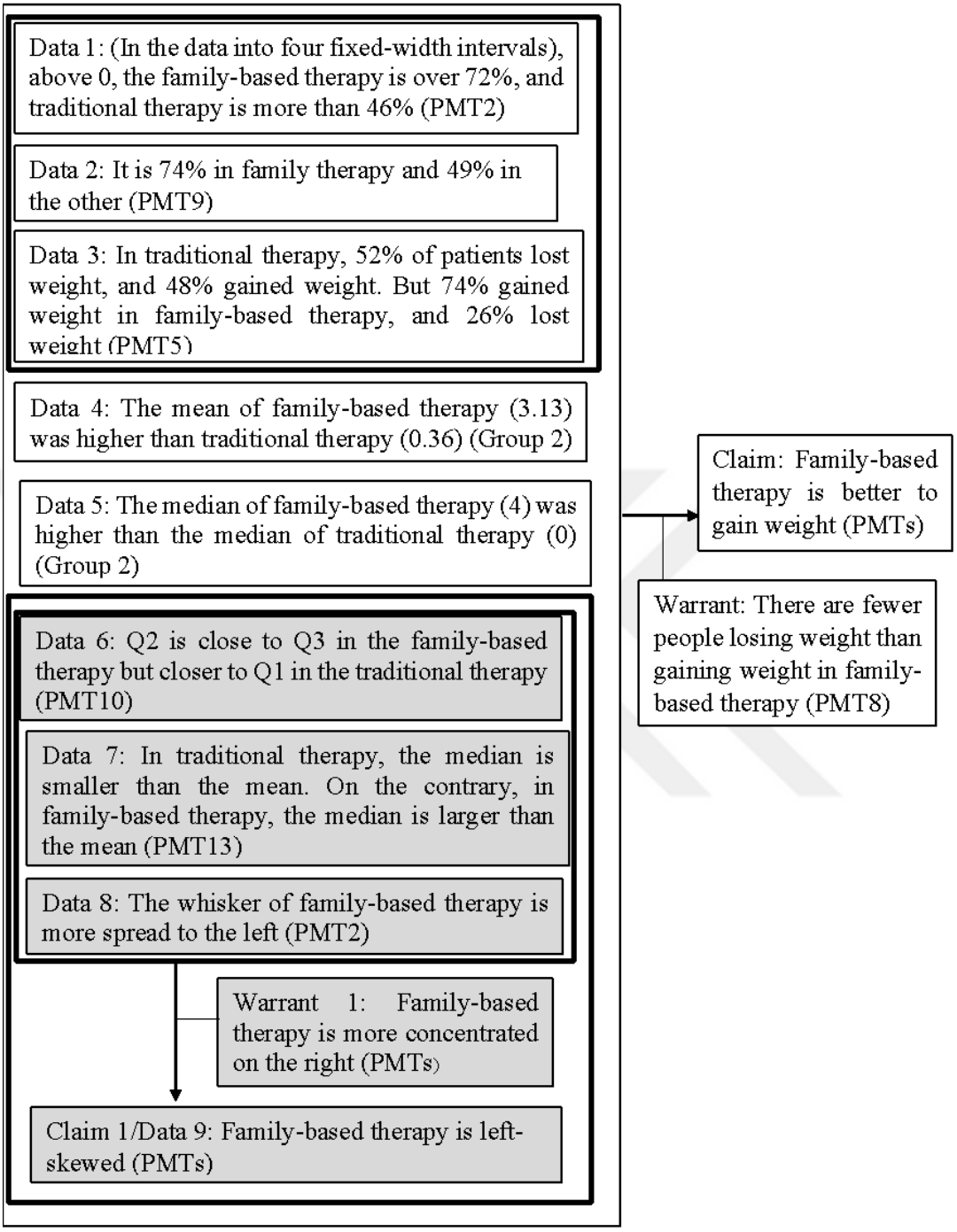


Figure 4.52. The argumentation scheme regarding the effect of the therapies on weight gain in the "anorexia" activity

The reasoning about the shape was also observed in the "recognizing the pattern in the data and supporting the pattern with different graphical representations-

ambulance" activity implemented in the last class session of the teaching experiment. The activity was the last in the teaching experiment. The main goal of the activity was to create graphs to investigate how the data was distributed by partitioning in different ways and reasoning proportionally. At the beginning of the activity, the PMTs talked about their experiences of being transported to the hospital by an ambulance, why they called an ambulance, incidents in which they had reached an ambulance at school or home, how long it took for an ambulance to arrive, the obligations of schools to call an ambulance in case of an accident in schools, the features that an ambulance should have, the importance of the response time of the ambulance called, etc. Meanwhile, the instructor mentioned a public school's contract with one of two ambulance companies, Acme or Lifeline. The school administration had narrowed their decision to the two companies based on their staff and equipment quality. The PMTs were expected to determine which had the better response times. Before the data was given to them, they were asked how it could be collected. They stated that they could look at customer reviews or call the companies. The instructor explained the data collection process in detail during the activity. The ambulance data was taken from response time records from a school district gathered last year about the two ambulance companies. In this period, 157 emergency calls were made to the Acme Ambulance and 278 to the Lifeline Ambulance. The PMTs predicted why the number of calls was different—the location of the ambulance company, the number of ambulances, safety, insurance status, and advertising. The activity data was offered in dot plots after the PMTs were familiarized with the context. The instructor asked them to interpret the two dot plots without statistics, measures, or calculations.

Some PMTs immediately said that the plots were both right-skewed. Even though they were interested in the shape of the distribution at first, it was important to understand what the graph was saying. For this, the instructor asked what the dots above specific values, like 6 or 8, and their frequency meant to understand what the data represented. For example, PMT5 interpreted the four points on the six-value on the x-axis in the Acme Ambulance distribution as follows. "Acme ambulance has

been four times in six minutes." The shapes of the distributions were then interpreted. Everyone in the class agreed that the distribution of Acme Ambulance was right-skewed. When explaining the shapes, PMT12 stated that the ambulance quickly arrived at the accident place because its data were concentrated on the left. There was disagreement about the shape of Lifeline Ambulance's distribution. Most PMTs argued that it was close to normal instead of right-skewed. More data was in Lifeline Ambulance, and the distribution looked almost leveled. Based on the interpretation, the PMTs, who said it was right-skewed, were convinced.

They all preferred Acme Ambulance as a better ambulance with a right-skewed distribution. When asked why, PMT4 explained, "Since there is more data on the left, the time is less on the left. The ambulance arrived in less time." As the ranges of the distributions from 6 to 22 minutes were equal, the conclusion of PMT4 based on the shape constituted valid reasoning. Although the sample numbers differed, when the distributions were evaluated within themselves, the values were piling up on the left side in the right-skewed distribution. Since the response time was desired to be at lower values according to the context, it was reasonable to prefer a right-skewed distribution rather than a normal distribution that spread almost evenly over the entire range. The PMTs worked in small groups to support the conclusion with different inscriptions.

The participant observer noted that variability was examined in Group 1. Since the ranges were the same in the distributions, PMT9 suggested finding an extra variability value for consistency. Due to the difference in the number of samples, they calculated MAD instead of SAD. The MAD value of Acme Ambulance was 3.07, and the MAD value of Lifeline Ambulance was 3.4. Once the MAD values were interpreted similarly, the interquartile ranges were compared in the box plot. It was 5.5 in Acme Ambulance and 6 in Lifeline Ambulance. Interquartile ranges were also interpreted as not being very different. While the distribution of the Lifeline Ambulance was described as normal, its variability was higher than the right-skewed Acme Ambulance distribution. It supported the TAS idea that *normality is not a criterion for variability*. While selecting appropriate statistics for the context,

comparing range, MAD, and interquartile range provided an example of the idea that *the measures of variability support each other*. When asked what conclusion they reached regarding variability in the class discussion, PMT2 explained, "There is an orientation almost equal to the mean because of the MAD. Their means are different. We prefer the one with the lower mean instead of looking at MAD. Whichever is shorter is important." Although they stated the mean for a low average, they did not make the comparison based on the mean.

In small groups, the PMTs concluded that Acme Ambulance responded faster by structuring the data differently in TinkerPlots. The box plot, in which they divided the data into four equal groups in TinkerPlots, was a common solution for all groups. When comparing quartiles, the groups focused on the median, interquartile range, or intervals where quartiles fall in the small groups. Group 3 was the first to share their solutions in class discussion. The group's solution was supported by other groups while comparing quartiles in different ways in class. It showed that everyone in the class actively understood and reacted to the others' solutions. Group 3 and Group 1 directly created a box plot in TinkerPlots to compare quartiles. However, Group 2 showed the box plot after partitioning the data into four equal groups in TinkerPlots. The following dialogue that the members of Group 3 started was observed in the class over graphs of Group 2 and Group 1 in Figure 4.53.

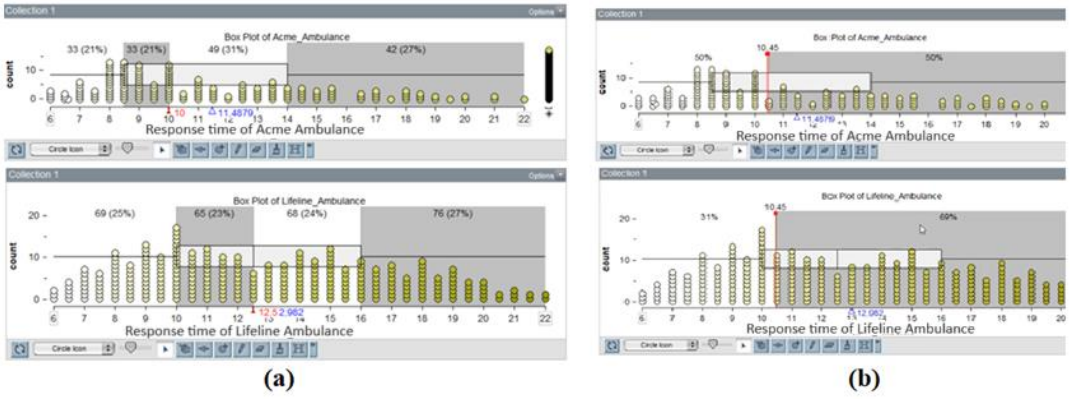


Figure 4.53. The graphs drawn by Group 2 (a) and Group 1(b) in TinkerPlots for the activity related to ambulance

PMT5. 50% is between 6 and 10 in the first and 6 and 12.5 in the second.

PMT3. Based on ten, 50% is on one side and 50% on the other.

PMT5. 50% on the top were below ten, and 25% on the bottom.

PMT12. Ten is on the border of the quartiles.

PMT10. As PMT5 said, we (Group 2) came to the same conclusion. 50% of Acme ambulances arrive in under ten minutes, but 25% of Lifeline ambulances arrive this time.

PMT2. 10 minutes may be an ideal time to protect the child from harm. When taken precisely above ten, it separates 42% and 58% at the above. Therefore, in our group (Group 1), we took a value like 10.45 as a cut point, not exactly above 10. (He shares their graph in Figure 4.53b). 50% of Acme ambulances arrive in under 10.45 minutes and 50% over. 31% of Lifeline ambulances are below this time, and 69% are above this time. Also, in the box plot, 75% of ambulances arrive in 14 minutes in the Acme Ambulance and 16 minutes in the other. Two minutes is significant for emergencies.

All groups were able to interpret the box plot partitioned into quartiles easily in TinkerPlots. The first solution method used by the PMTs was to compare the quartiles. When comparing the quartiles, the PMTs looked at the intervals in which the quartiles fell. For example, PMT5 investigated which of the two distributions had lower values in the range of the first 50% of the data. Since the minimum value of the two distributions was the same, she could conclude that the Lifeline Ambulance was more widespread. As can be understood from the explanations about 50% of the data made by PMT3, PMT5 compared the median of the distributions while comparing the intervals. Ten minutes was critical for both quartiles and the context. This value was the upper distribution's median and the lower one's first quartile. This comparison, which the other groups defended, was reasonable in favor of Acme Ambulance. While PMT2 mentioned the critical importance of ten minutes, he showed 13 dots with this value in the above distribution. They were not included in the first 50% of the data when ten minutes were taken as a cut-point. When ten minutes were taken as a reference, although it was the median of Acme Ambulance, the data could not be divided into half due to the distribution of the data. Therefore, they took 10.45 as a cut-point in their group, where values on the cut-point could be included in the first half (See Figure 4.53b). Lifeline Ambulance, which has a lower

ratio (31%) in this range, was not preferred, as it was essential to have more values below this value according to the context. At the end of the dialogue, PMT2 compared the third quartiles of the distributions, just as PMT5 compared the medians of the distributions. Since the starting points were the same, judging by the magnitude of the value corresponding to the third quartile was sufficient. The same rate, 75%, was spread over a narrower range on the upper one. In other words, the quartile in the narrower range was denser. This indicated using the TAS ideas that *the intervals allocated to quartiles have different widths*, and *the density is higher in the narrower intervals of the data partitioned into equal sizes*.

Regarding the interpretation of the box plot, the positions of the quartiles or the magnitudes of the measures of the center were examined in small groups, which were also shared in class discussions. While Group 1 evaluated the distribution as right-skewed due to the median of Acme Ambulance being close to Q1, Group 3 compared the distributions' means and medians. That is, they used the TAS idea that *the location of the measures of center judges the shape of the distribution.* For example, PMT3 in Group 3 said, "The one above is right-skewed, where the median is ten and the mean is 11.4. The mean is higher than the median. Both are very close to each other at the bottom, around 12. It is like a normal distribution." The PMTs in Group 2 interpreted that the Acme Ambulance was right-skewed based on the position of the quartiles and the size of the measures of the center. PMT10 in Group 2 said, "It is not like the weights in anorexia. We were looking for something close to Q3 because we were trying to gain weight there. Here, being closer to Q1 means the ambulance is faster." They reasoned what characteristics the distribution has according to the shape of the distribution. In these interpretations, the shape of the distribution gave information about the distribution's skewness due to extreme values, where the concentration was, and the location of the center based on the majority. In other words, how the data was distributed was explained by the shape of the distribution. This showed that the PMTs could consider the distribution as an entity.

In addition, they could also reason the distribution as a whole when they partitioned it into equal-width intervals in TinkerPlots. For example, the PMTs in Group 2, which shared their solution in the class, divided the distribution into six equal-width intervals in Figure 4.54. In TinkerPlots, they could determine both the frequencies and percentages within the intervals by only clicking on the "counts" button. They interpreted the differences in percentages rather than comparing frequencies across ranges because the sample sizes differed. They clearly stated that they could look at the percentages by emphasizing the sample sizes in the distributions.

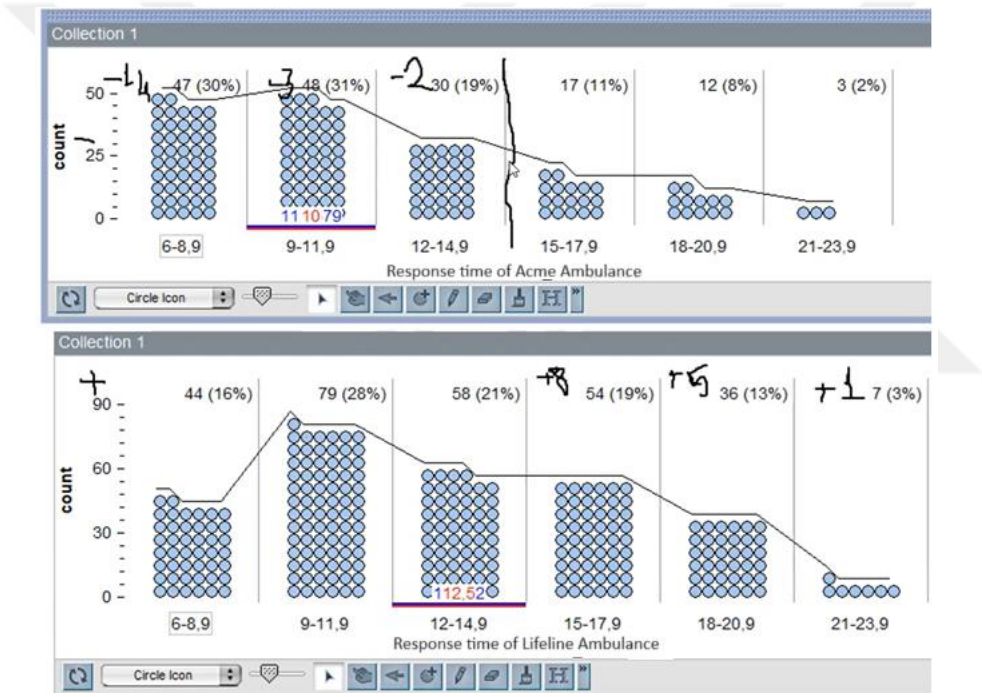


Figure 4.54. The data that Group 2 partitioned into equal-width intervals in TinkerPlots for the activity related to ambulance

> PMT10. I consider Acme Ambulance to be a minus and Lifeline Ambulance to be a plus. (She subtracted the below percentages from the above ones and wrote them on the graph.) 14, -3, -2, +8, +5 and +1.
>
> PMT9. For us, less time is more important. Where time is less is Acme.
>
> PMT10. This was something like a cut-point. Before here (She drew a line on 15, where the sign changed in the above distribution), it is Acme, then Lifeline.

> PMT9. We can take 15 as a reference value. It is very obvious. Acme Ambulance is faster.

When the differences in the percentages of data points in six equal-width intervals were examined, it was seen that the sign changed after three intervals. In Group 3, the PMTs divided the graphs into five equal-width intervals in TinkerPlots. They noticed that the sign changed after 12. When they took 12 as the cut point, they reached the result favoring Acme Ambulance. In class discussion, PMT5 stated, "60% of Acme ambulances arrived before 12 minutes, while 44% of Lifeline ambulances could arrive within this time." When Group 3 added up the differences, they discovered that the value was 0. Adding the differences in the percentages in the intervals meant taking the differences between the sample sizes, regardless of the distribution. There would be no difference when they took the ratios because they equalized the samples. Instead, comparing the proportions according to the cut-point led them to a reasonable conclusion about how the data was distributed. Focusing on different parts of the distributions, they generated additional data for their arguments in class discussion, as summarized in Figure 4.55.

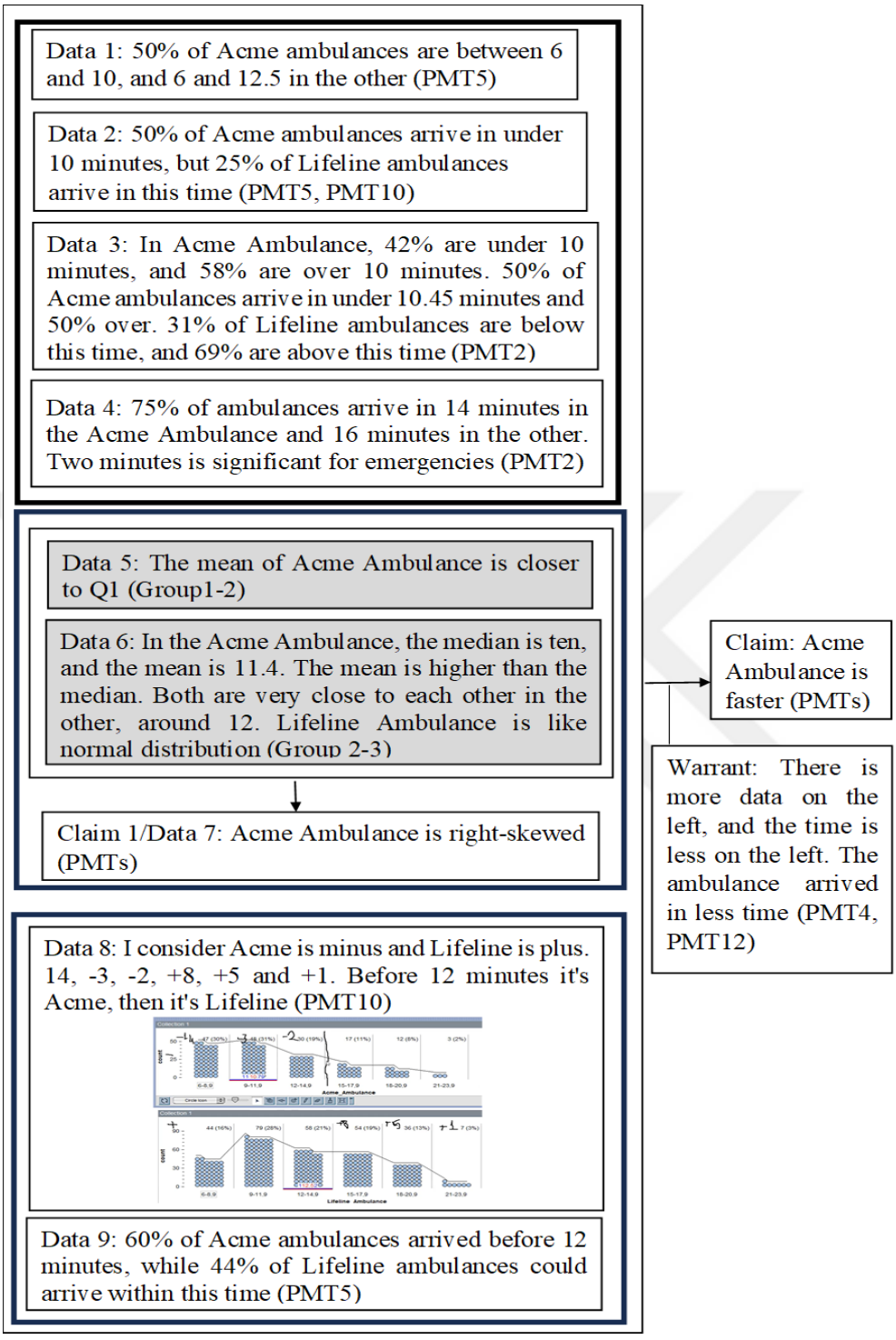


Figure 4.55. The argumentation scheme regarding the response time of the ambulances

As can be seen in Figure 4.55, the distributions were compared with different partitioning. Following the class discussion, the characteristics of these data structures were focused on. The instructor asked, "Which partitioning gives us more useful information about the data?" The PMTs stated that comparing the quartiles was more advantageous, and it was used when comparing distributions with unequal sample sizes at the end of the teaching experiment. When asked to explain why, PMT3 said, "The intervals differ. We know the same proportion of data in each interval, so we can predict that there is more data in the narrower interval. So, we can get more detailed results between intervals." It showed that PMT3 made sense that *the intervals allocated to quartiles have different widths*, and *the density is higher in the narrower intervals of the data partitioned into equal sizes*. Regarding the cut-point method, PMT2 stated that the intervals were not equal in the two groups, and interpretations could be made by comparing the ratios even if the sample sizes differed. Finally, it was asked what the distribution would look like when unequal sample sizes of distributions were equalized. PMT10 stated, "If we add an equal amount of data in the intervals in the distribution, the appearance of the distribution will not change." Distinctively, in activities where the number of samples was not equal in the teaching experiment, from the "comparing distributions with unequal sample sizes-backpacks-2" activity to the "recognizing the pattern in the data and supporting the pattern with different graphical representations-ambulance" activity, the PMTs assumed that the distribution did not change when proportioning by multiplying the frequencies in the intervals by a scalar factor to balance the number of samples. They acknowledged that growing the sample size by the same proportion does not change the distribution. It showed that the PMTs could explore the qualitative characteristics of distributions instead of the collections of data.

As a result, in the last two weeks of the teaching experiment, the PMTs were interested in comparing distributions with unequal sample sizes. To make a decision, they structured the data in dot plots in different ways, such as partitioning the data into intervals with equal width, partitioning the data into groups with equal numbers,

and partitioning the data into two groups based on a cut point. They reasoned proportionally in the frequencies in allocated partitionings. In this respect, the four TAS ideas regarding qualitative distribution characteristics emerged. They were 1) *The hill of the data set is in the bin with the most frequency*, 2) *The intervals allocated to quartiles have different widths*, 3) *The density is higher in the narrower intervals of the data partitioned into equal sizes*, and 4) *The location of the measures of center judges the shape of the distribution*. In all ideas, the PMTs focused on where the data was concentrated and concluded according to the shape of the distribution. Accordingly, the density analysis contributed to interpreting the shape of the distribution. In this context, reasoning about the shape of the distribution formed the basis for reasoning about the distribution.

To sum up, this study aimed to investigate the development of the PMTs' reasoning about distribution through a classroom teaching experiment based on RME and enriched by dynamic statistics software. It first focused on the changes in PMTs' reasoning about distribution before and after the teaching experiment to understand whether there was any improvement. For this, their reasoning levels were investigated based on statistical measures, data representations, shape and shifts in data, and proportional reasoning. After the teaching experiment, there was a noticeable improvement in reasoning levels, with a decrease in idiosyncratic, verbal, transitional, and procedural levels and an increase in the integrated process reasoning level. More specifically, before the teaching experiment, the PMTs used the aspects of distribution incorrectly and incompletely or reason based on a single aspect of the distribution without associating the aspects. However, after the teaching experiment, they started to see the distribution as a whole. In other words, more than half of PMTs at the integrated process level demonstrated a comprehensive understanding of distribution, interpreting center, spread, density, and skewness logically.

Since the teaching experiment occurred in a social classroom environment, CSPs emerging from TAS ideas were crucial in nurturing PMTs' reasoning about distribution. While examining the emergence of CSPs in conjunction with TAS ideas, the PMTs' reasoning levels when dealing with these ideas were also examined.

Thus, it was seen how the PMTs' reasoning development progressed while engaging in activities during the teaching experiment. All these examinations revealed how the PMTs' reasoning about the distribution developed throughout the classroom teaching experiment based on RME and enriched by dynamic statistics software. Table 4.7 below summarizes the emerging CSP, TAS ideas, and the PMTs' reasoning levels while dealing with these ideas.

Table 4.7 The emerged CSP, TAS ideas, and the PMTs' reasoning levels while dealing with these ideas

CSPs	**TAS ideas**	**Reasoning level**
CSP1. Reasoning about the average as the signal in a noisy process	If the sample sizes are the same, the mean and the total say the same thing	Transitional
	Average is the typical value representing the data	Transitional
	The mean distributes data items equally	Procedural
	The mean is responsive to outliers	Procedural
	The sum of the distances of scores above and below the mean is equal	Integrated process
CSP2. Measuring of variability in data	Normality is not a criterion for variability	Integrated process
	If the number of samples is different, the MAD is used rather than the SAD	Procedural
	The MAD is a predecessor of the standard deviation	Transitional
CSP3. Selecting appropriate statistics for the context	The measures of variability support each other	Transitional
	Connecting the measures of center with the measures of variability leads to more reliable results	Procedural
	Unlike the range, the interquartile range reduces the effect of outliers/extreme values	Integrated process

Table 4.7 (Continued)

CSP4. Exploring qualitative characteristics of collections of data points	The sample varies between extreme cases based on the specific value/mean	Integrated process
	The mean of the data set with high data points is high	Procedural
	The dispersed distribution is unreliable	Integrated process
	The data is concentrated where the mean is	Integrated process
CSP5. Exploring qualitative characteristics of distributions	The hill of the data set is in the bin with the most frequency	Integrated process
	The intervals allocated to quartiles have different widths	Integrated process
	The density is higher in the narrower intervals of the data partitioned into equal sizes	Integrated process
	The location of the measures of center judges the shape of the distribution	Integrated process

According to the table, five CSPs emerging from 19 TAS ideas were identified. They were 1) Reasoning about the average as the signal in a noisy process, 2) Measuring of variability in data, 3) Selecting appropriate statistics for the context, 4) Exploring qualitative characteristics of collections of data points, and 5) Exploring qualitative characteristics of distributions. The PMTs exhibited different levels of reasoning when dealing with the ideas within the scope of these CSPs. There was no linear progression between the levels of reasoning. In the first CSP, where they conceptualized the central tendency by taking variability into account, they were generally at the transitional or procedural reasoning level according to the focus on the center or variability of the data. They reached the level of integrated process reasoning in the idea that "*The sum of the distances of scores above and below the mean is equal*" as they interpreted how data was distributed. Similarly, in the second CSP, where they made sense of the variability measures, they were also at the integrated process reasoning level because they considered the distribution in the idea that "*Normality is not a criterion for variability,*" in which they could not

express the degree of variability according to the shape of the distribution. However, in the other ideas, they reasoned at the transitional or procedural level, depending on whether the variability measures or the center was involved. In CSP3, they considered multiple statistical measures to draw more accurate and reliable decisions for a statistical context. When they concluded that the variability measures supported each other, they reasoned at the transitional level because they focused only on the spread aspect of the distribution. However, they were at the procedural reasoning level when they also considered the center with the measures of variability. In the idea that "*Unlike the range, the interquartile range reduces the effect of outliers/extreme values*," they reasoned based on the skewness and density reflecting the distribution by comparing the variability measures, which demonstrated they were at the level of integrated process reasoning. In the last two CSPs, where they explored the qualitative characteristics of collections of data points and distribution, respectively, the PMTs were at the integrated process reasoning level, except the TAS idea that "*The mean of the data set with high data points is high.*" Describing the collection of data points, they understood the distribution of the data, and they commented on the spread and center aspects of the distribution as well as the density. In the last CSP, this understanding was based on the density in the distribution, which evolved into the shape of the distribution. That is, the PMTs reasoned about the distribution based on the shape of the distribution towards the end of the teaching experiment. The findings provided an understanding of how PMTs' reasoning about distribution improved throughout the teaching experiment based on RME and enriched by dynamic statistics software. Explicitly, the implementation of the HLT that gradually progressed by making sense of central tendency and variability, interpreting density in the data, and reasoning about the shape of the distribution in an integrated manner in the social learning environment contributed to the improvement of PMTs' statistical reasoning about distribution.

To sum up, the study's findings revealed that the PMTs' reasoning about the distribution developed throughout the classroom teaching experiment based on RME and enriched by dynamic statistics software. The PMTs reasoned about distribution

by reasoning about the average as the signal in a noisy process, measuring variability in data, selecting appropriate statistics for the context, exploring qualitative characteristics of collections of data points, and exploring qualitative characteristics of distributions throughout the teaching experiment.

CHAPTER 5

CONCLUSIONS AND DISCUSSION

The purpose of the study is to investigate the development of PMTs' reasoning about distribution as they participated in the classroom teaching experiment based on RME and enriched by dynamic statistics software. Concerning the purpose, the study's findings were organized according to the study's research questions. In the previous chapter, the findings related to the development in the PMTs' reasoning about distribution were presented based on their reasoning levels across the four skills required to reason about the distribution. Following this, the classroom practices that emerged throughout the classroom teaching experiment were reported. In line with the findings, in this chapter, the conclusions were discussed under the headings of reasoning skills to reason about distribution and the classroom statistical practices that emerged throughout the classroom teaching experiment. Next, based on the conclusions of this study, some implications and suggestions were offered to mathematics educators and curriculum developers who aim to develop statistical reasoning. Finally, the limitations of the study and suggestions for future research were addressed. In this context, this chapter comprised the reasoning skills to reason about distribution, the classroom statistical practices emerging throughout the classroom teaching experiment, implications of the study, limitations of the study, and suggestions for future research sections.

5.1 The Reasoning Skills to Reason about Distribution

In this study, the PMTs' reasoning levels before and after the classroom teaching experiment were examined to reveal the development in their reasoning. The study demonstrated that the PMTs' statistical reasoning skills, which were limited before participating in the classroom teaching experiment, improved after the teaching

experiment. Nevertheless, there was no difference in the PMTs' reasoning at the idiosyncratic and verbal levels. In other words, the PMTs did not generally misuse statistical concepts or fail to apply what they said. This may be due to their prior knowledge from previous statistics courses. More specifically, most of the PMTs had used one aspect of distribution (e.g., the center, spread, density, and skewness) at the transitional (28.6%) reasoning level or at least two aspects of distribution without associating them at the procedural (28.6%) reasoning level before the teaching experiment. In contrast, following the experiment, over half of them (57.2%) were at the integrated process reasoning level. That is, they made sense of the distribution and could confidently explain the statistical process. As a result, the PMTs performing at the transitional and procedural levels before the teaching experiment progressed to the integrated process reasoning level after the teaching experiment.

The shift in the PMTs' reasoning level occurred in each skill that they used when reasoning about distribution. In this regard, they employed the skills of reasoning about statistical measures, reasoning about representations of data, reasoning about shape and shifts in data, and proportional reasoning. Reasoning about statistical measures and reasoning about representations of data are among the types of statistical reasoning introduced by Garfield (2002) while reasoning about shape and shifts in data and proportional reasoning are encountered in the studies aiming to develop reasoning about distribution (Bakker, 2004; Bakker & Gravemeijer, 2004; Cobb, 1999; Gravemeijer & Cobb, 2006).

The findings of the study indicated that the PMTs could express two or more statistical measures, which revealed that they were at least at the procedural reasoning level. Furthermore, most of them (42.9%) were at the integrated process reasoning level because they were able to relate these measures. That is to say, they could determine appropriate statistical measures for the context of the statistical problem and explain what these measures meant in the context. In-depth questioning of the characteristics of the statistical measures in the challenging context helped the instructor overcome the problem of focusing on only one statistical measure. In this

context, the findings of this study do not support the previous research reporting that PMTs and teachers apply algorithms without conceptual understanding and interpreting the statistical measures in the context (Engledowl & Tarr, 2020; Groth & Bergner, 2006; Groth & Meletiou-Mavrotheris, 2018; Huey, 2011; Huey et al., 2018; Jacobbe, 2012; Landtblom & Sumpter, 2021; Leavy & O'Loughlin, 2006; Makar & Confrey, 2004; Peters, 2014; Shaughnessy, 2001; Silva & Coutinho, 2008). However, they performed similarly to the results in these studies before the teaching experiment. Therefore, it can be said that the classroom teaching experiment contributed to developing the PMTs' reasoning about statistical measures. In this regard, the study allowed the PMTs to go beyond the procedure and to logically justify the conclusions reached by statistical measures in statistics education.

Regarding reasoning about representations of data, in the study, the PMTs were able to consider different aspects of the distribution both before and after the teaching experiment. Differently, following the teaching experiment, they were able to integrate them and interpret the characteristics of the distribution, such as shape, center, and spread, in the given context. That is, the cluster at the procedural level (57.2%) before the teaching experiment progressed into the integrated process reasoning level at the end. In this sense, there was no considerable change in the PMTs' level of reasoning about representations of data. One reason may be the differences in the structures of the questions measuring this skill. In the test, Question 2, which requires interpreting the data in the histogram, and Question 3, which asks to evaluate the expressions related to the data in the bar graph, measure this skill. When the PMTs' reasoning levels in these questions were analyzed separately, their reasoning levels were generally at the integrated process level for Question 3 in both the pretest and posttest. In this context, pre-given reasonings in the tests may have directed the PMTs to the statement that represented the data appropriately. It showed that question structure can affect the way of reasoning, which needs to be investigated. However, the PMTs were successful in reasoning about the representations of the data as they could describe the characteristics of the distribution regardless of the question type. When reading the data in the

representations, they could classify particular attributes in the data instead of individual cases and perceive the data as a whole. In this regard, the participants of the study looked at the data as classifiers or an aggregate (Konold et al., 2015). Put differently, they could understand the data globally (Ben-Zvi & Arcavi, 2001).

Moreover, the study showed that the PMTs saw the data as a whole in the analysis of the reasoning about shape and shifts in the data that they used when comparing distributions with equal and unequal sample sizes. After the classroom teaching experiment, the proportion of the PMTs at the integrated process reasoning level was similar (57.2%). Before the teaching experiment, they compared the distributions at the transitional level based on one aspect of the distribution. As in the previous studies (Frischemeier & Biehler, 2018; Gökçe, 2019; Karatoprak, 2011; Kurt, 2016; Mercimek, 2013; Sorto, 2004; Yılmaz, 2019), the PMTs in the study were weak in statistical justification and reasoning on the characteristics of distributions before the teaching experiment. They focused on reaching a conclusion based on the hill or mean value of the distribution. On the other hand, after the teaching experiment, they exhibited a complete understanding of the distribution and interpreted their results in the context confidently. That is to say, the PMTs developed their reasoning about the distribution in this study. There are many possible reasons for the development of the PMTs' reasoning about distribution.

Firstly, the PMTs in the study used dynamic statistical software. In TinkerPlots, by structuring the data in different ways, such as partitioning into equal width intervals, partitioning into equal groups, partitioning into half, etc., they were able to identify the characteristics of the distribution (e.g., shape, center, and spread). As in Frischemeier and Biehler's (2018) study, TinkerPlots supported PMTs' reasoning in the process of comparing distributions with different characteristics. For example, in TinkerPlots, the PMTs were able to divide dot plots with both equal width and equal counts while comparing distributions with equal and unequal sample sizes. Examining both partitionings simultaneously prompted them to think about the relationship between frequency and scale in graphical representations and therefore the shape of the distribution.

Another reason for the development of the PMTs' reasoning about distribution can be the structure of the questions in the tests. Contrary to the skill of reasoning about representations of data, in the skill of reasoning about shape and shifts in data, the PMTs exhibited greater success in the posttest through their own inscriptions instead of evaluating the pre-given reasonings in the pretest. For example, in the last question of the posttest, they were asked to justify their summary quantitively by using statistical software. This requirement to structure the data presented a learning opportunity to increase the elements in comparing distributions and interpreting context, as suggested by Frischemeier and Biehler (2018).

Finally, the PMTs participating in the classroom teaching experiment studied activities that required the integration of different aspects of the distribution and looking at the whole distribution. At the beginning of the teaching experiment, they interpreted the center and variability aspects of the distribution in the context of the activity, but later, when graphical representations were used, they started to interpret the distributions by including the density and skewness aspects. That is to say, the activities that progressed gradually through the integration of aspects of the distribution helped the PMTs to develop reasoning about the distribution.

According to Cobb (1999) and Bakker (2004), proportional reasoning is necessary for statistical reasoning about distribution, particularly when comparing distributions with different sample sizes. While the three skills above (e.g., reasoning about statistical measures, reasoning about representations of data, and reasoning about shape and shifts in data) were used throughout the teaching experiment, proportional reasoning was planned to be used only in the learning activities in the last week of the experiment. In other words, the development of proportional reasoning was not the aim of the study. Interestingly, it was found that the PMTs' proportional reasoning also developed simultaneously as their statistical reasoning about distribution improved. Before the teaching experiment, the PMTs exhibited the behaviors of using additive reasoning instead of proportional reasoning at the idiosyncratic reasoning level or being aware of using proportional reasoning but incorrectly or incompletely at the verbal, transitional, and procedural reasoning

levels. However, most PMTs (71.5%) completely used proportional reasoning at the integrated process reasoning after the teaching experiment. In this context, the classroom teaching experiment contributed to developing the PMTs' proportional reasoning.

Not only the test results but also the PMTs' reasoning patterns throughout the teaching experiment showed that their proportional reasoning improved. The PMTs did not use incorrect reasoning, and their proportional reasoning improved over time, according to Copur-Gencturk et al.'s (2023) proportional reasoning levels. At the beginning of the study, they iterated the average into all parts of a data set based on homogeneity. Then, they considered the sizes of different units with the idea of "x elements for y elements" (Karagöz Akar, 2007; Ölmez, 2016) when performing operations such as multiplication and division. At the end of the study, they used multiplicative reasoning with relative frequency in the distributions with unequal sample sizes. That is, at the beginning of the teaching experiment, they were at the additive reasoning level, while towards the end, they used proportional reasoning due to the differing sample size distributions. In this respect, the findings of the study presented that researchers need to study statistical reasoning in detail within the multiplicative conceptual field that includes theoretically connected mathematical concepts such as multiplication, division, and proportional reasoning (Ezaki et al., 2023).

5.2 The Classroom Statistical Practices That Emerged Throughout the Classroom Teaching Experiment

The study's findings revealed that the PMTs' reasoning about distribution developed throughout the classroom teaching experiment based on RME and enriched by dynamic statistics software. A valuable finding of the study was the CSPs that emerged from the common ideas of the class while reasoning about distribution. The CSPs that emerged throughout the teaching experiment in this study were reasoning about the average as the signal in a noisy process, measuring variability in data,

selecting appropriate statistics for the context, exploring qualitative characteristics of collections of data points, and exploring qualitative characteristics of distributions. In other words, the PMTs dealt with these five CSPs, which nurtured their reasoning development in the social environment of the classroom. Table 5.1 shows the CSPs and the taken-as-shared interest or activities in the HLT that serve the emergence of these practices.

Table 5.1 The CSPs and the taken-as-shared interest/activities in the HLT

CSPs	**Taken-as-shared interest/Activity**
CSP 1: Reasoning about the average as the signal in a noisy process	1. Aggregating different ranked data-printer *Assignment 1:* What did we learn? 2. Estimating an unknown whole through the average- tracking a herd *Assignment 2*: Teaching the mean meaningfully (Uccellini, 1996) 3. Estimating an unknown part through the average-missing words
CSP 2: Measuring variability in data	4. Measuring variability in distributions with small sample sizes-soccer game
CSP 3: Selecting appropriate statistics for the context	5. Building arguments based on statistical measures that are appropriate for the context-batteries *Assignment 3:* Data invention
CSP 4: Exploring qualitative characteristics of collections of data points	6. Generating different samples in weight context-hot air balloons 7. Building arguments based on statistical measures that are appropriate for the context-watermelons 8. Understanding the shifts in hill, clump, or cluster of data with equal sample size-resting heart rate
CSP 5: Exploring qualitative characteristics of distributions	9. Comparing distribution with equal sample sizes-backpack 1 10. Comparing distribution with unequal sample sizes-backpack 2 *Assignment 4:* Understanding Center and Spread (Kader & Mamer, 2008) 11. Recognizing the pattern in the data and supporting the pattern with different graphical representations-anorexia 12. Recognizing the pattern in the data and supporting the pattern with different graphical representations-Ambulance

According to Table 5.1, the first two CSPs are a reflection of the activities implemented in the classroom because the topics of the statistical discourse in the corresponding activities were related to measures of central tendency and variability, respectively. In CSP 1, the PMTs reasoned the average taking into account variability, whereas in CSP 2 they were interested in variability measures. Regarding CSP 3, as the PMTs built arguments based on one or more statistical measures in each activity in the HLT while solving statistical problems, they had to decide which was more appropriate. In this context, although CSP 3 is associated with the fifth activity due to the taken-as-shared interest of the activity, it is the common reasoning way of the class that occurred in all activities in the HLT. This situation is also supported by the fact that the taken-as-shared interest of the seventh activity related to watermelons in the HLT is the same. I cited CSP 4 and CSP 5 from Cobb's (1999) study on the grounds that these CSPs fulfill the TAS ideas that the PMTs put forward when dealing with the activities in this study. In CSP 4, the PMTs compared individual data values with measures of central tendency and variability in data with small sample sizes to conclude the activities, whereas, in CSP 5, they compared distributions with large sample sizes from a holistic view of the data. Therefore, the PMTs' conclusions were based on characteristics of the distribution (e.g., the shape) beyond the data points.

Although the emergence of CSPs in the study appears in chronological order in Table 5.1, the TAS ideas that nurtured these practices and the CSPs occurred intertwined throughout the teaching experiment. In other words, the emergence of the CSPs based on TAS ideas was network-like, which supported the findings of Ayan Civak (2020) and Stephan and Rasmussen (2002). For example, the PMTs practiced selecting appropriate statistics for the context from the beginning of the teaching experiment. In other words, they were concerned with the CSP 3-Selecting appropriate statistics for the context while dealing with all other CSPs. This result was also coherent with the progression of the PMTs' reasoning levels in the 19 TAS ideas identified in this study. The reasoning levels of the PMTs in this study did not

progress linearly throughout the teaching experiment. That is, the levels did not progress at the idiosyncratic, verbal, transitional, procedural, and integrated process reasoning levels, respectively, as Garfield (2002) described. One unexpected finding was that the PMTs were not at the idiosyncratic reasoning level in which they misused statistical concepts or verbal reasoning level where they were incapable of applying what they expressed. That is to say, they were able to apply something they had learned in statistical problem contexts. In addition, in 4 out of 19 TAS ideas, the PMTs were at the transitional reasoning level, where they focused on a single dimension of the distribution. It showed that the PMTs were interested in at least two aspects of the distribution.

The PMTs in the current study tended to calculate the mean in all situations and encountered difficulties while determining and interpreting the appropriate measure for the given context of the statistical problem before the teaching experiment. During the teaching experiment, where challenging activities involving variability were studied, the PMTs could integrate the central tendency and variability measures rather than just focusing on them as independent constructs. In the first CSP, where they conceptualized the central tendency by taking variability into account, they interpreted the central tendency as a balance point, fair share, and typical value that reflects the variability in the data, as elucidated in prior studies (Konold & Pollatsek, 2002; Mokros & Russell, 1995; Uccellini, 1996). Another interesting finding of the current study was that the PMTs overcame uncertainty to predict the unknown by taking a specific value close to the truth, which Konold and Pollatsek (2002, p. 262) described as "the signal in a noisy process." Although the data seemed to be reduced to a numerical value corresponding to a central tendency measure, questioning the underlying reasons for determining the specific value prevented the loss of information about the data. In this sense, the finding of the study supported the assertion of using the term central tendency instead of center or location when representing the average of data, including variability (Konold & Pollatsek, 2002). This result provided insights into how to overcome the problem of PMTs focusing on the mean calculation without conceptual understanding (Groth & Bergner, 2006;

Landtblom & Sumpter, 2021; Leavy & O'Loughlin, 2006; Peters, 2014). In this respect, we conjectured that integrating center and variability in the social environment of the classroom facilitates the development of reasoning about distribution.

Regarding the second CSP, the PMTs not only measured the variability of the data but also explored the properties of variability measures by comparing them with each other. In this process, to reason about variability, the PMTs practiced the aspects of variation described by Wild and Pfannkuch (1999) and Reading and Shaughnessy (2004). After they noticed the variability in the activities they worked on, they measured, defined, described, and modeled the variability. In addition, they developed investigative strategies for reasoning variability in data, which is in line with the didactical phenomenological principle of RME (Freudenthal, 1983). These strategies included associating variability with the shape of the distribution and comparing different variability measures. These strategies prevented the PMTs in the study from confounding the variability measures observed in previous studies (Dubreil-Frémont et al., 2014; Turegun, 2011). The distribution of the data in the activities could lead to different variability measures being considered because different distributions had the same variability measure value. For example, in the soccer game activity, the SAD values of the distributions were the same, which required thinking of the MAD value. Another possible reason could be the paper assigned to the PMTs (Kader & Mamer, 2008). It introduces the concepts of the range, interquartile range SAD, MAD, standard deviation, and variance in the context of soccer game activity. The PMTs in the study had not yet read this paper when they were studying the data of the soccer games. However, they experienced the situation described in the paper in the classroom. They discovered the concepts of MAD and SAD in that class session because these concepts were not included in the curricula they had used until then (MoNE, 2018a, 2018b). They defined the variation based on middle values, extreme values, deviations from any anchor, or deviations from the central anchor throughout the teaching experiment, as in the study (Reading & Shaughnessy, 2004). When reasoning with the concepts of MAD

and SAD, they dealt with deviations from the central anchor. That is to say, they maintained the integration of the center and variability that emerged in the first CSP while reasoning about variation to describe how the data was distributed.

The PMTs in the study took into consideration the measures of variability as well as the center when selecting appropriate statistics for the context. They were interested in multiple statistical measures to make more accurate and reliable decisions for the statistical context. Even if they focused on one measure within a context, they were capable of presenting more than one piece of evidence using different measures to support their results. Since the activities that the PMTs studied during the teaching experiment were all based on the context, they practiced selecting appropriate statistics for the context from the beginning of the teaching experiment. In this process, they proceeded to integrate the center and variability. More specifically, while solving the statistical problem in the activities, the PMTs calculated both measures, one from the center and one from variability, and chose the one that was appropriate for the context. This finding also supported the conjecture of not separating the study of center and variability that emerged in the previous CSPs.

While exploring qualitative characteristics of collections of data points, the PMTs dealt with distributions consisting of a smaller number of data values in TinkerPlots, which was the dynamic statistical software used in the study. In this process, they focused on either particular data values, such as extreme values, outliers, and the mean, or a particular part of the distribution rather than the whole data. Therefore, they could not see the data in aggregate. Nevertheless, they were able to justify the qualitative characteristics of data values by considering all individual data values. In Cobb's (1999) study, 7th-grade students investigated the values above a certain value or closeness of values to each other. Differentially, in this study, the PMTs started to associate the part of the distribution with the whole because they generated data. As in Bakker's (2004) study, data invention enabled the PMTs in this study to look at other parts of the distribution. In this context, the "generating different samples in weight context-hot air balloons" activity had the significant advantage of enabling the PMTs to experience different distributions due to the lack of existing data. In this

activity, the PMTs produced hypothetical samples that were different from the resampling method, which used observed data or data-generating mechanisms in Madden's (2008) study. While generating data in this activity, the PMTs who had a history of the data reasoned in more detail with their own distributions. This finding confirmed Cobb's (1999) suggestion to experience the data collection process for better reasoning.

When dot plots start to be used in the activities in the HLT, the PMTs interpreted the spread, center, and shape of the distribution, which Garfield (2002) expresses as general characteristics of the distribution. That is, they considered the density in addition to the spread and center aspects of the distribution. As the number of samples in the activities increased, they described the data in dot plots according to the density. This process corresponded to the second stage in Bakker and Gravemeijer (2004), in which the density was interpreted in different data groupings in dot plots by increasing the sample size. As Bakker (2004) emphasized, increasing the number of samples with progressively sophisticated notions and diagrams led to reasoning about distribution depending on the density aspect. Concerning this, reasoning about density in data served as a transitional step in exploring qualitative characteristics of distribution rather than collection of data.

Depending on density, the PMTs reasoned about the shape of distribution so that they explored qualitative characteristics of distribution. This finding supported the conjecture of Bakker and Gravemeijer (2004, p. 149) that "reasoning with shapes forms the basis for reasoning about distributions." As in the previous studies (Bakker & Gravemeijer, 2004; Cobb, 1999; Cobb & Gravemeijer, 2006), the PMTs in this study performed reasoning about the distribution interpreting the shape. Toward the end of the teaching experiment, they were able to make sense of distribution, taking into account all aspects of distribution at the integrated process level. This indicated that the PMTs could reason about distribution as an overarching concept involving center, spread, density, and skewness. For example, they used measures of center to estimate the shape of the distribution. In other words, they developed what Bakker and Gravemeijer (2004) called the downward perspective. This finding revealed that,

like statisticians, the PMTs were able to look at the distribution as an organizing structure or a conceptual entity (Bakker & Gravemeijer, 2004; Cobb, 1999; Cobb & Gravemeijer, 2006).

According to Bakker and Gravemeijer (2004) and Cobb (1999), reasoning about the shape of the distribution based on the hill in distributions is the final step to developing reasoning about distribution. As suggested by Cobb (1999), this study used activities that required proportional reasoning and technological tools to reason about the shape of the distribution. In the process, the PMTs were able to perceive the distributions rather than the collection of data points using proportional reasoning based on the data partitioned in different ways, such as partitioning intervals with fixed-width or four equal groups that are the precursors of the histogram and box plots, respectively. In other words, proportional reasoning and using TinkerPlots were mediators in developing the PMTs' reasoning about distribution. Concerning this, TinkerPlots provided more learning opportunities by increasing the repertoire of comparison groups, as in Frischemeier and Biehler (2018). The ratios/percentages were easily calculated in the data partitioned differently in TinkerPlots. Moreover, the PMTs considered each part of the distribution and the whole, interpreting the relative frequencies in different intervals of the dot plots in the context of the activities. In other words, the PMTs with a global understanding of data (Ben-Zvi & Arcavi, 2001; Cobb, 1999) were able to identify a pattern or global trend in the data by comparing the relative frequencies in the parts of the distributions they divided.

In addition to using technology and proportional reasoning due to unequal sample sizes in the activities, the instructor's guidance in the classroom teaching experiment may be another factor in developing the PMTs' reasoning about distribution. The fact that the instructor asked about the similarities and differences in the dot plots in which the data were presented before moving on to data analysis in the activities may have prompted the PMTs to think about the shape of the distribution. Before analyzing the data in small group work, they discussed the shape of the distributions without any quantitative evidence. After that, they concluded by checking the consistency of the results in their data analysis with their previous conclusion based

on the shape of the distribution. Accordingly, in this study, reasoning about the shape of the distribution enabled the PMTs to see the distribution as a whole.

As a result, the findings of the study indicated the development of the PMTs' reasoning about the distribution. More specifically, the PMTs respectively made sense of data values in different representations, interpreted the density in different data groupings, and reasoned about the shape of the distribution depending on where the hill was in the distribution (Bakker & Gravemeijer, 2004; Cobb, 1999; Cobb & Gravemeijer, 2006). In this process, unlike these previous studies, the PMTs were more interested in the characteristics of the distribution in the early stages. In the first half of the classroom teaching experiment, they were more engaged in reasoning about statistical measures for four weeks. During the whole teaching experiment process, the PMTs gradually reasoned about statistical measures, data representations, shapes and shifts in the data, and proportionality. In Cobb's (1999) study, the 7th-grade students seemed to be doing multiplicative reasoning towards the end, whereas in this study, the PMTs used proportional reasoning from the beginning of the teaching experiment, although not intentionally. In other words, in this study, the PMTs practiced all the skills necessary for reasoning about distribution during the teaching experiment. This process resulted in the development of their reasoning about the distribution at the end of the experiment depending on these skills. Since this study concentrated on the development of reasoning about distribution as collective learning of the classroom community within the classroom's social setting, the CSPs discussed above were important to understand the development. The PMTs reasoned about distribution by reasoning about the average as the signal in a noisy process, measuring variability in data, selecting appropriate statistics for the context, exploring qualitative characteristics of collections of data points, and exploring qualitative characteristics of distributions throughout the teaching experiment.

5.3 Implications of the Study

The main concern of the study was the development of the PMTs' reasoning about distribution. Based on the conclusions of the current study, some implications were offered to mathematics educators and curriculum developers who aim to develop statistical reasoning. These implications can provide helpful information about how PMTs come to a better understanding of distribution.

In line with the study's results, the first implication was the creation of social learning environments that foster classroom practices in which PMTs can integrate the center, variability, density, and skewness. Previous studies demonstrated that the distribution concept has an important role in comprising central statistical ideas (Bakker & Gravemeijer, 2004; Cobb, 1999; Cobb & Gravemeijer, 2006). As these studies emphasized, this study showed that statistical concepts should be considered holistically in the data analysis process by integrating the idea of distribution instead of teaching the statistical content separately in the curricula (MoNE, 2018a, 2018b; NCTM, 2000). Contexts that encourage PMTs to establish the relationships between key statistical concepts can be included in classroom settings as opposed to objectivities that treat statistical concepts separately (Bakker & Gravemeijer, 2006). In this regard, challenging activities that promote classroom discussion and collaboration, such as those in this classroom teaching experiment, may be introduced in classroom environments. In such collaborative environments where social interactions occur between teacher and students or student and student, learners explain and justify their ideas to others and help them learn (Cobb & Yackel, 1996).

This study emphasizes the importance of the integration of dynamic statistics software into instruction. Many studies report the advantages of TinkerPlots in improving reasoning about distribution (Fielding-Wells, 2018; Frischemeier & Biehler, 2018; Hammerman & Terc, 2004; Koparan, 2019; Makar et al., 2011; Meletiou-Mavrotheris et al., 2009; Watson et al., 2020). Similarly, by using TinkerPlots, the PMTs in this study also demonstrated notable improvements in their

ability to reason about distributions. Since the PMTs had the opportunity to structure the data in dot plots in different ways, such as partitioning the data into groups with a specified interval width and partitioning the data into four equal groups, they were able to pay attention to different parts of the distribution. In this context, it is suggested that dynamic statistical software be used to develop reasoning about distribution.

This result highlights two important dimensions of the development of reasoning about distribution. The first one is data representation in statistics education. The PMTs employed various data representations, including bar graphs, dot plots, histograms, and box plots, to inscribe the data activities. Although the dot plot is not included in the curricula they use throughout their educational life (MoNE, 2018a, 2018b), it formed the basic logic of the histogram and box plot, which play an important role in reasoning about distribution. Given that the PMTs in this study reasoned about distribution using two important types of conventional inscriptions, histograms and box plots (Cobb, 1999), it would be beneficial to include histograms, box plots, and especially dot plots in the curriculum.

The other one is the pedagogical dimension. Although technology can serve as a valuable tool in facilitating the development of statistical reasoning abilities among PMTs, as Lee et al. (2014) and Biehler et al. (2013) emphasize, the use of technology alone is not enough. In addition to the use of technology, the implementation of instruction guides the development. In this study, the instructor's questions and the activities that required proportional reasoning contributed to the development. The instructional interventions prompted the PMTs to think about the shape of the distribution by enabling them to relate the whole to the part thanks to relativity. The PMTs used proportional reasoning from the beginning of the teaching experiment, although it was not intentional. They initially reasoned based on homogeneity using one-to-one matching and equal sharing interpretation of the mean. They then made multiplicative comparisons based on the relationships between the whole distribution and its different parts. That is, proportional reasoning was in the background of all the activities in the HLT. However, it came to the fore toward the end of the teaching

experiment when comparing distributions with unequal sample sizes, as can be seen in the taken-as-shared interest of the last three activities in the HLT. In this context, it is recommended to include proportional reasoning in the instruction to improve PMTs' reasoning about distribution.

The CSPs in this study emerged through the HLT enacted during the teaching experiment. The HLT in this study proceeded conceptualization of central tendency together with variability, integrating the center, spread, density, and skewness aspects of the distribution to describe its characteristics and reasoning about the shape of the distribution. The particular sequence for statistical reasoning has a structure including proportional reasoning. In this context, this study provides a route that can be followed when introducing the concept of distribution to preservice teachers. Although the preservice teachers can use the formal statistical knowledge they have learned in the statistics courses they have taken in their calculations, they struggle with statistical problems based on reasoning within a context. In this context, the preservice teachers who will teach statistics need statistics education in such a way that they will be able to direct their statistics courses to improve their students' reasoning. The PMTs were able to develop their reasoning about the distribution as guided in this study. Since teachers tend to teach the way that they have been taught (Alesandrini & Larson, 2002), if they learn statistical reasoning about distribution effectively, they can enable their students to construct the distribution concept in their future statistics teaching practice. Accordingly, mathematics educators can offer statistics courses in teacher education in a way that improves reasoning beyond procedural statistical knowledge. More specifically, the objectives of the course can be oriented to integrate key statistical concepts such as center, spread, and shape rather than just focusing on them as independent constructs. As in this study, the content of the courses may include tasks designed to generate data according to specific criteria and contexts, as well as to make judgments about data with different characteristics to enable PMTs to reason effectively.

5.4 Limitations of the Study and Suggestions for Future Research

This study focused on the development of the PMTs' reasoning about distribution throughout a classroom teaching experiment based on RME and enriched by dynamic statistics software. Although the study indicated some evidence of development in the PMTs' reasoning about distribution, it has some limitations. This study was completed in a single macro-cycle with the 14 PMTs who participated in the study. The participants were 3rd- and 4th-grade students studying in the Elementary Mathematics Education program at a public university in Ankara. In addition, they were willing to learn the statistical content of this study. Namely, the sample is not representative of the population in Turkey. The object of design research is not the generalization of findings but their transferability to different contexts, which refers to theoretical or analytical generalization (Bakker, 2019). In addition, this study was an adaptation of previous studies conducted with 7th-grade students (Bakker & Gravemeijer, 2004; Cobb, 1999; Cobb & Gravemeijer, 2006). Educational implications in these studies were revised according to the context of the current study. Although preservice teachers have different cognitive levels from the low-grade levels students, the educational implications for the 7th-grade students were appropriate for the PMTs who were the participants of the study as the PMTs were also weak in terms of statistical reasoning at the beginning of the study. In this regard, this adaptation helped nurture the PMTs' reasoning about distribution within the context of this study. To support the study's conclusions, replication in different contexts and with other participants is recommended.

The current study focused on the concept of distribution and its aspects (e.g., center, spread, density, and skewness) in terms of content. That is to say, the content is limited to exploratory data analysis, dealing with organizing, describing, representing, and analyzing data, as well as informal statistical inference, which does not address inferential statistics. This content is compatible with the curriculum (MoNE, 2018a) that the participants of the study, the PMTs, will teach in the future. However, there is a need to reveal the place of the concept of distribution, which is

the subject of the study and covers many key statistical concepts, in the curriculum. Therefore, future studies can examine the concept of distribution in school course materials, such as curricula and textbooks.

This study concentrated on the development of the PMTs' reasoning about distribution by disregarding the pedagogical dimension. While theories can be tested in controlled environments in education research, this study showed the applicability of the activities in the HLT in the natural classroom environment (Confrey & Lachance, 2000). Since the study is limited to the PMTs' subject matter knowledge about distribution, how they will transfer what they have learned in the study to real classroom environments requires further investigation. In this regard, it is suggested that teacher development experiments (Simon, 2000) concerning pedagogical and statistical development be conducted to evaluate the long-term effects of such experiments in school statistics education where PMTs interact with their students.

Finally, in terms of the theoretical framework on which the study is based, the emergent perspective, it was important to create a social classroom environment. However, this study was conducted online due to the COVID-19 pandemic. Accordingly, the online learning environment seemed to be a disadvantage. In addition, the researcher, the instructor, and the PMTs were unfamiliar with this type of learning environment. However, it was not as expected; they adapted very quickly and benefited from the features of the online learning platform, namely Zoom. The creation of breakout rooms and screen sharing especially enabled the PMTs to engage in activities as a class. Furthermore, since they were already eager to learn statistics, all PMTs actively participated in the instruction in the context of the study. Alternatively, future research may be conducted to assess the extent of the PMTs' individual reasoning development based on different theoretical frameworks.

REFERENCES

Alesandrini, K., & Larson, L. (2002). Teachers bridge to constructivism. *The Clearing House*, *75*(3), 118–121. https://doi.org/10.1080/00098650209599249

Aridor, K., & Ben-Zvi, D. (2018). Statistical modeling to promote students' aggregate reasoning with sample and sampling. *ZDM*, *50*(7), 1165-1181. https://doi.org/10.1007/s11858-018-0994-5

Ayan Civak, R. (2020). *The evaluation of mathematical practices in a seventh-grade classroom: Analyzing students' development of proportional reasoning* [Unpublished doctoral dissertation]. Middle East Technical University.

Bahar, G. (2016). *Measuring pre-service teachers' mathematics content knowledge and mathematics pedagogical content knowledge on ratio and proportion* [Unpublished Doctorate Dissertation]. Boğaziçi University

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tool* [Unpublished Doctorate Dissertation]. Utrecht University.

Bakker, A. (2019). *Design research in education: A practical guide for early career researchers*. Routledge.

Bakker, A., & Gravemeijer, P. E. (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). Springer Netherlands.

Bakker, A., & Hoffmann, M. H. (2005). Diagrammatic reasoning as the basis for developing concepts: A semiotic analysis of students' learning about statistical distribution. *Educational Studies in Mathematics*, *60*(3), 333-358. https://doi.org/10.1007/s10649-005-5536-8

Bannan, B. (2010). Integrative learning design framework: An illustrated example from the domain of instructional theory. In T. Plomp, & N. Nieveen (Eds.), *An introduction to educational design research* (pp. 153-169). SLO

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education*. National Council of Teachers of Mathematics.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. E., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology*, *25*(4), 527-547. https://doi.org/10.1080/0020739940250406

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. https://doi.org/10.1023/A:1013809201228

Ben-Zvi, D. (2004). Reasoning about data analysis. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 121-145). Kluwer Academic Publishers.

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. 3-15). Kluwer Academic Publishers.

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clement, A. J. Bishop, C. Keitel, J. Kilpatrick, J., & A. Y. L. Leung (Eds.), *Third international Handbook on Mathematics Education* (pp. 643-689). Springer.

Biehler, R., Frischemeier, D., Reading, C., & Shaughnessy, J. M. (2018). Reasoning about data. D. Ben-Zvi, J. Garfield, & K. Makar (Eds.), *International handbook of research in statistics education* (pp. 139–192). Springer.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions. *Journal of the Learning Sciences, 2*(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3). https://doi.org/10.1080/10691898.2002.11910677

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2008). The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education, 1*(1). https://doi.org/10.5070/T511000026

Chance, B., Garfield, J., & delMas, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Proceedings of the International Conference on Technology in Mathematics Education*, Auckland, NZ, 25-32. https://files.eric.ed.gov/fulltext/ED474052

Chicoine, D. (2004). Ignoring the obvious: A constructivist critique of a traditional teacher education program. *Educational Studies, 36*(3), 245-263. https://doi.org/10.1207/s15326993es3603_4

Clements, D. H. (2007). Curriculum research: Toward a framework for research-based curricula. *Journal for Research in Mathematics Education*, *38*(1), 35-70. https://doi.org/10.2307/30034927

Cobb, P. (1999). Individual and collective mathematical development: the case of statistical data analysis. *Mathematical Thinking and Learning, 1*(1), 5 -43. https://doi.org/10.1207/s15327833mtl0101_1

Cobb, P. (2000). Conducting teaching experiments in collaboration with teachers. In A. E. Kelly, & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 307-341). Routledge.

Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13. https://doi.org/10.3102/0013189X032001

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 375-395). Kluwer Academic Publishers.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801-823. https://doi.org/10.1080/00029890.1997.11990723

Cobb, P., Stephan, M., McClain, K., & Gravemeijer, K. (2001). Participating in mathematical practices. *Journal of the Learning Sciences, 10*(1&2), 113-163. https://doi.org/10.1207/S15327809JLS10-1-2_6

Cobb, P., & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational Psychologist, 31*(3-4), 175-190. https://doi.org/10.1080/00461520.1996.9653265

Cole, R., Becker, N., Towns, M., Sweeney, G., Wawro, M., & Rasmussen, C. (2012). Adapting a methodology from mathematics education research to chemistry education research: Documenting collective activity. *International Journal of Science and Mathematics.* https://doi.org/10.1007/s10763-011-9284-1

Collins, A. (1992). Toward a design science of education. In E. Scanlon, & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15-22). Springer.

CCSSM (2010). *Common core state standards for mathematics*. National Governors Association Center for Best Practices and the Council of Chief State School Officers. https://www.thecorestandards.org/wp-content/uploads/Math_Standards1.pdf

Confrey, J. (2006). The evolution of design studies as methodology. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 135-151). Cambridge University Press.

Confrey, J., & Lachance, A. (2000). Transformative Teaching Experiments through conjecture-drive research design. In A. Kelly & R. A. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp. 231-265). Lawrence Erlbaum

Copur-Gencturk, Y., Baek, C., & Doleck, T. (2022). A closer look at teachers' proportional reasoning. *International Journal of Science and Mathematics Education*, *21*(1), 113–129. https://doi.org/10.1007/s10763-022-10249-7

Council of Higher Education (2019). Higher education system in Turkey. CoHE. https://www.yok.edu.tr/Documents/Yayinlar/Yayinlarimiz/2019/Higher_Education_in_Turkey_2019_en.pdf

Cox, S. E. (2014). *Perceptions and influences behind teaching practices: Do teachers teach as they were taught?* [Unpublished master thesis], Brigham Young University.

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, *18*(5), 382-393. https://doi.org/10.5951/jresematheduc.18.5.0382

Çatman Aksoy, E., & Bostan, M. I. (2020). Seventh graders' statistical literacy: an investigation on bar and line graphs. *International Journal of Science and Mathematics Education*, 1-22. https://doi.org/10.1007/s10763-020-10052-2

delMas, R. C. (2002). Statistical literacy, reasoning, and thinking: A commentary. *Journal of Statistics Education*, *10*(2). https://doi.org/10.1080/10691898.2002.11910674

delMas, G., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58. https://doi.org/10.52041/serj.v6i2.483

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55-82. https://doi.org/10.52041/serj.v4i1.525

Denzin, N, K., & Lincoln, Y. S. (1994). Introduction: Entering the field of qualitative research. In N. K. Denzin & Lincoln, Y. S. (Eds.), *Handbook of qualitative research* (pp. 1-17). Sage Publications.

Dubreil-Frémont, V., Chevallier-Gaté, C., & Zendrera, N. (2014). Students' conceptions of average and standard deviation. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9), USA*. http://icots.info/9/proceedings/pdfs/ICOTS9_C142_DUBREILFREMONT.pdf

Engledowl, C., & Tarr, J. E. (2020). Secondary Teachers' Knowledge Structures for Measures of Center, Spread & Shape of Distribution Supporting Their Statistical Reasoning. *International Journal of Education in Mathematics, Science and Technology (IJEMST)*, *8*(2), 146–167.

Erath, K., Ingram, J., Moschkovich, J., & Prediger, S. (2021). Designing and enacting instruction that enhances language for mathematics learning: A review of the state of development and research. *ZDM–Mathematics Education*, *53*, 245-262. https://doi.org/10.1007/s11858-020-01213-2

Ezaki, J., Li, J., & Copur-Gencturk, Y. (2023). Teachers' Knowledge of Fractions, Ratios, and Proportional Relationships: the Relationship Between Two Theoretically Connected Content Areas. *International Journal of Science and Mathematics Education*, 1-21. https://doi.org/10.1007/s10763-023-10372-z

Fielding-Wells, J. (2018). Dot plots and hat plots: Supporting young students emerging understandings of distribution, center and variability through modeling. *ZDM*, *50*(7), 1125-1138. https://doi.org/10.1007/s11858-018-0961-1

Finzer, W. (2001). *Fathom Dynamic Statistics* (version 1.0) [Computer software]. Key Curriculum Press.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education* Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Reidel.

Freudenthal, H. (1991). *Revisiting mathematics education*. Kluwer Academic Publishers.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124-158. https://doi.org/10.2307/749671

Frischemeier, D., & Biehler, R. (2018). Preservice teachers comparing groups with TinkerPlots—an exploratory laboratory study. *Statistics Education Research Journal*, *17*(1), 35-60. https://doi.org/10.52041/serj.v17i1.175

Garfield, J. B. (1998). The statistical reasoning assessment: Development and validation of a research tool. *Proceedings of the Fifth International Conference on Teaching Statistics*, Netherlands, (Vol. 2, pp. 781-786). www.stat.auckland.ac.nz/~iase/publications/2/Topic6u.pdf

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3). https://doi.org/10.1080/10691898.2002.11910676

Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: issues challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 397–409). Kluwer Academic Publishers.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer Science & Business Media.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning*, *2*(1-2), 99-125. https://doi.org/10.1207/S15327833MTL0202_5

Garfield, J., & Gal, I. (1999). Teaching and assessing statistical reasoning. *Developing mathematical reasoning in grades K-12*, 207-219.

Gökçe, R. (2019). *Ortaokul matematik öğretmenlerinin istatistiksel akıl yürütmeye ilişkin alan ve pedagojik alan bilgilerinin incelenmesi* [Unpublished doctoral dissertation]. Pamukkale University.

Gravemeijer, K. (1994). Educational development and developmental research in mathematics education. *Journal for Research in Mathematics Education*, 443-471. https://doi.org/10.5951/jresematheduc.25.5.0443

Gravemeijer, K. (2002). Emergent Models as the basis for an instructional sequence on data analysis. In K. Gravemeijer, R. Lehrer, B. Oers, van, & L. Verschaffel (Eds.), *Symbolizing, modeling and tool use in mathematics education* Kluwer Academic Publishers.

Gravemeijer, K., & Cobb, P. (2006). Design research from a learning design perspective. In J. Van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 17–51). Routledge.

Gravemeijer, K., Cobb, P., Bowers, J., & Whitenack, J. (2000). Symbolizing, modeling, and instructional design. In P. Cobb, E. Yackel, & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design* (pp. 225–273). Lawrence Erlbaum Associates.

Gravetter, F., & Wallnau, L. (2015). *Statistics for the behavioral sciences* (10th ed.). Cengage Learning.

Gresalfi, M. S. (2015). Designing to support critical engagement with statistics. *ZDM*, *47*(6), 933-946. https://doi.org/10.1007/s11858-015-0690-7

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning*, *8*(1), 37-63. https://doi.org/10.1207/s15327833mtl0801_3

Groth, R. E., & Meletiou-Mavrotheris, M. (2018). Research on statistics teachers' cognitive and affective characteristics. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International Handbook of Research in Education* (pp. 327–355). Springer International Publishing.

Hall, L. D., Fisher, C., Musanti, S., & Halquist, D. (2006). What can we learn from PT3? *TechTrends*, *50*(3), 25-31. https://doi.org/10.1007/s11528- 006-76003-3

Hammerman, J. K., & Terc, A. R. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal*, *3*(2), 17-41. https://doi.org/10.52041/serj.v3i2.546

Hardiman, P. T., Well, A. D., & Pollatsek, A. (1984). Usefulness of a balance model in understanding the mean. *Journal of Educational Psychology*, *76*(5), 792-801. https://doi.org/10.1037/0022-0663.76.5.792

Heinz, K. R. (2000). *Conceptions of ratio in a class or preservice and practicing teachers* [Unpublished doctoral dissertation]. Penn State University.

Hershkowitz, R., & Schwarz, B. (1999). The emergent perspective in rich learning environments: Some roles of tools and activities in the construction of sociomathematical norms. *Educational Studies in Mathematics*, *39*(1), 149-166. https://doi.org/10.1023/A:1003769126987

Hofmann, T., Maxara, C., Meyfarth, T., Prömmel, A. (2014). Using the software FATHOM for learning and teaching statistics in Germany – A review on the research activities of Rolf Biehler's working group over the past ten years. In Wassong, T., Frischemeier, D., Fischer, P., Hochmuth, & R., Bender, P. (Eds.) *Mit Werkzeugen Mathematik und Stochastik lernen – Using Tools for Learning Mathematics and Statistics*. Springer Spektrum. https://doi.org/10.1007/978-3-658-03104-6_21

Huey, M. E. (2011). *Characterizing middle and secondary preservice teachers' change in inferential reasoning* [Unpublished doctoral dissertation]. University of Missouri.

Huey, M. E., Champion, J., Casey, S., & Wasserman, N. H. (2018). Secondary mathematics teachers' planned approaches for teaching standard deviation. *Statistics Education Research Journal*, *17*(1), 61-84. https://doi.org/10.52041/serj.v17i1.176

Jacobbe, T. (2012). Elementary school teachers' understanding of the mean and median. *International Journal of Science and Mathematics Education*, *10*(5), 1143-1161. https://doi.org/10.1007/s10763-011-9321-0

Jones, G. A., Langrall, C. W., Mooney, E.S., & Thornton, C. A. (2004). Models of development in statistical reasoning. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 97–117). Kluwer Academic Publishers.

Kader, G., & Mamer, J. (2008). Contemporary curriculum issues: Statistics in the Middle Grades: Understanding Center and Spread. *Mathematics Teaching in the Middle School*, *14*(1), 38. https://doi.org/10.5951/MTMS.14.1.0038

Karagöz Akar, G. (2007). Conceptions of between ratios and within ratios [Unpublished doctoral dissertation]. Pennsylvania State University.

Kaput, J. J., & West, M. M. (1994). Missing-value proportional problems: factors affecting informal reasoning patterns. In G. Harel & J. Confrey, (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 235-287). State University of New York Press.

Karatoprak, R. (2014). *Assessing preservice mathematics teachers' statistical reasoning* [Unpublished master thesis]. Boğaziçi University.

Konold, C., Higgins, T., Russell, S.-J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics*, *88*(3), 305–325. https://doi.org/10.1007/s10649-013-9529-8

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, *2*(1). https://doi.org/10.5070/T521000032

Konold, C., & Miller, C. (2011). *TinkerPlots*: *Dynamic data exploration* (Version 2) [Computer software]. Key Curriculum Press.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, 259-289. https://doi.org/10.5951/jresematheduc.33.4.0259

Koparan, T. (2019). Examination of the dynamic software-supported learning environment in data analysis. *International Journal of Mathematical Education in Science and Technology*, *50*(2), 277-291. https://doi.org/10.1080/0020739X.2018.1494861

Krummheuer, G. (1995). The ethnography of argumentation. In P. Cobb, & H. Bauersfeld (Eds.), *Emergence of mathematical meaning* (pp. 229-269). Lawrence Erlbaum.

Kurt, G. (2016). *Technological pedagogical content knowledge (TPACK) development of preservice middle school mathematics teachers in statistics teaching: a microteaching lesson study* [Unpublished doctoral dissertation]. Middle East Technical University.

Lamon, S. J. (1994). Ratio and proportion: Cognitive foundations in unitizing and norming. In G. Harel, & J. Confrey, (Eds.), *The development of multiplicative*

reasoning in the learning of mathematics (pp. 89-120). State University of New York Press.

Lamon, S. J. (1999). *Teaching fractions and ratios for understanding: Essential content knowledge and instructional strategies for teachers*. Lawrence

Lamon, S. J. (2007). Rational numbers and proportional reasoning: Toward a theoretical framework for research. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 629–668). Information Age Publishing.

Landtblom, K., & Sumpter, L. (2021). Teachers and prospective teachers' conceptions about averages. *Journal of Adult Learning, Knowledge and Innovation*, *4*(1), 1–8. https://doi.org/10.1556/2059.03.2019.02

Leavy, A. M. (2006). Using data comparison to support a focus on distribution: Examining preservice teacher's understandings of distribution when engaged in statistical inquiry. *Statistics Education Research Journal*, *5*(2), 89-114. https://doi.org/10.52041/serj.v5i2.502

Leavy, A., & O'Loughlin, N. (2006). Preservice teachers understanding of the mean: Moving beyond the arithmetic average. *Journal of Mathematics Teacher Education*, *9*(1), 53-90. https://doi.org/10.1007/s10857-006-9003-y

Lee, H., Kersaint, G., Harper, S., Driskell, S., & Leatham, K. (2012). Teachers' statistical problem solving with dynamic technology: Research results across multiple institutions. *Contemporary Issues in Technology and Teacher Education*, *12*(3), 286-307. https://www.learntechlib.org/primary/p/39310/

Lee, H. S., Kersaint, G., Harper, S. R., Driskell, S. O., Jones, D. L., Leatham, K. R., Argotti, R. L., & Adu-Gyamfi, K. (2014). Teachers' use of transnumeration in solving statistical tasks with dynamic statistical software. *Statistics Education Research Journal*, *13*(1), 25-52. https://doi.org/10.52041/serj.v13i1.297

Lehrer, R., Kim, M.-J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning*, *12*(3), 195–216. https://doi.org/10.1007/s10758-007-9122-2

Lesh, R., Post, T., & Behr, M. (1988). Proportional reasoning. In J. Hiebert, & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 93-118). Lawrence Erlbaum & National Council of Teachers of Mathematics, Inc.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Sage Publications.

Madden, S. R. (2008). *High school mathematics teachers' evolving understanding of comparing distributions* [Unpublished doctoral dissertation]. Western Michigan University.

Makar, K. (2014). Young children's explorations of average through informal inferential reasoning. *Educational Studies in Mathematics*, *86*(1), 61-78. https://doi.org/10.1007/s10649-013-9526-y

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*, 152–173. https://doi.org/10.1080/10986065.2011.538301

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). Kluwer Academic Publishers.

Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal*, *4*(1), 27-54. https://doi.org/10.52041/serj.v4i1.524

Mallows, C. (1998). The zeroth problem. *The American Statistician*, *52*(1), 1-9. https://doi.org/10.1080/00031305.1998.10480528

Martins, J. A., Estrada Roca, M. A., & Nascimento, M. M. (2014). Do you need to see it to believe it? Let's see statistics and geometry dynamically together! *European Journal of Science and Mathematics Education*, *2*(1), 39-52. http://hdl.handle.net/10459.1/57019

Meletiou-Mavrotheris, M., Paparistodemou, E., & Stylianou, D. (2009). Enhancing statistics instruction in elementary schools: Integrating technology in professional development. *The Montana Mathematics Enthusiast*, *6*(1 & 2), 57-78. https://doi.org/10.54870/1551-3440.1134

Mercimek, O. (2013*). Assessment of preservice mathematics teachers' knowledge for teaching statistics* [Unpublished doctorate dissertation]. Middle East Technical University.

Mickelson, W. T., & Heaton, R. M. (2004). Primary teachers' statistical reasoning about data. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 327–352). Kluwer Academic Publishers.

Miles, M. B., & Huberman, A. M. (1994). *An expanded sourcebook: Qualitative data analysis*. Sage Publications.

Ministry of Education Singapore (2012). Mathematics syllabus. Secondary one to four express course. Normal (academic) course. Singapore: Curriculum planning and development division.

Ministry of National Education, (2018a). *Mathematics course curriculum (Primary and secondary, 1, 2, 3, 4, 5, 6, 7 and 8th grades*). MoNE. http://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=191

Ministry of National Education (MoNE), (2018b). *Mathematics course curriculum (High school, 9, 10, 11 and 12th grades*). MoNE. http://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=191

Mintrop, R. (2016). *Design based school improvement: A practical guide for education leaders*. Harvard Education Press.

Mojica, G. F., Barker, H., & Azmy, C. N. (2019). Instrumented learning in a CODAP-enabled learning environment. In J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística*. https://digibug.ugr.es/bitstream/handle/10481/55215/mojica.pdf?sequence=1&isAllowed=y

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, *26*(1), 20-39. https://doi.org/10.5951/jresematheduc.26.1.0020

Moore, D. S., & McCabe, G. P. (2003). *Introduction to the Practice of Statistics* (4th ed.). W.H. Freeman.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. NCTM. https://www.nctm.org/standards2000/

Nieveen, N., & Folmer, E. (2013). Formative evaluation in educational design research. In T. Plomp & N. M. Nieveen (Eds.), *Educational Design Research*, (pp. 152-169). SLO. https://s2pnd-matematika.fkip.unpatti.ac.id/wp-content/uploads/2022/10/educational-design-research-part-a.pdf

Oikarinen, J. K., Järvelä, S., & Kaasila, R. (2014). Finnish upper secondary students' collaborative processes in learning statistics in a CSCL environment. *International Journal of Mathematical Education in Science and Technology*, *45*(3), 325-348. https://doi.org/10.1080/0020739X.2013.837519

Ölmez, İ. B. (2016). Two distinct perspectives on ratios: Additive and multiplicative relationships between quantities. *Elementary Education Online, 15*(1), 186–203. https://doi.org/10.17051/io.2016.94175

Ölmez, İ.B. (2018). Future Teachers' Use of Multiplication and Fractions When Expressing Proportional Relationships. In Strutchens, M., Huang, R., Potari, D., Losano, L. (Eds.) *Educating Prospective Secondary Mathematics Teachers*. ICME-13 Monographs. Springer, Cham. https://doi.org/10.1007/978-3-319-91059-8_13

Ölmez, İ. B. (2022). Preservice teachers' understandings of division and ratios in forming proportional relationships. *Mathematics Education Research Journal*, 1-25. https://doi.org/10.1007/s13394-022-00410-4

Patton, M.Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Sage Publications.

Peters, S. A. (2014). Developing understanding of statistical variation: Secondary statistics teachers' perceptions and recollections of learning factors. *Journal of Mathematics Teacher Education*, *17*, 539–582. https://doi.org/10.1007/s10857-013-9242-7

Pfannkuch, M. (2007). Year 11 students' informal inferential reasoning: A case study about the interpretation of box plots. *International Electronic Journal of Mathematics Education*, *2*(3), 149–167. https://doi.org/10.29333/iejme/181

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal*, *5*(2), 4–9. https://doi.org/10.52041/serj.v5i2.496

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). Kluwer Academic Publishers.

Plomp, T. (2010). Educational design research: An introduction. In T. Plomp & N. M. Nieveen (Eds.), *An introduction to educational design research* (pp. 9-35). SLO.

Rasmussen, C., & Stephan, M. (2008). A methodology for documenting collective activity. In A. E. Kelly, R. A. Lesh, & J. Y. Baek (Eds.), *Handbook of design*

research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching (pp. 195-216). Routledge.

Reading, C. & Canada, D. (2011). Teachers' Knowledge of Distribution, In C. Batanero, G. Burrill, and C. Reading (Eds.), Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study, (pp. 223-234), New York.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal*, *5*(2), 46-68. https://doi.org/10.52041/serj.v5i2.500

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). Kluwer Academic Publishers.

Rosidah, B. I., & Juniati, D. (2018). An analysis of statistical reasoning process of high school students in solving the statistical problem. *Second International Conference on Statistics, Mathematics, Teaching, and Research*, Indonesia, 1028. https://doi.org/10.1088/1742-6596/1028/1/012125

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3). https://doi.org/10.1080/10691898.2002.11910678

Saidi, S. S., & Siew, N. M. (2019). Assessing students' understanding of the measures of central tendency and attitude towards statistics in rural secondary schools. *International Electronic Journal of Mathematics Education*, *14*(1), 73-86. https://doi.org/10.12973/iejme/3968

Savin-Baden, M., & Major, C. (2013). *Qualitative Research: The essential guide to theory and practice*. Routledge.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Psychology Press.

Selmer, S. J., Bolyard, J. J., & Rye, J. A. (2012). Statistical reasoning over lunch. *Mathematics Teaching in the Middle School*, *17*(5), 274-281. https://doi.org/10.5951/mathteacmiddscho.17.5.0274

Shaughnessy, M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph & K. Carr (Eds.), *Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 6–22). Rotorua, NZ.

Shaughnessy, J. M. (2001). *Draft of grant proposal sent to the National Science Foundation.*

Shaughnessy, J. M. (2007). Research on statistics' reasoning and learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Information Age Publishing.

Shin, D. (2019). A framework for understanding how preservice teachers notice students' statistical reasoning about comparing groups. *International Journal of Mathematical Education in Science and Technology*, 1-22. https://doi.org/10.1080/0020739X.2019.1699968

Silva, C., & Coutinho, C. (2008). Reasoning about variation of a univariate distribution: A study with secondary mathematics teachers. In C. Batanero, G. Burril, C. Reading, & A. Rossman (Eds.) *Joint ICMI/IASE study: teaching statistics in school mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study*, *18*. https://iase-web.org/documents/papers/rt2008/T2P7_Silva.pdf

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, *26*(2), 114-145. https://doi.org/10.5951/jresematheduc.26.2.0114

Simon, M. (2000). Research on the development of mathematics teachers: The Teachers development experiments. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 335–360). Lawrence Erlbaum Associates.

Simon, M. (2020). Hypothetical learning trajectories in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 354–357). Springer.

Sorto, M. A. (2004). Prospective middle school teachers' knowledge about data analysis and its application to teaching [Unpublished doctoral dissertation]. Michigan State University.

Steffe, L. P. (1994). Children's multiplying schemes. In G. Harel & J. Confrey, (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 3- 39). State University of New York Press.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly &R. A. Lesh (Eds.), *Research design in mathematics and science education* (pp. 267-306). Lawrence Erlbaum Associates, Inc.

Stephan, M. (2020). Sociomathematical norms in mathematics education. *Encyclopedia of mathematics education*, 802-805.

Stephan, M., & Rasmussen, C. (2002). Classroom mathematical practices in differential equations. The Journal of Mathematical Behavior, 21(4), 459-490.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Treffers, A. (1987). *Three dimensions. A model of goal and theory description in mathematics instruction - The Wiskobas project*. Reidel Publishing Company.

Turan, N. (2017). *An examination of elementary and lower secondary level Japanese and Turkish mathematics curricula through quantitative and (co)variational reasoning in terms of the treatment of functional relationships* [Unpublished doctoral dissertation]. Boğaziçi University.

Turegun, M. (2011). *A model for developing and assessing community college students' conceptions of the range, interquartile range, and standard deviation* [Unpublished doctoral dissertation]. University of Oklahoma.

Uccellini, J. C. (1996). Teaching the mean meaningfully. *Mathematics Teaching in the Middle School*, *2*(2), 112-115. https://doi.org/10.5951/MTMS.2.2.0112

Van de Walle, J. A., & Karp, K. S. Bay Williams, J. M. (2019). *Elementary and middle school mathematics: Teaching developmentally* (10th ed.). Pearson Education.

Van den Akker, J. (2010). Curriculum design research. In T. Plompt, & N. Nieveen, An Introduction to Educational Design Research (pp. 37-49). Enschede: SLO.

Vergnaud, G. (1988). Multiplicative structures. In J. Hiebert, & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 141-161). Lawrence Erlbaum & National Council of Teachers of Mathematics, Inc.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Lawrence Erlbaum.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3-46. https://doi.org/10.52041/serj.v2i2.553

Watson, J., Fitzallen, N., English, L., & Wright, S. (2020). Introducing statistical variation in year 3 in a STEM context: Manufacturing licorice. *International Journal of Mathematical Education in Science and Technology*, *51*(3), 354-387. https://doi.org/10.1080/0020739X.2018.1562117

Wilburne, J. M., & Kulbacki, A. (2014). Connecting the "Missing Words" to the Common Core. *Mathematics Teaching in the Middle School*, *19*(7), 430-436. https://doi.org/10.5951/mathteacmiddscho.19.7.0430

Wild, C. J. (2006). The concept of distribution. *Statistics Education Research Journal, 5*(2), 10–26. https://doi.org/10.52041/serj.v5i2.497

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical inquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, *27*(4), 458-477. https://doi.org/10.5951/jresematheduc.27.4.0458

Yılmaz, N. (2019). *Öğretmen adaylarının istatistiği öğretme bilgilerinin öğretmenlik uygulaması temelli ders araştırmaları bağlamında incelenmesi* [Unpublished doctoral dissertation]. Hacettepe University.

Yolcu, A. & Haser, C. (2013). 8th grade students' statistical literacy of average and variation concepts. *Proceedings of Eighth Congress of European Research in Mathematics Education*, Turkey, 912-921. pdf of proceeding on ERME.site

Zieffler, A., Garfield, J., & Fry, E. (2018). What is statistics education? In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 37-70). Springer.

APPENDICES

A. Elementary Mathematics Education Curriculum

FIRST-YEAR

First Semester			Second Semester		
MATH111	Fundamentals of Mathematics	(3-0)3	MATH112	Discrete Mathematics	(3-0)3
MATH115	Analytic Geometry	(3-0)3	MATH116	Basic Algebraic Structures	(3-0)3
MATH117	Calculus I	(4-2)5	MATH118	Calculus II	(4-2)5
CEIT100	Computer Applications in Education	(2-2)3	BA100	Career Planning	(1-0)0
EDS200	Introduction to Education	(3-0)3	CEIT101	Introduction to Computer Programming	(2-2)3
ENG101	English for Academic Purposes I	(4-1)4	ENG102	English for Academic Purposes II	(4-1)4

SECOND-YEAR

Third Semester			Fourth Semester		
PHYS181	Basic Physics I	(4-2)5	PHYS182	Basic Physics II	(4-2)5
MATH219	Introduction to Differential Equations	(4-0)4	MATH201	Elementary Geometry	(3-0)3
STAT201	Introduction to Probability & Stat. I	(3-0)3	STAT202	Introduction to Probability & Stat. II	(3-0)3
EDS220	Educational Psychology	(3-0)3	MSE241	Introduction to Mathematics Education	(3-0)3
MSE221	Instructional Principles and Methods	(2-2)3	ENG211	Academic Oral Presentation Skills	(3-1)3
HIST2205	History of the Turkish Revolution I	(2-0)0	HIST2206	History of the Turkish Revolution II	(2-0)0
HIST2201	Principles of Kemal Atatürk I	(2-0)0	HIST2202	Principles of Kemal Atatürk II	(2-0)0

THIRD-YEAR

Fifth Semester			**Sixth Semester**		
MATH260	Basic Linear Algebra	(3-0)3	EDS304	Classroom Management	(3-0)3
MSE288	Community Service	(1-4)3	MSE329	Instructional Technology and Material Development	(2-2)3
MSE305	Assessment of Learning in Science and Mathematics	(3-0)3	MSE342	Methods of Teaching Mathematics in Middle Schools II	(2-2)3
MSE341	Methods of Teaching Mathematics in Middle Schools I	(2-2)3	TURK108	Turkish II	(2-0)2
			TURK306	Written Expression	(2-0)2
TURK107	Turkish I	(2-0)2		Restricted Elective	
TURK305	Oral Communication	(2-0)2		Elective	
	Elective				

FOURTH-YEAR

Fifth Semester			**Sixth Semester**		
MATH303	History of Mathematical Concepts I	(3-0)3	EDS416	Turkish Educational System and School Management	(3-0)3
MSE417	School Experience	(1-4)3	EDS424	Guidance	(2-2)3
MSE431	Research Methods in Education	(3-0)3	MSE418	Practice Teaching	(2-6)5
MSE441	Methods of Teaching Mathematics in Middle Schools III	(2-2)3		Elective	
				Elective	
ENG311	Advanced Communication Skills	(3-1)3			
	Restricted Elective				

B. Ethical Approval

UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr
www.ueam.metu.edu.tr

Sayı: 28620816 /

22 MART 2021

Konu : Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi : İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Prof.Dr. Mine IŞIKSAL BOSTAN

Danışmanlığını Yürüttüğünüz Nazlı AKAR'ın "Ortaokul Matematik Öğretmen Adaylarının Dağılım Kavramı Öğretimi İçin Bilgilerinin Geliştirilmesi" başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülmüş ve **081-ODTU-2021** protokol numarası ile onaylanmıştır.

Saygılarımızla bilgilerinize sunarız.

Dr.Öğretim Üyesi Ali Emre TURGUT
İAEK Başkan Vekili

C. Permission for Activities

Re: Distribution teaching

Cobb, Paul A göndericisinden 2020-12-31 01:04 tarihinde
Ayrıntılar Düz Metin

Durham Stats Unit.zip (~1,9 MB)

Dear Nazli,

We finished that work 15 years ago and it took me some time to find the attached folder.

Briefly, the teachers with whom we worked put together a set of instructional activities for the other math teachers in their school district. The fold contains the complete set of activities and associated teacher notes that they created with our assistance. I hope it will be useful to you.

With good wishes for 2021.

Paul

Re: Distribution teaching

koeno gravemeijer göndericisinden 2020-12-23 18:43 tarihinde
Ayrıntılar

Dear Nazli,
Feel free to use the learning trajectory and the tasks, and of course we (Paul Cobb and I) would like you to refer to our work if you do so.
With kind regards,
Koeno Gravemeijer

D. Consent Form (Turkish Version)

Bu araştırma, ODTÜ Fen Bilimleri Enstitüsü Matematik ve Fen Bilimleri Bölümü Matematik Eğitimi doktora öğrencisi Nazlı AKAR tarafından Prof. Dr. Mine IŞIKSAL BOSTAN danışmanlığında yürütülen bir tez çalışmadır. Bu form sizi araştırma koşulları hakkında bilgilendirmek için hazırlanmıştır.

Çalışmanın Amacı Nedir?
Araştırmanın amacı, ortaokul matematik öğretmen adaylarının dağılım kavramına ilişkin akıl yürütme gelişimini incelemektir.

Bize Nasıl Yardımcı Olmanızı İsteyeceğiz?
Araştırmaya katılmayı kabul ederseniz, sizden temel istatistik bilgilerinizi ölçen bir testi cevaplamanız ve bu testteki cevaplarınızla ilgili görüşmeye katılmanız beklenmektedir. Ardından kayıtlı olduğunuz derste size istatistik eğitimi ile ilgili eğitim verilecektir. Bu ders kapsamında, istatistiksel kavramların öğretimi ile ilgili proje oluşturmanız ve ders sürecinde size verilen çeşitli görevleri yerine getirmeniz beklenmektedir. Görüşmelerde sorulara verdiğiniz yanıtlarda ve dersler boyunca çevrimiçi öğrenme ortamında gerçekleştirilecek aktivitelerde ortaya çıkacak söylemlerde veri kaybı olmaması ve veri toplama sürecini hızlandırmak amacıyla görüşmeler ve ders süreci kayıt altına alınacaktır.

Sizden Topladığımız Bilgileri Nasıl Kullanacağız?

Araştırmaya katılımınız tamamen gönüllülük temelinde olmalıdır. Görüşmelerde sizden kimlik veya öğrenim gördüğünüz bölümü belirleyici hiçbir bilgi istenmemektedir. Cevaplarınız tamamıyla gizli tutulacak, sadece araştırmacılar tarafından değerlendirilecektir. Katılımcılardan elde edilecek bilgiler toplu halde değerlendirilecek ve bilimsel yayımlarda kullanılacaktır. Sağladığınız veriler gönüllü katılım formlarında toplanan kimlik bilgileri ile eşleştirilmeyecektir.

Katılımınızla ilgili bilmeniz gerekenler:

Çalışma, genel olarak kişisel rahatsızlık verecek sorular içermemektedir. Ancak, katılım sırasında sorulardan ya da herhangi başka bir nedenden ötürü kendinizi rahatsız hissederseniz cevaplama işini yarıda bırakıp çıkmakta serbestsiniz. Böyle bir durumda çalışmayı uygulayan kişiye, çalışmadan çıkmak istediğinizi söylemek yeterli olacaktır.

Araştırmayla ilgili daha fazla bilgi almak isterseniz:

Bu çalışmaya katıldığınız için şimdiden teşekkür ederiz. Çalışma hakkında daha fazla bilgi almak için ODTÜ öğretim elemanlarından Prof. Dr. Mine IŞIKSAL BOSTAN (E-posta: _____) ya da Nazlı AKAR (E-posta: _____) ile iletişim kurabilirsiniz.

Yukarıdaki bilgileri okudum ve bu çalışmaya tamamen gönüllü olarak katılıyorum.

(Formu doldurup imzaladıktan sonra uygulayıcıya geri veriniz).

Ad Soyad Tarih İmza

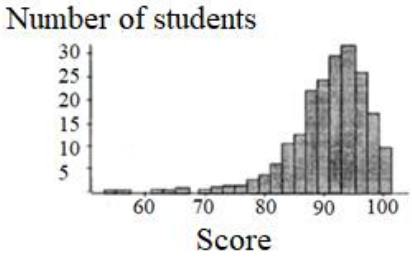
---/----/-----

E. Pretest

1) Below is given the total number of books read by 4th-grade students in different classes in a school during a semester. The principal wants to go on a trip with the better class in reading. With which class do you think the principal should go on a trip? Explain your answer with its reasons.

Class	The total number of books students read during a semester
4A	11, 13, 9, 8, 9
4B	7, 6, 12, 9, 9, 11
4C	10, 9, 9, 10, 7, 12
4D	11, 13, 10, 2, 12

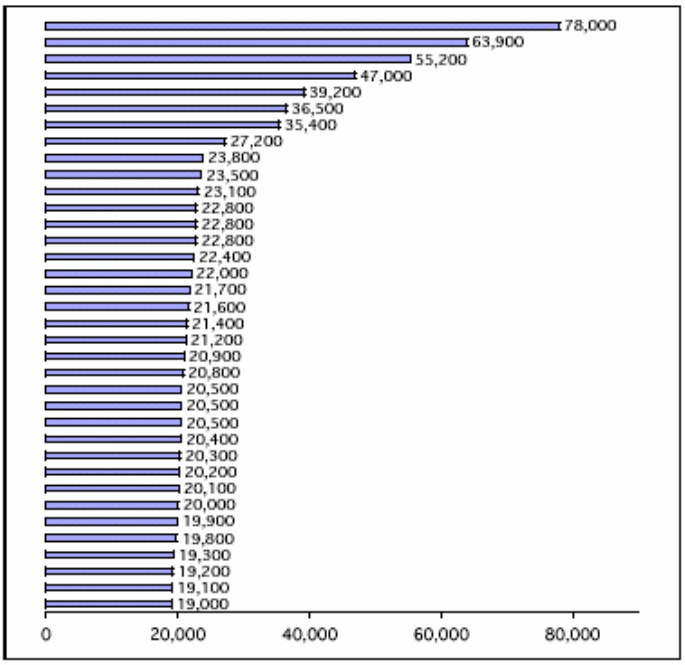
2)



The graph above shows the distribution of the scores of 238 eighth-grade students in a middle school in the mathematics exam. Mark TRUE, FALSE, or CAN'T TELL **for each of the following statements** according to the graph.

	TRUE	FALSE	CAN'T TELL
The distribution is skewed right.			
The median is greater than the mean.			
Exactly half of the schools are above average.			
More than half of the schools are above average.			
Less than half of the schools are above average.			
The exam is easy.			

3) The graph below shows salaries of 36 individuals who graduated last year from the Hurley School of Graphic Design. These are the salaries they earned the first year after graduating.



The School wants to publish this information in their newsletter to give current students an idea of how much they might expect to earn their first year working. But they don't have room to print the entire graph.

Part 1. Below are three possible summaries they are considering for their newsletter. Which summary do you think will give students the best sense of how much they might expect to each their first year working?

____ a. "The salaries for last year's graduates were quite varied and ranged all the way from a low of $19,000 to a high of $78,000."

____ b. "Last year's graduates earned an average (mean) salary the first year of $27,000. The highest salary was 78,000."

____ c. "The majority of last year's graduates made between $19,000 and $23,800 their first year. A few made considerably more than that, with one earning 78,000."

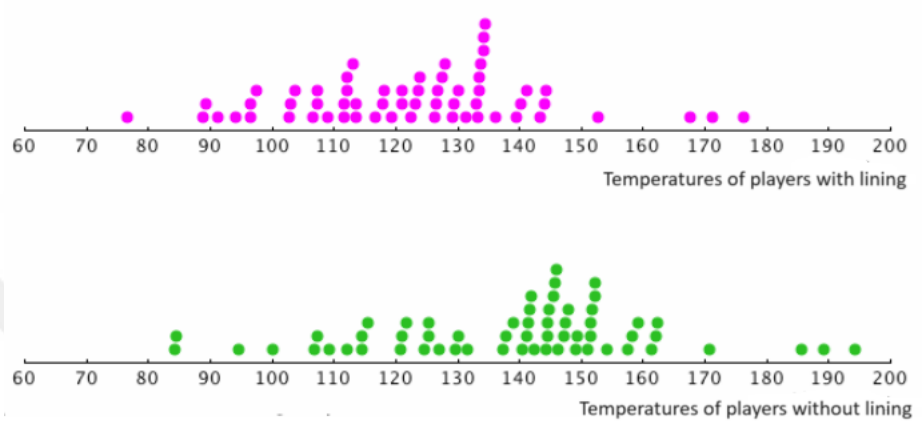
Part 2. Briefly say why you wouldn't choose the other two summaries.

I wouldn't choose summary ___ because ----------

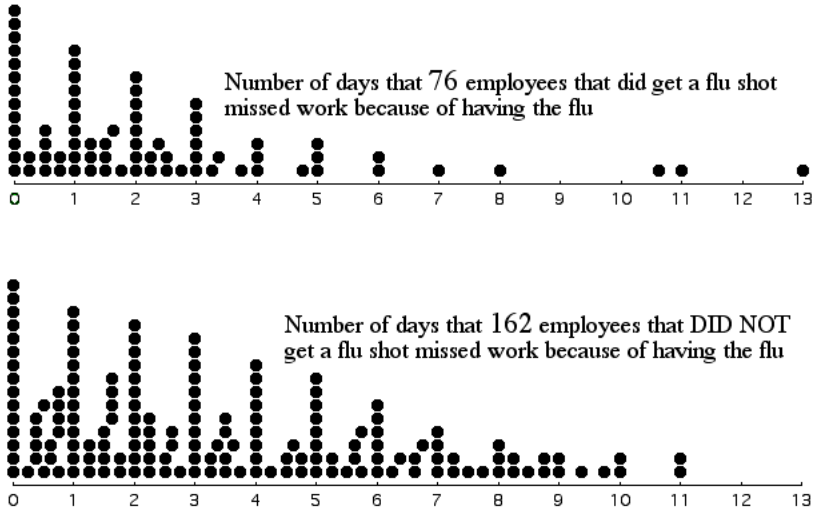
I wouldn't choose summary ___ because ----------

---

4) The National Football League is concerned about the health problems of players caused by heat exhaustion. They are doing research on the use of liners in helmets to reduce the amount of heat that players are exposed to. They conducted a test with two sets of 60 players, each of which participated in a two-hour practice. One group had liners and the other did not. At the end of the two hours, temperatures inside all helmets were recorded as shown in the data sets. While the helmets of the upper group have lining, the helmets of the other group do not have lining. Do you think the heat lining in the helmet reduces the amount of heat the players are exposed to? Explain your answer with its reasons.



5) The CEO of a company is considering providing free immunization flu shots for employees. But before making a decision about whether to buy flu vaccine, she wants to know if employees who have a flu shot will be absent from work for a fewer number of days. Currently, some workers are getting the shots by paying themselves, and others are not getting them. The CEO has gathered information on the number of days that each of the 238 Company employees was absent in a year due to the flu. She has separated the data in two groups, one with people who got a flu shot (76 employees) and another with people who did not get a flu shot (162 employees). The graphs below show these data. Each dot represents the number of days that one employee missed work because he or she had the flu. The first graph shows the data for the employees who got a flu shot, and the other shows the data for employees that did not get a flu shot. Look at the two graphs carefully.



The CEO has asked four different company executives to analyze the data in order to determine if the flu shots made a difference or not in the number of days that employees missed work. Read each of the analysis carefully and answer the questions.

A) As the table below shows, of the 28 people that had zero absentees 15 did not get a flu shot and 13 did. This indicates that flu shots were not effective because more people that were not immunized had zero absentees.

	Number of employees that had zero absentees
Employees that got a flu shot	13
Employees that did not get a flu shot	15
Total	28

Do you consider that this analysis would adequately inform the CEO about the data and help her make a decision?

Yes No

In the space below write a paragraph explaining your decision.

B) As the table below shows, 41% of the people who got a flu shot missed work one day or less, versus 26% of the people that did not get a flu shot. The flu shots seemed to have reduced the number of days that employees missed work.

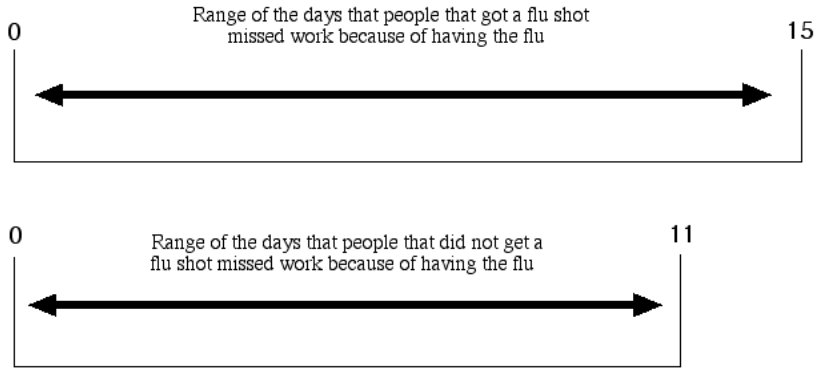
	Employees that missed work one day or less	Employees that missed work more than one day
Employees that got a flu shot	31 (41%)	45 (59%)
Employees that did not get a flu shot	42 (26%)	120 (74%)

Do you consider that this analysis would adequately inform the CEO about the data and help her make a decision?

Yes No

In the space below write a paragraph explaining your decision.

C) As the graphs below shows, the number of days that people who got a flu shot were absent from zero to fifteen, and the number of days that those who did not get the shots were absent ranged from zero to eleven. Viewed this way, the flu shots do not seem to have reduced the number of days that employees missed work.

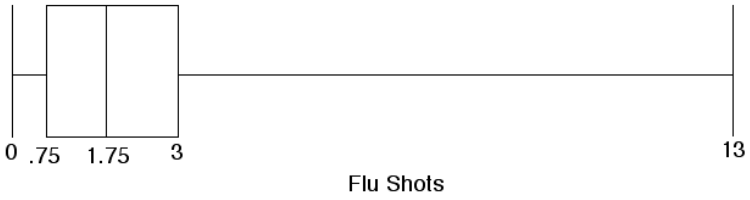


Do you consider that this analysis would adequately inform the CEO about the data and help her make a decision?

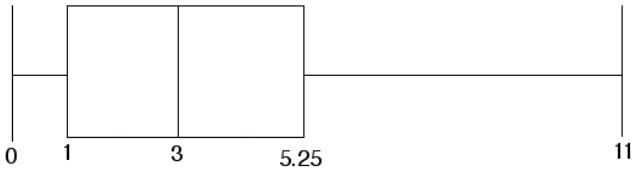
Yes No

In the space below write a paragraph explaining your decision.

D) As the graphs below show, the number of days that people who got a flu shot were absent are clustered between zero and three days, whereas the number of absences of people who did not take a flu shot are more spread out. Only 1/4 of the people who got a flu shot missed more than three days of work, whereas 1/2 of the people who did not get a flu shot were absent for more than three days. This indicates that the flu shots seemed to have reduced the number of days that employees missed work.



Flu Shots



No Flu Shots

Do you consider that this analysis would adequately inform the CEO about the data and help her make a decision?

Yes No

In the space below write a paragraph explaining your decision.

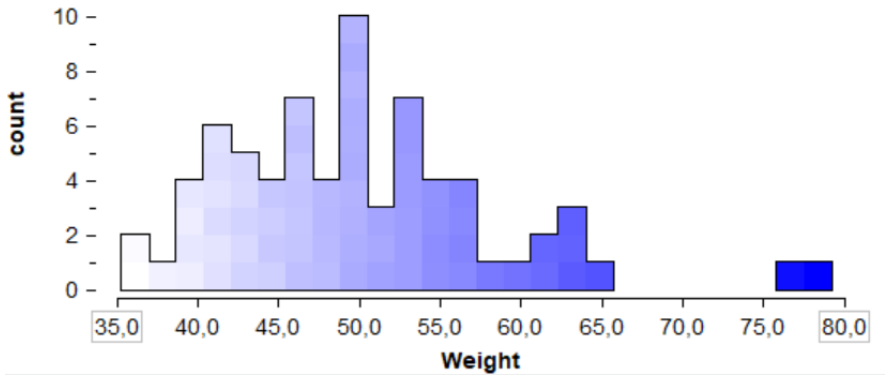
If you had to decide between the four statements, which one would you consider that would best inform the CEO about the data: A, B, C, or D? _____ (choose only one).

F. Posttest

1) Below is given the total number of books read by 4th-grade students in different classes in a school during a semester. The principal wants to go on a trip with the better class in reading. With which class do you think the principal should go on a trip? Explain your answer with its reasons.

Class	The total number of books students read during a semester
4A	11, 13, 9, 8, 9
4B	7, 6, 12, 9, 9, 11
4C	10, 9, 9, 10, 7, 12
4D	11, 13, 10, 2, 12

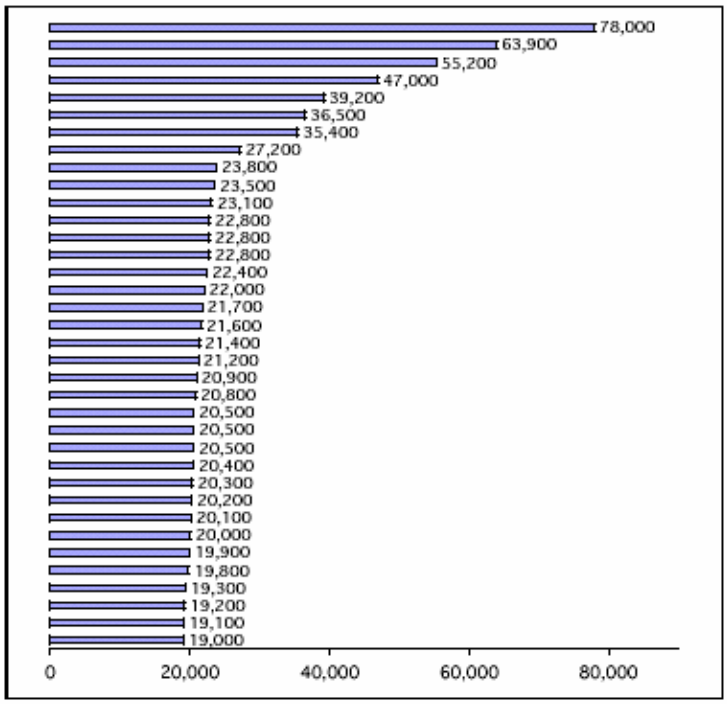
2)



The graph above shows the distribution of the weight of 71 eighth-grade students in a middle school. Mark TRUE, FALSE, or CAN'T TELL **for each of the following statements** according to the graph and explain reason of your choice.

	TRUE	FALSE	CAN'T TELL
The distribution is skewed right.			
Explanation:			
The median is greater than the mean.			
Explanation:			
Exactly half of the students are below the average.			
Less than half of the students are above average.			
Explanation:			

3) A college newspaper conducted a survey of college graduates who had finished their studies one year ago. These graduates were asked how much they earned during their first year working. The goal of the study was to collect and publish information that would help college students get a sense of how much money they could expect to earn once they graduated. The newspaper editors considered that this information could be especially useful for freshman and junior students when deciding which major to choose. The graph below shows the information that was collected on 36 graduates who had majored on graphic design. Each bar shows the amount of dollars that each of the 36 graduates earned in their first year working as graphic designers.



Part 1. Below are four possible ways of reporting the data shown in the graph to their readers. Which statement do you think will give student readers useful information?

____ a. "Our data shows that entering salary for a graphic designer ranges between $19,000 and $78,000 a year."

____ b. "Our data shows that the twenty-five percent of the people entering the field of graphic design that earned the least, earned between $19,000 and $20,200 in a year. The fifty percent in the middle earned between $20,300 and $23,500, and the twenty-five percent that earned the most earned between $23,800 and $78,000."

____ c. "Our data indicates that the average annual salary for a graphic designer entering the field was of $27,000."

____ d. "Our data shows that most people entering the field of graphic design made between $19,000 and $23,800 in their first year working."

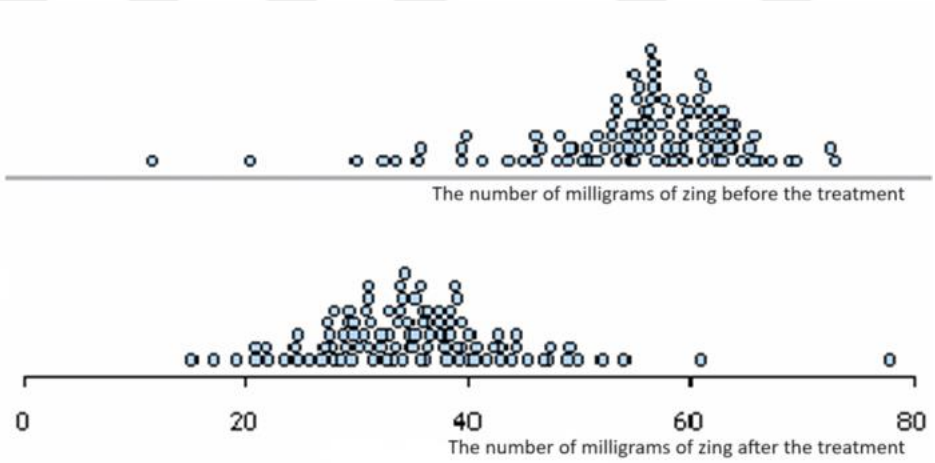
Part 2. Briefly say why you wouldn't choose the other three statements.
I wouldn't choose summary ___ because ----------

I wouldn't choose summary ___ because ----------

I wouldn't choose summary ___ because ----------

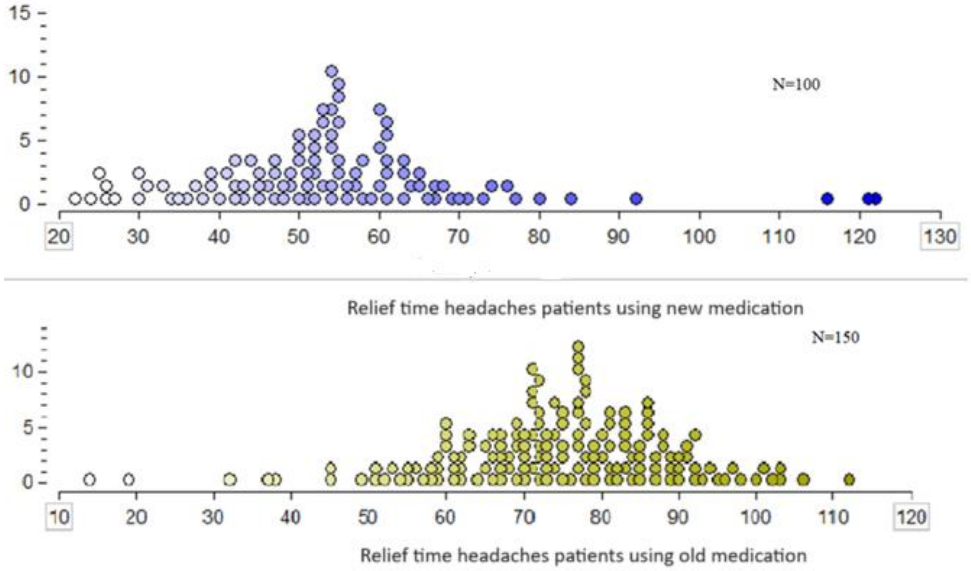
---

4) Many rivers and streams contain harmful substances. Zinc in a river, for example, can damage and kill a number of plants and animals, including the rare Black-footed ferret. The U.S. Geological Society is testing a method of removing zinc from rivers. They tested this treatment method by sampling from 100 different sites along a river known to have high levels of zinc. From each test site, the researchers collected exactly a liter of water and then measured the number of milligrams of zinc in that liter. They did this both before and then after using the new treatment method. The results are shown below.



Based on these results, do you think the new zinc treatment method is effective? Explain your answer with its reasons.

5) A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, they gave it to 100 people with headaches and timed how many minutes it took for the patient to report that the headache had gone. They compared the result from this test to previous results from 150 patients using the old formula under the exact same conditions. The results from both these clinical trials are shown below.



Part 1. Based on these results, write a short summary of what these data say about the effectiveness of the new treatment compared to the old. The summary is for the drug company who wants to decide whether to start marketing the new formula.

Part 2. Justify your summary quantitatively by using graphical representations. (*You can use TinkerPlots to create graphs. Data is given at the end of the test*.)

Part 3. Below are several statements about these results. For each statement, say whether you agree or disagree, and give a brief explanation for your answer.

a. You can't conclude anything from this study. In both groups some people got relief fairly quickly and others took a long time. So whichever formula you use, you can't predict how long the headache will last.	___I agree ___I disagree	Because…
b. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to 0 who took the new formula. Also, the worst result — near 120 minutes — was with the new formula.	___I agree ___I disagree	Because…
c. The average time for the new formula is lower than the average time for the old formula. I'd conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.	___I agree ___I disagree	Because…
d. I wouldn't conclude anything from these data. Because there are different numbers of patients in the two groups, there's no fair way to compare the two formulas.	___I agree ___I disagree	Because…

CURRICULUM VITAE

Surname, Name: Name: Akar, Nazlı

EDUCATION

Degree	Institution	Year of Graduation
MS	Balıkesir University Elementary Mathematics Education	2018
BS	Akdeniz University Elementary Mathematics Education	2014
High School	Elmalı Anadolu High School, Antalya	2010

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

Published Papers in International Indexed Journals

1. Akar, N. and Işıksal-Bostan, M. "The didactic transposition of quadrilaterals: the case of 5th grade in Turkey", International Journal of Mathematical Education in Science and Technology, 1-22 (2022)

2. Akar, N. and Dikkartın Övez, F. T. "An Anthropological Analysis of the Knowledge on Graphics within Middle School Mathematics", Journal of Education and Future, 9(13), 95–119 (2018).

3. Dikkartın Övez, F. T. and Akar, N. "Anthropological analysis of content knowledge of pre-service elementary mathematics teachers on graphs", Educational Research and Reviews, 13(18), 281–306 (2018)

4. Dikkartın Övez, F. T. and Akar, N. "The investigation of process of teaching function concept in an Adidactic learning environment", Pegem Eğitim ve Öğretim Dergisi, 8(3), 469–50 (2018)

Paper Presentations in International Conferences

1. Akar, N. and Işıksal-Bostan, M. "Preservice mathematics teachers reasoning about data analysis Balloon activity", Congress of The European Socıety for Research in Mathematıcs Educatıon (CERME, 2023), Budapeşte, Macaristan

2. Akar, N. and Işıksal-Bostan, M. "The knowledge to be taught a novice mathematics teacher plans to teach quadrilaterals in 5th grade", International Congress on Mathematical Education (ICME, 2021), Shanghai, Tailand

3. Akar, N. and Işıksal-Bostan, M. "Preservice middle school mathematics teachers statistical reasoning about distribution", Psychology of Mathematics Education (PME, 2021), Khon Kaen, China

4. Akar, N. "Öğretmen eğitiminde yaşam boyu öğrenme araştırmaları", Uluslararası Necatibey Eğitim ve Sosyal Bilimler Araştırmaları Kongresi (UNESAK 2018), Balıkesir, Türkiye

5. Dikkartın Övez, F. T. and Akar, N. "Ortaokul 7 ve 8 sınıf öğrencilerinin grafikler konusundaki hata ve kavram yanılgıları," Uluslararası Eğitim Bilimleri Kongresi (UEBK, 2018), Antalya, Türkiye.

6. Akar, N. and Dikkartın Övez, F. T. "Anthropological analysis of graphic content knowledge pre-service elementary mathematics teachers," International Conference on New Horizons (INTE, 2017), Berlin, Germany

7. Dikkartın Övez, F. T. and Akar, N. "Adidaktik Bir Öğrenme Ortamında Fonksiyon Kavramının Öğretimi", Uluslararası Eğitim Bilimleri Kongresi (UEBK, 2016), Antalya, Türkiye

Book Chapters

Şahin, H. B., Akar, N. and Fidan, Ü. "İzometrik ve Ortografik Çizimler". In A. N. Elçi, E. Bukova Güzel, B. Cantürk Günhan and E. Ev Çimen (Eds). Temel matematiksel kavramlar ve uygulamaları öğrenen ve öğretenler için, 605 -615 (2016). Pegem Akademi

Violin, Puzzle, Travelling, Computer Technologies, Movies