

Students' Reasoning about Variability in Graphs during an Introductory Statistics Course

Rachel Chaphalkar ^{1*}, Ke Wu ²

¹ University of Wisconsin – Whitewater, USA

² University of Montana, USA

* CORRESPONDENCE: ✉ chaphalr@uww.edu

ABSTRACT

Variation and variability are key concepts in K-16 statistics education. Prior research has investigated students' reasoning about variability in different contexts. However, there is a lack of research on students' development of understanding of variability when comparing distributions in bar graphs, dot plots, and histograms as they took an introductory college-level statistics course. This exploratory case study conducted three interviews with each of the ten participants through a four-month period, at the beginning, middle, and end of the course. The Structure of Observed Learning Outcomes (SOLO) taxonomy was used to analyze participants' responses. Results indicated that overall the group of participants demonstrated a stable understanding of variability over the semester (i.e. lack of improvement). However, when examining each student's reasoning, four types of reasoning development paths were found: improvement, lack of change, decline, and inconsistent. This study provides implications in teaching college introductory statistics course and recommendations for future research.

Keywords: statistical reasoning, college introductory statistics, distribution comparisons, qualitative case study

INTRODUCTION

Variation and variability are important concepts in K-12 and postsecondary statistics education (Guidelines for Assessment and Instruction in Statistics Education (GAISE) Pre-K-12 Report, 2007; GAISE College Report ASA Revision Committee, 2016; Common Core State Standards for Mathematics, 2010). Investigations on students' reasoning about variability have started to appear in the research literature in the past decade. Researchers have studied how students reason about measures of variability such as standard deviation (delMas & Liu, 2005) and variability in different contexts (e.g. descriptions of data, lists of data) (Watson, Kelly, Callingham, & Shaughnessy, 2003). Some studies have focused on students' conceptual understanding about variability in different graphical representations of data (e.g. Cooper & Shore, 2008, 2010; Kaplan, Fisher, & Rogness, 2010).

Hiebert and Lefevre (1986) defined conceptual knowledge as being knowledge that is rich in relationships. Students' conceptual knowledge and reasoning about variability in the context of different graphical representations of data has been examined in histograms (Cooper & Shore, 2008; delMas, Garfield, Ooms, & Chance, 2007; Kaplan, Gabrosek, Curtiss, & Malone, 2014), bar graphs (Cooper & Shore, 2010; Whitaker & Jacobbe, 2017), dot plots (Friel & Bright, 1995; Watson & Moritz, 1999), as well as changes over time (Ben-Zvi, 2004; Leavy & Middleton, 2011; Watson, 2001). Though these representations all graph a single variable, they are different in many aspects such as: (1) a histogram summarizes data whereas a dot plot demonstrates

Article History: Received 9 October 2019 ♦ Revised 12 December 2019 ♦ Accepted 16 December 2019

© 2020 by the authors; licensee Modestum Ltd., UK. Open Access terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>) apply. The license permits unrestricted use, distribution, and reproduction in any medium, on the condition that users give exact credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if they made any changes.

the exact data; (2) a bar graph is for a categorical variable whereas a histogram or a dot plot is for a numerical variable; and (3) the shape of a histogram or dot plot has meaning on the location of the data relative to measurements whereas the position of the bars in a bar graph can be rearranged by preference.

Conceptual Understanding of Variability in Histograms

Research has shown students have challenges in understanding variability presented in a histogram. At the college level, Cooper and Shore (2008) found that only 27% of students reasoned correctly when choosing which of two histograms had more variability. delMas et al. (2007) showed that only 47% of college students answered correctly about which of two histograms would have a higher standard deviation. Kaplan et al. (2014) investigated students' reasoning about histograms and discovered common misconceptions: (1) bar graphs and histograms are the same; (2) the x-axis of a histogram represents frequency; (3) histograms with flatter bins have less variability; and (4) histograms have a time component. Kaplan et al. (2014) found that each of these misconceptions was present in some proportion of college students at the beginning and the end of the course; the proportion of some of the misinterpretations increased at the end of the course; and some misunderstandings were found in 60% to 90% of students. Though not all of these constructs were focused on variability, they make it difficult to reason about variability in a histogram.

Conceptual Understanding of Variability in Bar Graphs

Although some studies have found that students confuse histograms with bar graphs (Cooper & Shore, 2010; Kaplan et al., 2014), few studies have focused on students' conceptual understanding of variability in bar graphs. College students often interpreted the bar graph as if it was a value bar chart, where each individual datum value is a bar instead of the bar representing the frequency of the category (Cooper & Shore, 2010). Students have also found the shape of the bar graph to be meaningful, such as a bell-shaped or skewed distribution, when the bars could have been displayed in any order (Whitaker & Jaccobe, 2017).

Kader and Perry (2007) investigated using unalikeability as a measure of variability for categorical variables when comparing whether every pair of data points is in the same category or in different categories. A graph with more categories tended to have a greater unalikeability than one with fewer categories; a graph with a more even distribution of categories had a greater unalikeability than a graph with more observations in fewer categories. For example, the set {A, B, C, D}, where each of the four values are different, has a greater unalikeability than the set {A, A, D, D}, where there are two of each type of value. Unalikeability can be used as a measure of variability in bar graphs, formalizing a general notion of how different data are from each other.

Conceptual Understanding of Variability in Dot Plots

Unlike histograms and bar graphs, dot plots do not summarize data. Although studies involving dot plots have revealed confusion students have about variability, such as appropriate sampling variability (e.g. Shaughnessy, Ciancetta, & Canada, 2004), no studies have shown specific conceptual constructs students have when reasoning about dot plots at the college level. One challenge that middle school students had was that they did not understand what the dots represented in the dot plot and instead they interpreted the dot plot as a value bar chart (Friel & Bright, 1995). Watson and Moritz (1999) found that students from grades 3 to 9 had difficulty comparing dot plots with different sample sizes, which required the use of proportional reasoning or a statistic such as the mean or median.

Changes in Reasoning about Variability

Several studies have looked at how students' reasoning progresses in understanding the center and variability in distributions over different length time periods, however, none looked directly at reasoning over time of college students. Ben-Zvi (2004) shared the progress of reasoning made by two Grade 7 Israeli students over a few class periods. At first, students commented about irrelevant information, then progressed to comparing two frequency tables, and later to using center and spread to make comparisons between two distributions. Leavy and Middleton (2011) studied five upper elementary and middle grades students, over a couple of months, through one-on-one teaching episodes focused on following their progress toward understanding the typicality of data. All students made some progress, although the youngest was unable to move past using the mode as a representative measure, and some of the five students put too strong a focus on the mean. Finally, Watson (2001) conducted a longitudinal study looking at third through ninth grade students' reasoning when comparing the center and spread of sets of two graphs. After three or four years without intervention outside of the regular school curriculum, 62% of students improved on the task, which

was attributed to general development. Ben-Zvi (2004) and Leavy and Middleton (2011) used instruction to help guide students' learning, whereas Watson (2001) attributed progression to general knowledge acquisition. All of these studies focused on the students, who were all pre-college and mostly pre-high school.

Existing research literature provided information on students' conceptual understanding and reasoning about variability in histograms, bar graphs, and dot plots. A few studies have looked at the development of students reasoning over time when comparing center and spread of distributions. However, there is a lack of research investigating the learning paths of the general cognitive understanding that college students have when reasoning about variability through an introductory statistics course. This study intends to fill this gap in the literature by answering the following research questions:

Does a typical college introductory statistics course enhance students' understanding of the concept of variation as it is displayed in graphical representations? What are the development patterns of students' understanding of variation throughout the course?

Background

This study took place at a mid-sized doctoral-degree granting university, with a primary focus on liberal arts education, in the Pacific Northwest region of the United States. The Introductory to Statistics course was offered from the Department of Mathematical Sciences. This 4-credit course consisted of three 50-minute lectures and one 50-minute lab section each week. There were 2 lecture sections, with nearly 250 students enrolled in each and approximately 25 students in each lab section.

The lectures were taught by two instructors (including the first author) who coordinated daily activities and lab section materials. Instructors went through the course pack, which was developed and revised by the statistics faculty who have taught this course many times. The course pack contained examples and notes, and added handwritten notes for each lecture. The lab sections were taught by four graduate teaching assistants and one adjunct faculty member. There were three exams spaced approximately evenly throughout the 16-week long semester. Both lecturers used an iClicker classroom voting tool throughout in order to motivate student attendance and use active learning in the lecture setting. Students also completed twelve worksheets in the lab sections, seven written one- to two- page homework assignments, and online homework assignments. Students' grades were mostly based on exams (69%), with homework, participation, and worksheets making up the rest of the grade (31%).

The class roughly followed the textbook, *Intro Stats*, (De Veaux, Velleman, & Bock, 2013) with the following main topics: exploratory data analysis, linear regression, data collection, randomness and basic probability, central limit theorem, confidence intervals, and hypothesis testing for one and two proportions and means. Although variability was specifically covered during exploratory data analysis, focusing on both measures (range, interquartile range (IQR), standard deviation) and conceptual understanding, the topics throughout the rest of the course contained measures of variability. For example, in hypothesis testing, the null was either rejected or failed to be rejected based on the size of the standard error. These concepts were also embedded in the histograms of sampling distributions created from simulations of repeated sampling of a population.

METHODS

This investigation was longitudinal because the research question was posed with regard to students' conceptual development over time (Creswell, 2013). A qualitative exploratory case study design was implemented because it allowed in-depth examination on reasoning about variability and the changes over time (Yin, 2017). Ten college students participated in the study. Data were collected through interviews. Analysis of data utilized the Structure of Observed Learning Outcomes (SOLO) taxonomy (Biggs & Collis, 1991). The SOLO taxonomy consists of five levels of increasingly sophisticated reasoning (pre-structural, uni-structural, multi-structural, relational and extended abstract). The structure of these levels is explained in greater detail in the data analysis section.

Participants

All students who took the Introduction to Statistics course were invited to participate in the study. A brief announcement describing the study was made in both lectures during the second week of class. Students who were willing to participate in the study filled out an online consent form and demographic information through Survey Monkey, an online survey system outside of the university online course system. On the consent form, students were also asked to indicate whether they would be willing to participate in a series of short

Table 1. Demographic Information on Interview Participants (with pseudonyms)

Participant	Gender	Lecture Section	Major	Class Standing	Age Range
Nicole	Female	2	STEM	Sophomore	18-25
Emily	Female	1	Business	Senior	18-25
Mark	Male	1	STEM	Senior	26-33
Josh	Male	2	Health	Senior	34-41
Megan	Female	2	Health	Sophomore	18-25
Allison	Female	2	STEM	Junior	34-41
Peter	Male	1	Business	Senior	>50
Brian	Male	1	STEM	Post-baccalaureate	26-33
Hannah	Female	2	Business	Freshman	18-25
Tim	Male	1	Health	Post-baccalaureate	26-33

interviews. A total of 82 students filled out the online consent form, of whom 39 students indicated they were willing to be interviewed. All students enrolled in the course (regardless of their participation in this study) were instructed to complete three online content-based surveys (see **Appendix A** for items on survey 1) throughout the course as part of their participation grade. Of the 39 students who were willing to be interviewed as indicated on the online consent form, 26 completed the first survey. To ensure representativeness of participants, a purposeful sampling strategy (Onwuegbuzie & Leech, 2007) was implemented with consideration of participants' characteristics such as gender, major, age, and their initial understanding of variability measured by the first survey. Twenty-one of the 26 students were invited to participate in three interviews over the course of the semester. Of the 21 interview invitations, 12 students scheduled the first interview. Of these students, one failed to attend the scheduled meeting and a rescheduled meeting. One other student only participated in the first interview and appeared to have stopped attending the course before the second interview, and did not respond to subsequent interview requests. The remaining ten students completed all three interviews. **Table 1** shows the demographic information on the ten participants.

Data Collection

Participants took three online content-based surveys and completed three interviews over the course of the semester. The first survey preceded the first interview. Subsequent interviews and surveys were grouped near each other but not in a particular order.

Surveys

This study focused on the changes in the learning of individual students' general (or overall) reasoning about variability, not just on a single type of graph. Thus, the survey was designed to assess student understanding of variability in graphs included bar graphs, dot plots, and histograms. These graphs provided multiple platforms where students' intuitive sense of variability could be explored. Items on the survey were adapted from prior studies (e.g. Watson, 2001; Cooper & Shore, 2008) and refined after a pilot study. Each survey contained ten items in which students compared two or three graphs to indicate which graph had greater variability. Through the three surveys, the contexts of the graphs changed, but the graphs remained fixed. Participants were asked to explain their answers to three survey items during each survey. See **Figure 1** for an example item. The complete first survey can be found in the **Appendix A**.

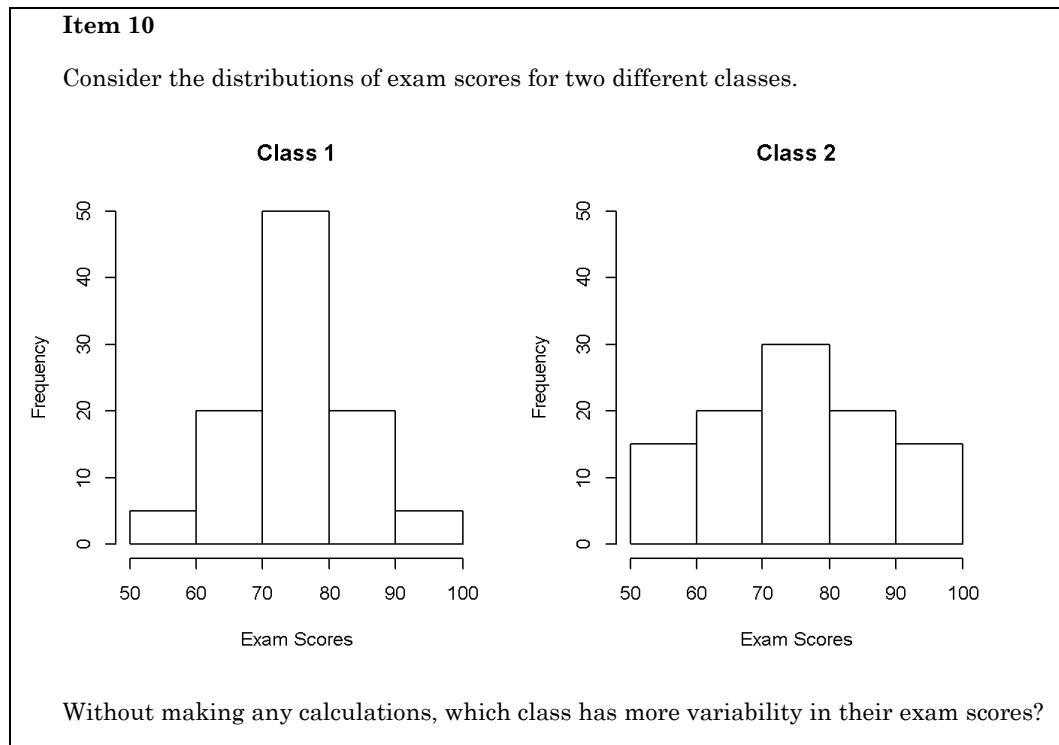


Figure 1. Item 10 was asked during each survey and the second and third interviews

Interview data

Ten students were interviewed three times over the course of the semester corresponding to each online survey, typically during the week before each exam. Interviews lasted between six and 23 minutes. All interviews were audio recorded and transcribed. During each interview, students were asked to explain their reasoning on five of the ten survey items. These questions were chosen first because there were interesting results from the survey items (either similar answers or different answers from most participants) or, secondly, students had not been asked to explain their responses on the survey. The interviewer gave participants positive feedback for explaining answers while attempting to not give feedback on the correctness of the answers until after all interviews were completed in order to understand participants' reasoning progression.

The descriptions of items and their inclusion in interviews can be found in **Table 2**. **Table 2** also summarizes the similarities and differences of the survey/interview items over the three surveys/interviews. For instance, Item 2 contained two dot plots, one with a larger range, and the second with a larger IQR. The item is embedded in the context of number of pets on survey 1 and survey 3 and the number of blood donations on survey 2. Participants were asked to explain their reasoning on this item on surveys 1 and 3 and during interviews 1, 2, and 3.

Table 2. Summary of survey and interview items, contexts, and explanation responses

Item	Graph type	Graph description	Item Context			Explanation S: Survey I: Interview
			Survey 1 and Interview 1	Survey 2 and Interview 2	Survey 3 and Interview 3	
1	Bar Graph	3 options	Blood types	Types of pets	Blood types	S1, I2
2	Dot Plot	Larger range vs IQR	Number of pets	Blood donations	Number of pets	S1, S3, I1, I2, I3
3	Dot Plot	Translations	Number of children	Number of final exams	Number of children	S2
4	Dot Plot	Mirror images	Number of bedrooms	Number of classes	Number of bedrooms	none
5	Histogram	Bimodal vs approximate z	Height	Arm span	Height	S3, I1, I2, I3
6	Histogram	Uniform vs approximate t	Running times	Lifespan of grandparent	Running times	I1, I2, I3
7	Histogram	Skewed vs approximate t	School and work hours	Hometown distance	Exam scores	I1, I3
8	Histogram	Bimodal vs uniform	Basketball game scores	Lifespan of pet	Basketball game scores	none
9	Histogram	Skewed vs uniform	Cost of groceries	Quiz score	Cost of groceries	S2, I1
10	Histogram	Approximate z vs t	Exam scores	Costume price	Minutes exercised	S1, S2, S3, I2, I3

Data Analysis

To answer the research questions, “Does a typical college introductory statistics course enhance students’ understanding of the concept of variation as it is displayed in graphical representations? What are the development patterns of students’ understanding of variation throughout the course?” a two-stage analysis was implemented. First, interview and survey explanation data were coded using the SOLO taxonomy to detect the characteristics of students’ reasoning at different SOLO levels (Biggs & Collis, 1991). Then descriptive statistics on students’ conceptual understanding over time (measured via three interviews) were conducted to identify the different ways students’ reasoning developed through the introductory statistics course.

Phase 1 analysis

The SOLO taxonomy was chosen based on its use in several research studies examining students’ responses to similar items, from which the survey/interview items in this study were adapted (Peters, 2011; Reading, 2004; Watson, 2009; Watson, Callingham, & Kelly, 2007; Watson, Collis, Callingham, & Moritz, 1995). The SOLO taxonomy had five levels (pre-structural, uni-structural, multi-structural, relational, and extended abstract), each based on the complexity of the argument. Complexity increased with each increase in level. This provided a structure to analyze the complexity of students reasoning about variability.

In the SOLO taxonomy, pre-structural responses indicated that students were unable to show a statistically meaningful understanding of variability. Uni-structural responses recognized only one relevant aspect of variability. Multi-structural responses showed several disjoint but relevant aspects of variability. Relational responses integrated several aspects of variability. Finally, extended abstract responses contained generalizations about variability. Detailed description of the characteristics of responses at each SOLO level and the literature supporting these SOLO level descriptions can be found in **Appendix B**. A summary of SOLO levels for understanding responses was provided in **Table 3**.

Table 3. Summary of SOLO taxonomy levels for conceptual understanding about variability in histograms, bar graphs, and dot plots

SOLO Level	Students' response descriptions when comparing variability in histograms, bar graphs, and dot plots
Level 1: Pre-structural	This graph has more variability because it has: <ul style="list-style-type: none"> • Bar heights that differ more or a peak • A larger mean • A larger range of y-values These graphs have the same amount of variability because they have: <ul style="list-style-type: none"> • Symmetry to themselves
Level 2: Uni-structural	This graph has more variability because it has: <ul style="list-style-type: none"> • More categories or more results different from the mean • A larger spread or range • Categories/bins with different numbers in each • A bimodal shape These graphs have the same amount of variability because they have: <ul style="list-style-type: none"> • The same number of responses in each category • The same spread or range
Level 3: Multi-structural	This graph has more variability because it has: <ul style="list-style-type: none"> • Heights of bars that differ less • More evenly spread out data • No peak These graphs have the same amount of variability because they have: <ul style="list-style-type: none"> • Bars with the same heights in different arrangements and the same range
Level 4: Relational	This graph has more variability because it has: <ul style="list-style-type: none"> • More data in tails or further from center/mean/median • A larger IQR, standard deviation, or variance These graphs have the same amount of variability because they have: <ul style="list-style-type: none"> • Shapes that are mirror images or opposites of each other • Differences to the mean that are equal (mean absolute deviations)
Level 5: Extended Abstract	Students demonstrated Level 4 reasoning and recognize the limitations of graphs or of measures of variability: <ul style="list-style-type: none"> • Without the data, it is impossible to tell which histogram has a larger IQR (or standard deviation) • In a dot plot, one graph might have a larger IQR while the other has a larger standard deviation.

Each SOLO level has distinct characteristics and a variety of types of responses that meet this level (see **Appendix B**). For example, students who responded at the uni-structural level (Level 2) recognized variation in one dimension, such as a larger range in a dot plot or histogram or a greater number of categories in a bar graph (Reading, 2004). Students often responded that a dot plot with a larger range was more variable regardless of the placement of the rest of the dots. In a bar graph, students reasoned that graphs that had the same number of categories had the same variability, regardless of the relative frequency of each category.

Phase 2 analysis

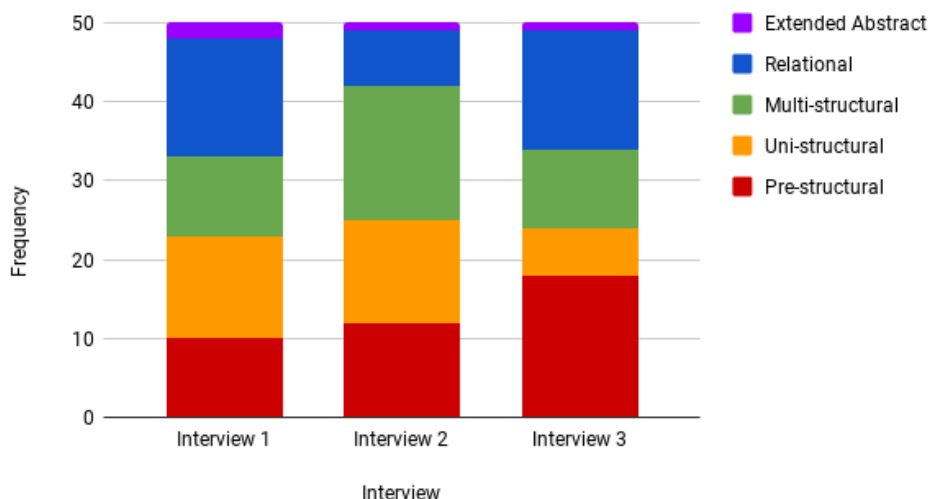
After all of the interview responses were coded, the researchers looked for patterns in the interview responses over the three time periods. Segmented bar graphs and relative frequency tables at both the overall level and at the individual student level were created. Comparisons among these graphs and tables (e.g. proportions of responses at each SOLO level) at the three interview time periods were used to identify patterns of students learning developments.

Interrater Reliability

Two researchers conducted coding using the SOLO taxonomy levels. Both researchers coded three randomly selected interviews and discussed their codes and resolved all the disagreements. Then they repeated the process. The raters' initial agreement on the SOLO levels was 84.2%. The two raters discussed and resolved all the disagreements.

RESULTS

This study investigated different ways students develop their conceptual understanding of variability when comparing graphical representations of distributions during a one-semester college introductory statistics course. The main findings are: (1) the overall development pattern of the 10 participants as a whole was lack of improvement; and (2) individual student's reasoning about variability over the course showed multiple trends: improvement, lack of change, decline, and inconsistent reasoning.



Graph 1. Segmented bar graph of all SOLO level responses of all participants by interview round

Conceptual Development Pattern of All Participants

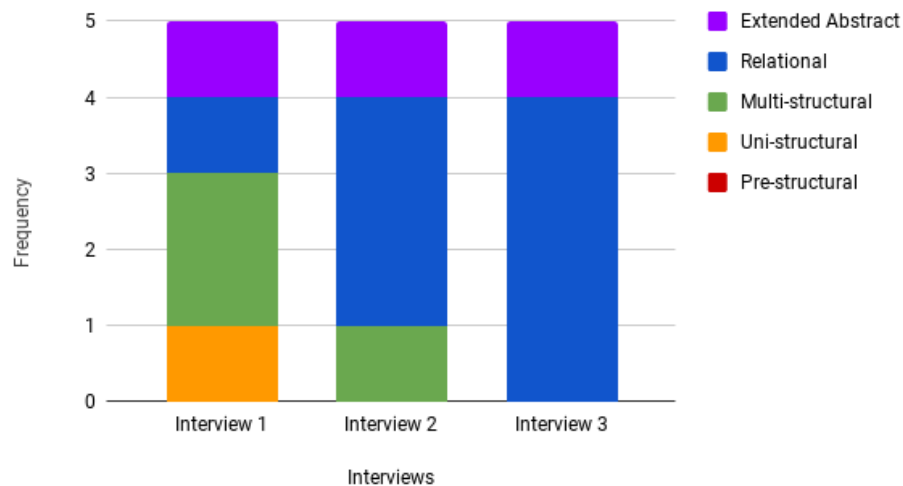
First the researchers investigated the changes in reasoning of all 10 participants over the course. With five item responses from each student for each interview there were a total of 50 item responses for each round of interviews. Each item response was coded a specific SOLO level. A segmented bar graph was generated using the SOLO levels of students' responses (**Graph 1**). Overall, the participants made little progress on their conceptual understanding of variability. There was an increase from the first interview to the third interview in the number of pre-structural responses and a decrease in the number of uni-structural responses, which alarmingly suggests that the students understood less at the end of the course. There was also a decrease in the relational responses in the second interview, however it returned to the original proportion in the third interview. This overall lack of improvement of the students prompted researchers to look at individual students to better understand if all of the students had the same experience.

Conceptual Development Patterns of Individual Participants

When looking at the changes in reasoning of individual students, four general themes emerged from their reasoning about variability when comparing graphs: improved ($n = 2$), decreased ($n = 2$), lack of change ($n = 3$), and inconsistent reasoning ($n = 3$). Each theme was described below.

Improved

Two participants, Josh and Megan (pseudonyms), overall improved their reasoning about variability over the course of three interviews. Megan's improvement was highlighted (see **Graph 2**).



Graph 2. Segmented bar graph of Megan’s SOLO level responses over three interviews

Megan had quite advanced responses, reaching the extended abstract level during each interview. Her improvement was through a reduction of the percentage of her responses at the lower SOLO levels, uni-structural and multi-structural, accompanied by an increase in responses at the relational level. The improvements made over the course of the interviews occurred on Items 2, 5, and 6.

On Item 2, the larger range versus larger IQR dot plot question, Megan gave a uni-structural response based on the graph’s range during the first interview: “Because it is spread out over a further period even though it has a lot more in the middle, I would go with group 1.” Then during the second interview, her response was at the relational level: “So it would be more spread out, away from the mean, I think.” She maintained this level of response during the third interview:

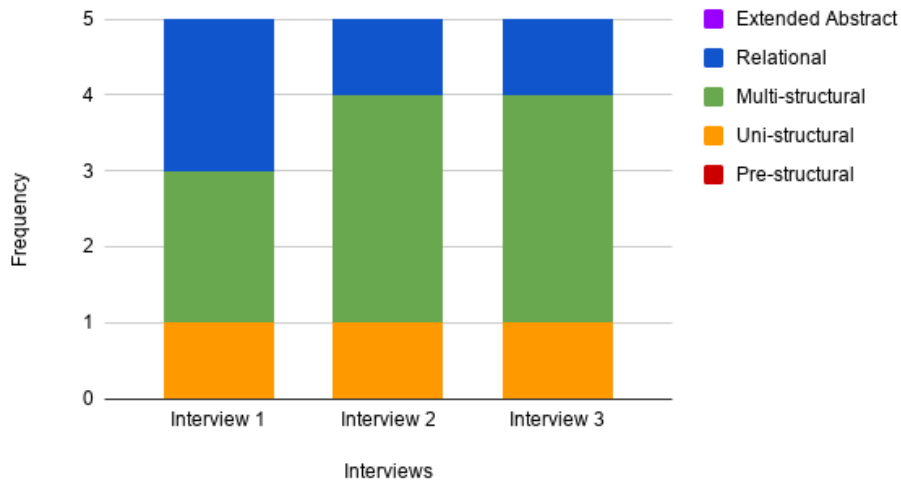
That’s what makes me wonder about group 2, that there’s more that aren’t right at the mean. Like this one has a lot right at the mean and not many far away, but they’re spread out overall. So that’s what kind of trips me up because I don’t know which one would have a bigger standard deviation without calculating it. Megan.

Even though she changed her answer back to the first graph as being more spread out and then focused on the standard deviation, which was larger for graph 1 (but not by a drastic amount), her reasoning was at the relational level.

On Item 6 (uniform and t -distributions), Megan went from a multi-structural response in the first interview to a relational response that became an extended abstract response in the second and third interviews when questioned further by the interviewer. In the first interview, she chose the uniform distribution as having more variability because “the bars were the same.” In the second and third interviews, she again chose the uniform distribution as having more variability than the t -distribution, but explained that if the bin width was changed to a smaller size, that might change the variability: “Yeah, you probably could because if these little bars were a different interval, it might change how they looked and that would make me change my mind.” Her extended abstract level of reasoning is shown by the foresight in seeing that she did not have the data to be able to determine how this might be different and change her overall decision on these graphs.

Lack of change

Three participants, Brian, Peter, and Allison, overall maintained their knowledge and reasoning about variability over the course of the three interviews. Brian continued to have responses at the uni-structural, multi-structural, and relational levels during all three interviews (see **Graph 3**).

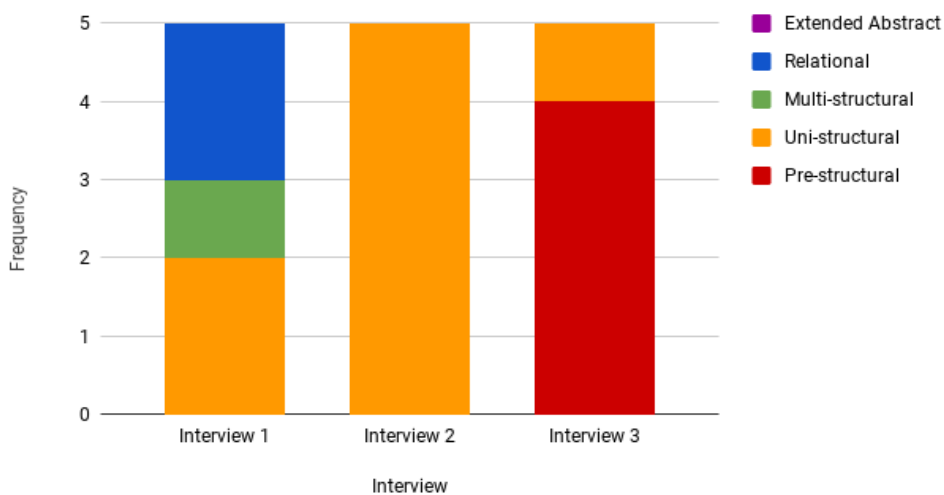


Graph 3. Segmented bar graph of Brian’s SOLO level responses over three interviews

Brian stayed consistent with his answer to Item 2 during all three interviews, basing his answer on the range of the two graphs and identifying the dot plot with the larger range as having more variability, a uni-structural response. On the items about histograms, he gave responses at the multi-structural level, almost always using the description of a graph being more evenly distributed as having more variability. During each interview, on at least one of the histogram items, Brian also had a response in the relational level. For example, in the second and third interviews, Brian had a relational response to Item 10 (*z*-distribution and *t*-distribution), comparing the distributions to determine which had more in the middle: “Because a lot more people are in group 1 are hogging up space in the middle, like 70 to 80 dollars on costumes, versus there’s a more even distribution of how money is spent in group 2.” Overall, Brian’s responses stayed relatively consistent during each interview.

Decreased

Two participants, Tim and Hannah, had overall decreases in the sophistication of their reasoning about variability over the course of three interviews. Tim’s changes were highlighted here (see **Graph 4**).



Graph 4. Segmented bar graph of Tim’s SOLO level responses over three interviews

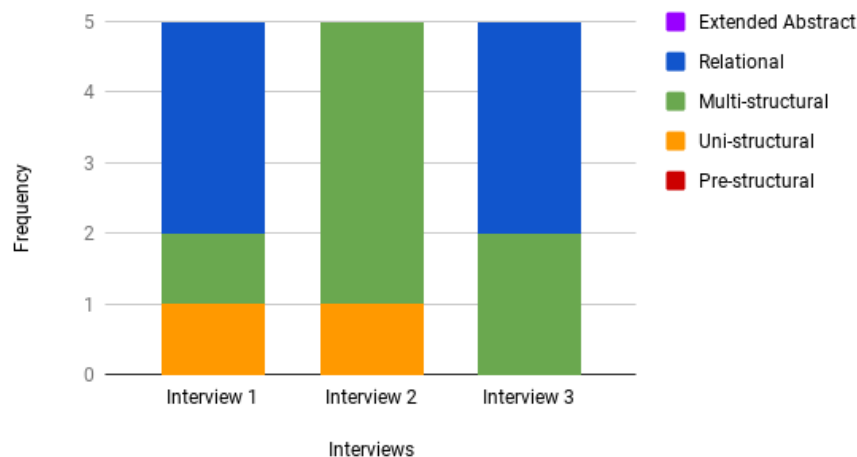
During Interview 1, Tim gave multiple uni-structural, multi-structural, and relational responses. During Interview 2, Tim ended his responses to each item with a uni-structural response, and by Interview 3, he gave a few uni-structural responses and mostly pre-structural responses. For example, on Item 5 (bimodal distribution and *z*-distribution), during the first interview, Tim chose Class 1 as having more variability because, “It’s just more numbers spread out over a greater range” and then explained that, “This one seems

to have a bunch of frequencies in the middle, and then it tapers off and this one has a lot of frequencies in the ends, and then it kind of drops down in the middle”, which was a relational level response. In Interview 2, he started by choosing the group with greater variability in the bar heights (a pre-structural response), then moved to describing that the other graph has more variability due to the “scattering” (a multi-structural response), before deciding the graphs have the same amount of variability due to having the same range (a uni-structural response). In the third interview, Tim started by stating that the variability was the same due to the same range (a uni-structural response) but ended with choosing the class with more variability in the bar heights was more variable (a pre-structural response).

A more direct decrease in responses was seen in Tim’s responses for Item 6 (uniform distribution and t -distribution). In the first interview, he gave a relational response. In the second interview, he gave a uni-structural response. In the third interview, he gave a pre-structural response. Tim did give consistent responses to Item 2 (larger range and larger IQR dot plots), each time basing his response on the range of the bar graph. However, overall, the quality of his answers decreased from quite strong to incorrect over the course of the three interviews.

Inconsistent

Three students’ responses were inconsistent over the three interviews. These students all had a decline in the sophistication of their descriptions from the first to the second interview, but then improved back to where they were at in the first interview during the third interview. Changes in reasoning of one student, Nicole, was highlighted here (see **Graph 5**).



Graph 5. Segmented bar graph of Nicole’s SOLO level responses over three interviews

During Nicole’s first interview, she responded at the uni-structural, multi-structural, and relational levels. During her second interview, she continued to respond at the uni-structural and multi-structural levels but did not reach the relational level. In her third interview, she responded at the multi-structural and relational levels. In addition, her reasoning on these questions was not consistent. For example, in Interview 2, Nicole’s response to Item 5 (bimodal and z -distributions) was at the multi-structural level, focusing on how evenly spread out the distribution was:

I’d say class 1 has more variability because in each of the categories, each of the different ranges, or boxes, there are more evenly spread out between all of them than in class 2 where most of the people have arm spans between 65 and 70 inches. Nicole.

During that same interview on Item 6 (uniform and t -distribution), she responded: “I’d say class 1 has more variability because all of the bars are the same height unlike in class 2” focusing her reasoning on the heights of the bars, which was a different type of response also at the multi-structural level. Her responses showed an inconsistent change in reasoning sophistication throughout the three interviews.

DISCUSSION

Prior research on variability has focused on students’ conceptual understanding on specific graphical representations of data such as histograms, dot plots, and bar graphs (Cooper & Shore, 2008, 2010; delMas et

al., 2007; Friel & Bright, 1995; Kaplan et al., 2014), as well as progress over time of K-12 students (Ben-Zvi, 2004; Leavy & Middleton, 2011; Watson, 2001). However, there was a lack of research examining college students' changes in the general cognitive reasoning about variability in graphs over time. This study used the SOLO taxonomy to investigate the following research questions: Does a typical college introductory statistics course enhance students' understanding of the concept of variation as it is displayed in graphical representations? What are the development patterns of students' understanding of variation throughout the course?

The major findings include: (1) the introductory statistics course in this study did *not* enhance students' understanding of the concept of variation as it is displayed in graphical representations; (2) there were mixed development patterns of reasoning of variation over a semester-long course; and (3) individual students showed multiple levels of sophistication of reasoning within a single interview. Discussion on the implications of each of these main findings follow.

Lack of Improvement in Reasoning of Overall Group of Students

When looking at the progress of the whole group of students, their overall reasoning did not change very much over the course. Proportions of multi-structural, relational, and extended abstract responses stayed relatively consistent, although there was a noticeable increase in pre-structural responses and a decrease in uni-structural responses over the three interviews. This lack of overall progress by students has also been seen in Zieffler and Garfield's (2009) longitudinal study in a semester-long college introductory statistics course. Their study focused on how students' reasoning about bivariate data changed throughout the course. The largest increase in scores happened between the beginning of the course and before starting to learn bivariate data (Zieffler & Garfield, 2009).

Why is there a lack of overall improvement of conceptual understanding on variability? One might argue that describing distributions often occurs at the beginning of a college level introductory statistics course and is often a topic that students bring prior knowledge and hence should not be heavily focused on. However, this study and other research show that there are many students who reason at low levels in the college-level introductory course (delMas et al., 2007, Cooper & Shore, 2008). The concept of variability continues to be important and are expanded upon later in the course when learning about sampling distributions. It is important for students to understand the variability within the distribution in order to understand the benefits of a larger sample size when comparing sampling distributions. Pfaff and Weinberg (2009) found students struggled to see the effect of the sample size on the sampling distribution and to choose an appropriate graph of a sampling distribution. Even post-calculus introductory statistics students struggled with comparing graphical representations of sampling distributions (Lunsford, Rowell, & Goodson-Espy, 2006). Thus it is important for instructors to help students cement their conceptual understanding of variability throughout the course in order to help students better understand sampling distributions, sampling variability, and statistical inference.

Mixed Patterns of Reasoning of Individual Students over a Semester-long Course

In addition to the lack of overall improvement on reasoning about variability of all participants, patterns of reasoning of individual students showed that, while some improved their reasoning, some made little change or were inconsistent, and others experienced large declines. Most students did not have a consistent increase in sophistication in their reasoning about variability when comparing distributions over the course.

Although most participants did not improve their reasoning about variability over the semester, it is important to explore possible factors associated with the two students who had improved reasoning. Examining the similarities of the participants in the improvement theme showed that both had health-related majors, but so did other students who did not make progress. Both took some form of pre-calculus and calculus, whereas the other students had not taken both of these courses, though they may have taken one or the other. Growth from students with a stronger mathematics background aligns with Garfield and Ahlgren's (1988) finding that lack of prerequisite mathematics skills and knowledge is part of the difficulty students have with learning statistics.

The finding that student learning does not always follow a simple path of improvement reminds statistics educators that it is necessary to intentionally design curriculum that helps students move along progressive pathways. It is important to remember that when new information and knowledge is presented to students, students may need to temporarily regress back to prior levels of conceptual understanding of the core ideas in order to figure out how to digest and integrate new concepts before they can reason at a higher level. Knowing

students in a course may have different paths in which to develop their reasoning about variability, instructors need to provide rich and multi-level tasks that facilitate all students' learning. These tasks should help students at all starting levels and lead them to more sophisticated levels of reasoning.

Multi-level Reasoning within a Single Interview

Another important result from this study is that individual students showed multi-level reasoning during a single interview when reasoning about variability in graphs. For instance, during the first interview, 40% of Tim's responses represented higher level (relational level by SOLO taxonomy) reasoning and understanding, 20% were at a middle level (multi-structural level), and 40% were at a lower level (uni-structural level). Although some of the items may have led to particular reasoning, a broad spectrum of graphs were given during each interview to lead to a general understanding of the conceptual structure of student reasoning. This study is the first to address individual students' reasoning about variability using the SOLO multi-level perspective. A few studies looked at student reasoning about variability through the SOLO taxonomy, finding a general increase in reasoning with the age of the student (Watson, 2001), however, researchers have not focused on reasoning at the individual student level apart from a learning trajectory (Ben-Zvi, 2004). Other studies have shown that students' geometric reasoning and degree of acquisition of knowledge can be at multiple levels simultaneously (Battista, 2011; Clements & Battista, 1992), though these studies used van Hiele levels instead of the SOLO taxonomy to classify reasoning levels. The multi-level reasoning of individual students at any moment shows the complexity of learning and has important implications to teaching.

Understanding that individual students reason about variability at multiple levels simultaneously helps instructors make pedagogical and assessment decisions that assist students in making progress on their reasoning. Educators should provide classroom activities and instructions that facilitate multiple levels of reasoning for individual students, and provide assistance for transitions from lower levels to higher levels of reasoning. These activities could involve rich discussions on the features of graphical displays and sensitivity of measures of variability.

Limitations/Future Research

This study took place at a single institution. Researchers contacted 21 purposefully selected participants to schedule interviews, however, only 12 students scheduled a first interview and only ten completed all three interviews. The interviews were conducted by someone participants saw as an instructor of statistics. This likely influenced their willingness and/or motivation to participate. Despite these limitations, analysis showed a diverse group of students participated.

Although researchers adapted survey items from existing research literature (e.g. Cooper & Shore, 2008; Watson, 2001), conducted a pilot study, and discussed these items with a statistician, the surveys and interview questions may need additional tests to establish validity. The pilot study consisted of a single survey taken by 66 students in the Introduction to Statistics course and short interviews with six students on six items. This provided feedback on the feasibility and appropriateness of the survey and interview questions. Adjustments were made to the survey and interview questions such as having participants only compare two dot plots at a time instead of four dot plots. After modifying the survey items, one of the researchers met with a statistician who had extensive experience teaching the course for over 30 years. Together they went through the survey questions and revised them appropriately to meet content validity.

From the results of this study, it appears that the concepts covered in this introductory statistics course do not increase all students' reasoning about variability, so it may be necessary to research the use of an intervention to support students in making improvements in their reasoning. This might look similar to the intervention in Watson's (2002) study where cognitive conflict was introduced to younger students, or may take a form more like the intervention in delMas and Liu's (2005) study on standard deviation which use technology to look at different scenarios in histograms. Future research can examine the progress of college students' reasoning about variation with a larger sample and testing of a hypothetical learning progression at the college level, perhaps similar to Leavy and Middleton's (2011) study of the upper-elementary and middle school levels. In addition, this study looked only at three types of graphs, looking at a wider range of graphs such as box plots, time series graphs, and scatterplots, may add additional knowledge to students' understanding of variability and how it changes.

CONCLUSIONS

In summary, this is the first study looking at the changes in students' conceptual understanding about variability during the beginning, middle, and end of a college introductory statistics course. The SOLO taxonomy was used to analyze students' responses at three interviews. Analysis provided detailed information on the different ways students reason as well as characteristics of students' reasoning at different SOLO levels. The findings that overall students lack improvement of their reasoning, individual students have multiple patterns (improvement, lack of change, decline, and inconsistent) of reasoning throughout the course, and that individual students can have multi-level reasoning during a single interview have important implications for teaching and research for college introductory statistics courses. With these highlighted results, we hope that instructors and researchers will develop and study classroom interventions to assist all students in learning to reason about variability in a sophisticated manner in college level introductory statistics courses. To facilitate positive changes in development of students' cognitive understanding of variability we recommend carefully designed curricula that allow for rich conceptual discussions. This should include extreme cases of variability within and between graphical displays to challenge students.

ACKNOWLEDGEMENTS

Thanks to Dr. Patterson for statistics consulting on this project, Dr. Disney for help with coding, Dr. Johnson for class time to ask for participation, and the participants who volunteered their time.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Rachel Chaphalkar – University of Wisconsin – Whitewater, USA.

Ke Wu – University of Montana, USA.

REFERENCES

- Battista, M. T. (2011). Conceptualizations and issues related to learning progressions, learning trajectories, and levels of sophistication. *The Mathematics Enthusiast*, 8(3), 507-570.
- Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, 3(2), 42-63.
- Biggs, J. B., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behaviour. *Intelligence: Reconceptualization and Measurement*, 57-76.
- Clements, D. H., & Battista, M. T. (1992). Geometry and spatial reasoning. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 420-464). New York, NY: NCTM/Macmillan Publishing Co.
- Common Core State Standards Initiative. (2010). *Common core state standards for mathematics*, Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.
- Cooper, L. L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education*, 16(2), 1-13. <https://doi.org/10.1080/10691898.2008.11889559>
- Cooper, L. L., & Shore, F. S. (2010). The effects of data and graph type on concepts and visualizations of variability. *Journal of Statistics Education*, 18(2), 1-16. <https://doi.org/10.1080/10691898.2010.11889487>
- Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.), Los Angeles, CA: Sage Publications.
- delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, 4(1), 55-82.

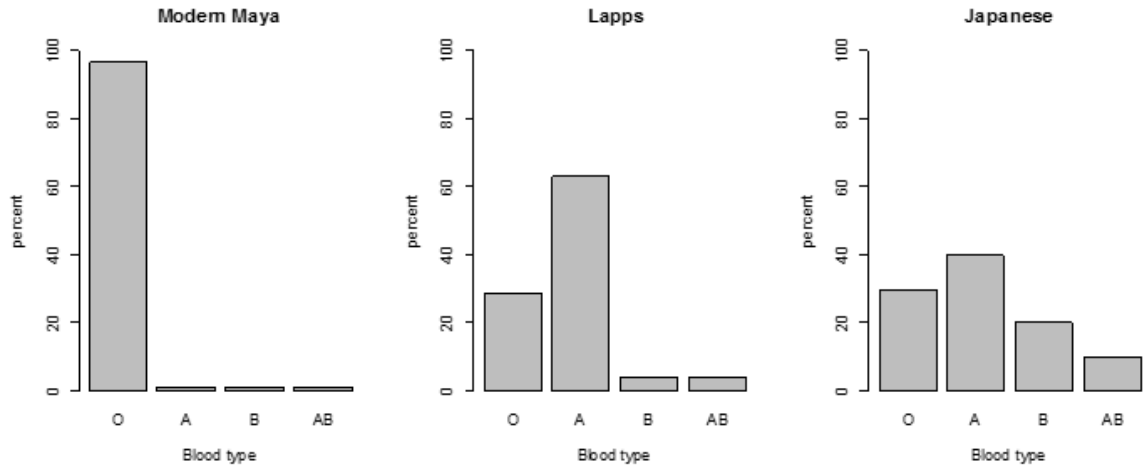
- delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2), 28-58.
- DeVeaux, R. D., Velleman, P. F., & Bock, D. E. (2013). *Intro Stats* (4th custom ed.), Boston, MA: Addison Wesley.
- Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K -12 curriculum framework*. Alexandria, VA: American Statistical Association.
- Friel, S. N., & Bright, G. W. (1995). Graph knowledge: Understanding how students interpret data using graphs. Paper presented at the *Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Columbus, OH. Retrieved from <https://files.eric.ed.gov/fulltext/ED391661.pdf>
- GAISE College Report ASA Revision Committee (2016). *Guidelines for assessment and instruction in statistics education college report 2016*. Retrieved from www.amstat.org/education/gaise
- Garfield, J., & Alhgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, 19(1), 44-63. <https://doi.org/10.2307/749110>
- Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1-27). Hilldale, NJ: Lawrence Erlbaum Associates, Inc.
- Jones, D. L., & Scariano, S. M. (2014). Measuring the variability of data from other values in the set. *Teaching Statistics*, 36(3), 93-96. <https://doi.org/10.1111/test.12056>
- Kadar, G. D., & Perry, M. (2007). Variability for Categorical Variables. *Journal of Statistics Education*, 15(2). <https://doi.org/10.1080/10691898.2007.11889465>
- Kaplan, J. J., Gabrosek, J. G., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education*, 22(2). <https://doi.org/10.1080/10691898.2014.11889701>
- Kaplan, J., Fisher, D., & Rogness, N. (2010). Lexical ambiguity in statistics: How students use and define the words: Association, average, confidence, random and spread. *Journal of Statistics Education*, 18(2), 1-22. <https://doi.org/10.1080/10691898.2010.11889491>
- Leavy, A. M., & Middleton, J. A. (2011). Elementary and middle grades students' constructions of typicality. *Journal of Mathematical Behavior*. 30(3), 235-254. <https://doi.org/10.1016/j.jmathb.2011.03.001>
- Lunsford, M. L., Rowell, G. H., & Goodson-Espy, T. (2006). Classroom research: Assessment of student understanding of sampling distributions of means and the central limit theorem in post-calculus probability and statistics classes. *Journal of Statistics Education*, 14(3). <https://doi.org/10.1080/10691898.2006.11910587>
- Onwuegbuzie, A. J., & Leech, N. L. (2007). A call for qualitative power analyses. *Quality & Quantity*, 41(1), 105-121. <https://doi.org/10.1007/s11135-005-1098-1>
- Peters, S. A. (2011). Robust understanding of statistical variation. *Statistics Education Research Journal*, 10(1), 52-88.
- Pfaff, T. J., & Weinberg, A. (2009). Do hands-on activities increase student understanding?: A case study. *Journal of Statistics Education*, 17(3). <https://doi.org/10.1080/10691898.2009.11889536>
- Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal*, 3(2), 84-105.
- Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004, July). Types of student reasoning on sampling tasks. In *Proceedings of the 28th International Group for the Psychology of Mathematics Education* (pp. 177-184), Vol. 4.
- Watson, J. M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics*, 47(3), 337-372. <https://doi.org/10.1023/A:1015158813656>
- Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics*, 51(3), 225-256. <https://doi.org/10.1023/A:1023622017006>
- Watson, J. M. (2009). The influence of variation and expectation on the developing awareness of distribution. *Statistics Education Research Journal*, 8(1), 32-61.
- Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, 37(2), 145-168. <https://doi.org/10.1023/A:1003594832397>

- Watson, J. M., Callingham, R. A., & Kelly, B. A. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning*, 9(2), 83-130. <https://doi.org/10.1080/10986060709336812>
- Watson, J. M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation*, 1(3), 247-275. <https://doi.org/10.1080/1380361950010303>
- Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, 34(1), 1-29. <https://doi.org/10.1080/0020739021000018791>
- Whitaker, D., & Jacobbe, T. (2017). Students' understanding of bar graphs and histograms: Results from the LOCUS assessments. *Journal of Statistics Education*, 25(2), 90-102. <https://doi.org/10.1080/10691898.2017.1321974/>
- Yin, R. K. (2017). *Case study research and applications: Design and methods*. Los Angeles, CA: Sage publications. <https://doi.org/10.1002/col.22052>
- Zieffler, A. S., & Garfield, J. B. (2009). Modeling the growth of students' covariational reasoning during an introductory statistics course. *Statistics Education Research Journal*, 8(1), 7-31.

APPENDIX A: SURVEY 1

Item 1

Consider the distributions of the blood types of three different ethnic groups.

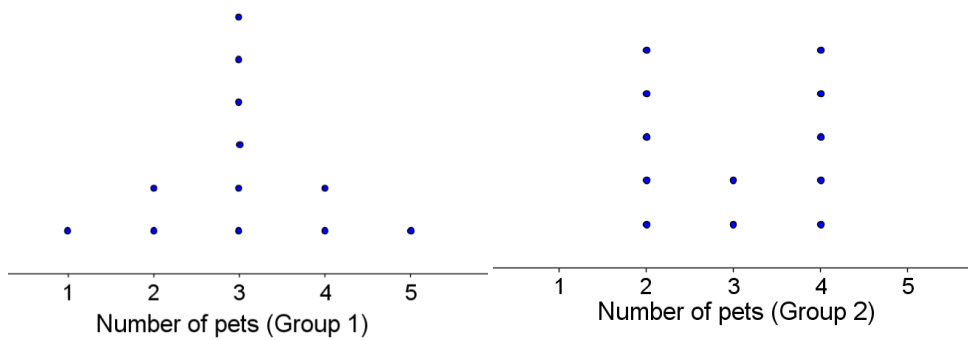


Without making any calculations, which group has the most variability in their blood types?

Which group has the least variability in their blood types?

Item 2

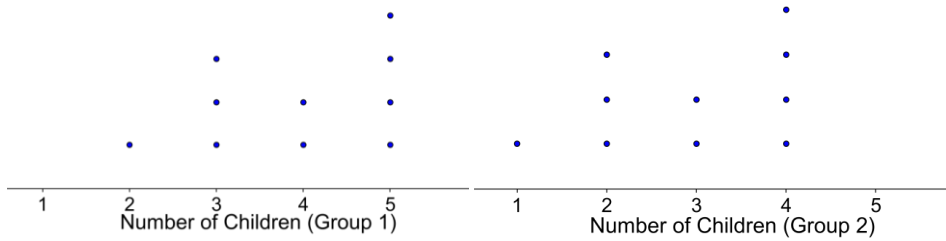
Two groups of people were asked how many pets they own. Their responses can be seen in the following dot plots.



Which group has more variability in the number of pets they own?

Item 3

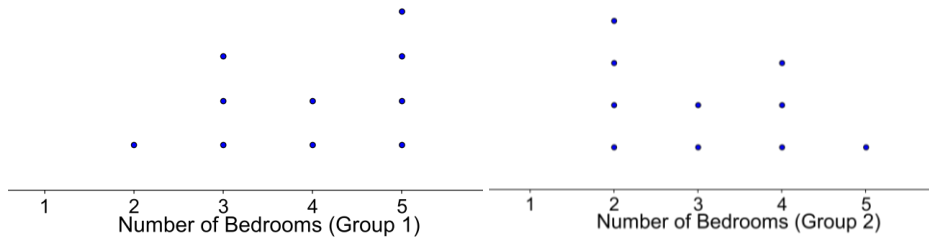
Two groups of people were asked how many children were in the family in which they were considered a child. Their responses can be seen in the following dot plots.



Which group has more variability in the number of children in a family?

Item 4

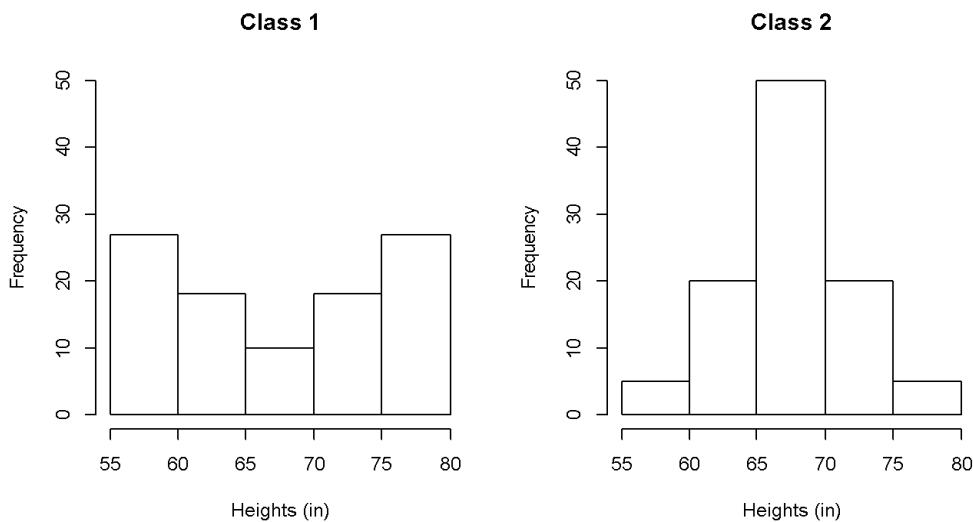
Two groups of people were asked how many bedrooms are in the house that they currently live in. Their responses can be seen in the following dot plots.



Which group has more variability in the number of bedrooms in the house?

Item 5

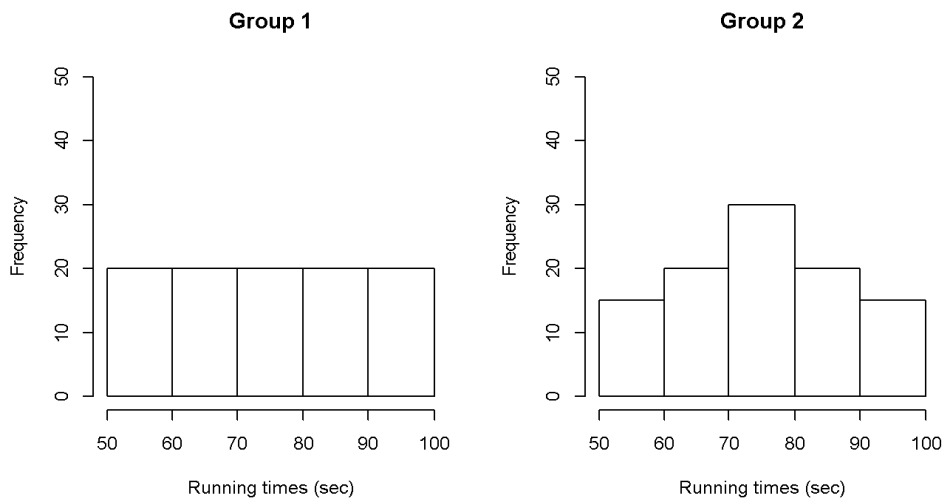
Consider the distributions of heights (in inches) for two different classes.



Without making any calculations, which class has more variability in their heights?

Item 6

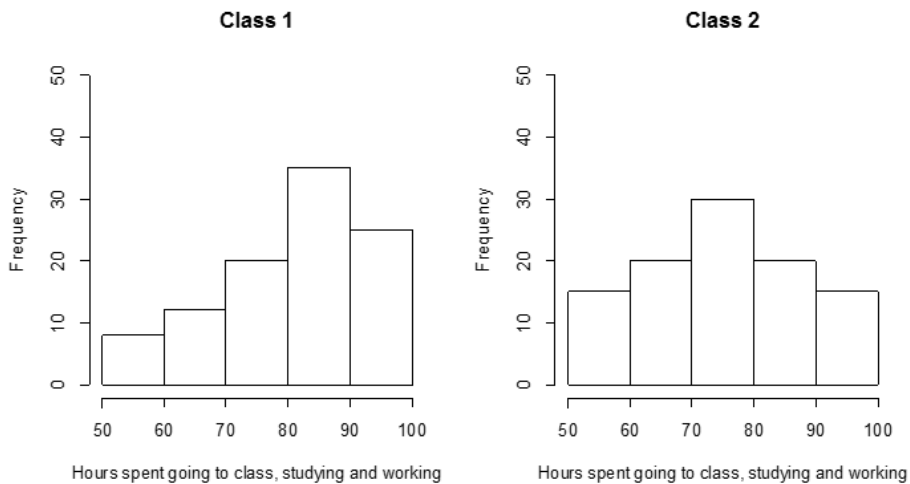
Consider the distributions of time (in seconds) it took two different groups of runners to run 400 meters.



Without making any calculations, which group has more variability in their time to run 400 meters?

Item 7

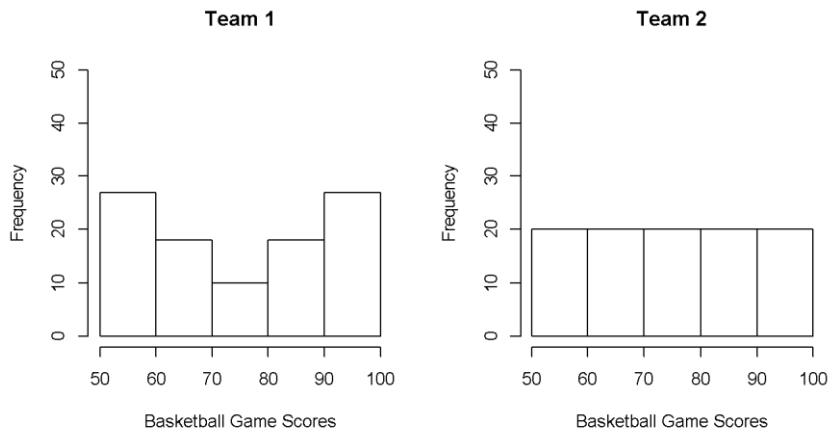
Consider the distributions of time (in hours) it took two different classes spent going to class, studying, and working in one week.



Without making any calculations, which class has more variability in their time spent going to class, studying, and working in one week?

Item 8

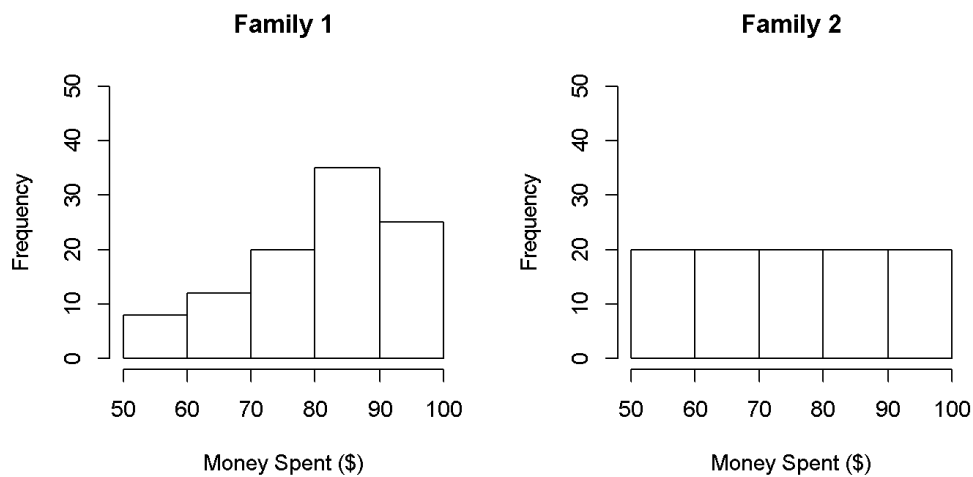
Consider the distributions of points scored in a game by two particular basketball teams over 8 years.



Without making any calculations, which team has more variability the points scored in a game?

Item 9

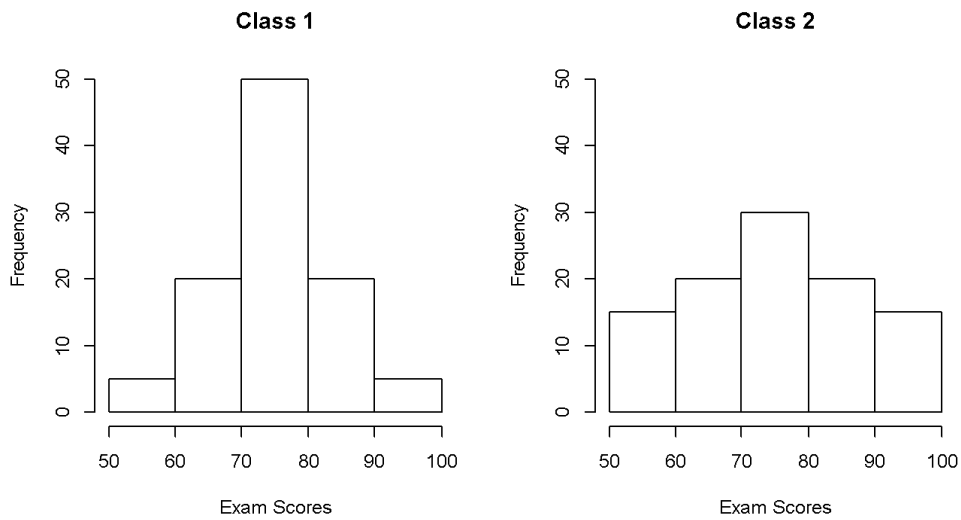
Consider the distributions of the amount of money spent on groceries each week for two particular families over nearly two years.



Without making any calculations, which family has more variability in the amount of money spent on groceries each week?

Item 10

Consider the distributions of exam scores for two different classes.



Without making any calculations, which class has more variability in their exam scores?

APPENDIX B: EXPANDED SOLO LEVEL RESPONSE DESCRIPTIONS

Characteristics of Pre-structural (Level 1) Responses

Students who responded at the pre-structural level recognized that there was variation in the displays but were unable to describe it in a statistically meaningful way. This often occurred by reasoning about variability as if the bar graph, dot plot, or histogram was a case-value chart; however, students did not necessarily interpret the graph as a case-value chart when asked in interviews. These students often thought that when the heights of the bars of a histogram, dot plot, or bar graph were more variable there was more variability in the display; however, the opposite is actually true. This is similar to a learning trajectory from Stage 1 - beginning from irrelevant information (Ben-Zvi, 2004).

Characteristics of Uni-structural (Level 2) Responses

Students who responded at the uni-structural level recognized variation in one dimension, such as a larger range in a dot plot or histogram or a greater number of categories in a bar graph (Reading, 2004). Students often responded that a dot plot with a larger range was more variable regardless of the placement of the dots. In a bar graph, students reasoned that graphs that had the same number of categories had the same variability, regardless of the relative frequency of each category.

Characteristics of Multi-structural (Level 3) Responses

Students who responded at the multi-structural level could identify two or more disjoint features of variation (Reading, 2004), such as in the horizontal and vertical directions, but were unable to put these ideas together. They responded to questions where the distributions had the same range by explaining that, although the histograms with bars that had similar heights had more variability, graphs with larger differences in heights had less variability; however, these students were unable to express the change in variability if the bars in a histogram were reordered. Similarly, students reasoned that dot plots with a more even distribution had more variability than a single value with many points, without mentioning the importance of these data points being in the center of the distribution. Finally, these students reasoned that bar graphs with bars of similar heights had more variability than bar graphs with bars of different heights (all of the bar graphs had four bars.)

Characteristics of Relational (Level 4) Responses

Students who responded at the relational level were able to make connections between the center of the data and the spread of the data for histograms and dot plots (Reading, 2004). This was either by discussing the proportion of data in the tails compared to the proportion of data in the center or by using the IQR or standard deviation. Both the IQR and standard deviation were measures of variability that took into account both center and spread of the distribution. However, students at this level were not able to recognize that, although one distribution could have a larger IQR, the second distribution could have a larger standard deviation such as in Question 2 (see **Appendix A**). In addition, students at this level were not able to notice that, without the actual data, histograms such as those in Question 8 (see **Appendix A**) could possibly either have the larger IQR or larger standard deviation. In bar graphs, a relational response would be represented by calculating or discussing which graph would have a greater measure of unalikeability.

It is also important to note that variance can be described as a measure of how far data are from each other, not just from the mean, so this measure of center does not need to be taken into account (Jones & Scariano, 2014). A response that gave this reasoning would also fall into the relational level.

Characteristics of Extended Abstract (Level 5) Responses

Students who responded at the extended abstract level were able to discuss how relational-level responses conflicted, or how certain information about the data could not be detected based on the graphs. Students who responded at this level were able to recognize when the IQR and standard deviation were in conflict, such as in Question 2 (**Appendix A**). They were also able to recognize that the histograms in Questions 6 through 10 did not necessarily lead to one or the other having a higher standard deviation or IQR. Either graph can have a smaller range, IQR, or standard deviation when extreme (but possible) examples are created.

<http://www.iejme.com>