Spring 5-12-2017

Assessing Students' Understanding of Variability and Graph Interpretation Through an Authentic Science Investigation

William M. Schlager
University of Maine - Main, wmschlager@gmail.com

ASSESSING STUDENTS' UNDERSTANDING OF VARIABILITY AND GRAPH INTERPRETATION THROUGH AN AUTHENTIC SCIENCE INVESTIGATION

By

William Schlager

B.S. Northland College, 2010

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Teaching

The Graduate School

The University of Maine

May 2017

Advisory Committee:

Molly Schauffler, Assistant Professor of Earth and Climate Science, University of Maine RiSE Center, Climate Change Institute, Advisor

Sarah Nelson, Associate Professor in the School of Forest Resources, Ecology and Environmental Sciences Program, University of Maine RiSE Center

Eric Pandiscio, Associate Professor Mathematics Education, College of Education and Human Development

ASSESSING STUDENTS' UNDERSTANDING OF VARIABILITY AND GRAPH INTERPRETATION THROUGH AN AUTHENTIC SCIENCE INVESTIGATION

By William Schlager

Thesis Advisor: Dr. Molly Schauffler

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Teaching
May 2017

This thesis research combined efforts of two existing projects at the University of Maine in collaboration with the Schoodic Institute, the Acadia Learning Snowpack Project and the Maine Data Literacy Project. The Snowpack Project provided a context to explore student learning of variability and graphing skills by gathering data on snowfall and accumulation throughout the winter and using the data to ask and answer a scientific question. The Maine Data Literacy Project provided a framework and instruments for assessing students' understanding of variability and graph interpretation skills.

The first goal of this research was to measure student learning about variability during the Snowpack Project. The study used a pretest posttest design and the multiple-choice *ASK-Var* assessment developed by the Maine Data Literacy Project. Data were first collected in January and May of 2015. When no differences were found, additional data from Snowpack Project students the following September and a separate group of seventh graders were analyzed to give a broader context.

The second goal of this research was to compare the multiple-choice *ASK-Var* assessment to an open-response assessment. This analysis used a correlation to measure how predictive success on the *ASK-Var* assessment was to success on the open-response assessment.

The third goal of thesis research was to describe what the results of both assessments revealed about student thinking around variability. This uses qualitative analyses to identify patterns in student thinking about histograms, box plots, and graph choice.

No quantitative differences were found between students before and after participating in the snowpack project, however there was some evidence suggesting that the high school Snowpack Project students did perform better than the seventh grade students. Data on the *ASK-Var* assessment and the open-response assessment correlated, but randomness under the surface suggested that there were skills being tested in the open-response assessment that were not being measured by the *ASK-Var* assessment. Finally, the qualitative analysis suggested that while students were generally able to read frequency plots, they sometimes inappropriately applied important context to their interpretations. The graph construction task revealed a split among students' ability to interpret their own graphs. Those who chose to display the data in frequency plots had a higher rate of success in accurately interpreting the results. This study offers insights into applications of the *ASK-Var* assessment and student thinking about graphing and variability.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Molly Schauffler for her sage advice, thoughtful perspectives, and abundant patience. I could not have made it through this process without her support and occasional goading. I would also like to thank my committee, Sarah Nelson and Eric Pandiscio, for their support along the way. They were always happy to help with anything I needed.

Several others deserve recognition for their integral roles in this process. I could not have done this without the support of the teachers who took time out of their busy school schedules to help me gather data. The guidance and support from the RiSE center research group, faculty, staff, and students was invaluable. I would especially like to thank Laura Millay for helping me process my data even if I showed up at 4:30, Beth Byers-Small for guiding me through challenges in the classroom, and Adam Rodgers for his analytical tools. This work was supported financially by the National Oceanic and Atmospheric Administration B-WET program and the Davis Family Foundation.

A special thank you goes to the members of the Rubric Taskforce, Rachel Martin and Marina Van der Eb, who brought sanity to the countless hours of rubric development and scoring open-response questions and for all of the support and encouragement throughout the rest of process. Finally, three thirds of the Rubric taskforce is on board.

Finally, I want to thank my wife for her mental, physical, and editorial support throughout the process. I love you and could not have done it without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CoCoRaHS	Community Collaborative Rain, Hail, and Snow Network
EDA	Exploratory Data Analysis
ISI	Informal Statistical Inference
MDLP	Maine Data Literacy Project
NGSS	Next Generation Science Standards

1 INTRODUCTION

The modern world is data-driven. Data are used to sell everything from cars to college educations, report the news, and advocate for important policy changes in government. Citizens who are able to understand and interpret what those data mean are in a better position to make informed decisions than those who are not data literate.

A data literate person has the skills to collect, organize, and summarize data in a logical manner. He or she can use that information to answer a question or make an informed decision that demonstrates an understanding of limitations inherent to the data set and/or its presentation. A data literate person understands that a mean alone may not represent a set of data well and that the variability of a data set may be lost in a bar graph. These skills are essential for all citizens, not just professionals who work with data like scientists and business professionals.

A solid conceptual foundation in key statistical ideas such as variability and graphing will give students the tools they need to make sense of the data they will be exposed to in everyday life and eventually learn advanced analytical techniques. Understanding variability is considered by Garfield and Ben-Zvi (2005) to be an inherent characteristic of any sample, and idea that is fundamental to understanding statistics. In order to make sense of data collected from that population, a student must have the skills to summarize them, and test them to determine whether a meaningful difference exists.

1.1 Project Setting

This thesis research combined efforts of two existing projects at the University of Maine in collaboration with the Schoodic Institute, the Acadia Learning Snowpack Project and the Maine Data Literacy Project. The Snowpack Project provided students with an opportunity to design a study and collect and interpret data about local snowpack in collaboration with scientists. The Maine Data Literacy Project provided a framework and instruments for assessing students' understanding of variability and graph interpretation skills.

1.1.1 The Snowpack Project

The Snowpack Project is a student-teacher-scientist partnership among Schoodic Institute educators, scientists from the University of Maine Climate Change Institute, US Geological Survey, Maine Sea Grant, the National Weather Service, and middle and high school science classes. The students collected data on snowfall and snowpack in Maine's three climate zones for the Community Collaborative Rain, Hail, and Snow Network (CoCoRaHS) database. This is a source of important data for the scientists studying snow and an opportunity for students to be involved in and learn about research. The project was designed for students develop their own research questions and use their data to answer them. A series of lessons on snowpack, data, and variability were developed to go along with the field work, but implementation was flexible and varied from classroom to classroom.

How teachers implemented the Snowpack Project in their classrooms was flexible, however there were some commonalities. Instructional support provided by the Snowpack Project included a professional development workshop in the summer, teaching skills necessary to gather snowpack data and a set of instructional resources that supported the project's instructional goals. These resources included six units that covered topics such as background information on snow, writing scientific questions, carrying out field investigations, and communicating research results. In addition, discussions with participating teachers revealed that they all considered data literacy to be an essential component of their science curriculum and invested time in teaching the subject throughout the year.

In the classroom, students were typically introduced to the project by discussing types of research questions they might ask, measurements they could take, and establishing at least one plot site (though often two or more) in which to gather data. Data collection began in January after winter break or at the onset of snowpack, and it continued until the snow melted in the spring. Required measurements for the CoCoRaHS database included snowpack depth, new snowfall depth, and snow-water equivalent from a level open site, however some classes collected more extensive data like snow temperature, and snow depth on hills or under tree cover. In the spring, the students analyzed the data that helped answer their question, and presented their findings to their peers. Presentations varied but were typically a poster or slide show presentation.

1.1.2 The Maine Data Literacy Project

The Maine Data Literacy Project (MDLP) is a partnership between the University of Maine and the Schoodic Institute that is working to understand how students think and learn about data and graphs, and to develop tools and best practices for teaching data literacy.

One of the MDLP's initiatives developed the Assessment of Student Knowledge of Variability (*ASK-Var*), a 32 question multiple-choice assessment instrument designed to identify variability concepts and graph interpretation skills that students understand and those that require more attention (See Appendix A)(Zoellick, Schauffler, Flubacher, Weatherbee, & Webber, 2016). The instrument was developed through an iterative process to verify that it tests the concepts and skills identified as important by its authors and successfully predicts how well students apply their understanding of variability and frequency plots to draw inferences when comparing two groups.

1.2 Overview of Study

This study was conducted to gain insight into student learning in the Snowpack Project and the applicability of the *ASK-Var* assessment instrument in a new setting. The study was designed with two distinct parts. The first consisted of a pretest/posttest assessment design looking for growth in understanding of graphing and variability through the *ASK-Var* post-assessment in the context of the Snowpack Project. The second part used an open-response instrument along with the *ASK-Var* assessment to test for correlation between the two assessments and explore student thinking.

Specifically, three research questions were investigated:

1. To what degree do students participating in the inquiry-based Snowpack Project improve their understanding of graphing and variability by the end of the project?
2. To what extent are student scores on the open-response assessment aligned with how they perform on the *ASK-Var* assessment?
3. What can be learned about how students thought about variability and graphing from the assessments in the study?

This thesis describes a study of student learning about graphing and variability while participating in the Snowpack Project. Chapter Two provides an overview of data literacy and the importance of variability. Chapter Three reviews literature on data literacy in the classroom, challenges to integrating data literacy into the science classroom, and how textbooks support instructors in teaching these concepts. Chapter Four describe the research setting, the assessments that were used, how they were implemented, and how the data were analyzed. Chapter Five describes the analysis and results, and Chapter Six discusses the significance of those results in terms of the three research questions. Finally, Chapter Seven summarizes the key finding and suggests avenues of further investigation.

2 CONCEPT OVERVIEW

This concept overview defines data literacy in the context of this thesis, and describes the importance of exploratory data analysis to learning to think statistically. It focuses on variability and graphing skills as key components of data literacy.

2.1 Data literacy

Data literacy describes a set of skills that allow people to interact with data and graphs in an informed, responsible way. It enables people to transform data into useful evidence by asking questions of the data, processing those data, generating graphs that help answer the question, and using the data to make an argument that considers variability. Data literate people can also evaluate statistical arguments and graphical representations prepared by others. At the center of all of these skills is the ability to think about data as an aggregate and consider variability.

Scientists ask questions. When addressing data literacy, a question needs two characteristics: it must be something the data can answer; and it must be a statistical question. A statistical question is one that considers variability. Rather than asking “How long was the game last night?” a statistical question would ask “How long is a typical game?” It is asking about a summary of a group of games rather than a fact about a single one.

Summarizing data in graphs is a powerful skill, and different graph types highlight different features of a data set. The statistical question will determine the best graph types and generate appropriate graphical representations that help answer the question. Questions about comparing groups or variability are best represented by frequency plots like dot plots, histograms, and boxplots because they display variability.

Finally, data literacy involves connecting data to its context to create a logical argument. This is how evidence is born, but it is only useful when it is considered with respect to variability. Data literacy is most potent when can use the inherent uncertainty of a dataset to rationally generalize beyond the data.

2.2 Variability

Understanding what variability is and how to work with it is essential for data literacy because it is inherent to populations, and it is central to statistical understanding. Variability is the inherent differences that exist among individuals in a populations (ex. the heights of a class of 3rd graders), differences over a period of time or across space (ex. the temperature in January in Orono, ME), or in repeated measure of a single thing (ex. different students using balances to mass the same object). Mathematically, it is the shape and spread of the distribution of data around its center.

Groups of measurements are often summarized with a single value. For example, the average height of a third grade class might be 55 inches. This value was calculated using all of the values in the class, but it hides the variability. Displaying the entire distribution in a graph is important for visualizing the variability. Accounting for the variability in a sample leads to more informed and nuanced decisions.

There are two common ways to describe variability. The more common way is mathematically. It is common to report values like range and standard deviation. When developing a conceptual understanding of variability, however, it is also useful to learn to use informal language to describe the shape and spread of a set of data. Informal language is especially helpful in developing conceptual understanding of variability in young students before concepts like mean, median, mode, and standard deviation are introduced.

2.3 Exploratory data analysis and informal statistical inference

Exploratory Data Analysis (EDA), a term coined by John W Tukey in 1977, refers to a way of describing data and informally looking for patterns and relationships in them. A lot can be learned about a data set before applying quantitative statistical tools by thinking critically about it and studying graphical representations. For example, a bimodal distribution would be hidden by a mean or median, but would be obvious in a histogram of the data.

EDA allows statisticians to apply the tools of their trade more deliberately, and it is a way for students to think critically about the principles underlying statistical analysis. In the education world, informal statistical inference (ISI) is a common EDA strategy. ISI provides a framework for younger students to reason about data as an aggregate and make appropriate claims that consider variability without needing advanced math skills (Bakker & Derry, 2011). Learners using ISI are able to critically evaluate statistical tools rather than just apply algorithms (Ainley, Pratt, & Nardi, 2001). However, data analysis and statistics are frequently taught as quantitative endeavors where the only objective is to memorize procedures. Students learn to calculate summary values like mean, median, and mode; range; and standard deviation but don’t understand their significance on a conceptual level (Bakker & Derry, 2011). A student could use ISI to look at a distribution of data and decide whether mean, median, or mode is most appropriate as a summary measure.

Using ISI, students learn to apply statistical concepts in the context of a problem. Makar and Rubin (2009) identified four concepts that were critical to inferential reasoning. These concepts included the ability to articulate a claim in terms of uncertainty and variability, make generalizations about a group using aggregate properties, recognize a tendency that “went beyond the data,” and connect data and reasoning to create evidence. The following is an example of a claim using inferential reasoning: The home team is probably a slightly better batting team than the away team. Even though they have a lower team batting average, the away team has two batters in their line-up that have very high batting averages skewing the data. A better measure of center in this case would be median, and the home team has a higher median than the away team.

Classrooms that encourage these concepts assist students in constructing conceptual understanding of data analysis. Students are able to use their prior understandings to construct statistical principles in context. Makar (2014) describes how a class of young students (aged 10 and 11) began seeing statistical questions as having two possible types of answers. They believed that either the data sample represented the population perfectly, or they believed that the variability in the population made it impossible to make any predictions about another class. By the end of the inquiry-based activity, the author found the students gained an understanding of data as an aggregate and a command of probabilistic language that allowed them to communicate their prediction and its uncertainty. This shows that even with few math skills, young students can understand and apply important statistical concepts.

3 LITERATURE REVIEW

The two primary goals of this study were to investigate how students learned about variability concepts and graphing in classrooms involved in the Snowpack Project and to investigate how the *ASK-Var* assessment tool measured changes over the course of the project with that group of students. This literature review explores the following questions to support these goals:

1. What data literacy skills are students expected to demonstrate in middle and high school?
2. How can statistical thinking be integrated into science learning?

3.1 What data literacy skills are students expected to demonstrate in middle and high school?

The essence of data literacy is the application of statistical principles to derive meaning from data. It is defined here as the ability to turn data into evidence that can be used to answer a question or defend a position. To apply this definition, students must be able to consider a question asked of data, display the data in a way that helps answer that question, interpret the display to extract new relevant information, and answer the question using evidence from the data (Roth, Bowen, & Masciotra, 2002).

Statistics is essentially the study of variability, and the ability to consider variability in all data-based decisions is essential for a data literate person (Konold, Higgins, Russell, & Khalil, 2015). Variability is the center, shape and spread of a distribution of data. When considering statistical questions, the answers and insights do not come from any individual datum, but are emergent properties of the data as a whole (Konold et al., 2015). Visualizing and describing variability is key to mastering the skills associated with data literacy: asking relevant questions, choosing appropriate representations or graphs, interpreting the representation, and constructing a complete argument using the evidence (Garfield & Ben-Zvi, 2005; Wild & Pfannkuch, 1999).

Science practices and data literacy go hand in hand and have been part of the discussions among academics and policy-makers for decades (S. Brown & Melear, 2007; Project 2061, 1993; Rutherford & Ahlgren, 1991). These principles are embedded throughout science and math national learning standards and even in English language arts to some extent (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010; NGSS Lead States, 2013).

Benchmarks for Science Literacy (1993) was published by the American Association for the Advancement of Science to support states in developing standards for science education, and it was the national guiding document for educators until 2012. The *Benchmarks* present a view of science that is consistent with science as a set of practices in the first chapter, "The Nature of Science," emphasizing three sections: The Scientific Worldview, Scientific Inquiry, and The Scientific Enterprise. The chapter describes how the process of science occurs *in situ*, but the rest of the document offers little support for teachers and curriculum developers wanting to integrate those ideas into the classroom, a common weakness of science texts (Morris, Masnick, Baker, & Junglen, 2015).

The math-focused portions of the benchmarks suffer from similar shortcomings to the science portions. Data literacy concepts are included, but they are not integrated into the science benchmarks. Understanding variability, referred to as uncertainty, is neither central to nor well-developed in the benchmarks despite being widely regarded as essential to data literacy (Bakker, 2004; Garfield & Ben-Zvi, 2005; Gould, 2004; Moore, 1997; Reading, 2004; Wild & Pfannkuch, 1999). In the lower grades, the benchmarks primarily describe variability as how likely it is something will happen and focus on central tendency (Project 2061, 1993). The upper level benchmarks do refer to the key components of data literacy including asking questions, collecting and organizing data, representing data in tables and graphs, interpreting the data, and communicating the results, but they lack specific focused support for teachers trying to teach these complex ideas (Project 2061, 1993, p. 228).

The central problem is not that teaching data literacy and the nature of science are incompatible with the *Benchmarks*, but by segregating the math and science skills, they do not emphasize essential transdisciplinary nature of data literacy (Vahey, Yarnall, Patton, Zalles, & Swan, 2006). They also lack guidance for teachers who may have little experience working with data and conducting authentic scientific inquiry, in integrating authentic research into their classrooms.

The National Research Council's document, *A Framework for K-12 Science Education* (2012), addresses many of the previous critiques and was a guiding framework for how data literacy should be integrated into science education and the development of the Next Generation Science Standards (NGSS) (NGSS Lead States, 2013). The NGSS are composed of three interconnected components: practices, crosscutting concepts, and disciplinary core ideas. Practices (Figure 1) are the activities in which scientists engage when investigating a phenomenon and generating new knowledge. Crosscutting concepts (Figure 1) are a set of ideas that inform scientific thinking and help students engage with new scientific ideas in a rigorous way. These are ideas like "Patterns" and "Systems and system models" which can be found across scientific disciplines. Disciplinary core ideas are the content the students are expected to learn in each of four areas: physical sciences; life sciences; earth and space sciences; and engineering, technology and applications of science. Performance expectations integrate these three components and divide them into actionable pieces.

Because components of data literacy such as data collection and interpretation through graphs are integral to the practices and crosscutting concepts, they are explicitly included in the performance expectations. This approach is intended to model an authentic science process with explicit support for teachers in integrating reasoning with quantitative data into science class. The middle school performance expectation MS-PS3-1 reads "Construct and interpret graphical displays of data to describe the relationships of kinetic energy to the mass of an object and to the speed of an object on energy.” This is a clear example of how the NGSS integrate data literacy skills into the other content and skills**.** (NGSS Lead States, 2013).

Scientific and Engineering Practices
1. Asking questions and defining problems
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations and designing solutions
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

Crosscutting Concepts
1. Patterns
2. Cause and effect: Mechanism and explanation
3. Scale, proportion, and quantity
4. Systems and system models
5. Energy and matter: Flows, cycles, and conservation
6. Structure and function
7. Stability and change

Figure 1. Next Generation Science Standards (NGSS) Practices and Crosscutting Concepts

Of the eight practices in the NGSS, one specifically refers to data and six are closely related (Figure 1). Practice 4, “Analyzing and interpreting data,” integrates opportunities to work with data into the Disciplinary Core Ideas. By nesting practices under each performance expectation, the NGSS can help teachers take advantage of opportunities to work with data in ways that they might not have recognized in the past.

Through the NGSS practices, elementary standards plant the seeds of data literacy as early as kindergarten. Students are expected to begin looking at information and gathering data, asking questions, and displaying data in tables and graphs (See Appendix E)(NGSS Lead States, 2013). These standards introduce practices essential to data literacy and lay the groundwork for more advanced skills in the future. A student in third grade would begin to address these standards by asking what a typical third grader’s height would be, as in Makar (2014). The activity got students asking questions about heights of their whole class, collecting data, and using graphs and tables that showed the variability of their dataset.

Data literacy becomes a focus in the performance expectations for middle and high school students. Students continue using graphs to display data and ask and answer questions with them, but in new and more sophisticated ways. At this level the practice “Using mathematics and computational thinking” introduces opportunities to use more quantitative analyses such as interquartile range and graphical representations that consider variability in data like boxplots, dot plots, and histograms.

The Common Core Math Standards (CCMS) complement the NGSS. CCMS introduce data in kindergarten by graphing and comparing frequencies of objects in different groups. By fifth grade students are collecting data and displaying them in dot plots and bar graphs. In addition they are introduced so some basic analyses such as categorizing, comparing group size, and calculating range and mean.

The concept of variability in data is introduced in the sixth grade math standards. Students are introduced to the idea of statistical questions and visually how data are distributed along a number line using frequency plots. Because statistics is fundamentally the study of variability, these sixth grade standards are keystone concepts for future understanding of data literacy concepts.

The seventh and eighth grade statistic and probability standards build on the sixth grade standards but with more sophisticated advanced ideas. Students learn the significance of sampling populations and to consider variability in comparing groups. They are also introduced to probability and comparing two variables with scatter plots. In high school, students continue to work with the frequency plots introduced in middle school and are introduced to quantitative measure of variability like standard deviation. They develop the skills to apply their understanding of variability to make inferences about a population from a sample (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010).

Support for data literacy principles is even found in the Common Core English Language Arts Standards. Middle school students are expected to make claims and use data and evidence to defend them. Because the math, science, and language arts standards all support data literacy skills in different ways, they encourage a context-rich transdisciplinary perspective of data literacy (Vahey et al., 2006). In addition, standards that support data literacy start as early as kindergarten and build on each other year after year, giving students time to process these complex ideas (Garfield & Ben-Zvi, 2005).

In middle and high school, students can begin engaging with data using frequency plots and quantitative summaries as they did in the Snowpack Project. In this project, students measured new snowfall and total snow depth, asked statistical questions of those or related data, and presented their findings to their peers in a professional presentation.

3.2 How can statistical thinking be integrated into science learning?

3.2.1 Missing Concept: Visualizing, describing, and interpreting variability in data

Data are transformed into evidence by identifying patterns (Wild & Pfannkuch, 1999). This is achieved through a variety of mathematical calculations or graphical representations that summarize the data. The focus here will be on visualizing variability in data through graphical representations including box and whisker plots, dot plots, and histograms. Each of these types of plots shows distribution shape, center, and range with varying degrees of precision.

One challenge in learning to recognize variability in data is not seeing datasets as aggregates. In their study of elementary, middle, and high school students, Konold et al. (2015) identified four "loosely hierarchal" perspectives held by students for inscribing or interpreting data: pointers, case valuers, classifiers, and aggregators. From the least developed "pointer" perspective, the inscription is used to reference the event from which the data were collected, while the most developed "aggregators" are able to identify emergent properties of the dataset. While each of these perspectives has its value, a data literate student must be able to use the aggregate perspective. In another study, middle school students in Israel were directed to come up with a question about name lengths in Israel and America. They began by focusing on irrelevant features of individual data such as the number of names beginning with “Mc” rather than comparing name lengths in the two countries. The irrelevant feature obscured the aggregate differences which the students were unable to identify the key features until they received assistance from their teacher (Ben-Zvi, 2004).

Another challenge of transforming data into evidence is being able to describe the variability in a dataset. Bakker (2004), Meletiou-Mavrotheris and Paparistodemou (2015), and others have argued that using informal language to describe variability helps build conceptual understanding in students. Bakker (2004) also found graph types and pedagogical techniques that deemphasize the individual data points (ex. distribution of data represented by a smooth curve) may help students overcome the less sophisticated data perspectives (ex. pointer, case values, and classifiers) and see aggregate properties. When reasoning about graphs that showed the distribution shape students were able to discuss skew and slope without being distracted by specific cases.

According to Roth et al. (2002), there are three hierarchical levels from which people perceive graphs. In the first, termed segmenting inscriptions, the reader is attempting to make sense of the graph piece by piece, and context is generally ignored in favor of constructing a coherent understanding of the graph itself. For example, a student describing a boxplot by only listing the range, median, and quartiles without incorporating the significance of those values would be interpreting the graph by segmenting inscriptions. In the second, termed hermeneutical reading, the reader takes the idea the graph was conveying and relates it to a broader context. This step requires background knowledge of the graph content, so even skilled graph readers may struggle with unfamiliar fields of study. The third level is termed transparent reading. This occurs when both the graphical representation and the content are familiar, and the reader is able to describe the setting and background of the situation as it relates to the graph. This was primarily observed in professionals who were looking at graphs they had constructed.

Roth's hierarchy is reflected in graph interpretation strategies at different levels of data literacy. Those not trained in science may lack both fluency with the graphical medium and the contextual material in which to ground it, and so may interpret graphs by segmenting inscriptions; piece by piece. This has been observed in secondary students, college science majors, pre-service science teachers, and graduates with BS and MS degrees who are not working as scientists (Bowen & Roth, 2003; Roth et al., 2002; Roth, McGinn, & Bowen, 1998; Roth & McGinn, 1998). As a result, the information conveyed by the graph is limited reducing the value of the representation.

Fluency in graph interpretation is essential for describing and interpreting variability in graphs. Scientists are able to engage in what the graph represents rather than the graph itself and easily move between the graph and the physical event it is describing (Roth et al., 2002). Interpreting a graph requires integrating both the technical aspects of the graph and the physical phenomenon it describes, which is what scientists do to construct in their minds the story the graph is telling (Bowen & Roth, 2003). This level of interpretation would be described as hermeneutical or transparent reading and is ultimately the goal for students working with data in science class (Roth et al., 2002).

3.2.2 Missing pedagogy: authentic science learning

The word "science" refers to both a body of knowledge and a set of practices employed by scientists. These practices include asking questions, making observations, gathering data, creating theories and models, generating hypotheses, and thinking critically about each stage of the process (*National Science Education Standards*, 1996; NGSS Lead States, 2013). It is an iterative process where new solutions beget new questions, and the direction of inquiry is defined by the investigator (National Research Council (N. R. C.), 2012).

Science content can support students in learning statistical thinking and data literacy by providing essential and meaningful context (Wild & Pfannkuch, 1999; Wu & Krajcik, 2003). Statistical thinking which includes graph interpretation, merges the data (numbers) with the real-world phenomena they represent (Reading, 2004). Using data from a topic being studied in science class to practice graph interpretation may ease the cognitive load and allow students to focus their mental resources on the graph interpretation (Kirschner, Sweller, & Clark, 2006). Even experts benefit from familiar content. Roth et al. (2002) found that scientists lose some graphing fluency when presented with unfamiliar content graphed in familiar ways or familiar content graphed in unfamiliar ways.

An authentic learning environment is not a particular activity or pedagogy, but rather an "emergent property of a dynamic system of learning" that is created by the participants; students, teachers, and scientists (Rahm, Miller, Hartley, & Moore, 2003). Authentic learning environments are responsive to the participants and involve activities similar to those of professionals (J. S. Brown, Collins, & Duguid, 1989; van Eijck & Roth, 2009). Authentic science experiences provide context for engaging in scientific practices such as collecting and analyzing data, asking statistical questions, and generating appropriate graphs to help answer those questions, which allows students to access them when presented with novel problems (Herrington & Oliver, 2000).

Successful authentic learning environments offer at least two major advantages for students learning to think statistically and interpret graphs. First, students are invested in the work they are doing (Gibson & Chase, 2002). The work has some significance beyond the classroom or the grade, and the students care about the quality of the data and the outcomes of the project much like a professional scientist. This investment on the part of the students improves both learning outcomes and engagement in the subject. Gibson and Chase (2002) found long-term positive effects on student attitude towards science after short two-week inquiry-based summer science camp in middle school.

Second, authentic learning environments provide an open-ended context in which to interpret graphs scientifically. The process of collecting data and taking measurements helps the students understand the physical event being represented graphically. In addition, integrating the math and the science content helps students understand the math concepts and how to apply them (Bowen & Roth, 2003; Roth, 1996). The combination of the math background and science concepts are the two key ingredients that allow scientists to fluently engage with graphs (Roth, 1996)

3.2.3 Insufficient support from textbooks

Textbooks are more than just guides or supplemental resources; they frequently play a dominant role in determining the focus of the class both in content and practice (Banilower et al., 2013; Binns, 2013; Chiappetta & Fillman, 2007; Morris et al., 2015; Valanides, Papageorgiou, & Rigas, 2013). The 2012 National Survey of Science and Mathematics Education reported that "Textbooks appear to exert substantial influence on instruction, from the amount of class time spent using the textbook to the ways teachers use them to plan for and organize instruction" (Banilower, 2002) The same report found that among middle and high school teachers only 62% and 70% respectively reported doing hands-on laboratory activities at least once a week, 54% and 58% respectively reported having students graph and or analyze data, and 23% and 18% respectively reported engaging their classes in project-based learning activities. Since textbooks are so widely used, a well-constructed text could improve pedagogy in data literacy, however they often do not align well with the contemporary standards or pedagogy (Budiansky, 2001; Hubisz, 2003; Stern & Roseman, 2004). For example, the nature of science is frequently presented in the traditional view where it is a linear experimental process rather than an iterative process with multiple modes of investigation (Binns, 2013; Hubisz, 2003).

Available research on textbooks indicates that they do not provide enough support for data literacy (Binns, 2013; Morris et al., 2015; Valanides et al., 2013). In a survey of 20 middle school science texts, Morris et al. (2015) found that of 731 activities analyzed only 2.5% included opportunities to record data, and there was little support within those activities for how to record those data. The team also reported that only 3% of data analysis activities provided step-by-step instructions, and none of them provided explanations of why a particular analysis was chosen. Despite data literacy being the focus of the study, neither graph construction nor variability were addressed directly by the authors. Another analysis of middle school physical science texts included critiques of graphing activities that encouraged the use of more real data and data collection, but data and graphing were absent in its concluding suggestions to teachers, authors, or publishers (Hubisz, 1998). These were the only studies found that addressed data literacy directly even though others identified data collection, analysis, and interpretation as important in their introductions (Park & Lavonen, 2013; Valanides et al., 2013). It appears that science textbooks and researchers are not adequately supporting data literacy instruction in the classroom.

This thesis used the *ASK-Var* multiple-choice assessment to measure what students participating in the inquiry-based Snowpack Project learned about graphing and variability. It is important to develop tools to measure students’ ability to visualize, describe, and interpret variability in data while interventions like the Snowpack Project use authentic projects and data to improve support and pedagogy for teaching these important skills.

4 METHODS

4.1 Research setting

The first goal of this study is to understand to what extent students improve their understanding of graphing, variability, and data literacy in general in the context of the Snowpack Project. The second is to see how well the multiple-choice Assessment of Student Knowledge of Variability (*ASK-Var*) (Zoellick et al, 2016) predicts their scores on an open-response assessment with questions that are relevant to concepts encountered in the Snow-pack Project. The third goal is to identify ways that students engaged with the snowpack data. The questions addressed are:

1. To what degree do students participating in the inquiry-based Snowpack Project improve their understanding of graphing and variability by the end of the project?
2. To what extent are student scores on the open-response assessment aligned with how they perform on the *ASK-Var* assessment?
3. What can be learned about how students thought about variability and graphing from the assessments in the study?

Four volunteer teachers were solicited for this project from a pool of 17 teachers who participated in the Snowpack Project. Their students (n=150) responded to a multiple-choice assessment (*ASK-Var*) as part of their Snowpack Project activity. Of those, 16 students taught by two of the teachers also took the open-response assessment developed for this study. Three of the teachers and 142 students were in a public school setting while one teacher and eight students were in an alternative school that focused on experiential learning. The majority of the students (n=134) were in a required science class while 16 were in elective classes (Table 1).

Table 1. Summary of study classroom characteristics. Total number of students and number of students per class are estimates because they changed throughout the semester, with students transferring classrooms or schools.

Teacher code	Class title	School Type	Students per class	Total students	Grade
1	Earth Systems Science	Public	~14	~65	9th
2	Earth Systems Science	Public	~14	~65	9th
3	Geology and Natural History of Passamaquoddy Bay	Private	8	8	9th-12th
4	Introduction to Scientific Research	Public	8	8	10th-12th

4.2 Measures and scoring

This study employed two instruments to measure students' understanding and skills. The first was the *ASK-Var* assessment, a multiple-choice assessment of graphing and variability skills developed by Zoellick et al. (2016) as part of the Maine Data Literacy Project. The second was a series of open-ended questions that required students to interpret data relevant to snowpack and winter weather.

4.2.1 Multiple-choice assessment

This study used a near-final version of the *ASK-Var* assessment developed by the Maine Data Literacy Project (Zoellick et al, 2016). It consists of 32 questions with four options for each response (see Appendix A). The three distractors for each question were chosen from known misconceptions so that teachers could use the responses to not only identify topics their students do not understand, but could also identify the misconceptions they hold.

The *ASK-Var* assessment questions were developed by the Maine Data Literacy Project to target concepts that related to variability found in the Common Core Standards for Mathematics in middle and high school. It was refined through an iterative process where questions were revised based on initial responses from a group of students outside the study, and it included questions that covered a range of difficulties and topics related to graphing and variability. The objective was to create an assessment that specifically targeted graph interpretation skills and understanding of variability in data with minimal noise from confounding factors that might affect a student's score such as reading ability.

The Rasch analysis (described in more depth in section 4.4.1) was used by the MDLP to develop the *ASK-Var* assessment, and it was used in this study to check the fit of the assessment for the participants in this thesis. The version of the *ASK-Var* assessment used in this thesis was very close to the final published version; three questions were removed and one was added (Zoellick et al, 2016).

The *ASK-Var* assessment was administered in 2015 in participating Snowpack Project classrooms through an internet-based survey platform (SurveyMonkey) that made implementation and data retrieval simple and reliable. The students responded to the assessment during class using devices provided by the school. Students were each given no more than one class period to complete the assessment, which ranged from 40 to 80 minutes across the schools. Assessments were administered by the normal classroom teacher as part of regular instruction.

Responses were scored using an R script, coding 0 for incorrect responses and 1 for correct ones. In order for a student's response to be counted, 75% of the questions had to be answered. For respondents who met this threshold, blank responses were considered incorrect if any questions further along in the test were answered, assuming that the student skipped those questions because they did not know the correct answers. If questions at the end of the test were not answered, it was assumed that the student did not have time to finish, and the blank questions were not counted against the final score.

4.2.2 Open-response assessment

The open-response assessment was developed specifically for this study to measure students' abilities to apply their data literacy skills to an open-ended problem without the help of multiple-choice options. It was written using three datasets related to climate, temperature, and snow topics relevant to the Snowpack Project. Questions included a pair of box plots, a histogram, and a graph construction activity (See Appendix B). The questions required students to independently generate a graphical representation of a dataset and describe and interpret graphs in the context of open-ended questions. These applied skills are difficult to test directly in a multiple-choice format, which provides a limited number of options of which one is correct.

The open-ended assessment was revised after reviewing responses from a trial group of 10th grade biology students unrelated to the Snowpack Project. The final version of the test had nine questions about three different data scenarios, with data represented in either graphs or tables. In the first two scenarios, students were asked to describe and interpret the data displayed in two box plots (Questions 2 and 3) and a histogram (Questions 4-7). For the third scenario, students constructed a graph from a provided data table to address a driving question and used it to answer the remaining three specific questions (Questions 8-10).

Scenario 1 depicted the average monthly high temperatures for two different fictitious towns in a pair of box plots and asked students to compare the temperature regimes. The students were asked to use the graph to describe the similarities and differences between the climates of the two towns and explain how those similarities and differences might affect someone living in each place.

Scenario 2 measured the students’ ability to interpret a histogram showing data of past events to make predictions about the future. The graph depicts the date of the first snowfall of the year in Orono, Maine from 1995 until 2014. The students were asked to describe what the heights of the graphs represent, describe the variability in the graph, predict when the snowfall would occur next year, and explain the evidence from the graph that supported their response. Questions associated with both of these graphs assessed students’ ability to read and interpret box plots, and reason about variability in the data.

Scenario three presented students with a table of data of the length of growing seasons in weeks for towns in two fictitious counties. Students were asked to graph the data in a way that would help them answer the question, choosing an appropriate type of graph. After drawing their graphs, students were asked make a claim about whether the two counties had similar or different growing seasons and explain how evidence from their graph supported their claim. Questions associated with this activity assessed students' ability to choose an appropriate graph, read it, and interpret its meaning in terms of its variability.

The rubric used to score the open-ended responses was developed by a team of three graduate students (Appendix C). It was initially written using a template from the Maine Data Literacy Project and revised based on preliminary student responses from a group of students unrelated to the Snowpack Project. A few minor final clarifications were made to the rubric prior to final scoring, and all questions were graded using the final rubric.

The final rubric specified criteria for four levels of response: does not meet expectations (1), partially meets expectations (2), mostly meets expectation (3), or meets expectation (4) (Appendix C). Each question was identified as addressing one of these four categories: graph description, graph interpretation, graph mechanics, and graph interpretation. Because each question was unique, each one was assigned a customized rubric with specific criteria for that question and a small list of example responses. Question 7, the graph construction task, had a slightly different organization. Responses were scored for two different aspects of graph construction: choice of graph type and graph mechanics.

Two participating teachers volunteered to give their 16 students the open-response assessment. It was administered electronically via SurveyMonkey, with the exception of the graph construction task, which was done with paper and pencil then scanned and submitted via email. Emailed responses were matched with to the corresponding electronic assessment by a student code assigned by the teacher. Students were allowed one class period to complete the assessment. Responses were scored by the same team that assisted with the rubric development to ensure maximum reliability among scores. Because of the small number of participants to the open-response assessment (n=16), all responses were scored by all team members. Questions were scored by each person, and then all of the scores for that question were compared. When disagreements of scores arose, the team referred back to the rubric and previous similar responses. Disagreements were resolved through discussion until unanimity was reached among scorers for every score.

4.3 Implementation of assessments

The multiple-choice assessment was administered to students in participating classrooms twice. Once in January of 2014-2015 (n=182), and again to the same students in May of 2015 (in April or May, n=122), for a set of 98 paired pre-post assessment scores (once absentees and incomplete responses were removed from the dataset). It was administered third time in September of the following school year (September 2015, n=101) with a different group of ninth grade students who were unpaired. The January and May assessments were originally intended to be a pretest/posttest design, as most of the activities for the Snowpack Project did not begin until January. A preliminary analysis of responses, contrary to expectation, showed no difference between the January and May assessment scores. Interviews with the participating teachers indicated that they had all started data literacy instruction early in the year and made it a focus of their class with the Snowpack Project being a culmination of the year's data literacy work rather than the central feature. In light of this information, a third round of testing was added the following fall measure a group of presumably similar students' understanding of variability concepts at the beginning of the year. The structure of this study design is diagramed in Figure 2.

Further, to attempt to check to see if the assessment would detect a difference between students at markedly different grade levels, scores from students in the Snowpack Project were compared with a group of seventh grade students from different schools and outside the Snowpack Project.

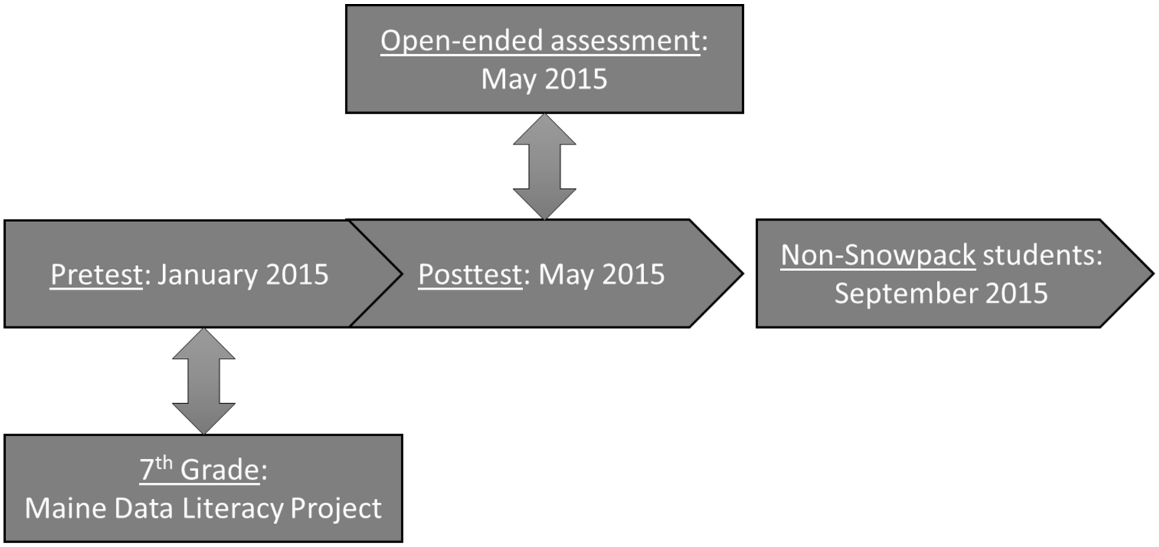


Figure 2. Diagram of assessments timing. Arrows show comparisons between groups.

4.4 Data analysis

4.4.1 Rasch analysis of the *ASK-Var* assessment to determine "fit" of this assessment for this sample of students

The Rasch analysis is an analytical tool used to measure the difficulty and unidimensionality of an assessment. Rasch analysis gives each question a difficulty score based on how respondents performed on that item. It is graphed on the Y axis using logit values with larger positive numbers being more difficult and larger negative number being easier. Zero represents the level of difficulty where 50% of the respondents would be predicted to answer correctly and 50% would be predicted to answer incorrectly.

A unidimensional assessment measures only one particular skill and is identified by the Rasch analysis as "fit." An analogy could be made using height and weight. A unidimensional measurement only measures one dimension, for example height or weight, not a summary of the two. Body mass index is not unidimensional as it combines height and weight into a single value. An example of the shortcomings of a bi-dimensional measure like this can be seen when body builders with very little fat but a lot of muscle mass register as obese according to their body mass index.

Items that fall outside the threshold of +/- 2 infit t statistic units on the X axis do not fit the unidimensional model (See Bond & Fox, 2001 for a detailed discussion of the Rasch model). A lack of fit could be caused by a number of factors including confusing wording, too much or too little background knowledge, or unfamiliar vocabulary (Bond & Fox, 2001). This thesis used the Rasch analysis to measure how well the *ASK-Var* assessment “fit” this sample of students (item fit) in terms of their understanding of the concepts addressed by the questions, and not reading level or some other construct.

Item fit is used to describe the likelihood that an assessment item is answered correctly by students with an ability measure greater than or equal to the difficulty measure of the item. Ability measure is based on the number of questions the student answered correctly, while item difficulty is based on the number of students that answered that item correctly. For example a student who answered 50% of the questions correctly would receive an ability score of zero, and an item that was answered correctly by 50% of the students would receive a difficulty score of zero.

Rasch analysis also converts ordinal-level raw percentage scores into interval-level data on a logit scale (Figure 3). This means that the intellectual ability required to move one unit on the logit scale is the same no matter where it may fall in the range. This differs from raw percentage scores because the intellectual growth required to move from 40% to 50% is less than the intellectual growth required to move from 85% to 95%. When assessment items or persons are plotted on the logit scale, the space between data points becomes comparable, much like comparing differences in temperature. Interval-level data makes comparisons of students’ abilities and analysis of item difficulty much more powerful because we can now describe how much more difficult one item is from another or how much more able one student is from another. These logit values estimate abilities of students and difficulties of assessment items.

One of the limitations of the Rasch analysis data is that each measuring instrument is graduated differently based on the group of people who took the assessment and the assessment itself. While comparisons of logit scores within a dataset are flexible and powerful for comparisons within that sample, comparisons between datasets are more limited. Comparing logit score in two different Rasch analyses would be like comparing distances measured with two different rulers with unknown graduations.

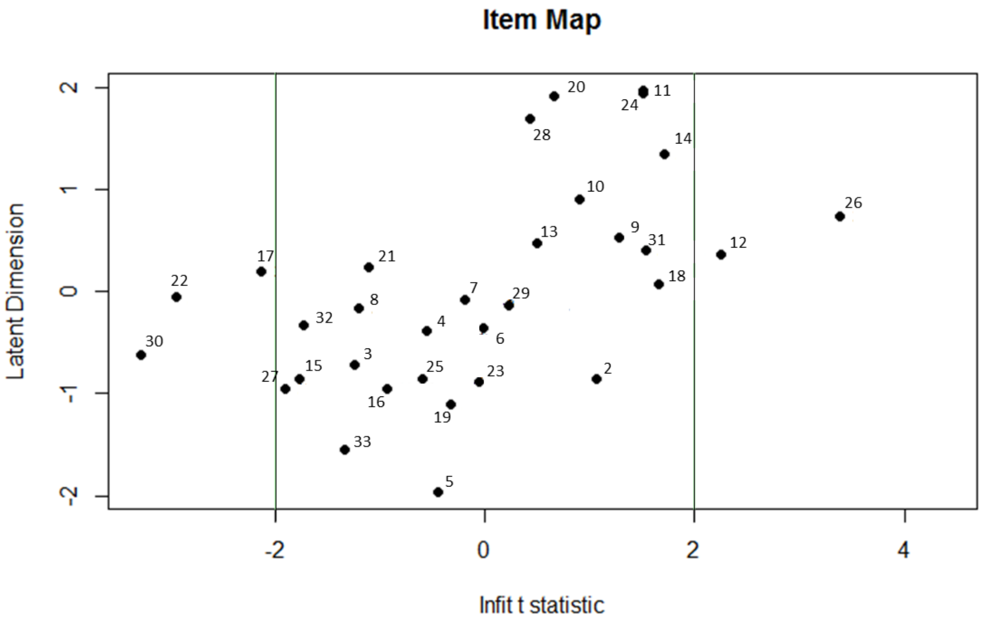


Figure 3. A Rasch item map visually showing the distribution of assessment items across the two Rasch dimensions.

Despite the limitations outlined above, the Rasch analysis data are useful for characterizing the multiple-choice assessment and for investigating the first research question of this thesis. It was used here to verify whether the assessment is an appropriate tool for measuring the participating students' understanding of variability in data.

4.4.2 Analysis of the pre and post *ASK-Var* assessments

The *ASK-Var* assessment data were used in answering research questions one, two, and three. Each pair of pre-posttest scores were analyzed by the whole test and broken into four conceptual categories. These four concept categories were: Variability Concepts, Interpret

Meaning, Read Graphs, and Language. Items in the Variability Concepts category were identified as primarily assessing a student's ability to describe variability and identify it in different graphical and verbal contexts. For items categorized as Interpret Meaning, students were asked to evaluate interpretations of graphs and choose the best analysis statement. For items categorized as Read Graph, students were asked to pull information form the graphs provided. For items categorized as Language, students were asked to define and use key vocabulary words. Figure 4 shows examples of questions in each category. The concept categories were included to identify if any subset of knowledge looked different from the others or the assessment as a whole. Summary statistics and t-tests were performed with Microsoft Excel 2013 and IBM SPSS 16 to see if there were pre-post gains.

To answer Question 1, data were compared January to May (Did students show any gains before and after the project?), September to May (Is there any "proxy" evidence that students might have scored a lower at the beginning of the year prior to any instruction in data literacy?), and seventh grade to high school (January) (Does the assessment pick up differences between high school and middle school students?).

4.4.2.1 To what degree do students participating in the inquiry-based Snowpack Project improve their understanding of graphing and variability by the end of the project?

Paired data from January and May were compared with paired sample t-tests to identify any changes that might have occurred during the spring semester. The t-tests were performed for the whole test and for each of the conceptual categories.

The January and May responses were compared to identify if students changed their responses, and if so, how? The stability analysis was used to identify questions or concepts where students might be guessing, and shifts to or from responses that would offer insight into the students' learning. Comparison of pre and posttest scores were analyzed in two different ways.

Variability concepts

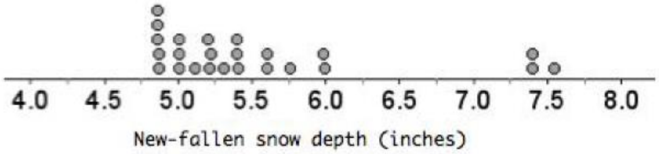
10. Which set of data has the greatest variability?

□ 1, 1, 2, 4, 8, 12
□ 6, 3, 7, 2, 5, 4
□ 2, 3, 4, 4, 7, 8
□ 10, 12, 12, 13, 13, 14

Interpret Meaning

Below are the depths of new-fallen snow measured at 24 sites following a snowstorm. Use this graph to answer the next question.

28. Which of the following statements about the data presented in the snow-depth graph is correct?



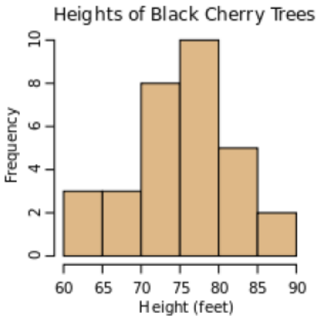
□ The median snow depth will be greater than the mean snow depth.
□ The mean snow depth will be greater than the median snow depth.
□ The mean snow depth will be the same as the mode.
□ The mode is located in the cluster of points between 5 and 6.

Read Graphs

Below is a histogram of the heights of 31 black cherry trees. Use this graph to answer the next three questions.

19. Which height range occurs most frequently among all of the trees?

□ 60 to 65 feet
□ 70 to 75 feet
□ 75 to 80 feet
□ 85 to 90 feet



Language

2. What is the best description of the “median” value in a data set?

□ The middle point in the data set
□ The value in the data set that occurs most frequently
□ The sum of the values divided by the number of items
□ The largest value in the data set

Figure 4. Examples of *ASK-Var* questions from each conceptual category.

The first response stability analysis took a coarse-grained look at a summary of all responses from all students on the whole test and in each conceptual category. Each pair of responses was grouped into one of four categories. Students' with incorrect responses in January and May were coded 1, and correct responses in January and incorrect response in the May were coded 2. Responses that moved from incorrect to correct were coded 3, and responses that were correct both times were coded 4 (Table 2). Resulting scores showed net shifts in response correctness for the whole assessment and for groups of questions.

Table 2. Change analysis code interpretation

January Response	May Response	Code	Interpretation
Incorrect	Incorrect	1	The concept was not learned (a guess) or a new misconception was introduced
Correct	Incorrect	2	A new misconception was introduced
Incorrect	Correct	3	A new correct concept was learned
Correct	Correct	4	The concept was already known

The second response stability analysis looked at changes in the distribution of students of answer choices from each question. This helped identify shifts in thinking at the question level, and it exposed changes from one incorrect response to another that were not reflected in final scores.

4.4.2.2 Comparison between Snowpack Project students and other student groups

Multiple-choice assessment scores were also compared between May and September by whole assessment and conceptual categories. These data were not paired, so independent sample t-tests were performed to identify significant differences in the means of the two samples.

As with the September to May group comparison, summary statistics and independent sample t-tests were used to compare middle school and high school groups (January) to see if the assessment could detect a difference between the two datasets. The t-tests were performed for the whole assessment and the four conceptual categories.

4.4.3 Analysis of the open-response assessment

The open-response assessment results were compared to the May multiple-choice assessment results to see how the skills and abilities from the *ASK-Var* assessment translated to the open-response assessment. To summarize the open-response scores, frequencies of rubric scores 3 or 4, “Mostly meets expectation” or “Meets expectation,” were calculated for each student and that value was correlated to the Rasch ability value for the same student to test for a correlation between to two assessments. In other words, to what extent were *ASK-Var* scores predictive of open-response scores?

Students’ *ASK-Var* scores were also compared to open-response scores on a question by question basis to identify where the *ASK-Var* assessment was not discriminating well compared to the open-response scores. This was a way of correlating degree of success on a single open-response question to score on the *ASK-Var* assessment (see Table 6 on page 46). Open-response answers were flagged when a student scored relatively well on the whole *ASK-Var* assessment and relatively poorly on the open-response assessment question (ex. A student scored 88% on the *ASK-Var* assessment and a 2 on open-response Question 5.).

4.4.4 Qualitative analysis of patterns in student response

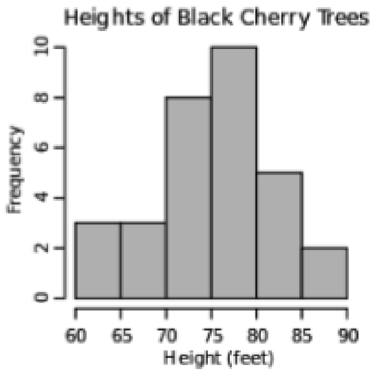
In the qualitative stage of analysis for this thesis, three topics of interest were identified from the open-response and multiple-choice assessments: histogram interpretation, boxplot interpretation, and graph choice. These topics emerged from examining student responses.

Six questions from the two assessments (*ASK-Var* Questions 19, 20, and 21 and open-response Questions 4, 5, and 6) were identified as assessing students’ ability to interpret data represented in histograms (Figure 5, Figure 6). For brevity, questions from the *ASK-Var* assessment will be labeled as AV (ex. AV19) and questions from the open-response assessment will be labeled OR (ex. OR4). These questions asked about histograms in three different ways. Questions AV19, AV21, OR4, and OR6 identified a feature of a histogram or asked the student to identify a feature of a histogram and interpret it in terms of the real-world phenomena that it represented (representation to reality). Question AV20 identified a feature of the real world and asked the students to identify the portion of the graph that represented it (reality to representation). And question OR5 asked students to describe the variability of the dataset represented in the graph (variability). Responses to question OR5 were also coded into three different groups. Group 1 included responses that did not address variability in any way, group 2 responses began to address variability but only mentioned a measure of center or the spread, and group 3 responses described variability in terms of a measure of center and the spread (Appendix D). The question characteristics can be found in Table 7, and the questions can be found in Appendix A and B.

Responses to OR2 and OR3 from the open-response assessment were used to analyze boxplot interpretation (Appendix B). In reading the responses to the two questions, one key idea was pulled from each. Rubric scores were also considered in the analysis.

Trends in students' choice of graph type were identified using OR8 and OR10 (Appendix B). Responses to OR8, the graph construction task, were grouped by two dimensions; each student graph was classified as either a frequency plot or not a frequency plot and as a graph where the groups being compared were graphed together or where the groups being compared were graphed separately.

Below is a histogram of the heights of 31 black cherry trees. Use this graph to answer the next three questions.



19. Which height range occurs most frequently among all of the trees?

□ 60 to 65 feet □ 70 to 75 feet
□ 75 to 80 feet □ 85 to 90 feet

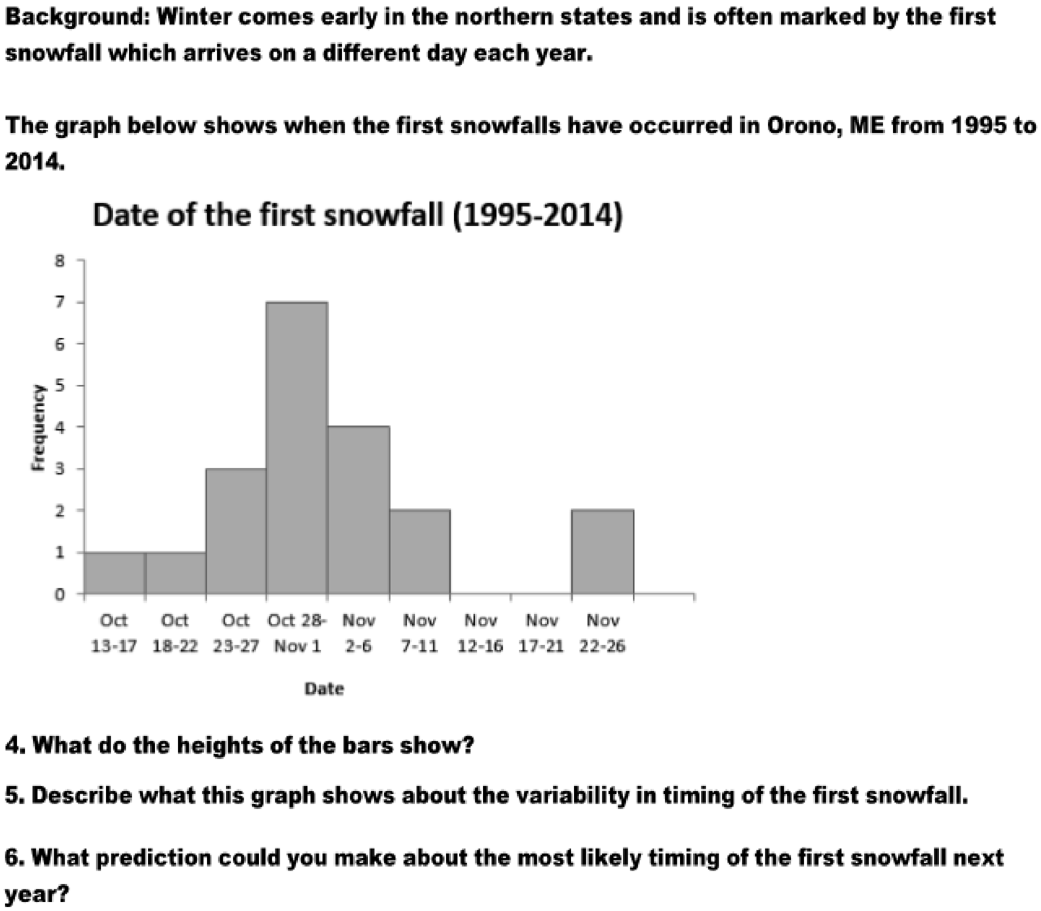
20. How many trees are in the tallest group of trees?

□ Two □ Three □ Eight □ Ten

21. What does the height of the tallest column mean in this histogram?

□ The number of trees that are 10 feet tall
□ The number of trees that are the tallest in the group
□ The total number of trees measured
□ The trees in this height group occurred most often

Figure 5. *ASK-Var* Questions 19, 20, and 21.



Background: Winter comes early in the northern states and is often marked by the first snowfall which arrives on a different day each year.

The graph below shows when the first snowfalls have occurred in Orono, ME from 1995 to 2014.

Date of the first snowfall (1995-2014)
Frequency
0
1
2
3
4
5
6
7
8
Oct 13-17
Oct 18-22
Oct 23-27
Oct 28-Nov 1
Nov 2-6
Nov 7-11
Nov 12-16
Nov 17-21
Nov 22-26
Date

4. What do the heights of the bars show?

5. Describe what this graph shows about the variability in timing of the first snowfall.

6. What prediction could you make about the most likely timing of the first snowfall next year?

Figure 6. Open-response Questions 4, 5, and 6.

5 RESULTS

This study investigated two aspects of data literacy among high school students: understanding of variability and interpretation of data distributions. The first used the *ASK-Var* assessment to address the first research question, "To what degree do students participating in the inquiry-based Snowpack Project improve their understanding of graphing and variability by the end of the project?" The second question compared the *ASK-Var* assessment results to open-response assessment results to address the second research question, "To what extent are student scores on the open-response assessment aligned with how they perform on the *ASK-Var* assessment?" This chapter describes results, beginning with a check into the validity and reliability of the instruments used.

5.1 Assessment of validity and reliability

5.1.1 *ASK-Var* assessment

The *ASK-Var* assessment was previously shown to be a valid tool for describing a group of middle school and early high school students' understanding of variability with a different group of students. The Rasch analysis was performed on the data in this study to verify that the assessment would work as predicted (Zoellick et al., 2016).

The Rasch analysis data used in this section are displayed in four scatter plots that characterize two dimensions of each assessment item (i.e. question) (Figure 7). These plots are used to understand the distribution of the assessment items and people across the variables and identify specific questions that don't fit well. In other words, did all of the questions actually assess the students' understanding of variability?

The Rasch item plots (Figure 7 a-d, page 30) are evidence that the assessment is appropriate for all three groups of students. The assessment only has one underfit item (Infit t >2) for the high school January group and the middle school group and two for the highs school May group while the high school September group has none. The distribution of item difficulties on the Y axis indicates that the difficulty in all groups is a reasonable range from about +2 to -3 and the questions are evenly distributed throughout with no large gaps, indicating that relatively small improvements in ability should be reflected in assessment score.

5.1.2 Open-response assessment

The first iteration of the open-response assessment was given to an unrelated group of tenth grade students. In grading the responses, ambiguities in the rubric and questions were identified and modifications were made to address them. Problems with question clarity were identified when student responses did not address the intended target of the question, and problems with rubric clarity were identified when disagreements arose among the graders or when the rubric could not accurately score a reasonable response. The final scores on the 16 open-response assessments were deemed sufficiently reliable by 100% agreement among three scorers as determined by the scoring group.

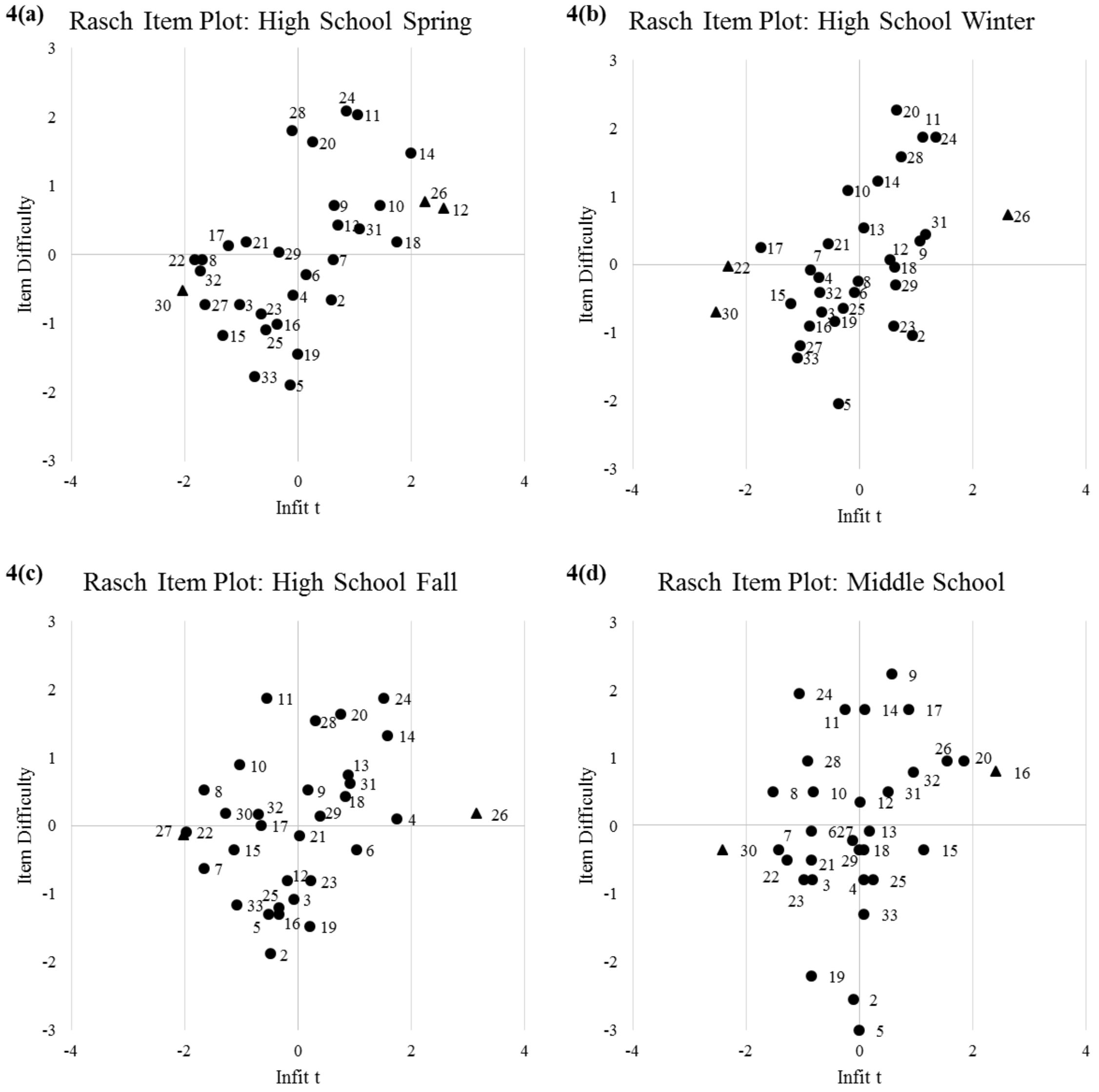


Figure 7 (a-d). Rasch item plots for snowpack students, September group, and middle school group. The Y axis represents item difficulty from easy (-3) to difficult (+2). The X axis represents fit where low values (<-2) are overfit and high values (>2) are over fit. Fit describes how predictable the responses are to the item with overfit items being more predictable than expected and underfit items being less predictable than expected.

5.2 Research question 1: To what degree do students participating in the inquiry-based Snowpack Project improve their understanding of graphing and variability by the end of the project?

5.2.1 January ("pre") versus May ("post") performance: Did Snowpack Project students score better on *ASK-Var* at the end of the project?

Paired scores collected from high school classrooms in January of 2015 and the May of 2015 were compared to identify changes in data literacy skills that may have occurred during the Snowpack Project. The underlying hypothesis was to find that after engaging in the project students would demonstrate improved understanding of variability as measured by their total *ASK-Var* scores at the end of the project.

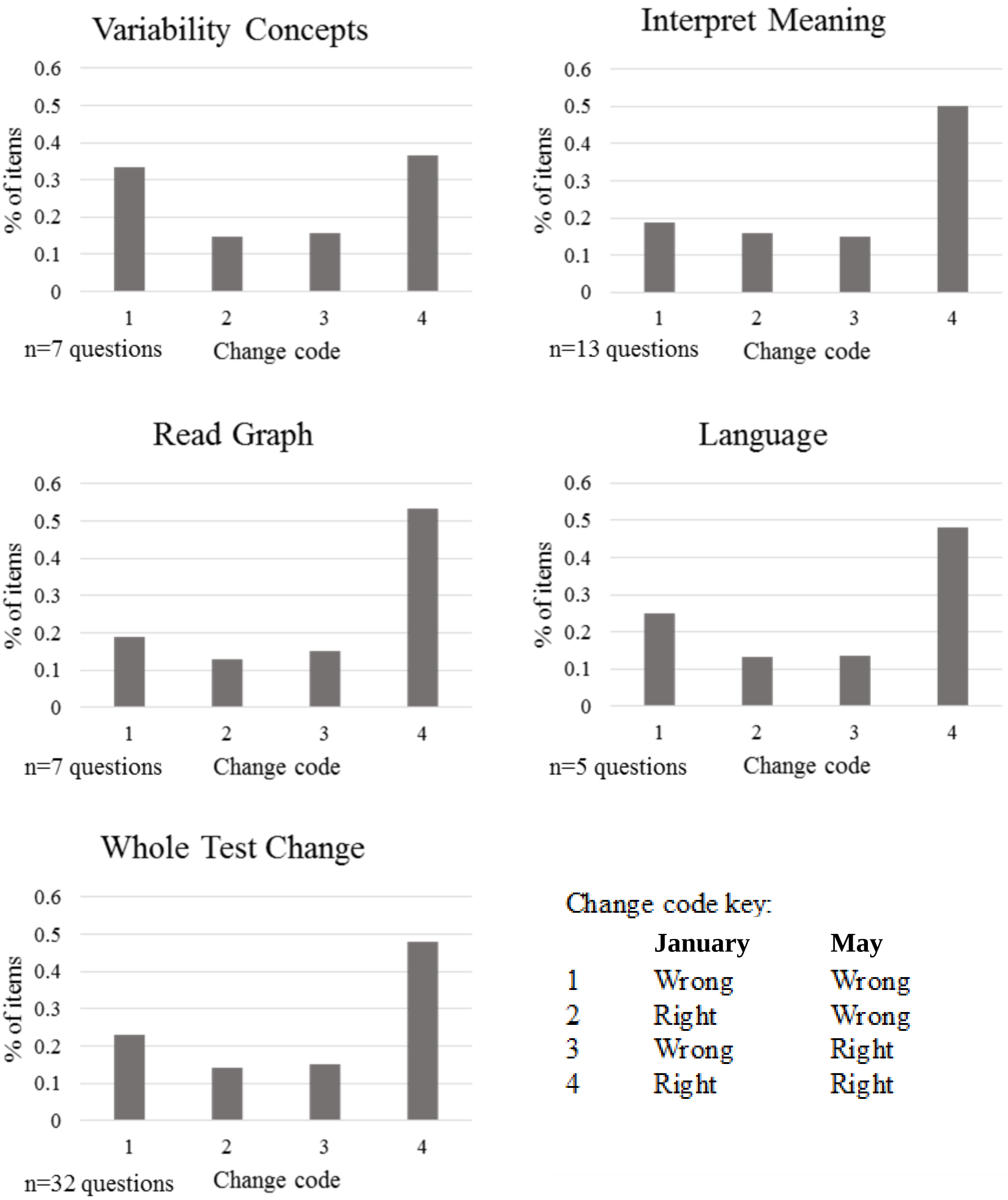
Table 3. Summary statistics for the January and May high school *ASK-Var*. None of the comparisons pre to post tested as significant (P<0.05).

(n=98)	**Whole Test (32 questions)**		**Variability Concep (7 questions)**		**Interpret Meaning (13 questions)**		**Read Graphs (7 questions)**		**Language (5 questions)**	
	January	**May**	**January**	**May**	**January**	**May**	**January**	**May**	**January**	**May**
Mean	0.62	0.63	0.51	0.52	0.61	0.62	0.66	0.65	0.66	0.68
SD	0.16	0.17	0.21	0.21	0.20	0.20	0.24	0.24	0.19	0.20
t	0.684		0.415		-0.429		1.144		0.197	
P (2-tailed)	0.495		0.679		0.669		0.255		0.844	

Initial inspection of the paired "pre-post" scores revealed no significant changes in mean score of all 98 pairs over the whole test or between any of the four topic areas (Table 3). With no change observed between the means of the January and May tests, the data were reanalyzed in three different ways to look at stability of responses and identify patterns in how responses changed (page 29). For example, if a large proportion of students shifted from correct responses in January to incorrect responses in May on questions related to a particular concept such as interpretation histograms, perhaps a new misconception was taught.

Results of the first response stability analysis are displayed in Figure 8, which includes a graph of scores on the whole test (Figure 8 e, page 40) and one for each topic area (Figure 8 a-d, page 40). The bars represent the proportion of total responses that fell into each of four categories of response change from January to May: incorrect to incorrect (code 1), correct to incorrect (code 2), incorrect to correct (code 3), correct to correct (code 4) (Figure 8). All graphs show a similar pattern. Code 4 represented the largest proportion of the responses in all four topic areas and for the entire test. The next largest proportion in all five cases was code 1. Codes 2 and 3 each represented about the same proportion of each topic area and the smallest proportions of the whole group.

The topic “Variability Concepts” followed these general trends, however, codes 1 and 4 represented more similar proportions of the population than in the other groups suggesting that this topic was initially more difficult for students than the other conceptual areas (Code 1), but students also learned similar amounts (Code 4). Codes 2 and 3 remained similar to each other and the codes 2 and 3 in other topic areas. This means that students likely started with less knowledge of Variability Concepts as assessed by the *ASK-Var* assessment but showed similar rates of misconceptions introduced (code 2) and knowledge gained (code 3) as other topic areas.



Change code key:

	January	**May**
1	Wrong	Wrong
2	Right	Wrong
3	Wrong	Right
4	Right	Right

Figure 8. Graphs characterizing changes in the paired *ASK-Var* responses from January to May on the whole assessment (e) and for groups of questions (a-d).

5.2.2 September versus May performance: How did *ASK-Var* scores in May compare with a new group of incoming student the following September?

The fact that there were no differences between the January and May assessments raised the possibility that students had already learned the content in the first semester. In interviews with the teachers, all four reported spending significant time on data literacy throughout the year starting in September 2015. To measure difference in student abilities at the beginning of the year compared to the end, the two ninth grade teachers whose students comprised a majority of the January/May sample gave the *ASK-Var* to their new students in September of the following year (2015). Scores of the high school students collected in September and May of 2015 were compared. The assumption was the new students would not have learned the data literacy concepts yet, and might be a proxy for the snowpack students at the beginning of the year.

Descriptive statistics and independent sample t-tests were calculated for the May and September assessments (September n=101 students, May n=98 students). No differences were observed between May and September means for the whole assessment scores or for any of the conceptual categories (Table 4). Mean scores across conceptual categories were similar to the January and May responses, with Variability Concepts scores being slightly lower than the other three. The mean score for Variability Concepts was 0.50 while the mean score for Interpret Meaning, Read Graphs, and Language were 0.61, 0.66, and 0.68 respectively. Results of the t-tests must be considered with caution because confounding variables such as differences in educational experiences, gender ratios, and socioeconomic backgrounds were not formally accounted for.

Table 4. Comparison between September and May *ASK-Var* results.

(Sept n=101) **(May n=98)**	**Whole Test** **(32 questions)**		**Variability Concept** **(7 questions)**		**Interpret Meaning** **(13 questions)**		**Read Graphs** **(7 questions)**		**Language** **(5 questions)**	
	Sept	**May**	**Sept**	**May**	**Sept**	**May**	**Sept**	**May**	**Sept**	**May**
Mean	0.62	0.63	0.50	0.52	0.61	0.62	0.66	0.65	0.68	0.68
SD	0.17	0.17	0.24	0.21	0.24	0.20	0.19	0.24	0.17	0.20
t	-0.602		-0.676		-1.314		-0.785		2.211	
P (2-tailed)	0.548		0.500		0.191		0.434		0.028	

5.2.3 High school versus middle school students: Can the *ASK-Var* pick up group differences?

The middle school data were included to see if the assessment was capable of detecting differences between two groups with a greater difference in age and education. It was expected that the high school students in the Snowpack Project would score higher than a group of middle school students outside the project simply because they have more learning experience in school and, being older, are more cognitively developed.

The mean score on the whole assessment and conceptual categories are summarized in Figure 9. The median scores for the whole assessment, Language, Interpret Meaning and Read Graphs was between 61% and 66% while the median score for Variability Concepts was somewhat lower at 51%. Comparisons between the Snowpack Project students' and middle school students' *ASK-Var* scores revealed statistically lower scores among middle school students for the whole assessment and in all three conceptual categories except Language. The mean score on the whole test for the middle school group was only 49% with a standard deviation of 17%. The seventh graders' performance was also analyzed based on the four conceptual categories introduced earlier. For Language, Interpret Meaning and Read Graphs, the students had mean scores of 54%, 52%, and 54% respectively; the mean score for Variability Concepts was 38% (Table 5, Figure 9).

Table 5. Summary statistics for seventh grade and high school (January) *ASK-Var* assessment (P=0.05)

(n=33)	**Whole Test (32 questions)**		**Variability Concept (7 questions)**		**Interpret Meaning (13 questions)**		**Read Graphs (7 questions)**		**Language (5 questions)**	
	Seventh Grade	**High School**	**Seventh Grade**	**High School**	**Seventh Grade**	**High School**	**Seventh Grade**	**High School**	**Seventh Grade**	**High School**
Mean	0.49	0.62	0.38	0.51	0.52	0.66	0.54	0.66	0.54	0.61
SD	0.17	0.16	0.24	0.21	0.16	0.24	0.29	0.19	0.21	0.20
t	-3.740		-2.810		-2.263		-4.215		-1.704	
P (2-tailed)	<0.001		0.007		0.028		<0.001		0.094	

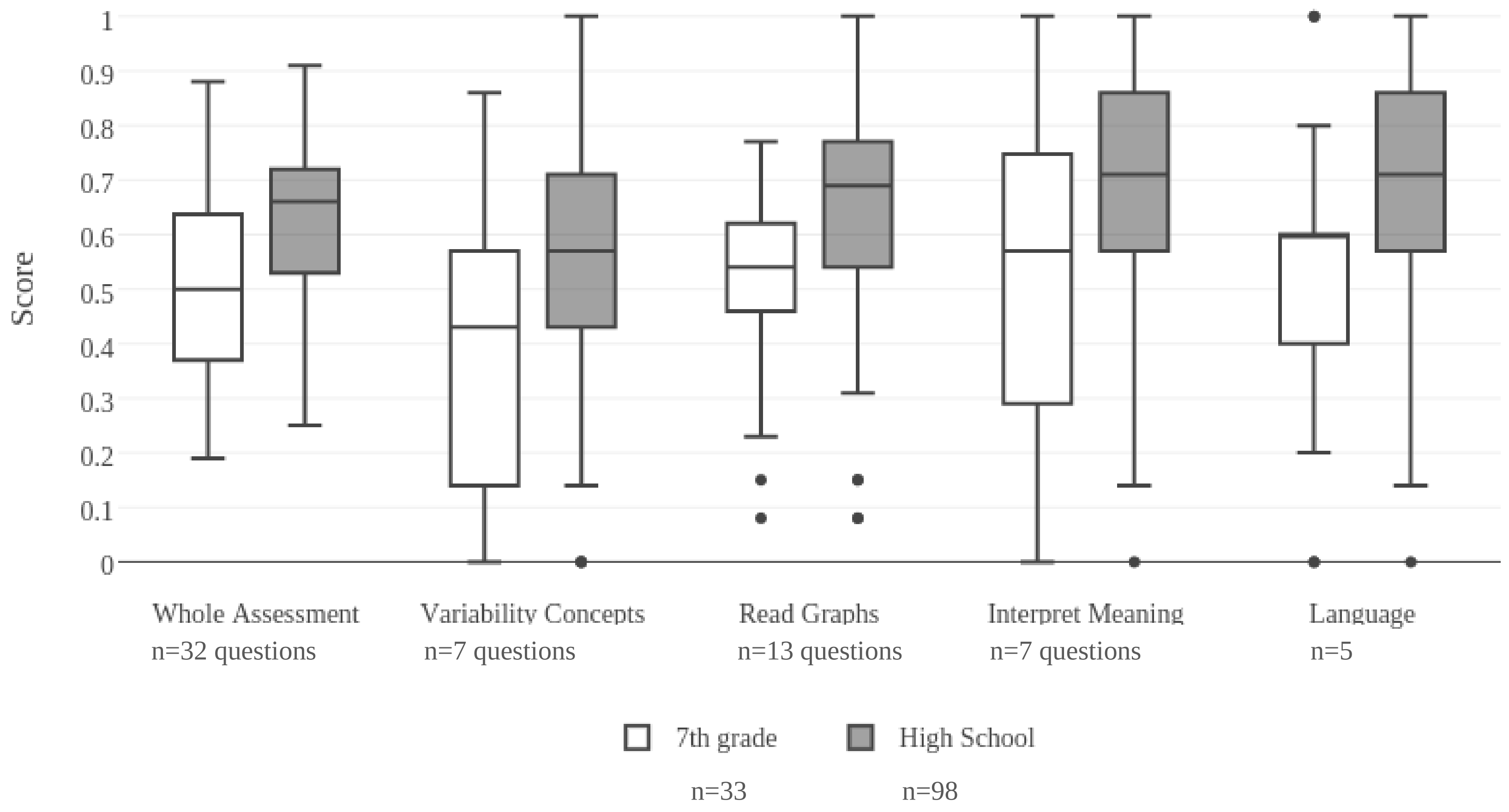


Figure 9. *ASK-Var* whole assessment scores and by conceptual category. The high school group includes responses collected in January. Middle school responses were collected sometime during the same winter from middle school students from different school districts by the Maine Data Literacy Project.

5.3 Research Question 2: To what extent are student scores on the open-response assessment aligned with how they perform on the *ASK-Var* assessment?

The open-response assessment results were compared to the May *ASK-Var* assessment results to evaluate how the two would correlate and identify interesting patterns. Analysis of the open-response assessment results revealed patterns in some concept areas and a lack of pattern in others.

Students' scores on the *ASK-Var* assessment were positively correlated to the open-response assessment (R^2=0.37). The correlation was calculated between the number of items scored as "Mostly Meets Expectation" or better (3 or 4 on the rubric) and the Rasch ability estimate (Figure 10).

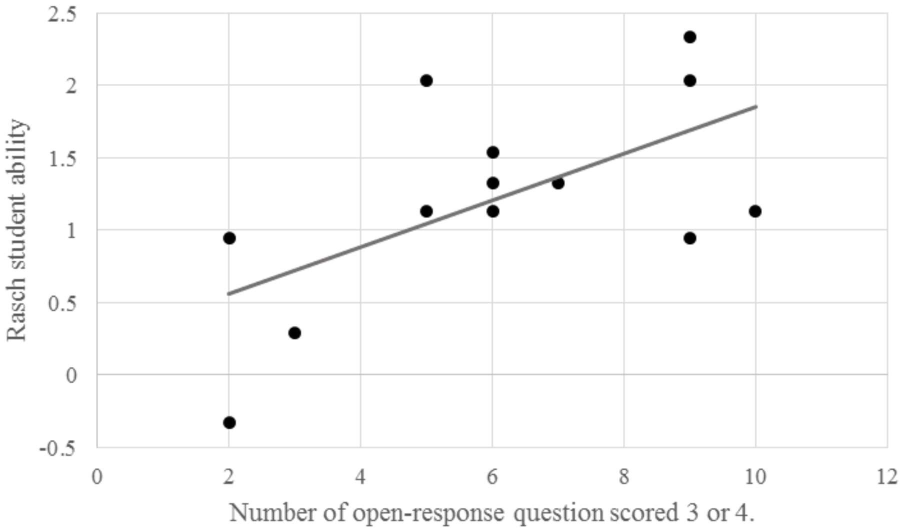


Figure 10. Correlation between total multiple-choice and open-response scores.

When observing patterns of responses on the open-response assessment for one question across the sample of students, there was no strong correlation. In nearly all cases, some students with higher abilities as measured by the *ASK-Var* assessment scored poorly on and open-response question, while students with lower abilities sometimes scored higher on the same question (Table 6).

Table 6. Sample of open-response question scores compared to total score on the *ASK-Var*. The responses are ordered by multiple-choice score; highest to lowest. Shaded cells indicate examples of students with high multiple-choice scores and low open-response scores (*Italics*), and students with low multiple-choice scores and high open-response scores (**Bold**).

Student Code	Multiple-Choice Score	Sample Open-Response Scores			
		Q2	**Q3**	**Q4**	**Q5**
T3_S_4	0.88	3	*1*	4	*2*
T3_S_7	0.84	4	2	4	*2*
T4_S_10	0.84	*2*	*1*	3	4
T3_S_8	0.78	4	2	*2*	1
T3_S_1	0.75	3	2	3	2
T3_S_2	0.75	2	2	4	1
T3_S_5	0.72	3	2	4	**4**
T4_S_13	0.72	3	**3**	3	**4**
T4_S_15	0.72	3	**3**	3	1
T3_S_3	0.69	**4**	2	**4**	2
T4_S_9	0.69	2	1	2	1
T4_S_12	0.56	**3**	2	3	2
T4_S_14	0.44	1	2	3	2

5.4 Qualitative analysis and observations

5.4.1 Interpretation of histograms

Six questions from the two assessments (*ASK-Var* Questions 19, 20, and 21 and open-response Questions 4, 5, and 6) were identified as assessing students' ability to interpret data represented in histograms. For clarity questions from the *ASK-Var* assessment will be labeled AV (ex. AV19) and questions from the open-response assessment will be labeled OR (ex. OR4).

The questions asked about histograms in three different ways. (1) A feature of the histogram was identified or students were asked the student to identify a feature of the histogram and interpret it in terms of the real-world phenomena that it represented (representation to reality) (AV 19 & 21, OR 4 & 6). (2) A feature of the real world was identified and students were asked to identify the portion of the graph that represented it (reality to representation) (AV 20). (3) Students were asked to describe the variability of the dataset as represented in the graph (variability) (OR 5).

Table 7. Summary of assessment questions about histograms. For the open-response assessment the count of number of correct responses represents students scoring 3 or 4 on the rubric. Representation to reality refers to questions that ask the student to interpret a feature of a graph and describe what it represents in reality. Reality to representation refers to question that ask the student to find how a feature of reality is represented in a graph. Variability refers to questions that focus on identifying and describing variability.

Assessment	**Question**	**Category**	**# of Correct Responses (n=13)**
Open-response	4	Representation to reality	11
	5	Variability	2
	6	Representation to reality	12
ASK-Var	19	Representation to reality	13
	20	Reality to representation	7
	21	Representation to reality	11

Responses to Question OR5 were also coded into three different groups. Group 1 included responses that did not address variability in any way, Group 2 began to address variability but only mentioned a measure of center or the spread, and Group 3 described variability in terms of a measure of center and the spread (Appendix D). The question characteristics can be found in Table 7, and the questions can be found in Appendix B.

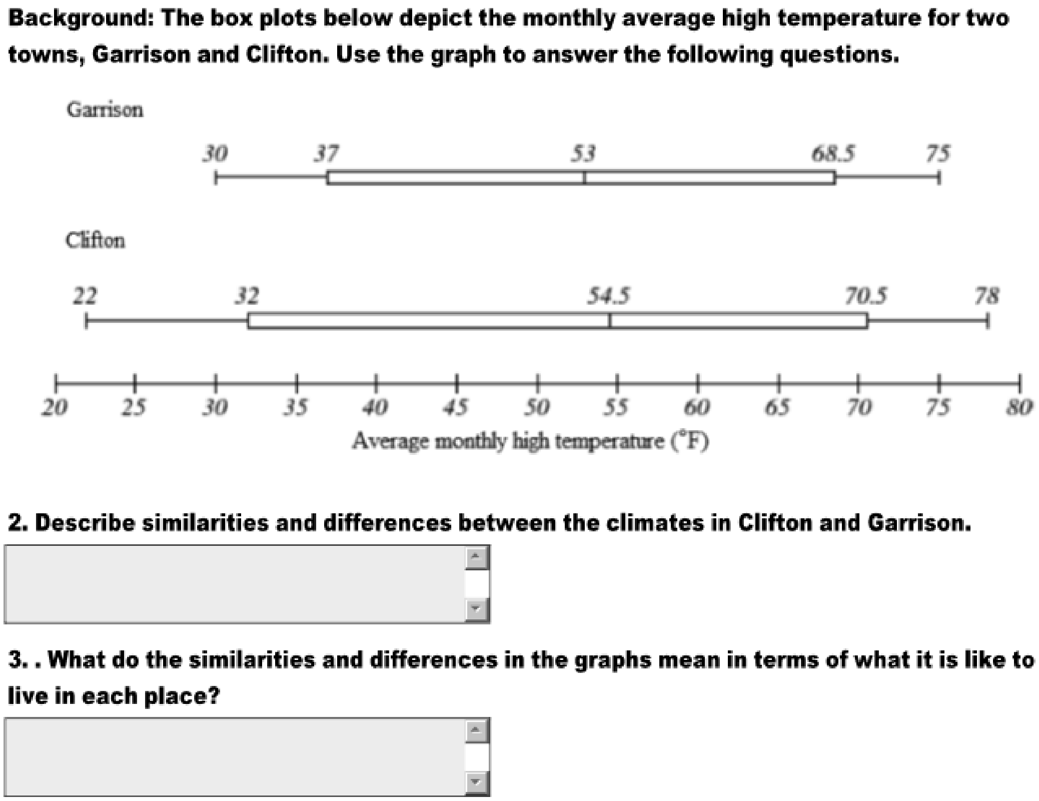
Table 8. Summary of histogram questions and scores. Scores are from the 13 paired samples of responses to the open-response and *ASK-Var* assessments administered in May of 2015.

Question Code	**Question Text**	**Class Summary Score (% correct)**
AV19	Which height range occurs most frequently among all of the trees?	100
AV20	How many trees are in the tallest group of trees?	54
AV21	What does the height of the tallest column mean in this histogram?	85

Question Code	**Question Text**	**Rubric Score**			
		4	**3**	**2**	**1**
OR4	What do the heights of the bars show?	5	6	2	0
OR5	Describe what this graph shows about the variability in timing of the first snowfall?	3	0	6	4
OR6	What prediction could you make about the most likely timing of the first snowfall next year?	10	2	1	0

5.4.2 Interpretation of boxplots

Two questions on the open-response assessment asked students to engage with data through boxplots (Figure 11). Question 2 asked students to asked students to describe the similarities and differences in the climate in two fictitious towns from data graphed in two boxplots. Question 3 asked students to describe how those similarities or differences might affect life in each town. Rubric scores were used to group responses according to group success rate (Figure 12), and themes from the responses were identified. One unifying theme from Questions 2 and 3 emerged. In Question 2, responses could be divided into two categories; responses that pointed to multiple concrete markers in the boxplots (median, quartiles, and whiskers) and those that did not. Responses to Question 3 could also be divided into two groups; those that correctly considered the importance of seasonal variation in comparing the variability of the two towns and those that did not (Figure 14).



Background: The box plots below depict the monthly average high temperature for two towns, Garrison and Clifton. Use the graph to answer the following questions.

2. Describe similarities and differences between the climates in Clifton and Garrison.

3. . What do the similarities and differences in the graphs mean in terms of what it is like to live in each place?

Figure 11. Open-response Questions 2 and 3 with the provided graph and context.

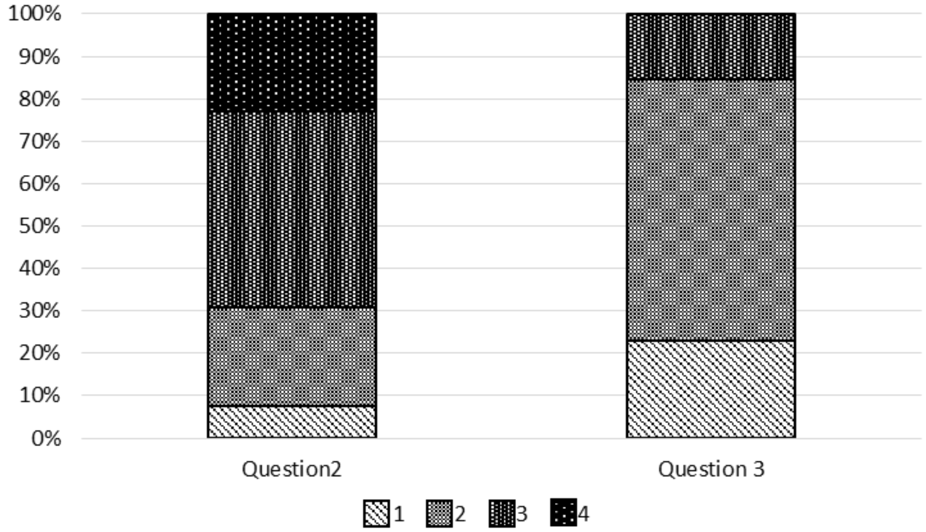


Figure 12. Summary of rubric scores on open-response Questions 2 and 3.

Considered seasonal temperature variation

> *"The climates are similar but Clifton seems to hotter hottest days and colder coldest days"*

Did not consider seasonal temperature variation

> *"Clifton's temperature is much more variable so it may be harder to predict the weather. They both are within the same range of temperatures from the 20's/30's to the 70's."*

Figure 13. Examples of two types of student responses from open-response question 2.

Identified concrete markers on the boxplot

> *"The Median of the two is similar but the min and max are further to either extreme in clifton than garrison"*

Did not identify concrete markers on the boxplot

> *"Clifton is more variable then Garrison and gets colder."*

Figure 14. Examples of two types of student responses from open-response question 3.

5.4.3 Graph choice

Analysis results of the graph construction task, Question 8 in the open-response assessment, show clear disparities of graph choice between the two classrooms (Figure 15). Students' choice of graph type in Question 8 varied distinctly by teacher. It also showed 8 out of 13 responses were some kind of frequency plot and 6 out of 13 graphed the comparison groups separately.

Question 10 asked students to explain the evidence in their graph that supported their claim (from Question 9) about which county had the longer growing season. Responses including rubric scores and full-text responses for Question 10 were compared with students' graph choices (Question 8) (Figure 16).

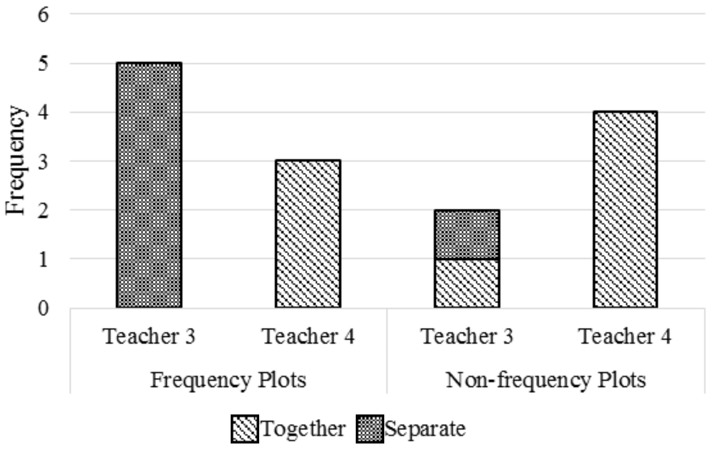


Figure 15. Classification of student-constructed graphs by type and data organization.

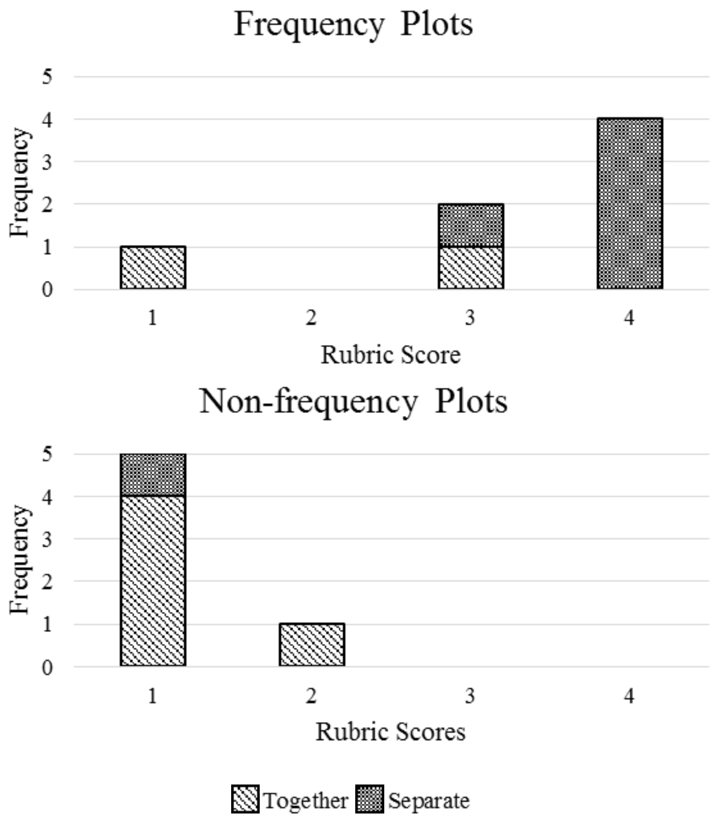


Figure 16. Summary of open-response graph construction and interpretation questions. "Together" denotes students that graphed the groups of data being compared in one group. "Separate" denotes students that graphed the groups of data being compared in two groups.

6 DISCUSSION

The objective of this thesis is threefold: to describe the degree to which students participating in the inquiry-based Snowpack Project improve their understanding of graphing and variability by the end of the project, to describe the extent to which open-response assessment results aligned with the *ASK-Var* assessment results, and to describe how students approached graphing and variability. To meet the first objective, the *ASK-Var* assessment was analyzed with the intent of identifying and measuring the content students learned and the areas in which there was no change. The second objective was met by analyzing posttest *ASK-Var* results and concurrent open-response assessment results to look for correlation. Finally, patterns that emerged from the open-response and *ASK-Var* assessment were explored qualitatively to shed light on how students thought about graphing and variability.

6.1 Is there evidence that students' understanding of variability improved after engaging in the Snowpack Project?

The *ASK-Var* assessment was initially administered twice: once in January before the Snowpack Project began (pretest) and once in May upon completion of the students' final presentations (posttest). When the pretest and posttest results showed no significant differences, the assessment was given again the following September to see if there were any detectable differences between a new cohort of students at the beginning of the year and the Snowpack Project students on the posttest given in May (non-Snowpack). When differences were not detected, Snowpack Project pretest results were compared with results from a group of seventh grade students from the Maine Data Literacy Project study to determine if the assessment could differentiate between groups with a greater disparity in age and education. Only the seventh grade total scores appeared different from the other three assessments (pretest, posttest, and non-Snowpack students). The Snowpack Project students (January and May 2015 scores) outperformed the seventh graders on the *ASK-Var* assessment, evidence that the assessment is able to detect differences between groups. The *ASK-Var* assessment appears to describe students' understanding of variability concepts on a coarse scale. Only the largest differences in experience between high school Snowpack Project students and seventh graders were detected in the *ASK-Var* scores.

The *ASK-Var* assessment was developed by the Maine Data Literacy Project as a potential formative assessment for teachers to "measure students' progress in learning to think about data aggregations and variability" (Zoellick et al., 2016). During the development process, the Rasch analysis was used to focus the questions on the central construct identified as "understanding variability as a property of data aggregations" (Zoellick et al., 2016). The *ASK-Var* authors recommend the assessment be used to characterize a group's abilities rather than to assess individuals, due to error values at the high and low extremes. The Rasch analysis was also used to verify that the assessment captured the full range of ability in each group, indicated by evenly distributed scores.

The Rasch item plots produced by the assessment data in this study showed characteristics of a good fit (Figure 7 on page 37). The items ranged in difficulty from accessible to most students to challenging for most students, and in every case (pretest, posttest, non-Snowpack, and seventh grade), no more than two assessment items were underfit or overfit indicating an assessment that was focused and fit the study group.

Though there was no clear pattern of gain or loss in students' understanding of variability as measured by the *ASK-Var* assessment, the Rasch data suggests it is a good measure of understanding of variability for these students. Several factors might have contributed to a lack of pattern in gains or losses pre and post: (1) Wild and Pfannkuch (1999) suggest that variability concepts and graph reading skills require more time and focused instruction to learn than the five months between the pretest and posttest, (2) the multiple-choice assessment may not have detected smaller changes in student understanding that did not move responses from incorrect to correct, or (3) the final assessment of this study was timed too close to the end of the school year to capture students' peak skills.

Describing and interpreting variability graphed in distributions such as box plots or histograms, requires students to perceive aggregate properties of datasets as the key features rather than focusing on individual points in the dataset. These skills are developed over years with repeated exposure and specific, targeted instruction (Konold et al., 2015). It is possible that the *ASK-Var* assessment did not detect learning gains because big shifts in conceptual understanding simply take more time than this study allowed.

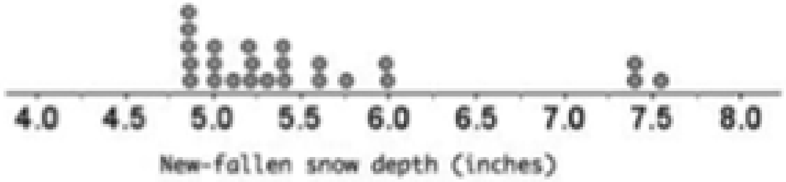
The Snowpack Project is an opportunity for students to engage in authentic scientific practices. It would be possible for a teacher to participate in the project without specific instruction on variability or even data literacy, and still engage his or her students in a rich experience learning about scientific process. Considering teachers' limited time resources and the number of diverse learning opportunities the Snowpack Project provided, including conducting scientific investigations and data collection and management, it would have been easy to neglect specific instruction on reasoning about variability, the subject measured by the *ASK-Var* assessment.

That said, the Snowpack Project provided opportunities for teachers and students to improve their content and data literacy skills. As part of the Snowpack Project, students designed and carried out an investigation to answer a scientific question of their choice about snowpack and created graphs from the data to help answer it. These skills are emphasized in NGSS practices 1, 3, 4, 7, and 8 and CCMS including 6.SP.A.2, and 6.SP.B.4 (Figure 1on page 28, Appendix E). To assist the teachers in teaching the content and skills, the Snowpack Project provided data-rich "mini lessons" on topics like data organization, presentation, and interpretation and professional development in data literacy. It also facilitated sharing current and past snow data from across the state among participating schools.

In addition to the broad focus of the Snowpack Project, reasoning about variability is traditionally not a focus of science class. However, since understanding of variability in the context of climate change is one of the primary goals of the Snowpack Project, performance on the assessments may also suggest that students need more direct instruction in variability concepts than the Snowpack Project teachers are currently offering. Since prioritizing data literacy instruction was left up to the teachers, simply providing access to professional development and lessons on variability and graphing may not be enough to get them to invest the requisite time and energy to improve students' understanding and skills.

Many of the most difficult question on the *ASK-Var* assessment required students to identify, describe, and/or synthesize and apply knowledge of variability to a graphical context at an eighth grade level. These include questions like 28 and 31 where students are presented with graphs and asked to determine how the distribution of the data might affect the differences between the mean, median, and mode. For example, Question 28 (Figure 17) shows a dot plot with a right skewed distribution and asks student to choose the correct statement from options like "The mean snow depth will be greater than the median" and "The mean snow depth will be the same as the mode." Answering these questions correctly requires student to be able to read

Below are the depths of new-fallen snow measured at 24 sites following a snowstorm. Use this graph to answer the next question.



28. Which of the following statements about the data presented in the snow-depth graph is correct?

□ The median snow depth will be greater than the mean snow depth.
□ The mean snow depth will be greater than the median snow depth.
□ The mean snow depth will be the same as the mode.
□ The mode is located in the cluster of points between 5 and 6 inches.

Figure 17. Question 28 from the *ASK-Var* assessment. It is an example of a question asking students to apply mean, median, and mode to a graphed dataset.

and interpret data represented in dot plots, apply their understanding of variability to the graph, and understand how variability can affect different measures of center.

Students were asked to define variability from three different perspectives in Questions 8, 11, and 14. They chose the best definition of variability using *non-technical* words with options like “The center of the group of values” and “How clumped or scattered the values are along a number line” in Question 8. They chose the set of *technical* words that best described variability in Question 11 from the following options: Range, center, distribution; mean, median, mode; group size and skew; and minimum and maximum values (see Appendix A). Finally, they picked the description of the data set with the most variability in question 14. Options include “All of the values are different – there are no repeats” and “The values are the most spread out from the middle.” Answering these questions required students to know the definition of variability and apply that definition in three different contexts. Scores were calculated from binary information on each question, right or wrong, and so did not detect shifts in student thinking that may have been more correct but did not shift multiple-choice responses all the way from wrong to right. This type of scoring can miss a lot of valid but subtle shifts in understanding that may be taking place. In addition, the spring assessment was administered in late May near the end of the school year. State testing had been completed and summer vacation was a few weeks away. Test fatigue could have reduced students’ ability to focus and reduced performance on the assessment.

In the final stages of the Snowpack Project, students all participated in individual or small group projects in which they analyzed snow data with respect to a scientific question. Examining 24 of those final projects representing 40 students showed that the students were largely focused on questions that considered variability. Most students asked at least one question of the data directly related to variability, and, with only three exceptions, all of the students who asked these questions used frequency plots in their presentations. This suggests that even though the *ASK-Var* assessment did not show gains, the Snowpack Project still provided students with opportunities to practice important data literacy skills in the context of the project.

6.2 Are *ASK-Var* and open-response scores correlated?

Carefully designed multiple-choice assessments can be useful indicators of student knowledge (Savinainen & Scott, 2002). They take less time for teachers to score than open-ended assessments do, but open-ended assessments can offer deeper insights into student thinking. The second objective of this thesis was to find out how well students' scores on the *ASK-Var* multiple-choice assessment correlated with their scores on an assessment with open-ended questions involving interpretation of weather-related data and frequency plots.

A subset of students who took the May *ASK-Var* assessment concurrently took the open-response assessment (n=13). The Rasch item difficulty score was used to represent student performance on the *ASK-Var* assessment, and the number of items in which a student scored a three or four on the rubric ("Mostly Meets Expectation" or better) represented student performance on the open-response assessment (See Appendix C to reference rubric). The strength of the weak positive correlation (R^2=0.37) is limited by the small number of participants and the small number of items in the open-response assessment.

Despite the correlation, none of the open-response questions except one (Question 6) discriminated well on its own relative to the *ASK-Var* assessment (Table 6 on page 21). While the whole open-response assessment *did* discriminate between higher and lower achievers on the *ASK-Var* assessment as measured by the correlation, on most questions one or two individuals received unexpectedly high or low open-response scores (see Table 6 on page 45).

Two explanations are possible: (1) the unexplained variability in the correlation may show that some of the open-ended questions assessed different kinds of knowledge or skills than the *ASK-Var* assessment or (2) there was confusion in the wording of the open-ended questions. The former seems likely because open-response questions require students to apply additional skills such as constructing a graph or writing an explanation without prompts whereas multiple-choice questions simply require students to choose among four possible responses. Writing skills in particular are absent in multiple-choice assessment responses but essential to open-responses assessments; the variability observed in scores could have reflected challenges students had with articulating their ideas rather than challenges with the ideas themselves.

The unevenly distributed open-response scores relative to the *ASK-Var* scores could also be attributed to questions with unclear wording or confusing expectations (Figure 10). The open-response assessment was written for this thesis and was not as thoroughly vetted as the *ASK-Var* assessment was. It underwent only one round of revisions with real student responses, and no responses were collected from students outside the snowpack project with the final version of the assessment prior to collecting data from Snowpack Project students. While no evidence was collected that could clearly disentangle the influences of the different set of skills required to complete the open-response assessment and the potentially unclear expectations on the open response assessment on assessment scores, I suspect both were contributing factors.

6.3 What do results reveal about how students think variability?

A deeper look at patterns in students' responses to specific *ASK-Var* multiple-choice and the open-response questions revealed three interesting observations related to how students interpret histograms, how they choose a graph type, and how they interpret box plots.

6.3.1 Interpretation of histograms

Both assessments had three questions that presented a distribution of data in a histogram. The *ASK-Var* questions (referred to as AV19, AV20, and AV21) asked students to interpret a histogram of heights of black cherry trees (Figure 5, page 34; Figure 6, page 34). The open-response questions (referred to as OR4, OR5, and OR6) asked students to interpret a graph of the dates of the year's first snowfall in Orono, Maine (Figure 6, page 34). The differences in performance on these questions asking student to interpret histograms in different ways suggests that while students do understand histograms in simpler settings, they are less comfortable thinking about them in more complex ways (Table 8, page 46).

AV19, AV20, and AV21 asked students to interpret the meaning of the heights of the bars of a histogram in three different ways. Questions AV19 and AV21 asked students to interpret features of the histogram and describe what they represented in the real world, and they performed well (13/13 correct and 11/13 correct respectively). Question AV20 asked students to consider a feature of the real world and find where it was represented in the histogram. Performance on this question was lower (7/13 correct).

The wording in question AV19 guides interpretation by using the words "height" and "frequency", both of which are found on the axis labels, and question AV21 identifies a specific feature, the tallest column, for students to interpret. These aids focus attention on key features in the graph. In addition, both questions asked students to look at the graph and describe the physical phenomenon it represents (the trees) such that the students were applying a simplified model to a more complex reality. The test writers did the challenging work of identifying and highlighting the information required to answer the question, and the students simply had to interpret it correctly.

But on question AV20 students had to identify the number of trees in the tallest group of trees. The question requires student to think about the complex reality that the graph represents, identify the important information, and apply it to an abstract model (the graph). All of the students who answered this question incorrectly referenced the tallest column rather than the group of the tallest trees. The students who answered Question AV20 incorrectly approached the problem in the same way they approached the other two, by first looking at the model then describing what it reflected concretely.

In the open-response data, Questions OR4, OR5, and OR6 show a similar pattern with a slightly different topic. The questions ask students to describe what the height of the columns in a histogram represent (OR4), describe the variability displayed in the graph (OR5), and use the graph to make a prediction (OR6). Similar to AV19 and AV21, OR4 required a straight-forward interpretation of a feature of the histogram. A successful response did not require the students to go beyond the information provided on the page, and, as expected, performance on this item was high with 11/13 students scoring “Partially Meets Expectation” or better on the rubric (3 or 4).

Students also performed well on OR6 with 12/13 scoring “Partially Meets Expectation” (3) or better, but this question is more complex. Such high success was surprising at first because making a prediction seems like a very different and more challenging task than does interpreting heights of bars on a given histogram as with AV19 and AV21 and OR4. However the two lines of questioning are similar. Like AV19, AV21, and OR4, OR 6 asked students to look at the graphical representation and apply it to the real world (make a prediction). Performance on this item may have also been helped by the students’ familiarity with the content which has been shown to improve graph interpretation (Roth et al., 2002).

Open-response Question 5 (OR5) posed a different challenge from the other five histogram questions. It asked the students to describe what the graph showed about the variability of the timing of the first snowfall. Success on this question required all of the histogram reading skills demonstrated in the other five questions plus an ability to interpret variability. Scores on OR5 were lowest among the histogram questions with only 2/13 students scoring a 3 or 4 on the rubric.

Student responses on this question varied in ways that the rubric did not discriminate between, so, to look more closely for patterns in how students were thinking about the question, their responses were coded into one of three groups: those who did not describe variability (Group 1), those who described variability using only the measure of center or the spread (ex. range) (Group 2), and described variability using both the measure of center and the spread of the data (Group3). While many of the responses’ codes and rubric score aligned, Figure 18 is an example of a student response that not. The coded results show that while almost half of the students did not describe variability at all, one third addressed both a measure of center and the spread of the data (Figure 19).

The high rate of success on the other five histogram questions suggests that describing variability was not a challenge for these students because they did not understand significance of the heights of the bars of a histogram. Other barriers such as vocabulary or a conceptual understanding of variability were more likely holding them back.

> *"The first snowfall happens around ranges usually around the middle of October to the middle of November but has been as late as the end of November in some years."*

Figure 18. Open-response to Question 5 from student T3_S_1. The response scored a 2 with the rubric and a 3 with the code.

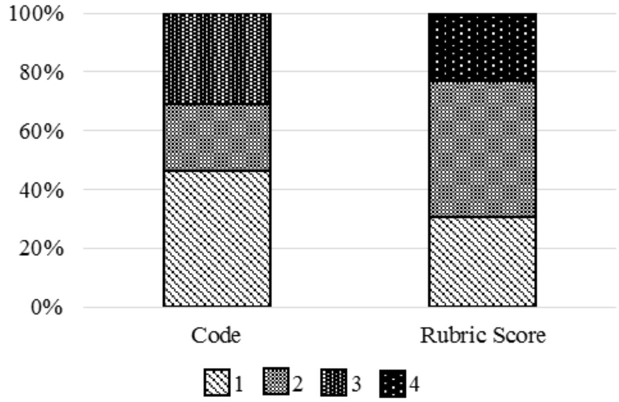


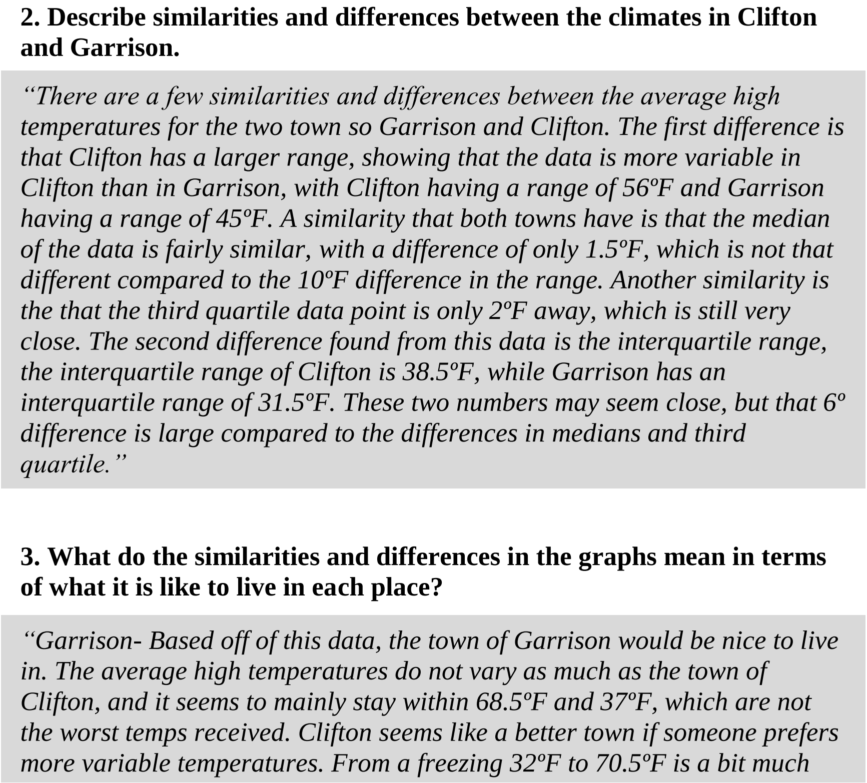
Figure 19. Graph of code groups and rubric scores for open-response Question 5.

These examples demonstrate that students have a general grasp of what histograms represent and how to read them. In addition, they can take information from a histogram and apply it to the real world situation that it represents, but they struggle when asked to go the other way around, from the real world to the histogram. Also, describing and applying the concept of variability is a big challenge for many of these students.

6.3.2 Interpretation of boxplots

Most students (9/13) were able to interpret the box plot in open-response Question 2 to describe similarities and differences between the climates in two fictitious towns, Clifton, and Garrison. The students identified the markers on the boxplots (median, quartiles, and maximum and minimum values) and compared them between graphs. All students but two, even those who did not answer the question correctly, referenced at least one of the reference points on a boxplot. Having these concrete markers may be an advantage of boxplots over other frequency plots for students who are first learning to describe and interpret distributions and variability because, unlike histograms and dot plots, box plots give students concrete reference points that can be used to discuss the data.

Open-response Question 3 asked students to interpret the boxplot to describe what the similarities and differences in the graphs meant in terms of what it was like to live in each place (Figure 11 on Page 47). Even students who described the box plot accurately and completely did not address the fact that the graph was showing a year of seasonal variation in temperature (Figure 20).



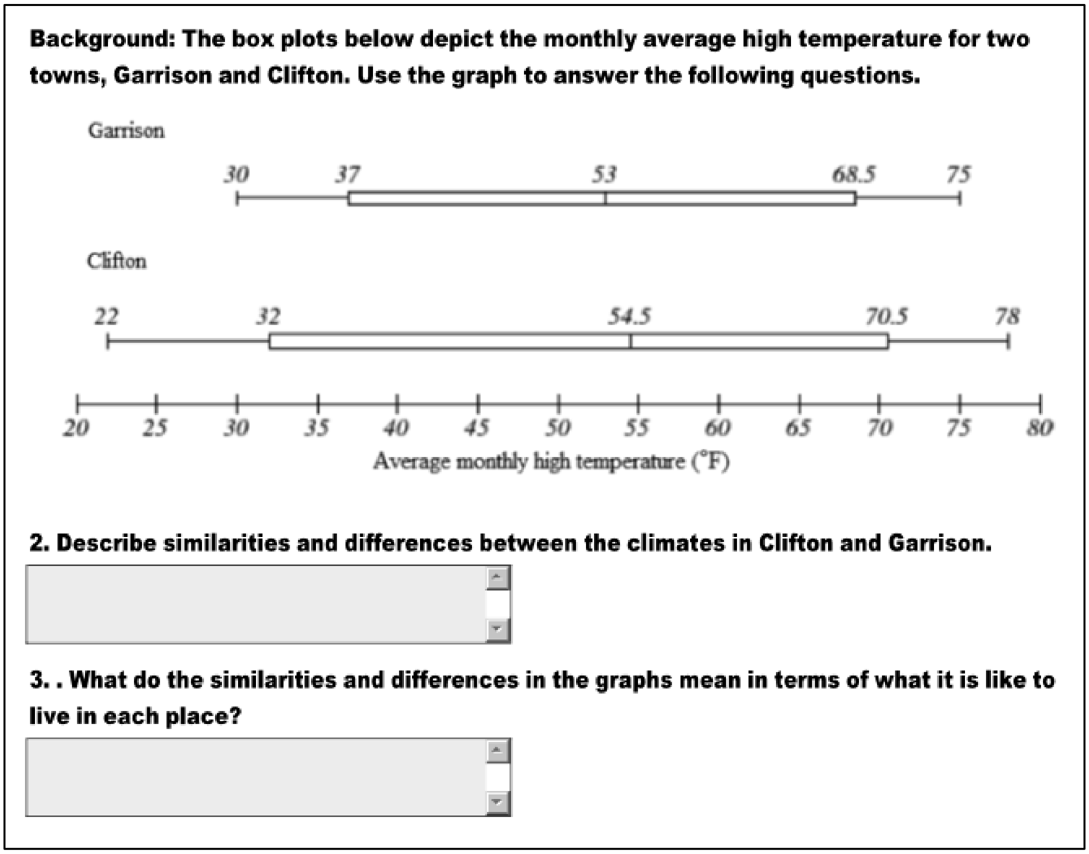
2. Describe similarities and differences between the climates in Clifton and Garrison.

“There are a few similarities and differences between the average high temperatures for the two town so Garrison and Clifton. The first difference is that Clifton has a larger range, showing that the data is more variable in Clifton than in Garrison, with Clifton having a range of 56ºF and Garrison having a range of 45ºF. A similarity that both towns have is that the median of the data is fairly similar, with a difference of only 1.5ºF, which is not that different compared to the 10ºF difference in the range. Another similarity is the that the third quartile data point is only 2ºF away, which is still very close. The second difference found from this data is the interquartile range, the interquartile range of Clifton is 38.5ºF, while Garrison has an interquartile range of 31.5ºF. These two numbers may seem close, but that 6º difference is large compared to the differences in medians and third quartile.”

3. What do the similarities and differences in the graphs mean in terms of what it is like to live in each place?

“Garrison- Based off of this data, the town of Garrison would be nice to live in. The average high temperatures do not vary as much as the town of Clifton, and it seems to mainly stay within 68.5ºF and 37ºF, which are not the worst temps received. Clifton seems like a better town if someone prefers more variable temperatures. From a freezing 32ºF to 70.5ºF is a bit much

Figure 20. Response to open-response Questions 2 and 3 from student T3_S_7. It is an example of a student who described the boxplot completely (Q2), but did not consider seasonal variation in the interpretation (Q3).



Background: The box plots below depict the monthly average high temperature for two towns, Garrison and Clifton. Use the graph to answer the following questions.

2. Describe similarities and differences between the climates in Clifton and Garrison.

3. . What do the similarities and differences in the graphs mean in terms of what it is like to live in each place?

Figure 21. Open-response Questions 2 and 3 with context.

Many students focused on how easy it would be to predict the temperature in each town rather than interpreting the wide box to mean hotter summer days and colder winter days in Clifton (Figure 21). For example, student T3_S_2 said “Clifton's temperature is much more variable so it may be harder to predict the weather. They both are within the same range of temperatures from the 20's/30's to the 70's.” While the ability to predict the outcome of an event is related to the variability of the data, it is not important for Question 3. To interpret implication of a wide range and wide interquartile range for actual seasonal weather, students needed to bring in outside knowledge about seasonal temperature variation and apply that to their graph interpretation. Only two students included a discussion of summer and(or) winter temperatures in their response. The low performance on Question 3 appears to have less to do with ability to interpret box plots quantitatively than the ability to apply that interpretation of a real-world context in terms of question being asked. It may suggest that ability to interpret graphed data *in its context* is as important as ability to mechanically read a graph, and that using data from a familiar context may help students with interpretation.

6.3.3 Choice of graph type

The graph construction task in the open-response assessment (Question 8) asked students to construct a graph of the data provided to help them decide which of the two fictional counties, Jones and Highland, had the longer growing season (Figure 21). An ideal graph to help answer this would be a frequency plot such as a boxplot or a dot plot, and it would plot data for each county separately (Figure 22). Students' responses were thus classified by whether or not a frequency plot was used to graph the data and whether or not the students divided the data into two categories to compare them (Figure 16 on Page 50).

Students who drew frequency plots described variability when comparing groups more than those who did not draw frequency plots. However a number of other factors in this study also aligned with the students' ability to accurately explain how evidence supports their claim. All but two of the students who used frequency plots and all of the students who graphed the two counties separately came from one teacher's classroom who had engaged in at least one year of professional development in data literacy before joining the Snowpack Project. The other teacher had only received professional development in data literacy through the Snowpack Project. This observation suggests that measurable improvement in students data literacy skills may result only after extended professional development and, for students, longer classroom exposure to practice with data than a few months of an authentic science project.

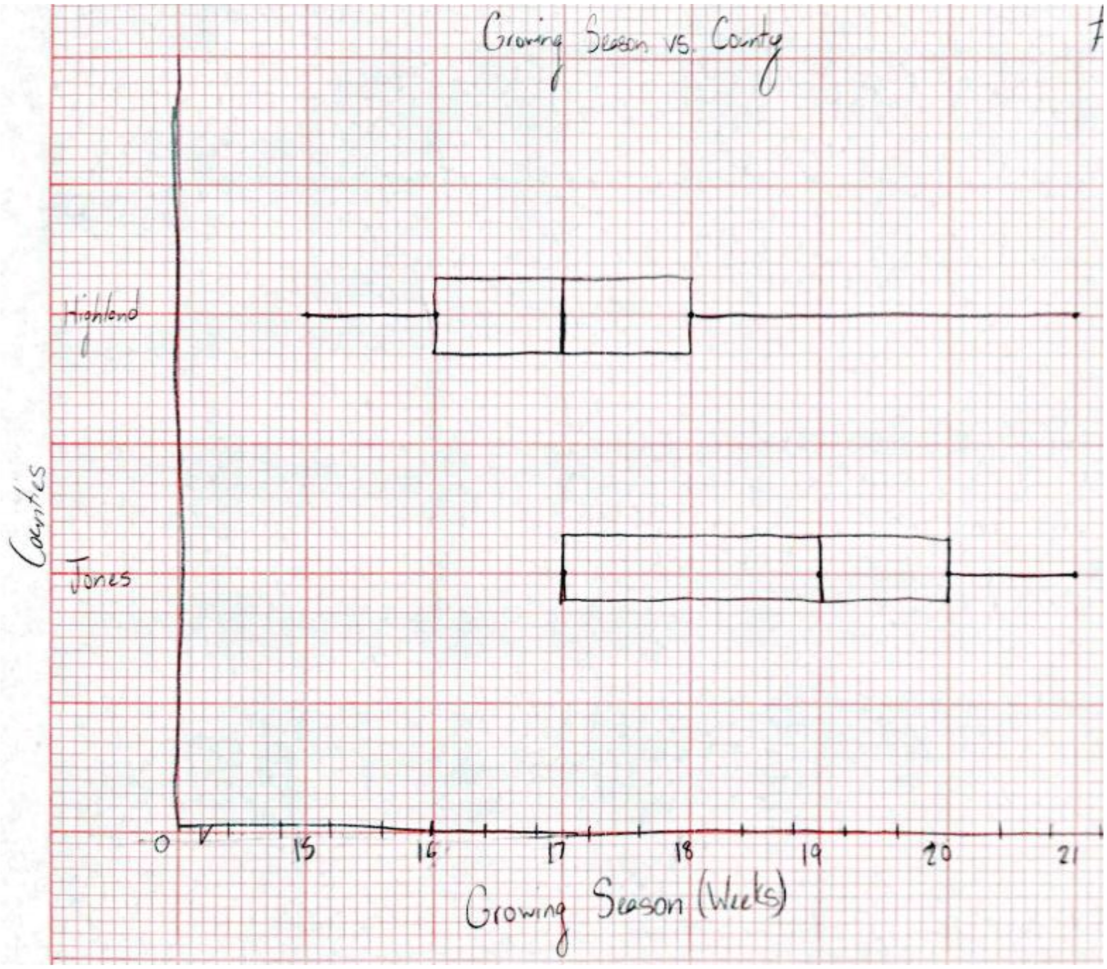


Figure 22. Frequency plot created by student T3_S_7.

7 CONCLUSIONS

7.1 Key findings

The data collected by the *ASK-Var* assessment did not register any gains in understanding of variability and graphing for participants of the Snowpack Project though it did detect a difference between the high school Snowpack Project students and a group of middle school students. The *ASK-Var* assessment may be of limited value for measuring change over a short period of time as applied in this study, but it still appears to be a useful instrument for characterizing a group of students' understanding of variability concepts and related graphing skills.

In every instance, performance on the Variability Concepts category was the lowest of the four conceptual categories. This reinforces the idea that students struggle with understanding variability.

Evidence from the Rasch analysis of the *ASK-Var* data and the correlation between the open-response and *ASK-Var* results suggest the *ASK-Var* assessment is a valid assessment tool for measuring understanding of variability concepts. However, it did not detect changes in student understanding of graphing and variability over the course of the Snowpack project. There were several confounding variables that may have contributed to the result including the year-end timing of the assessment, the assessment's ability to discriminate between small changes in student conceptions, and insufficient time for teaching or learning about variability to take hold.

7.2 Application of the study to classrooms

The *ASK-Var* assessment can be used to point teachers towards areas of weakness in the class and target specific instruction throughout the year to address the class' most challenging concepts. However, it is important to keep in mind that there may be a lot of important skills that might not be measured. The open-response assessment is still a useful tool to help students practice written communication skills, proper data representation, and independent (unprompted) data interpretation.

The large standard deviations in the *ASK-Var* assessment results indicate that the groups of students in the sample have a wide range of skills and abilities. It is important for teachers to keep this in mind when planning lessons on variability and include options for differentiating the lesson for a wide range of skills.

Science class is an excellent opportunity to use math skills as problem-solving tools that support more authentic learning experiences. Supporting students in using their math skills in new ways requires science teachers to understand the math content and pedagogy. Like students, teachers benefit from ongoing professional development and support like the Snowpack Project and the Maine Data Literacy Project provides.

Finally, this study described some challenges students had with graph interpretation. Regarding the histogram example, students fell short when applying a physical phenomenon like tree heights to a graph despite demonstrating the ability to interpret the graph the other way around (ex. describing what the graph showed about tree heights). In the boxplot example, few students applied an understanding of seasonal variation to their graph interpretations. These examples suggest that context is essential to thorough graph interpretation, and students need to be supported in applying context even if it is familiar.

7.3 Future research

The Snowpack Project offered students opportunities to engage in science in two primary ways: experience in authentic scientific practices and opportunity to learn a variety of data literacy skills by working with real data. Considering the diversity of potential benefits, I would suggest expanding the scope of the research questions to encompass project evaluation and consider incorporating qualitative methods like interviews and records of how time was spent in the classroom into future work. Using a mixed methods study design could capture a much wider range of benefits and offer new insights into the functioning of the assessment itself. It would be interesting to investigate not only the types of content learned or neglected by Snowpack Project participants, but also the changes in attitudes towards science and confidence in engaging with novel data outside the classroom.

There are opportunities to continue exploring relevant applications of the *ASK-Var* assessment with larger sample sizes over longer periods of time. It would produce more statistically robust data and come from a progression of age groups to investigate how thinking about graphing and variability may change on a year-to-year basis.

A final potential direction is conducting a qualitative study of how the Snowpack Project affects students thinking and attitudes about variability, graphing, and scientific practices.

REFERENCES

Ainley, J., Pratt, D., & Nardi, E. (2001). Normalising: children's activity to construct meanings for trend. *Educational Studies in Mathematics*, *45*(1–3), 131–146.

Bakker, A. (2004). Reasoning about shape as a pattern of variability. *Statistics Education Research Journal*, *3*(2), 64–83.

Bakker, A., & Derry, J. (2011). Lessons from Inferentialism for Statistics Education. *Mathematical Thinking and Learning*, *13*(1–2), 5–26. https://doi.org/10.1080/10986065.2011.538293

Banilower, E. R. (2002). *2000 National Survey of Science and Mathematics Education: Status of High School Physics Teaching*. Retrieved from http://2000survey.horizon-research.com/reports/high_physics.php

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. A., Campbell, K. M., & Weis, A. M. (2013). Report of the 2012 National Survey of Science and Mathematics Education. *Horizon Research, Inc.* Retrieved from http://eric.ed.gov/?id=ED548238

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42–63.

Binns, I. C. (2013). A Qualitative Method to Determine How Textbooks Portray Scientific Methodology. In M. S. Khine (Ed.), *Critical Analysis of Science Textbooks* (pp. 239–258). Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-94-007-4168-3_12

Bond, T., & Fox, C. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Second). Mahwah, NJ: Routledge.

Bowen, G. M., & Roth, W. (2003). Graph interpretation practices of science and education majors. *Canadian Journal of Science, Mathematics and Technology Education*, *3*(4), 499–512. https://doi.org/10.1080/14926150309556585

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*, *18*(1), 32–42. https://doi.org/10.3102/0013189X018001032

Brown, S., & Melear, C. (2007). Preservice teachers' research experiences in scientists' laboratories. *Journal of Science Teacher Education*, *18*(4), 573–597.

Budiansky, S. (2001). The Trouble with Textbooks. *ASEE Prism*, *10*(6), 24–27.

Chiappetta, E. L., & Fillman, D. A. (2007). Analysis of Five High School Biology Textbooks Used in the United States for Inclusion of the Nature of Science. *International Journal of Science Education*, *29*(15), 1847–1868. https://doi.org/10.1080/09500690601159407

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *SERJ Editorial Board*, 92.

Gibson, H. L., & Chase, C. (2002). Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. *Science Education*, *86*(5), 693–705. https://doi.org/10.1002/sce.10039

Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal*, *3*(2), 7–16.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational Technology Research and Development*, *48*(3), 23–48.

Hubisz, J. (1998). *Review of Middle School Physical Science Texts*. New York, NY: Glencoe/McGraw-Hill.

Hubisz, J. (2003). Middle-school texts don't make the grade. *Physics Today*, *56*(5), 50–54. https://doi.org/10.1063/1.1583534

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist*, *41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

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

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., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.

Meletiou-Mavrotheris, M., & Paparistodemou, E. (2015). Developing students' reasoning about samples and sampling in the context of informal inferences. *Educational Studies in Mathematics*, *88*(3), 385–404. https://doi.org/10.1007/s10649-014-9551-5

Moore, D. S. (1997). New Pedagogy and New Content: The Case of Statistics. *International Statistical Review / Revue Internationale de Statistique*, *65*(2), 123. https://doi.org/10.2307/1403333

Morris, B. J., Masnick, A. M., Baker, K., & Junglen, A. (2015). An Analysis of Data Activities and Instructional Supports in Middle School Science Textbooks. *International Journal of Science Education*, *37*(16), 2708–2720. https://doi.org/10.1080/09500693.2015.1101655

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington D.C.: National Governors Association Center for Best Practices, Council of Chief State School Officers. Retrieved from http://commoncore.aetn.org/training/ccssm-2nd/CCSSM-2nd-Handout.pdf

National Research Council (N. R. C.). (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, D.C.: National Academies Press.

National Science Education Standards. (1996). Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/4962

NGSS Lead States. (2013). Next generation science standards: For States, By States. Retrieved from http://link.springer.com/article/10.1007/s10972-014-9379-y

Park, D.-Y., & Lavonen, J. (2013). An Analysis of Standards-Based High School Physics Textbooks of Finland and the United States. In M. S. Khine (Ed.), *Critical Analysis of Science Textbooks* (pp. 219–238). Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-94-007-4168-3_11

Project 2061. (1993). *Benchmarks for Science Literacy*. Oxford University Press.

Rahm, J., Miller, H. C., Hartley, L., & Moore, J. C. (2003). The value of an emergent notion of authenticity: Examples from two student/teacher-scientist partnership programs. *Journal of Research in Science Teaching*, *40*(8), 737–756. https://doi.org/10.1002/tea.10109

Reading, C. (2004). Student Description of Variation While Working with Weather Data. Statistics Education Research Journal. Retrieved from https://e-publications.une.edu.au/vital/access/manager/Repository/une:190?to=&query2=&field1=creator&conjunction2=AND&query1=reading&field2=Text&from=&source=Advanced&conjunction1=AND&query3=&field3=Text

Roth, W.-M. (1996). Where IS the Context in Contextual Word Problem?: Mathematical Practices and Products in Grade 8 Students' Answers to Story Problems. *Cognition and Instruction*, *14*(4), 487–527.

Roth, W.-M., Bowen, G. M., & Masciotra, D. (2002). From thing to sign and "natural object": Toward a genetic phenomenology of graph interpretation. *Science, Technology & Human Values*, *27*(3), 327–356.

Roth, W.-M., & McGinn, M. K. (1998). Inscriptions: Toward a Theory of Representing as Social Practice. *Review of Educational Research*, *68*(1), 35. https://doi.org/10.2307/1170689

Roth, W.-M., McGinn, M. K., & Bowen, G. M. (1998). How prepared are preservice teachers to teach scientific inquiry? Levels of performance in scientific representation practices. *Journal of Science Teacher Education*, *9*(1), 25–48.

Rutherford, F. J., & Ahlgren, A. (1991). *Science for All Americans*. Oxford University Press.

Savinainen, A., & Scott, P. (2002). The Force Concept Inventory: a tool for monitoring student learning. *Physics Education*, *37*(1), 45.

Stern, L., & Roseman, J. E. (2004). Can middle-school science textbooks help students learn important ideas? Findings from project 2061's curriculum evaluation study: Life science. *Journal of Research in Science Teaching*, *41*(6), 538–568. https://doi.org/10.1002/tea.20019

Vahey, P., Yarnall, L., Patton, C., Zalles, D., & Swan, K. (2006). Mathematizing middle school: Results from a cross-disciplinary study of data literacy. In *Annual Meeting of the American Educational Research Association, San Francisco, CA*. Retrieved from http://www.academia.edu/download/41242182/Mathematizing_middle_school_Results_from20160114-3111-14f2rzp.pdf20160115-19908-kgtg8j.pdf

Valanides, N., Papageorgiou, M., & Rigas, P. (2013). Science and Science Teaching. In M. S. Khine (Ed.), *Critical Analysis of Science Textbooks* (pp. 259–286). Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-94-007-4168-3_13

van Eijck, M., & Roth, W.-M. (2009). Authentic science experiences as a vehicle to change students' orientations toward science and scientific career choices: Learning from the path followed by Brad. *Cultural Studies of Science Education*, *4*(3), 611–638. https://doi.org/10.1007/s11422-009-9183-8

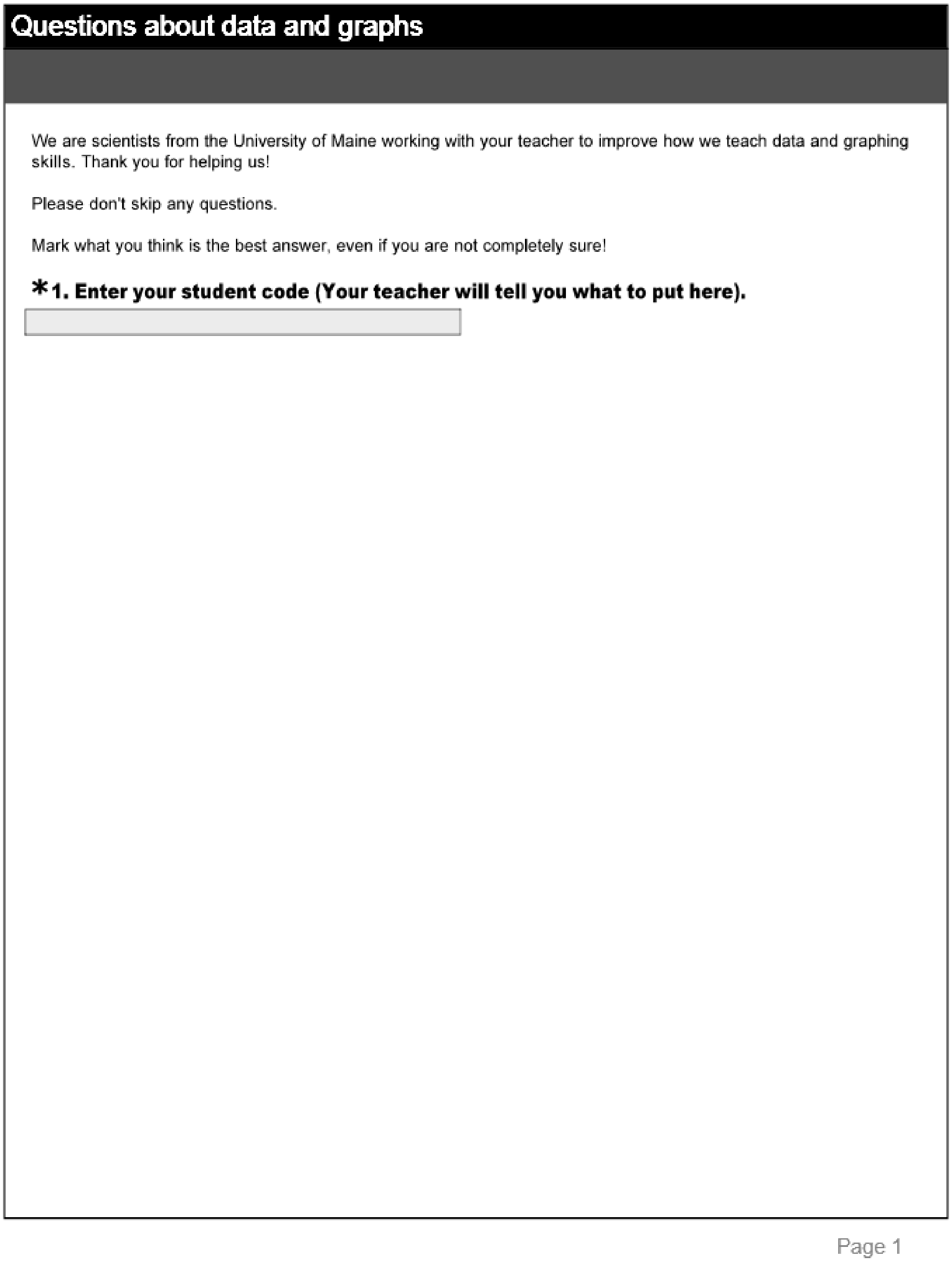
Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248.

Wu, H.-K., & Krajcik, J. (2003). *Inscriptional Practices in Inquiry-based Classrooms: How do Seventh Graders Construct and Interpret Data Tables and Graphs?*

Zoellick, B., Schauffler, M., Flubacher, M., Weatherbee, R., & Webber, H. (2016). Data Literacy: Assessing Student Understanding of Variability in Data. Retrieved from https://www.researchgate.net/profile/Bill_Zoellick/publication/301802243_Data_Literacy_Assessing_Student_Understanding_of_Variability_in_Data/links/5728e0fc08ae057b0a033a02.pdf

APPENDICES

Appendix A: *ASK-Var* assessment



Questions about data and graphs

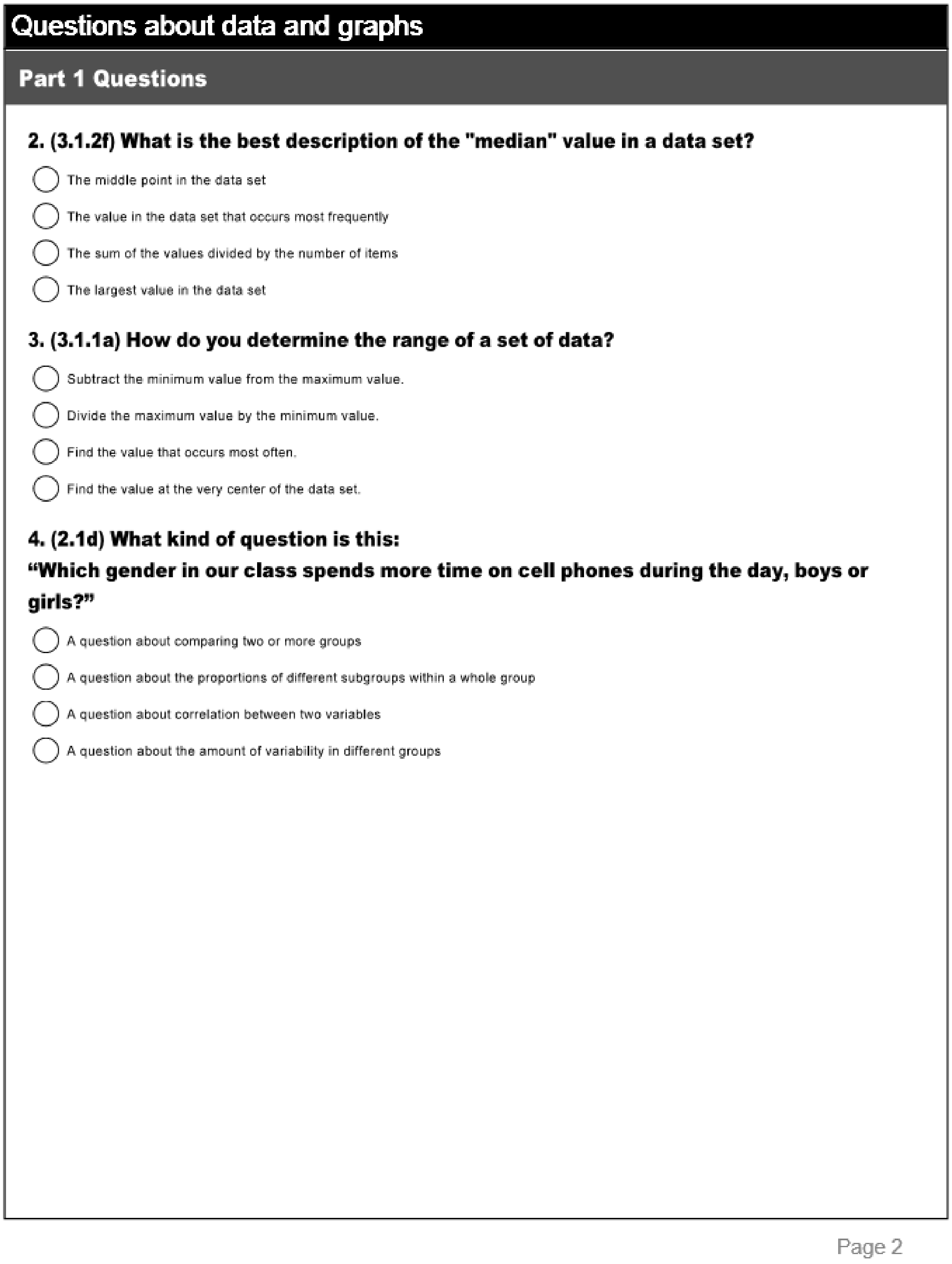
We are scientists from the University of Maine working with your teacher to improve how we teach data and graphing skills. Thank you for helping us!

Please don't skip any questions.

Mark what you think is the best answer, even if you are not completely sure!

***1. Enter your student code (Your teacher will tell you what to put here).**

Page 1



Questions about data and graphs

Part 1 Questions

2. (3.1.2f) What is the best description of the "median" value in a data set?

○ The middle point in the data set

○ The value in the data set that occurs most frequently

○ The sum of the values divided by the number of items

○ The largest value in the data set

3. (3.1.1a) How do you determine the range of a set of data?

○ Subtract the minimum value from the maximum value.

○ Divide the maximum value by the minimum value.

○ Find the value that occurs most often.

○ Find the value at the very center of the data set.

4. (2.1d) What kind of question is this:
“Which gender in our class spends more time on cell phones during the day, boys or girls?”

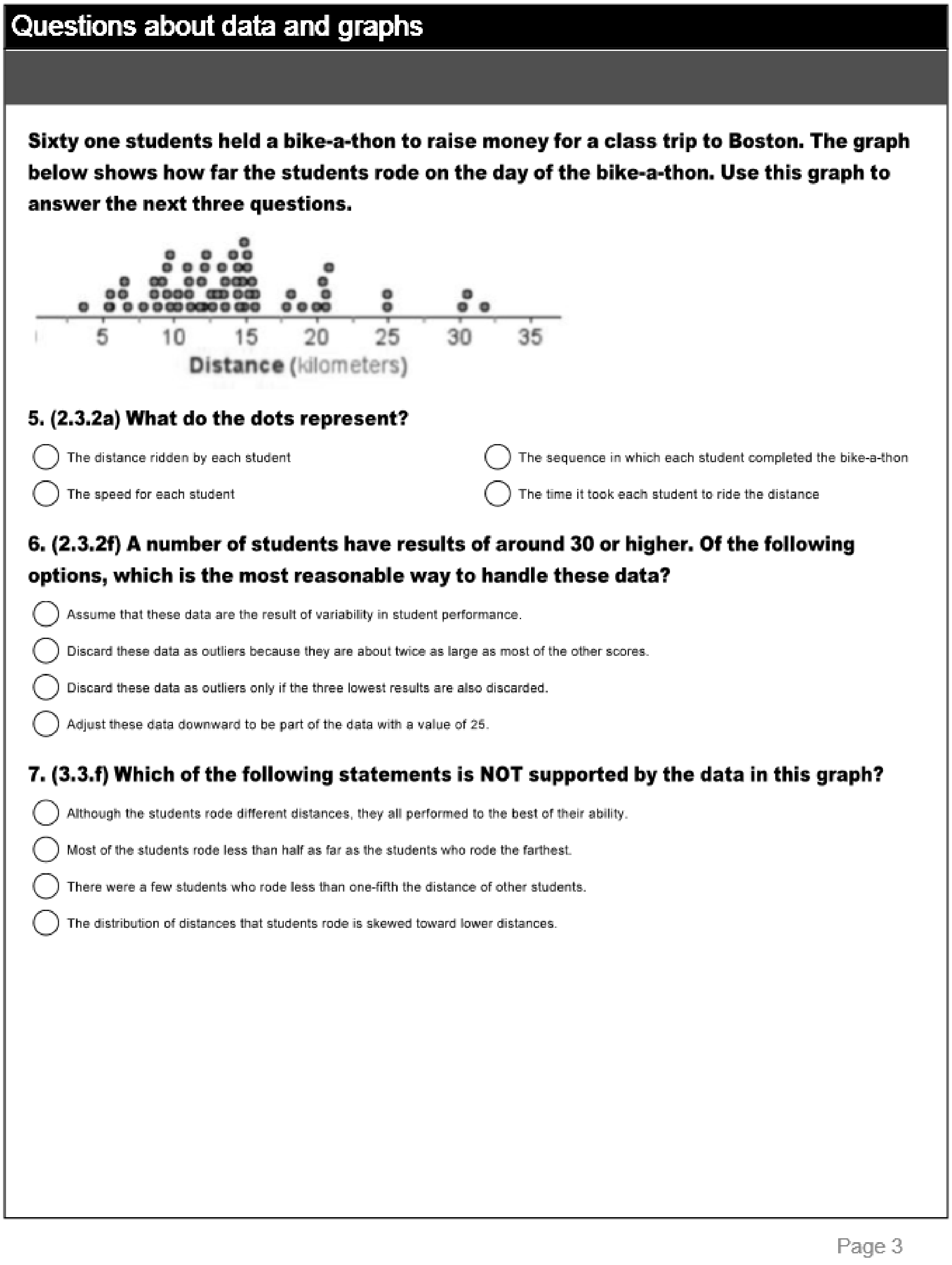
○ A question about comparing two or more groups

○ A question about the proportions of different subgroups within a whole group

○ A question about correlation between two variables

○ A question about the amount of variability in different groups

Page 2



Questions about data and graphs

Sixty one students held a bike-a-thon to raise money for a class trip to Boston. The graph below shows how far the students rode on the day of the bike-a-thon. Use this graph to answer the next three questions.

5. (2.3.2a) What do the dots represent?

○ The distance ridden by each student

○ The sequence in which each student completed the bike-a-thon

○ The speed for each student

○ The time it took each student to ride the distance

6. (2.3.2f) A number of students have results of around 30 or higher. Of the following options, which is the most reasonable way to handle these data?

○ Assume that these data are the result of variability in student performance.

○ Discard these data as outliers because they are about twice as large as most of the other scores.

○ Discard these data as outliers only if the three lowest results are also discarded.

○ Adjust these data downward to be part of the data with a value of 25.

7. (3.3.f) Which of the following statements is NOT supported by the data in this graph?

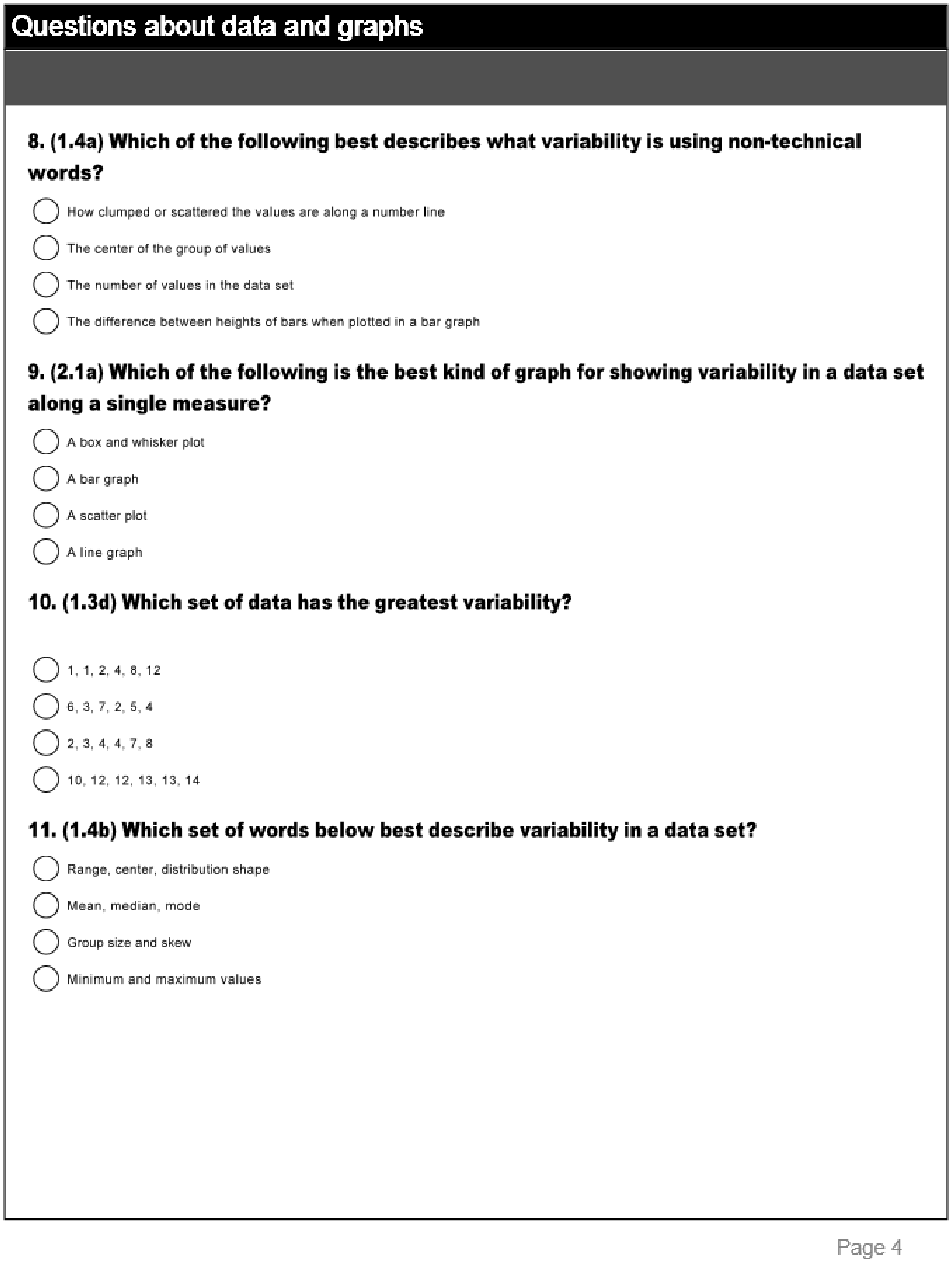
○ Although the students rode different distances, they all performed to the best of their ability.

○ Most of the students rode less than half as far as the students who rode the farthest.

○ There were a few students who rode less than one-fifth the distance of other students.

○ The distribution of distances that students rode is skewed toward lower distances.

Page 3



Questions about data and graphs

8. (1.4a) Which of the following best describes what variability is using non-technical words?

○ How clumped or scattered the values are along a number line

○ The center of the group of values

○ The number of values in the data set

○ The difference between heights of bars when plotted in a bar graph

9. (2.1a) Which of the following is the best kind of graph for showing variability in a data set along a single measure?

○ A box and whisker plot

○ A bar graph

○ A scatter plot

○ A line graph

10. (1.3d) Which set of data has the greatest variability?

○ 1, 1, 2, 4, 8, 12

○ 6, 3, 7, 2, 5, 4

○ 2, 3, 4, 4, 7, 8

○ 10, 12, 12, 13, 13, 14

11. (1.4b) Which set of words below best describe variability in a data set?

○ Range, center, distribution shape

○ Mean, median, mode

○ Group size and skew

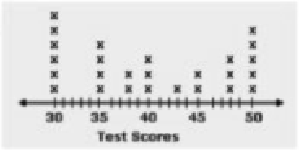
○ Minimum and maximum values

Page 4

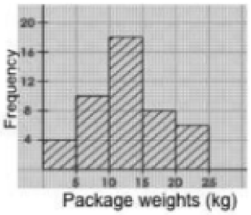
Questions about data and graphs

Use these four graphs to answer the following question.

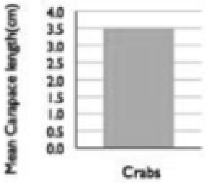
A



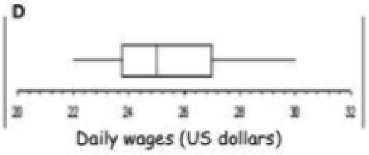
B



C



D



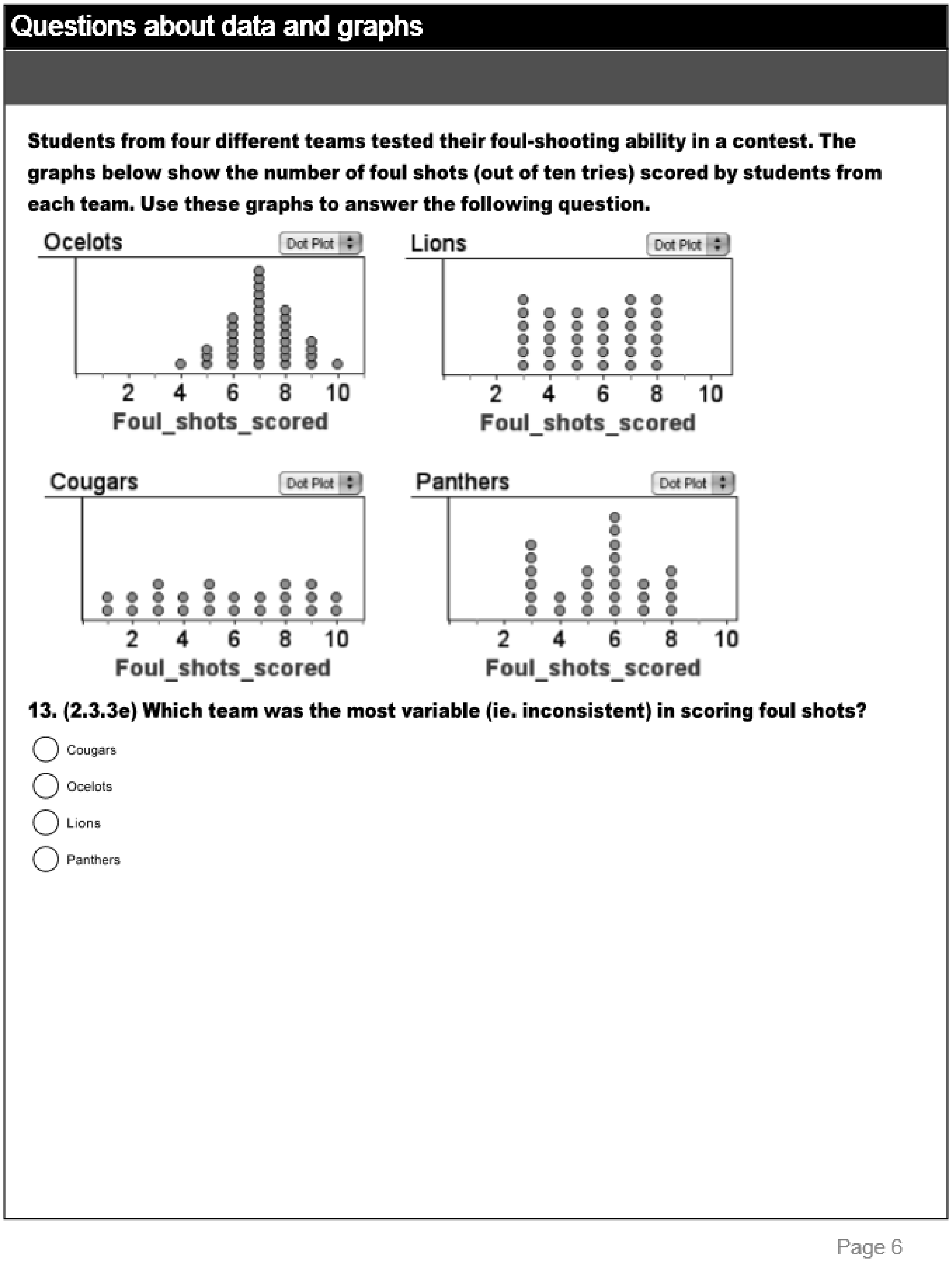
12. (2.1c) Which of the graphs above does NOT show the variability in a single group?

○ A

○ B

○ C

○ D



Questions about data and graphs

Students from four different teams tested their foul-shooting ability in a contest. The graphs below show the number of foul shots (out of ten tries) scored by students from each team. Use these graphs to answer the following question.

13. (2.3.3e) Which team was the most variable (ie. inconsistent) in scoring foul shots?

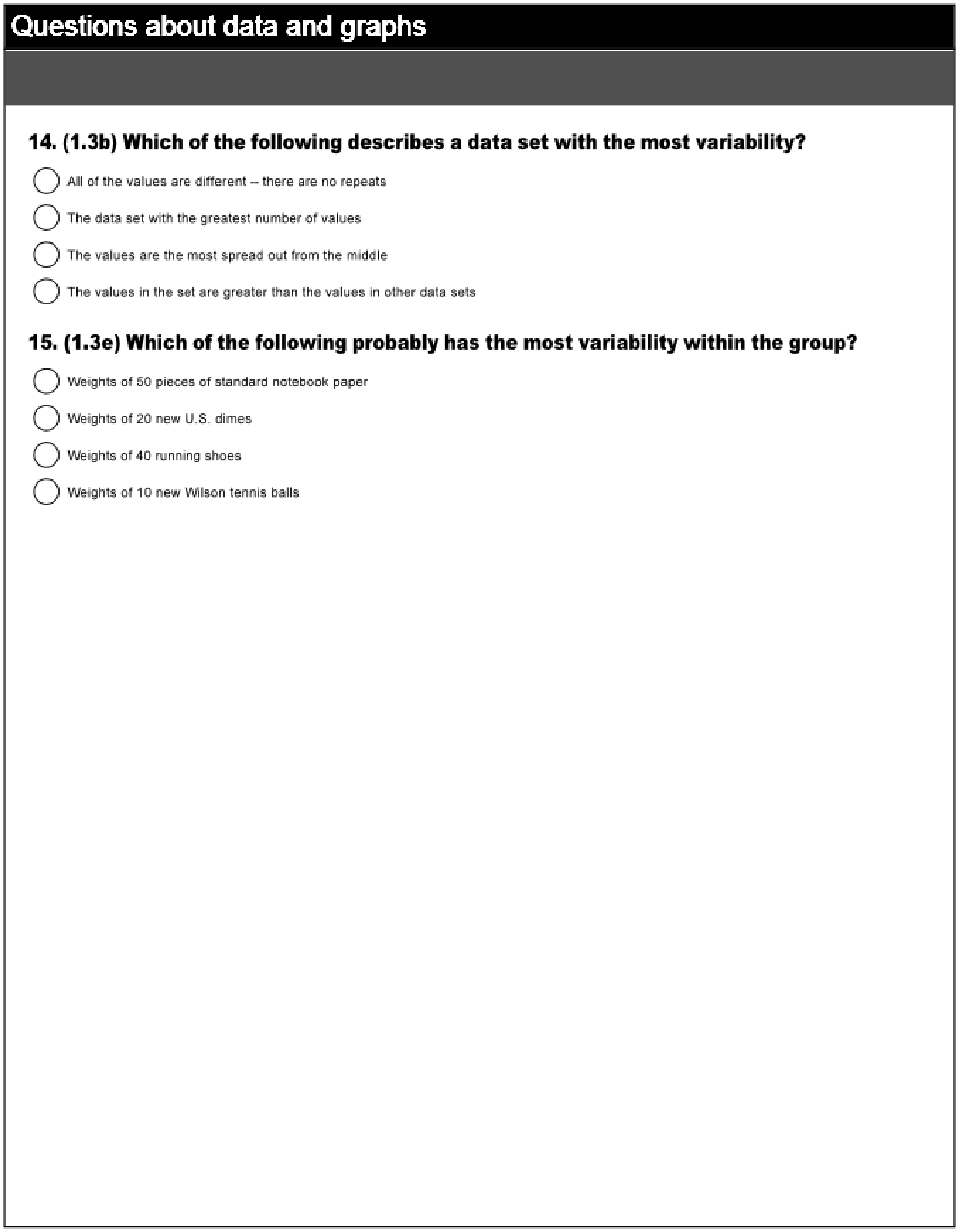
○ Cougars

○ Ocelots

○ Lions

○ Panthers

Page 6



Questions about data and graphs

14. (1.3b) Which of the following describes a data set with the most variability?

○ All of the values are different – there are no repeats

○ The data set with the greatest number of values

○ The values are the most spread out from the middle

○ The values in the set are greater than the values in other data sets

15. (1.3e) Which of the following probably has the most variability within the group?

○ Weights of 50 pieces of standard notebook paper

○ Weights of 20 new U.S. dimes

○ Weights of 40 running shoes

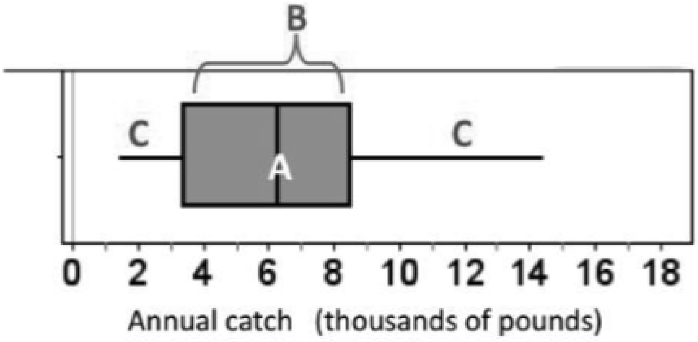
○ Weights of 10 new Wilson tennis balls

Page 7

Questions about data and graphs

The graph below shows the total pounds of elvers (baby eels) caught by Maine fishermen each year from 1995 to 2011 (in thousands of pounds). Use this graph to answer the next three questions.

Maine elvers caught per year from 1995 to 2010



16. (2.5.2 a) What does the vertical line, labeled "A", represent?

○ The range of the middle half of the data points
○ The annual catch for most fishermen
○ The mean of the data set
○ The median of the data set

17. (2.5.2c) What does the box, labeled "B", represent?

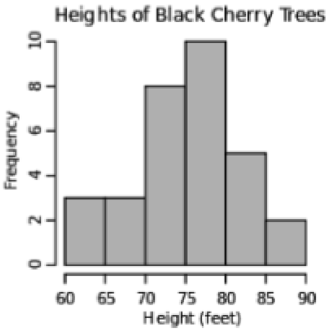
○ The median number of fish caught
○ The range of the middle half of the data points
○ The range of all but a few data points
○ The range of pounds of fish caught yearly

18. (2.5.2d) What do the horizontal lines, labeled "C", represent?

○ The range of all but a few of the data points
○ The range of pounds of fish caught yearly
○ The median number of fish caught each year
○ The range of the middle half of the data set

Questions about data and graphs

Below is a histogram of the heights of 31 black cherry trees. Use this graph to answer the next three questions.



19. (2.4.2a) Which height range occurs most frequently among all of the trees?

○ 60 to 65 feet ○ 70 to 75 feet ○ 75 to 80 feet ○ 85 to 90 feet

20. (2.4.2b) How many trees are in the tallest group of trees?

○ Two ○ Three ○ Eight ○ Ten

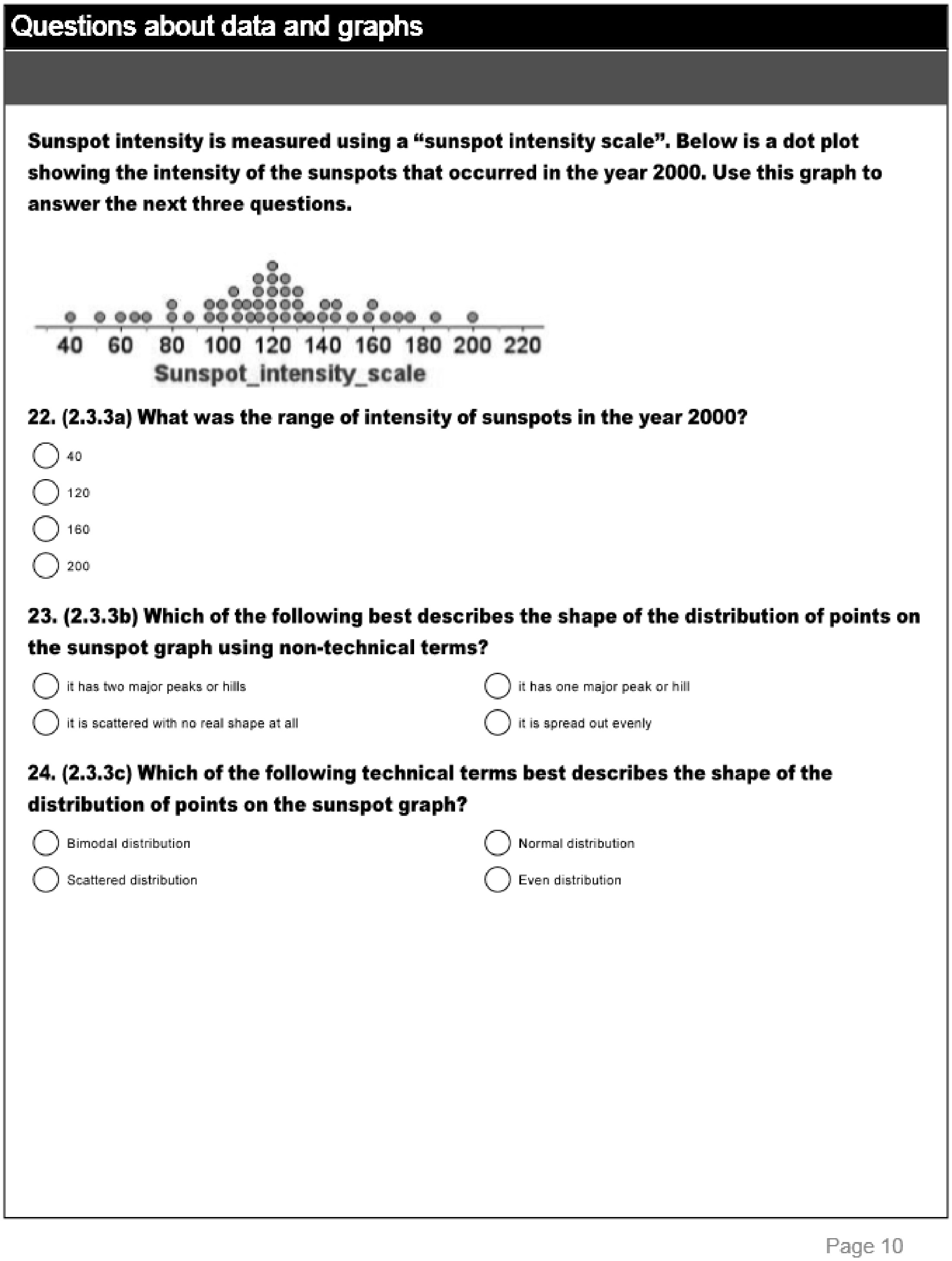
21. (2.4.2c) What does the height of the tallest column mean in this histogram?

○ The number of trees that are 10 feet tall

○ The number of trees that are the tallest in the group

○ The total number of trees measured

○ The trees in this height group occurred most often.



Questions about data and graphs

Sunspot intensity is measured using a "sunspot intensity scale". Below is a dot plot showing the intensity of the sunspots that occurred in the year 2000. Use this graph to answer the next three questions.

22. (2.3.3a) What was the range of intensity of sunspots in the year 2000?

○ 40

○ 120

○ 160

○ 200

23. (2.3.3b) Which of the following best describes the shape of the distribution of points on the sunspot graph using non-technical terms?

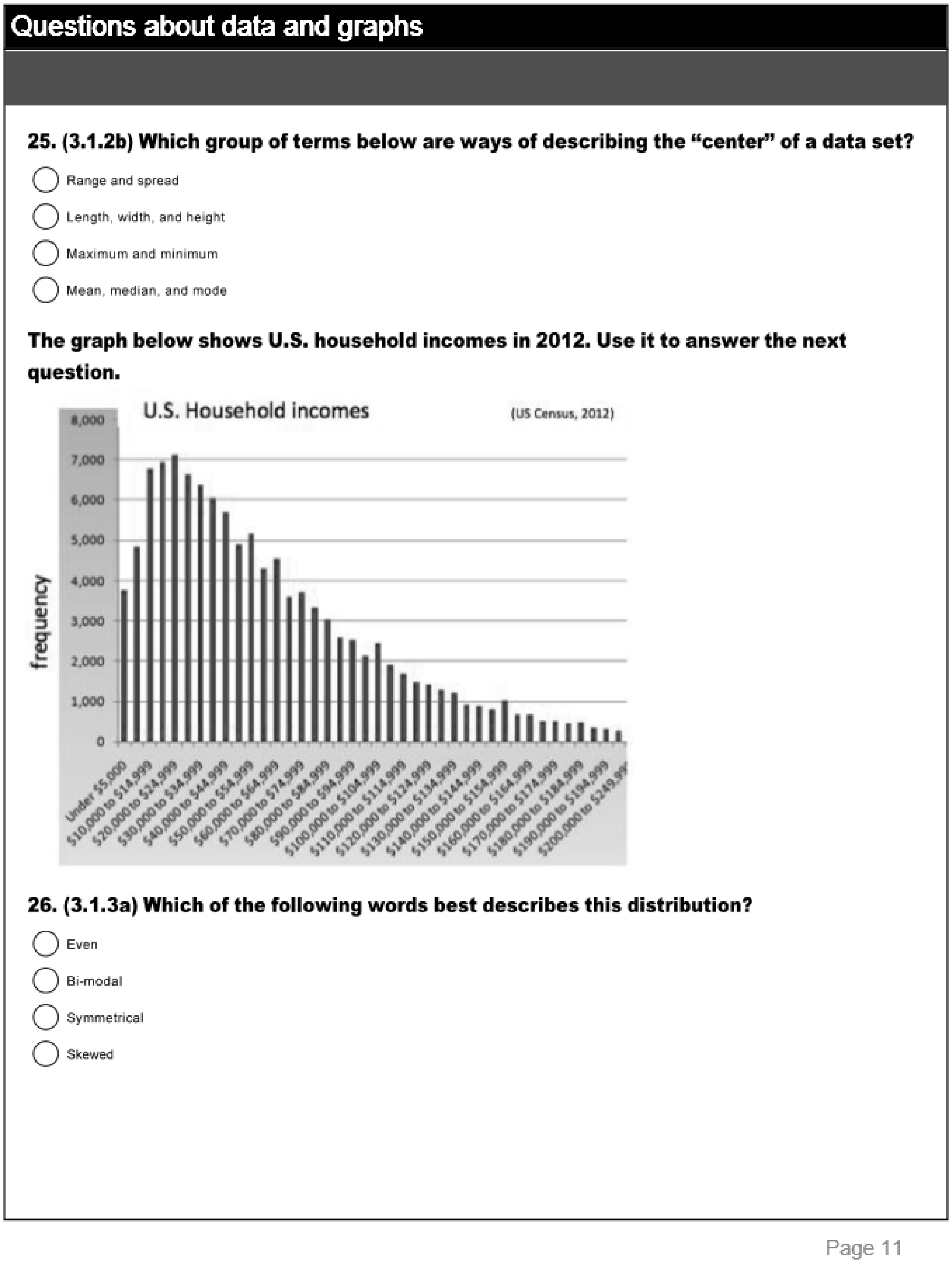
○ it has two major peaks or hills ○ it has one major peak or hill

○ it is scattered with no real shape at all ○ it is spread out evenly

24. (2.3.3c) Which of the following technical terms best describes the shape of the distribution of points on the sunspot graph?

○ Bimodal distribution ○ Normal distribution

○ Scattered distribution ○ Even distribution



Questions about data and graphs

25. (3.1.2b) Which group of terms below are ways of describing the "center" of a data set?

○ Range and spread

○ Length, width, and height

○ Maximum and minimum

○ Mean, median, and mode

The graph below shows U.S. household incomes in 2012. Use it to answer the next question.

26. (3.1.3a) Which of the following words best describes this distribution?

○ Even

○ Bi-modal

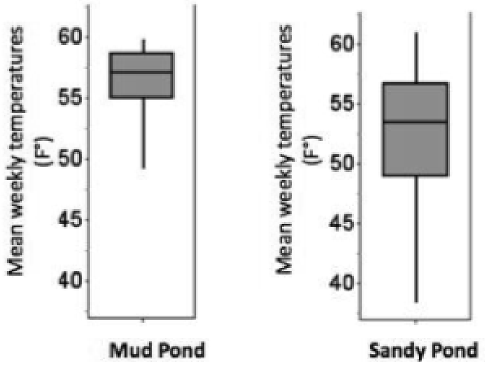
○ Symmetrical

○ Skewed

Page 11

Questions about data and graphs

Below are two graphs showing mean weekly water temperatures for two ponds from April to October. Use these graphs to answer the next question.



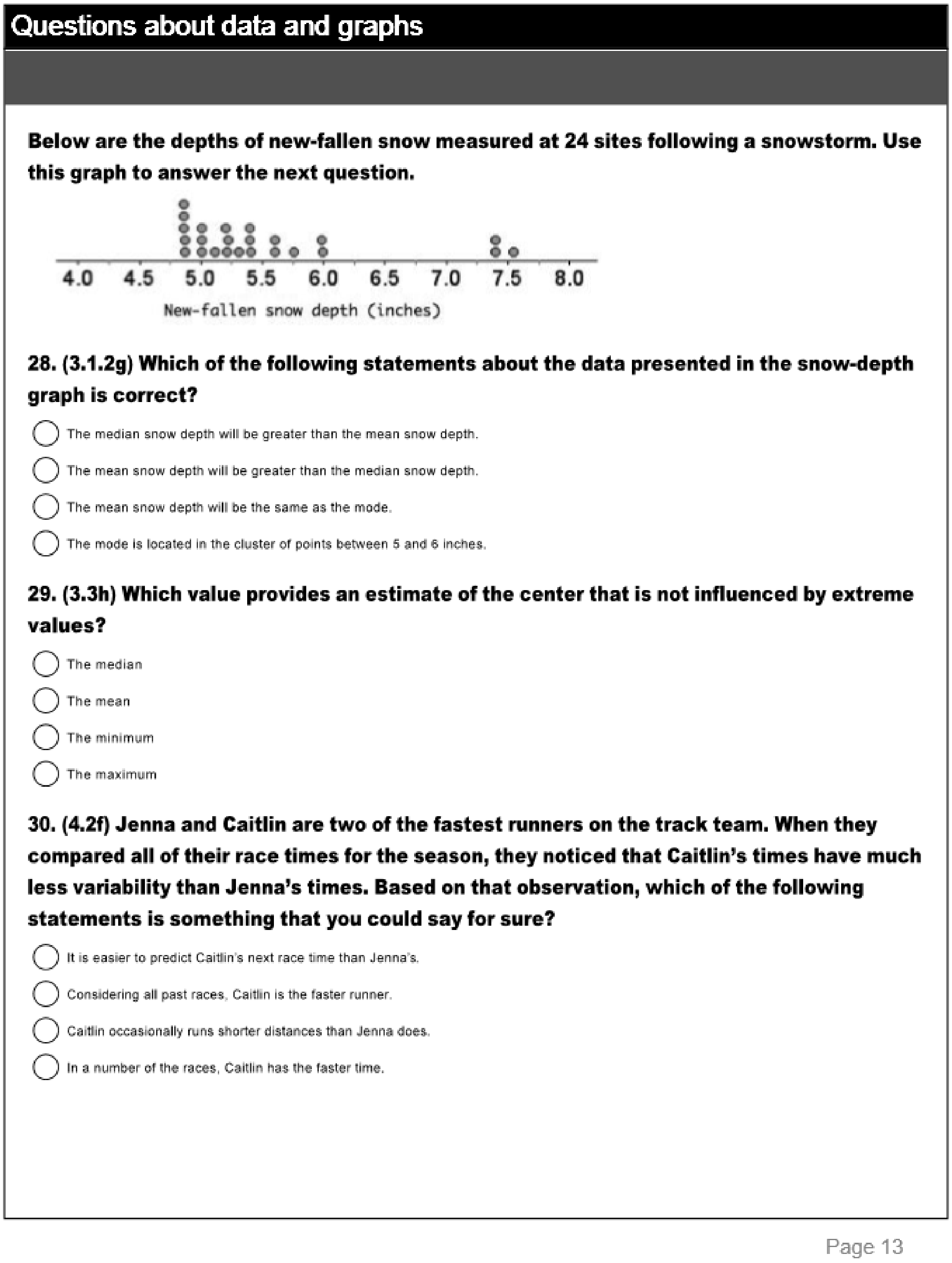
27. (4.2g) Which of the following claims is supported by the graphs?

○ Temperature is more variable in Mud Pond than in Sandy Pond.

○ Temperature is more variable in Sandy Pond than in Mud Pond.

○ Sandy Pond is always warmer than Mud Pond.

○ Mud Pond is always warmer than Sandy Pond.



Questions about data and graphs

Below are the depths of new-fallen snow measured at 24 sites following a snowstorm. Use this graph to answer the next question.

28. (3.1.2g) Which of the following statements about the data presented in the snow-depth graph is correct?

○ The median snow depth will be greater than the mean snow depth.

○ The mean snow depth will be greater than the median snow depth.

○ The mean snow depth will be the same as the mode.

○ The mode is located in the cluster of points between 5 and 6 inches.

29. (3.3h) Which value provides an estimate of the center that is not influenced by extreme values?

○ The median

○ The mean

○ The minimum

○ The maximum

30. (4.2f) Jenna and Caitlin are two of the fastest runners on the track team. When they compared all of their race times for the season, they noticed that Caitlin's times have much less variability than Jenna's times. Based on that observation, which of the following statements is something that you could say for sure?

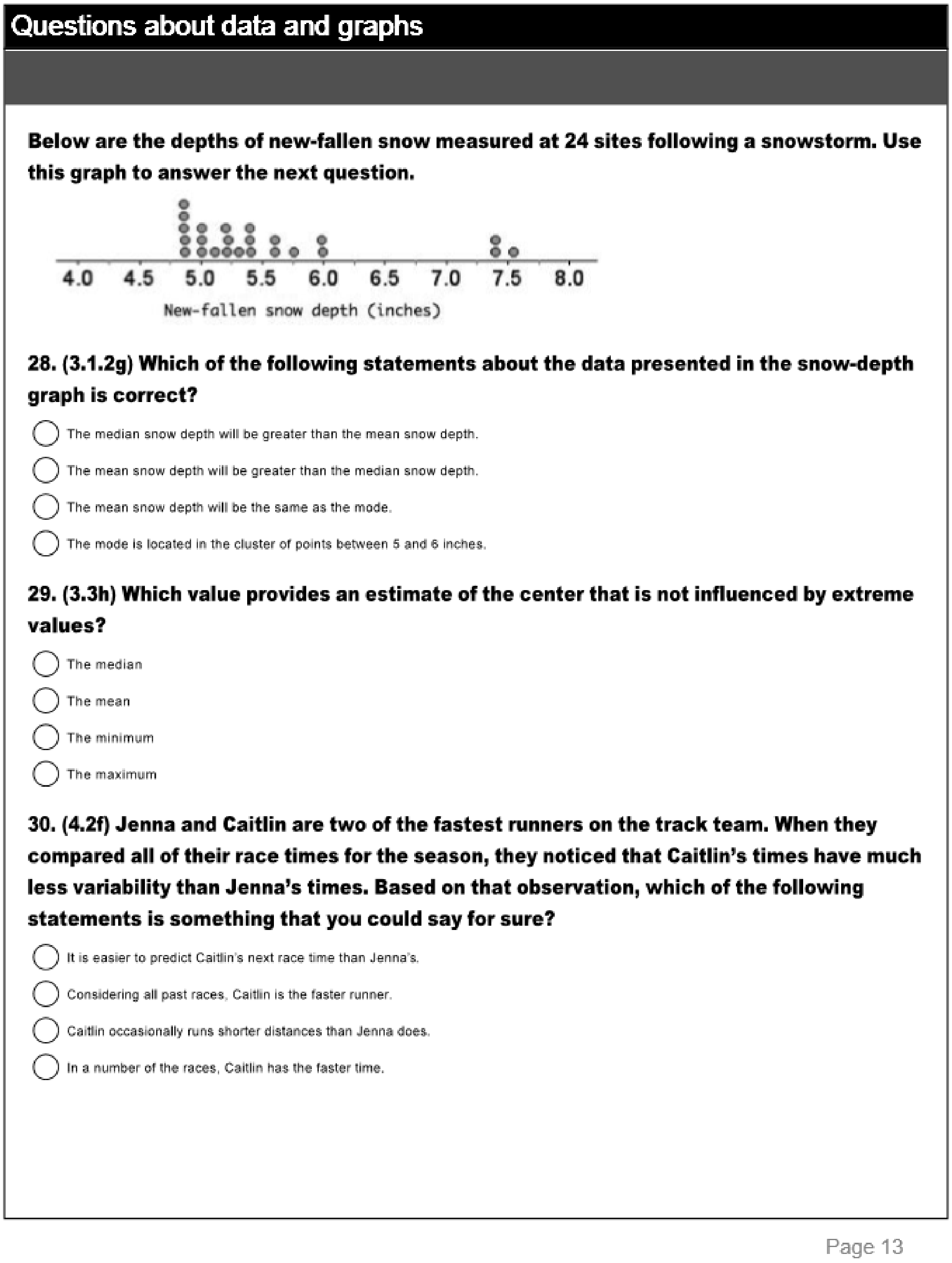
○ It is easier to predict Caitlin's next race time than Jenna's.

○ Considering all past races, Caitlin is the faster runner.

○ Caitlin occasionally runs shorter distances than Jenna does.

○ In a number of the races, Caitlin has the faster time.

Page 13



Questions about data and graphs

Below are the depths of new-fallen snow measured at 24 sites following a snowstorm. Use this graph to answer the next question.

28. (3.1.2g) Which of the following statements about the data presented in the snow-depth graph is correct?

○ The median snow depth will be greater than the mean snow depth.

○ The mean snow depth will be greater than the median snow depth.

○ The mean snow depth will be the same as the mode.

○ The mode is located in the cluster of points between 5 and 6 inches.

29. (3.3h) Which value provides an estimate of the center that is not influenced by extreme values?

○ The median

○ The mean

○ The minimum

○ The maximum

30. (4.2f) Jenna and Caitlin are two of the fastest runners on the track team. When they compared all of their race times for the season, they noticed that Caitlin's times have much less variability than Jenna's times. Based on that observation, which of the following statements is something that you could say for sure?

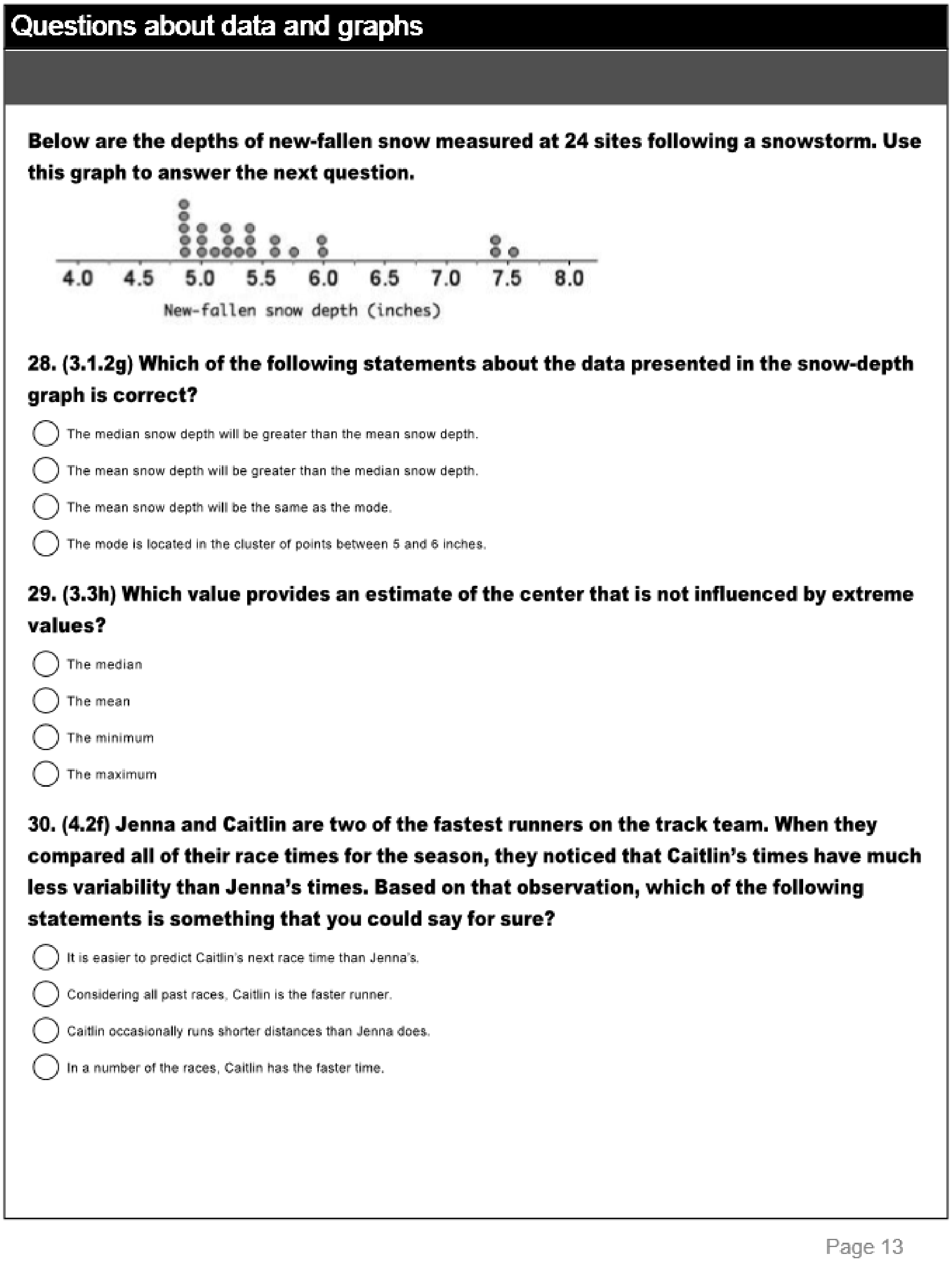
○ It is easier to predict Caitlin's next race time than Jenna's.

○ Considering all past races, Caitlin is the faster runner.

○ Caitlin occasionally runs shorter distances than Jenna does.

○ In a number of the races, Caitlin has the faster time.

Page 13



Questions about data and graphs

Below are the depths of new-fallen snow measured at 24 sites following a snowstorm. Use this graph to answer the next question.

28. (3.1.2g) Which of the following statements about the data presented in the snow-depth graph is correct?

○ The median snow depth will be greater than the mean snow depth.

○ The mean snow depth will be greater than the median snow depth.

○ The mean snow depth will be the same as the mode.

○ The mode is located in the cluster of points between 5 and 6 inches.

29. (3.3h) Which value provides an estimate of the center that is not influenced by extreme values?

○ The median

○ The mean

○ The minimum

○ The maximum

30. (4.2f) Jenna and Caitlin are two of the fastest runners on the track team. When they compared all of their race times for the season, they noticed that Caitlin's times have much less variability than Jenna's times. Based on that observation, which of the following statements is something that you could say for sure?

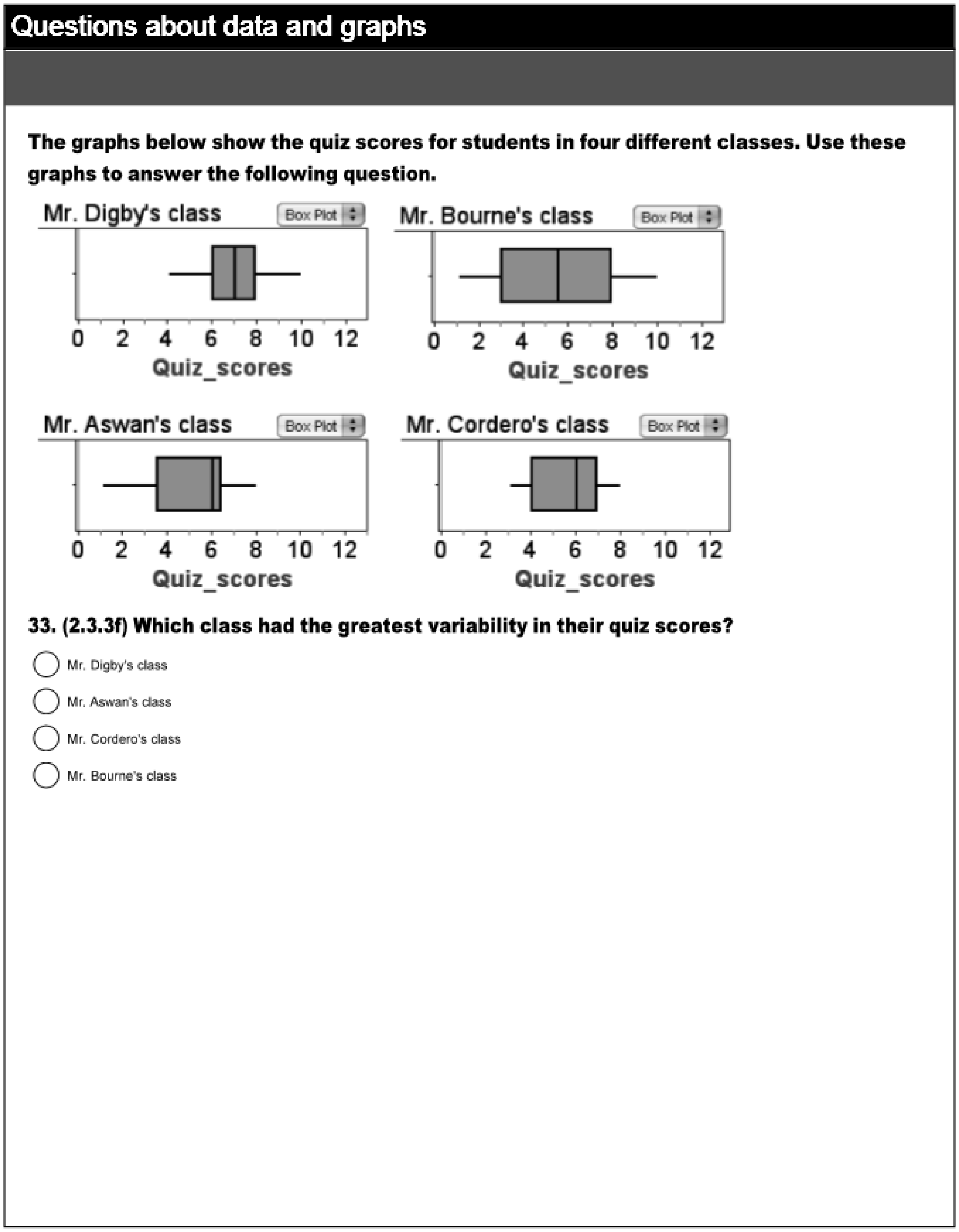
○ It is easier to predict Caitlin's next race time than Jenna's.

○ Considering all past races, Caitlin is the faster runner.

○ Caitlin occasionally runs shorter distances than Jenna does.

○ In a number of the races, Caitlin has the faster time.

Page 13



Questions about data and graphs

The graphs below show the quiz scores for students in four different classes. Use these graphs to answer the following question.

33. (2.3.3f) Which class had the greatest variability in their quiz scores?

○ Mr. Digby's class

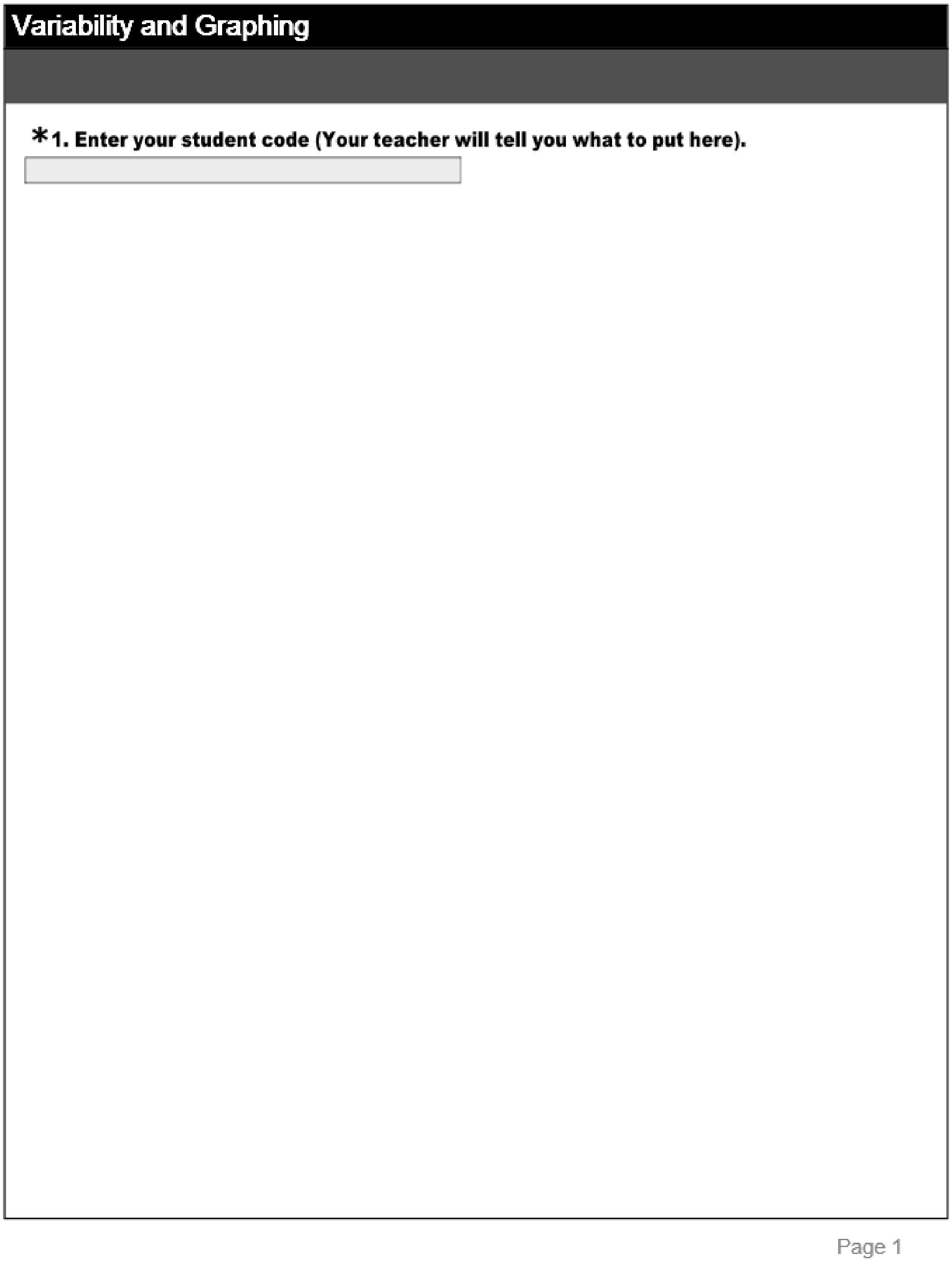
○ Mr. Aswan's class

○ Mr. Cordero's class

○ Mr. Bourne's class

Page 16

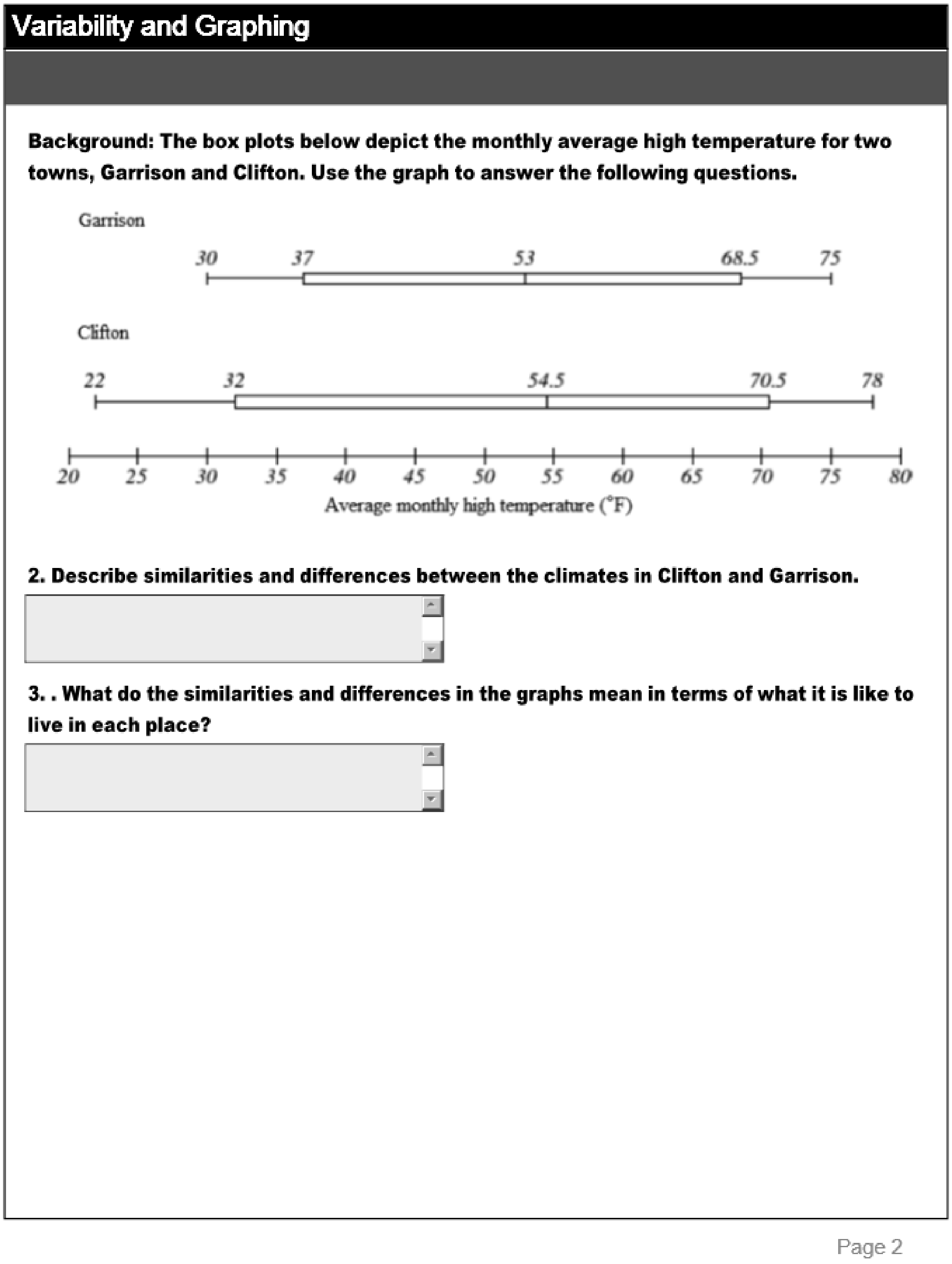
Appendix B: Open-response assessment



Variability and Graphing

*1. Enter your student code (Your teacher will tell you what to put here).

Page 1



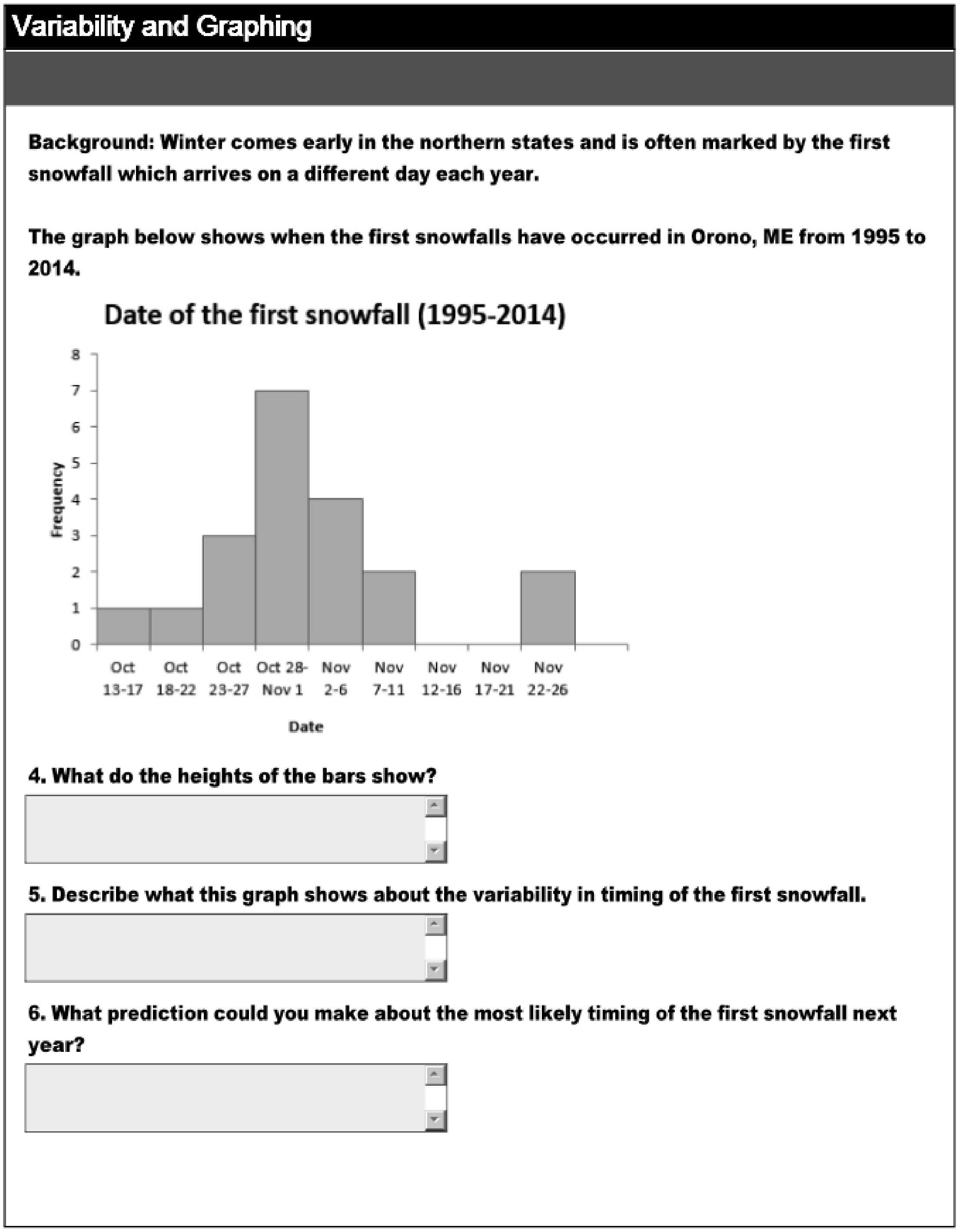
Variability and Graphing

Background: The box plots below depict the monthly average high temperature for two towns, Garrison and Clifton. Use the graph to answer the following questions.

2. Describe similarities and differences between the climates in Clifton and Garrison.

3. . What do the similarities and differences in the graphs mean in terms of what it is like to live in each place?

Page 2



Variability and Graphing

Background: Winter comes early in the northern states and is often marked by the first snowfall which arrives on a different day each year.

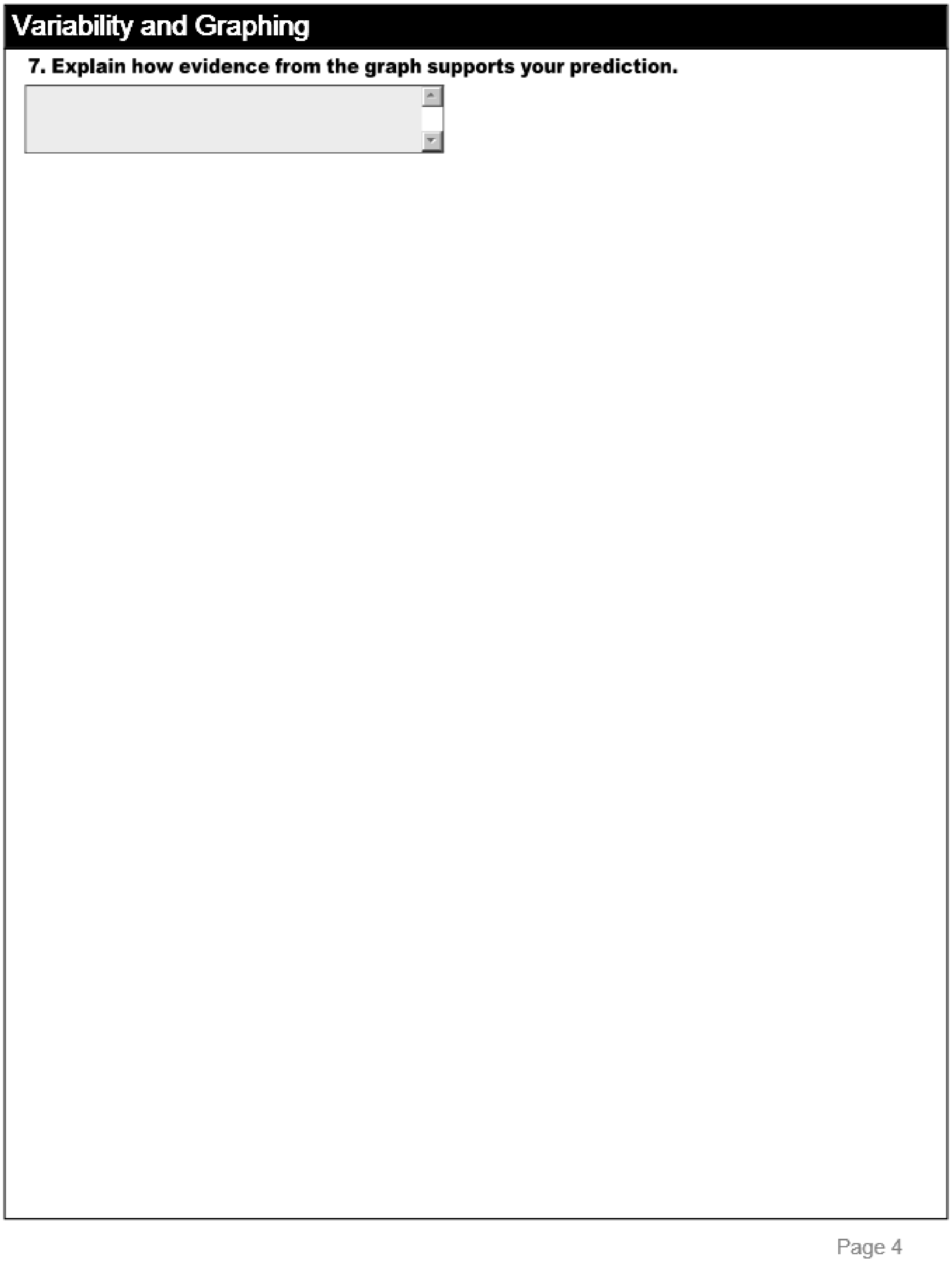
The graph below shows when the first snowfalls have occurred in Orono, ME from 1995 to 2014.

4. What do the heights of the bars show?

5. Describe what this graph shows about the variability in timing of the first snowfall.

6. What prediction could you make about the most likely timing of the first snowfall next year?

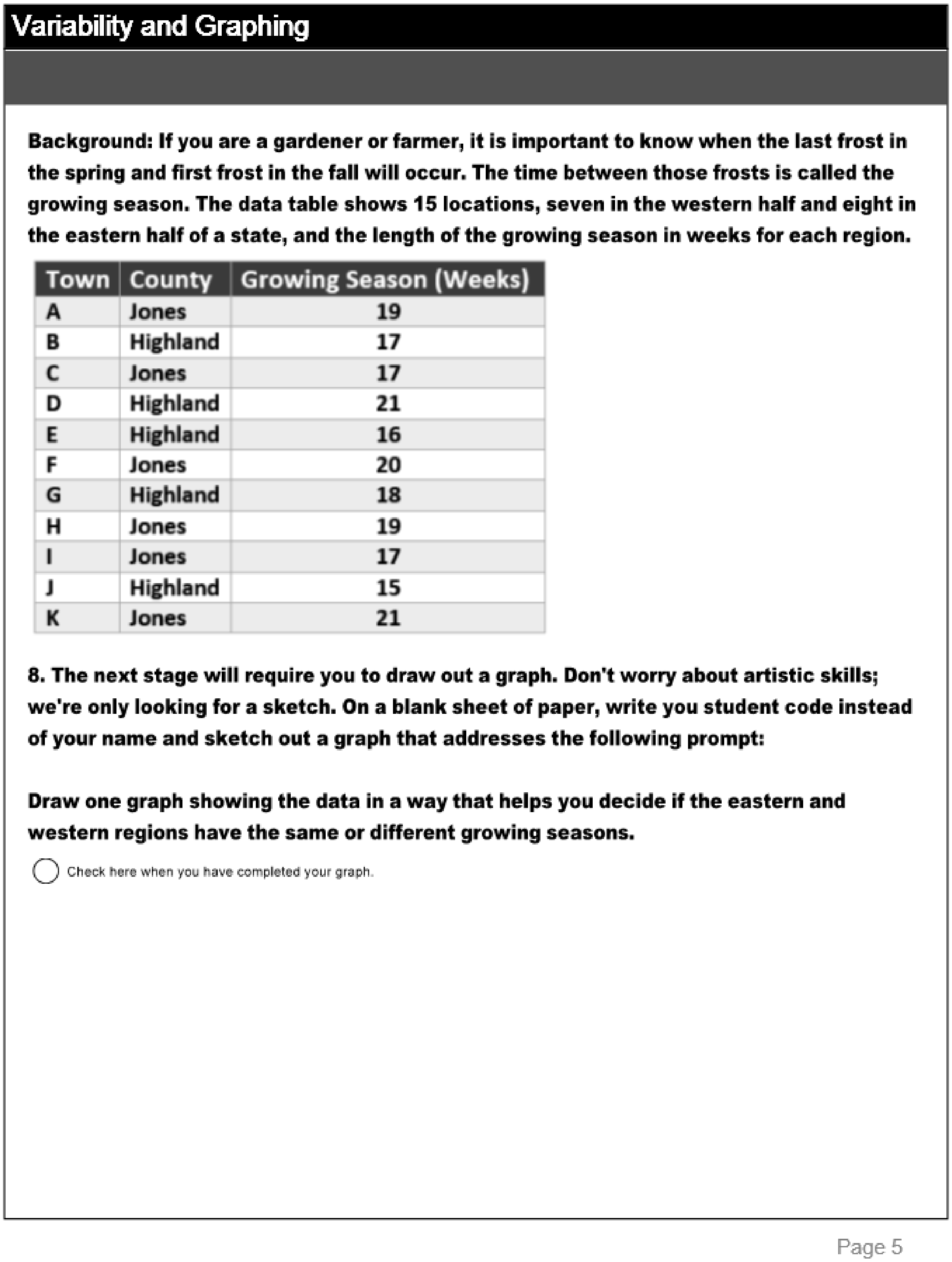
Page 3



Variability and Graphing

7. Explain how evidence from the graph supports your prediction.

Page 4



Variability and Graphing

Background: If you are a gardener or farmer, it is important to know when the last frost in the spring and first frost in the fall will occur. The time between those frosts is called the growing season. The data table shows 15 locations, seven in the western half and eight in the eastern half of a state, and the length of the growing season in weeks for each region.

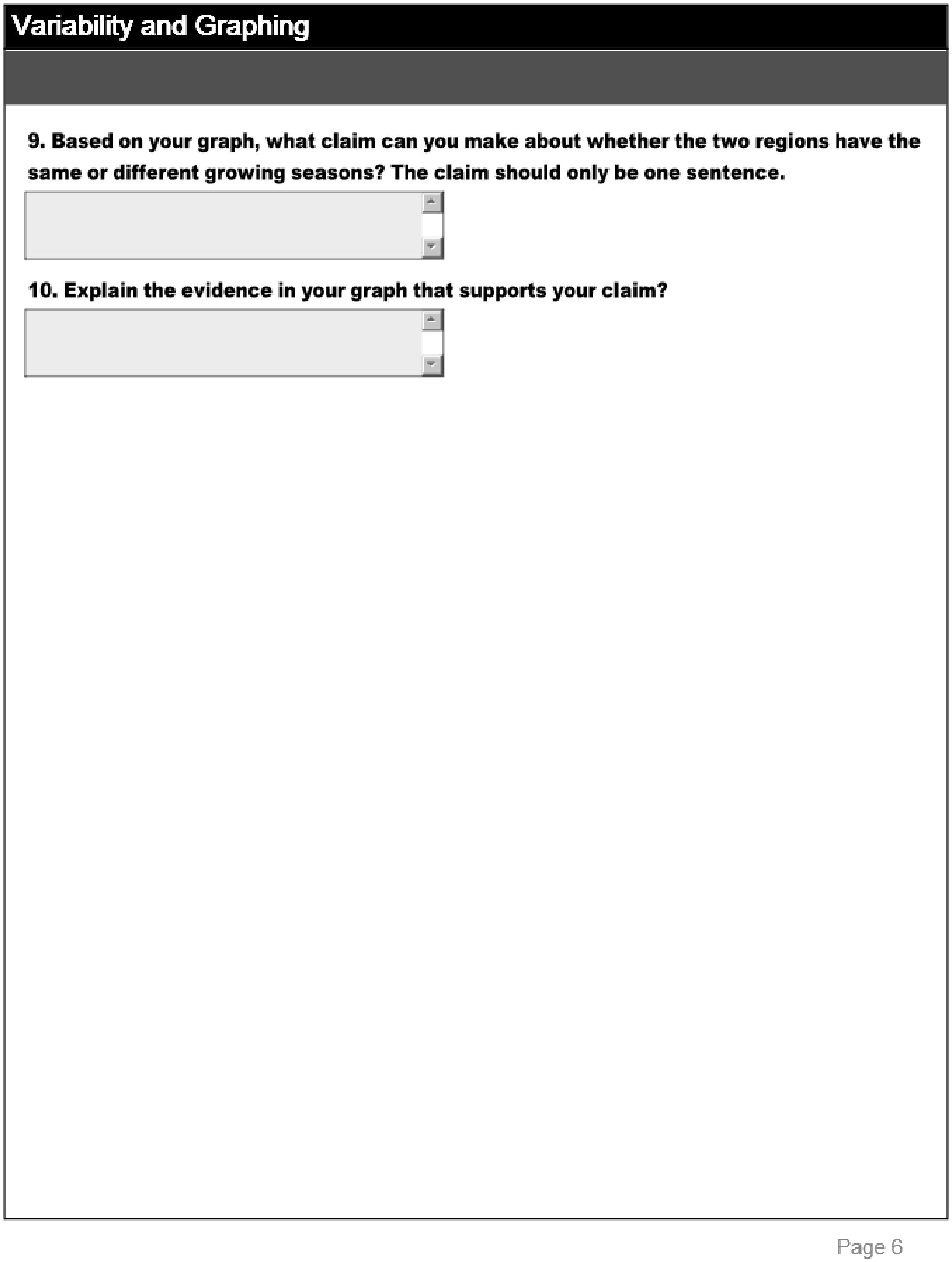
Town	County	Growing Season (Weeks)
A	Jones	19
B	Highland	17
C	Jones	17
D	Highland	21
E	Highland	16
F	Jones	20
G	Highland	18
H	Jones	19
I	Jones	17
J	Highland	15
K	Jones	21

8. The next stage will require you to draw out a graph. Don't worry about artistic skills; we're only looking for a sketch. On a blank sheet of paper, write you student code instead of your name and sketch out a graph that addresses the following prompt:

Draw one graph showing the data in a way that helps you decide if the eastern and western regions have the same or different growing seasons.

○ Check here when you have completed your graph.

Page 5



Variability and Graphing

9. Based on your graph, what claim can you make about whether the two regions have the same or different growing seasons? The claim should only be one sentence.

10. Explain the evidence in your graph that supports your claim?

Page 6

Appendix C: Open-response rubric

2. Describe similarities and differences between the climates in Clifton and Garrison.				
Graph description	**4** **Meets expectation** **(Mastery)**	**3** **Mostly meets expectation**	**2** **Partially meets expectation**	**1** **Does not meet expectation**
Answers question	___Response accurately and completely answers the question using quantitative data. ___The response includes a description of extremes and a measure of center.	___Response accurately and completely answers question, but it does not specifically reference the data. The response is not quantitative. ___A complete answer must include at least one similarity and one difference.	___ Response is only partially correct or partially answers the question, but it does relate to the data. ___Only one similarity or difference may be stated. ___Response may or may not be statistical or quantitative.	___No similarities or differences mentioned. ___No summary measures are identified.

4. Meets expectation

The median average monthly temperature in Clifton is only 1.5°F higher than the median temperature in Garrison. The hottest average temperature Clifton is three degrees warmer than in Garrison, and the coldest average monthly temperature in Clifton is 8°F colder than in Garrison. The spread of the interquartile range in Clifton is larger than in Garrison with Clifton's being both warmer and colder than Garrison's.

3. Mostly meets expectation

The median temperatures in Clifton is similar to Garrison and Clifton has a bigger range. (or "Clifton has a larger range than Garrison.")

The Median of the two is similar but the min and max are further to either extreme in clifton than garrison

2. Partially meets expectation

Both Clifton and Garrison have an interquartile range roughly between 30 and 70. Clifton's weather is more variable than Garrison's because it has a bigger range and interquartile range.

It gets hotter and colder in Clifton.

The average is the same and the outliers are different.

1. Does not meet expectation

They are both too cold.

They are the same temperature.

Lengths of the graphs

3. What do the similarities and differences in the graphs say about what it would be like to live in each place?				
Graph interpretation	**4** **Meets expectation** **(Mastery)**	**3** **Mostly meets expectation**	**2** **Partially meets expectation**	**1** **Does not meet expectation**
Answers question	___Response accurately and completely answers the question and uses quantitative data.	___Response accurately and completely answers question, but it does not specifically reference the data. The response is not quantitative. ___To completely answer the question the student must refer to the central tendency and the extremes.	___ Response is only partially correct or partially answers the question, but it does relate to the data. ___The student may only reference the center or the extremes but not both.	___ Response does not answer the question.

4. Meets Expectation

Residents of Clifton would experience hotter summers and colder winters than residents of Garrison, however the temperatures would on average be similar.

Residents of Clifton would experience hotter summers and colder winters than residents of Garrison, and the temperatures on average would be a little warmer.

3. Mostly meets expectation

Clifton has hot summers and cold winters.

2. Partially meets expectation

They are the same.

Garrison is warmer.

1. Does not meet expectation

I would rather live in Garrison.

Clifton is harder to live in

Clifton has more extreme weather.

4. What do the heights of the bars show?				
Graph mechanics	**4** **Meets expectation** **(Mastery)**	**3** **Mostly meets expectation**	**2** **Partially meets expectation**	**1** **Does not meet expectation**
Answers question	___Response describes the bars as the frequency of the first snowfall of the year for a particular range of dates.	___Response describes the bars as the frequency of the first snowfall of the year for a particular range of dates.	___Response describes the bars as representing a frequency, but of something other than first snowfalls.	___Response does not describe the bars as representing a frequency**.**

4. Meets Expectation

The heights of the bars represent how many times the first snowfall of the year has occurred on the range of dates for the bar between 1995 and 2014.

3. Mostly meets expectation

The heights of the bars show how often it snowed for the first time that year.

2. Partially meets expectation

The heights of the bars represent how often it snows each day.

The heights of the bars represent how often it snows.

1. Does not meet expectation

The heights of the bars represent how much snow there is.

The heights of the bars represent how much

The heights of the bars represent how much snow fell.

5. Describe what this graph shows about the variability in timing of the first snowfall.

Graph description	4 Meets expectation (Mastery)	3 Mostly meets expectation	2 Partially meets expectation	1 Does not meet expectation
Answers question	___Response accurately and completely answers the question and uses quantitative data.	___Response accurately and completely answers question, but it does not specifically reference the data. The response is not quantitative (specific dates or date range). ___The response references extremes and a measure of center	___ Response is only partially correct or partially answers the question, but it does relate to the data. ___The student may only reference the center or the extremes but not both.	___ Response does not answer the question.

4. Meets Expectation

This distribution of the first snowfalls is normal with two extreme values occurring on November 22-26th. The range is from October 13th to November 26th and the median date range is October 28th through November 1st. The first snowfall fell most frequently in the last week of October, but it fell as early as October 13th and as late as November 26th some years.

3. Mostly meets expectation

The date of the first snow can vary by over 1 month. Very early and very late first snows are rarer.
The average high temperature is similar. The low temperatures are different.

2. Partially meets expectation

The mode for this graph is October 28th through November 1st.

1. Does not meet expectation

The data goes up then it goes down again.
Most of the dates are October
The dates are all in groups of five.
A lot of the snow happened in late October.

6. What prediction could you make about the most likely timing of the first snowfall next year?

Graph interpretation	4 Meets expectation (Mastery)	3 Mostly meets expectation	2 Partially meets expectation	1 Does not meet expectation
Answers question	___Student makes a clear and accurate claim based on the data.	___Student makes a claim that is mostly clear and accurate but is missing something or is difficult to understand.	___Student makes a claim, but it is unclear or unrelated to the data.	___A claim is not made.

4. Meets Expectation

The first snowfall will likely occur between October 28 and November 6.

The time of the snowfall next year may be around October 28 to November 26.

3. Mostly meets expectation

The first snowfall will be around the end of October.

The first snowfall will happen between October 13 and November 26.

2. Partially meets expectation

It will rise even more because of the patterns in the previous year

The snow fall will probably be low, because there had already been a large snow fall recently.

Since it first snowed on October and now we're in November it will most likely snow in October again because it was so long ago. With the way the pattern is, it would eventually snow in May or August which is clearly a summer month. So October would be reasonable.

The timing was early.

1. Does not meet expectation

I think the frequency will decrease.

October 28- November 1 is the heaviest. Maybe be the same or a little more.

It will be the same

7. Explain how evidence from the graph supports your prediction.

Graph interpretation	4 Meets expectation (Mastery)	3 Mostly meets expectation	2 Partially meets expectation	1 Does not meet expectation
Answers question	___Response accurately and completely answers the question using quantitative data.	___Response accurately and completely answers question, but it does not specifically reference the data. The response is not quantitative.	___ Response is only partially correct or partially answers the question, but it does relate to the data.	___ Response does not answer the question.

4. Meets Expectation

The first snowfall happened between those dates 11 of the last 20 years, so it is most likely to happen then next year.

That is when most of the first snowfalls happened in the past.

These dates have the highest frequency of being the date of the first snow fall.

The graph shows that the highest frequency of first snowfalls occurs between those few days.

3. Mostly meets expectation

That is when the largest amount of data is.

It was done in a bell shaped curve where the peak was late October in the graph.

2. Partially meets expectation

It usually happens then.

The data states that the most frequent snowfalls happen around the last days of October.

1. Does not meet expectation

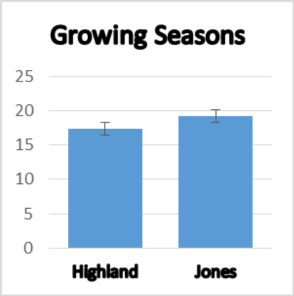
It will rise even more because of the patterns in the previous year

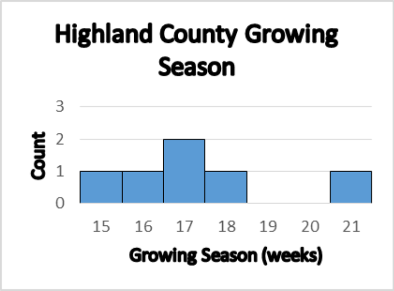
It happened then.

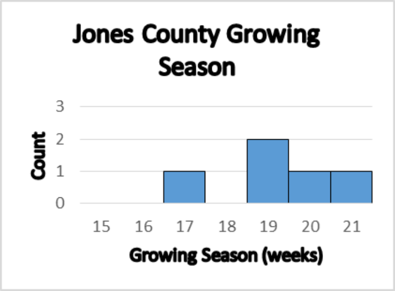
8. Graph construction task

Graph construction	**4** **Meets expectation** **(Mastery)**	**3** **Mostly meets expectation**	**2** **Partially meets expectation**	**1** **Does not meet expectation**
Graph type	Graph type is ideal -Boxplot -Histogram -Dot plot -Bar graph with error bars	Data are resolved into groups and summarized in a way that allows comparison -Bar graph of averages -Stem and leaf plot	Data are displayed in a way that one *could* figure out the answer to the question with effort -bar graph resolved into groups	The type of graph does not represent data in a way that helps answer the question -Pie chart -line graph -Bar graph not resolved into groups
Graph mechanics	All elements are present & reasonably clear _Graph is overall neat & legible _Axes are drawn & labeled _Axes scales are clear & correct _Data are plotted accurately _Legend is present, if needed ** Title not required **Do not judge based on graph choice	All elements are present but may be poorly executed (inconsistent scale on the axes, non-descriptive or inaccurate labels). The graph is not "incorrect".	One of the elements under "Meets" may be missing or incorrect.	More than one of the elements under "Meets" are missing.

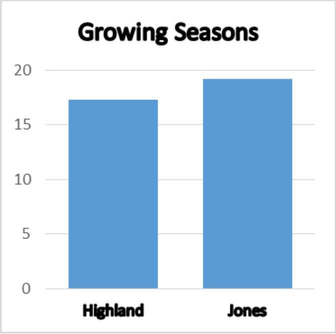
4. Meets Expectation



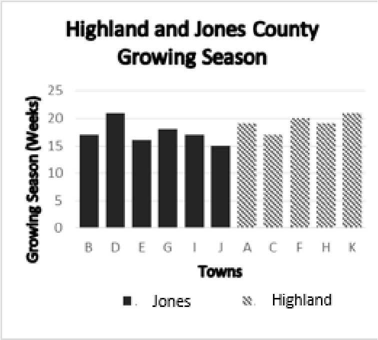




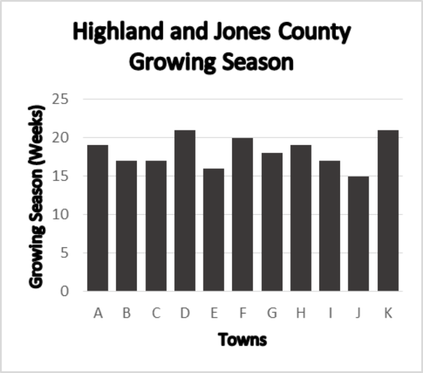
3. Mostly meets expectation

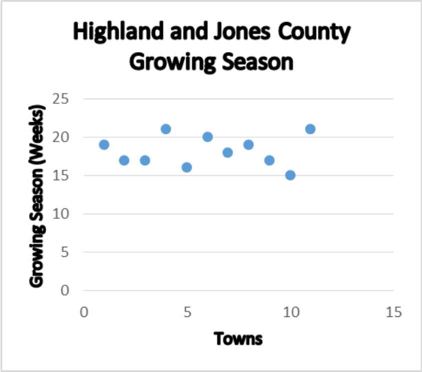


2. Partially meets expectation



1. Does not meet expectation





9. Based on your graph, what claim can you make about whether the two regions have the same or different growing seasons? The claim should only be one sentence.				
Graph interpretation **(Claim about the question and reasoning)**	**4** **Meets expectation** **(Mastery)**	**3** **Mostly meets expectation**	**2** **Partially meets expectation**	**1** **Does not meet expectation**
Answers question	___ A specific claim is made about how the two regions are similar or different.	___A claim is made about similarities or differences in the two regions but it does not describe HOW they are similar or different.	___A claim is made, but it is not supported by the graph.	___A claim is not made.

*Success on any of these items is not contingent on graph choice.

4. Meets Expectation
Both regions have roughly the same growing season
East Has the better growing season.
The east has a slightly longer growing season
I believe that the two regions growing seasons are pretty similar.

3. Mostly meets expectation
They have completely different growing seasons, the west has shorter seasons than the east.
Basically the same

2. Partially meets expectation
Variability is the same
Different because they all took different weeks

1. Does not meet expectation
The temperature is a bit different
They vary a lot. Ups and downs.
The two regions are only a few weeks apart

10. Explain the evidence in your graph that supports your claim?

Graph interpretation	4 Meets expectation (Mastery)	3 Mostly meets expectation	2 Partially meets expectation	1 Does not meet expectation
Answers question	___The reasoning used to arrive at the claim is logical and explained clearly.	___The reasoning used to arrive at the claim is consistent with the graph but it is unclear or too general.	**___The reasoning is incomplete** or is inconsistent with the graph.	___Reasoning is incoherent or unrelated to the data.
Statistical Concepts	___ A measure of center (mean, median, mode) as well as some additional evidence like quartiles, range, or skew is used in the reasoning.	___Only evidence that considers variability like quartiles, range, or skew is used in the reasoning. No measure of center (mean, median, mode) is used.	___Only evidence that does not consider variability like mean, median, mode is used in the reasoning. Variability is not considered (quartiles, range, or skew).	___No evidence is offered.

4. Meets Expectation

The bars are all about the same height and the means are only X weeks apart.

The means in each region is outside the interquartile range of the other region so the growing seasons are different.

3. Mostly meets expectation

This evidence supports my claim because most of the bars are around the same height. There's not much of a range.

They're all around the same numbers

2. Partially meets expectation

Supports it by the number of growing weeks.

Average numbers support my statement.

Because the wests growing season is later

They are all about the same

There is one who has only 15 weeks as to one who has 21 weeks.

They look almost the exact same

1. Does not meet expectation

The east numbers are close together

The eastern towns growing season increases then decreases then increases again. The western sides decreases then increases.

Appendix D: Coded responses to open-response assessment Question 5

Student Code	Open-response Question 5: Describe what this graph shows about the variability in timing of the first snowfall.	Code
T3_S_1	The first snowfall happens around ranges usually around the middle of October to the middle of November but has been as late as the end of November in some years.	3
T3_S_2	The graph shows that it is most variable from October 23 to November 11.	1
T3_S_3	This graph shows that the variability of the first snowfall ranges from the middle of October to end of November.	2
T3_S_4	The graph shows that it is more likely to snow for the first time in late October to early November. The variability of the first snowfall is not great as it will generally stick to the pattern of late October into early November.	2
T3_S_5	Based on this graph, the first snowfall could possibly occur anytime from early-mid October to even late November. However, most frequently is occurs between October 28 and November 1.	3
T3_S_7	The graph shows that there is a large variability of the first snowfall occurring between October 13 and November 11. It also shows that in Orono, Maine, there have been no snowfall from November 12 to November 21.	2
T3_S_8	The graph shows that the variability in timing of the first snowfall is not incredibly variable.	1
T3_S_9	There was only one instance of the snow falling in 13-17	1
T3_S_10	it ranges from October 13th to November 26th and most of the storms happening on October 28th to	3
T3_S_12	there is some variability but within only a few weeks	1
T3_S_13	They vary from October to November but are usually in mid- to late october.	3
T3_S_14	that its more likely in beginning of november	1
T3_S_15	It happened once in between Oct 13 and Oct 17	1

Code	Code Description
1	No variability description
2	Variability description is incomplete (Addresses spread or central tendency, not both)
3	variability description complete (Addresses spread and central tendency)

Appendix E: Relevant Next Generation Science Standards and Common Core State Standards

ELEMENTARY STANDARDS (K-5)

Science

K-ESS2-1

Use and share observations of local weather conditions to describe patterns over time.

K-2-ETS1-1

Ask questions, make observations, and gather information about a situation people want to change to define a simple problem that can be solved through the development of a new or improved object or tool.

3-ESS2-1

Represent data in tables and graphical displays to describe typical weather conditions expected during a particular season.

3-PS2-3

Ask questions to determine cause and effect relationships of electric or magnetic interactions between two objects not in contact with each other.

5-ESS1-2

Represent data in graphical displays to reveal patterns of daily changes in length and direction of shadows, day and night, and the seasonal appearance of some stars in the night sky

Math

CCSS.MATH.CONTENT.K.MD.B.3

Classify objects into given categories; count the numbers of objects in each category and sort the categories by count.

CCSS.MATH.CONTENT.1.MD.C.4

Organize, represent, and interpret data with up to three categories; ask and answer questions about the total number of data points, how many in each category, and how many more or less are in one category than in another.

CCSS.MATH.CONTENT.3.MD.B.3

Draw a scaled picture graph and a scaled bar graph to represent a data set with several categories. Solve one- and two-step "how many more" and "how many less" problems using information presented in scaled bar graphs. *For example, draw a bar graph in which each square in the bar graph might represent 5 pets.*

CCSS.MATH.CONTENT.4.MD.B.4

Make a line plot to display a data set of measurements in fractions of a unit (1/2, 1/4, 1/8). Solve problems involving addition and subtraction of fractions by using information presented in line plots. *For example, from a line plot find and interpret the difference in length between the longest and shortest specimens in an insect collection.*

CCSS.MATH.CONTENT.5.MD.B.2

Make a line plot to display a data set of measurements in fractions of a unit (1/2, 1/4, 1/8). Use operations on fractions for this grade to solve problems involving information presented in line plots. *For example, given different measurements of liquid in identical beakers, find the amount of liquid each beaker would contain if the total amount in all the beakers were redistributed equally.*

MIDDLE SCHOOL STANDARDS (6-8)

Science

MS-PS3-1

Construct and interpret graphical displays of data to describe the relationships of kinetic energy to the mass of an object and to the speed of an object.

MS-LS2-1

Analyze and interpret data to provide evidence for the effects of resource availability on organisms and populations of organisms in an ecosystem.

MS-ESS3-2

Analyze and interpret data on natural hazards to forecast future catastrophic events and inform the development of technologies to mitigate their effects.

MS-ETS1-3

Analyze data from tests to determine similarities and differences among several design solutions to identify the best characteristics of each that can be combined into a new solution to better meet the criteria for success

Math

CCSS.MATH.CONTENT.6.SP.A.1

Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers. *For example, "How old am I?" is not a statistical question, but "How old are the students in my school?" is a statistical question because one anticipates variability in students' ages*.

C**CSS.MATH.CONTENT.6.SP.A.2**

Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.

CCSS.MATH.CONTENT.6.SP.B.4

Display numerical data in plots on a number line, including dot plots, histograms, and box plots.

CCSS.MATH.CONTENT.7.SP.A.1

Understand that statistics can be used to gain information about a population by examining a sample of the population; generalizations about a population from a sample are valid only if the sample is representative of that population. Understand that random sampling tends to produce representative samples and support valid inferences.

CCSS.MATH.CONTENT.7.SP.A.2

Use data from a random sample to draw inferences about a population with an unknown characteristic of interest. Generate multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions. *For example, estimate the mean word length in a book by randomly sampling words from the book; predict the winner of a school election based on randomly sampled survey data. Gauge how far off the estimate or prediction might be*.

Draw informal comparative inferences about two populations.

CCSS.MATH.CONTENT.7.SP.B.3

Informally assess the degree of visual overlap of two numerical data distributions with similar variabilities, measuring the difference between the centers by expressing it as a multiple of a measure of variability. *For example, the mean height of players on the basketball team is 10 cm greater than the mean height of players on the soccer team, about twice the variability (mean absolute deviation) on either team; on a dot plot, the separation between the two distributions of heights is noticeable*.

CCSS.MATH.CONTENT.7.SP.B.4

Use measures of center and measures of variability for numerical data from random samples to draw informal comparative inferences about two populations. *For example, decide whether the words in a chapter of a seventh-grade science book are generally longer than the words in a chapter of a fourth-grade science book*.

Investigate chance processes and develop, use, and evaluate probability models.

CCSS.MATH.CONTENT.7.SP.C.5

Understand that the probability of a chance event is a number between 0 and 1 that expresses the likelihood of the event occurring. Larger numbers indicate greater likelihood. A probability near 0 indicates an unlikely event, a probability around 1/2 indicates an event that is neither unlikely nor likely, and a probability near 1 indicates a likely event.

CCSS.MATH.CONTENT.7.SP.C.6

Approximate the probability of a chance event by collecting data on the chance process that produces it and observing its long-run relative frequency, and predict the approximate relative frequency given the probability. *For example, when rolling a number cube 600 times, predict that a 3 or 6 would be rolled roughly 200 times, but probably not exactly 200 times*.

CCSS.MATH.CONTENT.7.SP.C.7

Develop a probability model and use it to find probabilities of events. Compare probabilities from a model to observed frequencies; if the agreement is not good, explain possible sources of the discrepancy.

CCSS.MATH.CONTENT.8.SP.A.1

Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.

English Language Arts

CCSS.ELA-LITERACY.WHST.6-8.1.A

Introduce claim(s) about a topic or issue, acknowledge and distinguish the claim(s) from alternate or opposing claims, and organize the reasons and evidence logically.

CCSS.ELA-LITERACY.WHST.6-8.1.B

Support claim(s) with logical reasoning and relevant, accurate data and evidence that demonstrate an understanding of the topic or text, using credible sources.

HIGH SCHOOL STANDARDS (9-12)

Science

HS-ESS2-2

Analyze geoscience data to make the claim that one change to Earth's surface can create feedbacks that cause changes to other Earth systems

HS-LS3-3

Apply concepts of statistics and probability to explain the variation and distribution of expressed traits in a population.

BIOGRAPHY OF THE AUTHOR

William Schlager was born in Ames, IA. He graduated from Platteville High School in Platteville, WI in 2006. He attended Northland College in Ashland, WI where he graduated in 2010 with a B.S. in Biology. After graduating, he spent four years traveling and working seasonal ecology and forestry jobs for universities and state and federal agencies. He is a candidate for the Master of Science in Teaching degree from The University of Maine in May 2017.