International Journal of Science and Mathematics Education (2023) 21:1243–1264
https://doi.org/10.1007/s10763-022-10293-3

Undergraduate Students' Use of Primitive Notions When Reasoning About Variability

Oguz Koklu[1] · **Jennifer J. Kaplan**[2]

Received: 17 November 2021 / Accepted: 14 May 2022 / Published online: 10 June 2022

Abstract

Previous research has shown that students form fragmented understandings about variability even after instruction and that these primitive notions function as obstacles to a robust understanding of variability. To investigate the effect of disruptions to students' primitive notions of variability, we examined undergraduate students' reasoning about variability when distributions of quantitative datasets to be compared (a) have equal ranges, (b) do not include extreme values, and (c) have approximately the same number of different values. We designed homework questions and collected student responses to these questions from a large number of students attending an introductory statistics course. We also reorganized these questions into interview tasks and conducted task-based interviews with students. The results showed that students either did not address variability in these narrowly framed situations or provided limited, ambiguous, or inconsistent responses. The findings suggest that students typically think about variability in terms of range's colloquial meaning and how different the values are from each other. Students' reasoning was often contingent upon the particular and more prevalent characteristics of the distributions on which they were asked to work. Future work should explicate ways to utilize students' available conceptions of variability in teaching the concept's normative meaning.

Keywords Intuitive notions of variability · Reasoning about variability · Statistics education research · Undergraduate statistics education

Variability is a core concern in statistical investigations, and teaching and learning the concept is central in statistics instruction (Bargagliotti et al., 2020). Although reasoning

✉ Oguz Koklu
kokluoguz@gmail.com

Jennifer J. Kaplan
jennifer.kaplan@mtsu.edu

[1] Department of Mathematics and Science Education, Marmara University, Istanbul, Turkey

[2] Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN, USA

about variability is crucial for understanding and practicing statistics, it is multifaceted and often complex for students (American Statistical Association [ASA], 2016; Çatman Aksoy & Işıksal Bostan, 2021; National Governors Association Center for Best Practices [NGA Center] & Council of Chief State School Officers [CCSSO], 2010). The lack of attention to variability impedes the understanding of other key statistical ideas such as distribution, sampling, and inference (Biehler et al., 2018; Makar, 2016).

Previous studies (see, e.g. Garfield et al., 2007; Lann & Falk, 2003; Reid & Reading, 2008) have shown that developing the normative meaning of variability is challenging. Students often rely on their preexisting "basic notions of variability" (Pingel, 1993, p. 71). These notions are often reflected in students' actions as (a) comparing only ranges across datasets, (b) focusing only on individual values (usually the extreme ones), and (c) exploring the extent to which observations in a distribution are repeated. In brief, students' overreliance on their preexisting primitive notions of variability and difficulty in attending to how data values cluster around a central value make students' reasoning about variability an overwhelming task.

Because primitive notions often result in limited attention to the core meaning of variability and could function as obstacles to the robust understanding of variability, it is essential to investigate how students reason when their primitive notions become insufficient to address variability. Therefore, in this study, we investigated how undergraduate students reason about variability when their primitive notions are neither applicable nor fruitful because of the particular characteristics of a given distribution or dataset. We sought answers to the following research questions:

How do undergraduate students reason about variability when the datasets or distributions to be compared have (a) equal ranges? (b) no extreme observations? and (c) approximately the same number of different values?

Literature Review

The Concept of Variability

The normative statistical meaning of variability, as traditionally described in textbooks (see, e.g. Bock et al., 2012) refers to the measures that indicate how data values typically deviate from a center value. Two approaches can be used to describe variability in the case of a univariate quantitative variable (Jones & Scariano, 2014): The first approach is based on the distance between two individual data values as in the case of range and interquartile range (IQR) (Jones & Scariano, 2014). The second approach, which is more representative of statistical norms, is based on the average distances of the observations in a dataset from a centrally located value, usually the mean of the observations (Jones & Scariano, 2014). In this approach, variability is a characteristic of a distribution (Ciancetta, 2007), and each value in the given distribution contributes to the property in different weights. The widely used formal measure is the standard deviation (SD) (Bock et al., 2012).

These two approaches used in evaluating variability, the distance between two points and the average distance of the points from the center, emphasize different characteristics of distributions. Although using range and IQR is practical if a dataset

has only a few values and if there is no clustering in the dataset, they are less helpful if a distribution is approximately bell-shaped (Bock et al., 2012). On the other hand, although measures of variability such as SD and mean absolute deviation (MAD) prioritize information on how observations differ from the mean, the use of these measures is less useful in skewed distributions because the mean and SD are easily distorted. To conclude, students need to recognize the difference between these distinct approaches (i.e. measures) to variability and build skills that help students to select and employ the most appropriate approach under particular conditions.

Students' Understanding of Variability

Students, including the ones at the tertiary level, hold a limited understanding of variability, often fragmented, incomplete, or even contradictory (Garfield et al., 2007; Lann & Falk, 2003; Loosen et al., 1985; Pingel, 1993; Reid & Reading, 2008). Even after exposure to statistical instruction, many students rely on individual informal methods (Jones & Scariano, 2014). In this study, naive understandings of variability are called *primitive* notions of variability. The most common primitive notions that are listed in pertinent literature are centering variability arguments on the ideas that a distribution will indicate more variability if it (a) has a greater range, (b) has (more) extreme values, or (c) has (more) diverse values.

Has a Greater Range

Students tend to equate variability with range and thus focus only on range differences among the distributions under investigation (Ciancetta, 2007; Lann & Falk, 2003; Shaughnessy, 2007). Lann and Falk (2003) found that the greater proportion of first-year university students employed range than any other single measure of variability (such as MAD, IQR, and SD) when asked to intuitively compare the variability of given pairs of small raw datasets. Garfield et al. (2007) and Lann and Falk (2003) found students to be easily distracted by the differences in range values between datasets, which in turn, discourages students from investigating variability further. In their hierarchy of consideration of variation framework, Reid and Reading (2008) categorized the use of range only as an indication of a weak consideration of variation.

Has (More) Extreme Values

When asked to reason about variability, students, including undergraduate students and some prospective teachers, narrow their attention to individual data points only (Garfield et al., 2007; Reid & Reading, 2008; Shaughnessy, 2007), usually to the extreme values or outliers. In their study with undergraduate students, Reid and Reading (2008) found that students used extreme values and the shape of the distribution to establish their reasoning about within-group variation. Focusing only on individual data points precludes students from considering the overall characteristics of distribution, thereby providing an insufficient assessment of variability (Garfield & Ben-Zvi, 2005; Lann & Falk, 2003; Shaughnessy, 2007).

Has (More) Diverse Values

Students often assess variability based on how much the values in the given quantitative dataset or distribution differ from each other (Lann & Falk, 2003; Loosen et al., 1985). Loosen et al. (1985) investigated 154 undergraduate students' intuitive ways of understanding variability of quantitative variables and found that students usually base their choice on how much the values differ from each other. Similarly, Lann and Falk (2003) found that students claimed recurrences of the same value in a dataset as indicative of less *heterogeneity*, thereby, an implication of smaller variability. In brief, equating variability to "diversity" outlines this primitive notion.

Our review of the literature suggests that students often follow strategies that are similar to their intuitions. Little is known, however, about how students reorganize their reasoning when their primitive notions remain contradictory or ineffectual. For example, we found no research study in which students could not use range—because the given datasets to be compared had equal ranges—when exploring variability. Investigating students' reasoning in a more focused and constrained way, as outlined in the research questions, may help diagnose other not-reported primitive reasoning mechanisms students use and examine how those mechanisms could lead to a more formal understanding of variability.

Methodology

This study uses an explanatory mixed-methods design to investigate how undergraduate students reason about variability in data sets when comparing datasets or distributions constrained by primitive notions of variability. In this section, we first describe the context of the study. Next, we introduce the data collection procedures. We then explain the analytical frameworks and the analysis of data.

Study Setting: Introductory Statistics Course

The subjects in the study were all students in a multi-section, 4-credit, undergraduate statistics course offered each semester at a large research university located in the Southeastern part of the USA. Course topics included data collection, study design, descriptive statistics, confidence intervals and hypothesis testing for one-sample and two-sample proportions and means, correlation and simple linear regression, two-way tables, and the chi-square test. The course included five tests, twenty homework assignments, and ten computer lab assignments. The tests constituted 80% of the final grade of the course, and the remaining 20% was distributed between homework and computer lab assignments in differing weights. Students were required to use a course assignment platform called WebAssign (https://webassign.net/) to take the tests, lab assignments, and homework questions. Students received slightly different assignments from each other, and the WebAssign performed the grading automatically.

Data Collection

The data associated with this study were a large body of students' responses to homework questions in which the students were asked to choose which of the data sets or distributions had more variability and provide written justifications of their choice. Video-taped interviews with students provide a second source of data. Data collection methods and the recruitment of the interview participants were approved by the Institutional Review Board (IRB) of the institution where the study was conducted. In the following sections, we explain the details of both data resources and the data collection.

Homework Questions

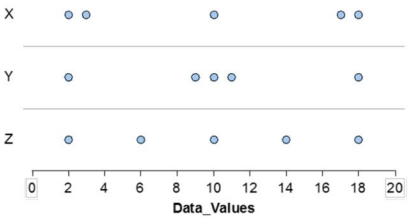
The first step of the data collection involved gathering student responses to questions embedded in the online homework assignments. Results from five homework questions are reported. We designed these homework questions (Fig. 1) to elicit students' ways of reasoning when the datasets to be compared have equal ranges, have no extreme values, or have an equal number of data values.

The data were collected from all of the students who submitted their responses to the assignments in the spring semester. The concept of variability was introduced in the second week of the semester. Question 2 appeared on a homework question the following week. Subsequent questions appeared on homework assignments in weeks 4 (question 3), 5 (question 5), 7 (question 6: Shoes), and week 10 (question 7). A summary of the features of the questions and the number of responses is provided in Table 1. Total enrollment for the course was around 1200 students, so the last column in the table suggests that the total number of responses was often high. The reason questions 3 and 7 have lower totals was that these questions had multiple versions and only one version of the questions was appropriate for the study.

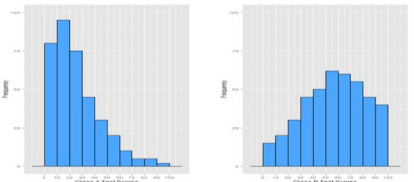
Interviews

We incorporated interviews into the research study by recruiting students for a series of task-based interviews during the semester students were enrolled in the course. We developed interview tasks by modifying the homework questions. For the recruitment of interview participants, the teaching assistants forwarded the recruitment letter and the approval of the study by the Institutional Review Board (IRB) to the enrolled students with the contact information if they wanted to volunteer. Each interview was video recorded with a camera pointed at the paper on which the interview participant put her work. We piloted the interview tasks with three students during the spring semester and revised the tasks. Then, two students were recruited for interviews during the summer and four students were recruited during the fall semester, but only the four students from fall were included in the analyses. Overall, we worked with nine students for the interviews. Our preliminary analysis of the two students'

Q2. Which of the datasets depicted in the graph has the least variability?

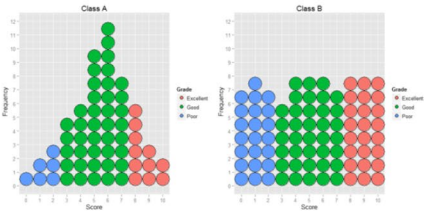


Q3. Compare the variability of the distributions of test scores for Class A and Class B.



Q5. Two datasets are given as follows:
(1st dataset) 3 3 3 40 40 40 77 77 77
(2nd dataset) 10 11 12 13 40 40 70 75 84
Without calculating, determine which dataset has more variability. Explain the reason.

Q7. For which class, A or B, are the scores more variable (i.e. have the higher standard deviation)?



Q6. Which group has more variability?

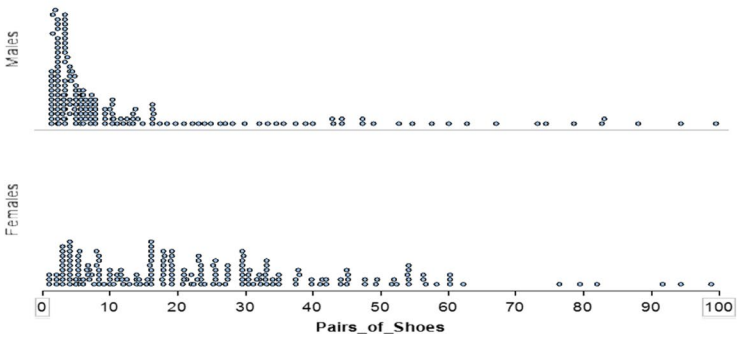


Fig. 1 Homework questions

interview data, which were not included in this paper, did not seem to provide additional information derived from the four participants presented here. The data achieved saturation from the in-depth analysis of the reasoning of these four participants.

Table 2 provides information about interview participants (real names replaced with pseudonyms) whose data are reported in this paper. Only Karen needed a third interview to complete all of the tasks. The schedule for the fall semester was identical to the spring semester, which meant that the students had a chance to work on the aforementioned homework questions before they were interviewed. All of the interviews were conducted by the first author. It was reiterated during the interviews that participants' ways of reasoning were more essential than whether or not their responses were correct.

Table 1 Description of the homework questions

Question	Data features	Presentation (# distributions)	Multiple-choice component?	Total response
2	Equal range, equal number of data points, same number of data values	Dot plots (3)	Yes (cannot determine)	1072
3	Equal range, no extreme values, different shape	Histograms (2)	No	314
5	Equal range, equal number of data points, different number of data values	Raw data (2)	No	1065
6	Equal range and no extreme values	Dot plots (2)	Yes (same/cannot determine)	987
7	Equal range and same number of values	Dot plots (2) Histograms (2)	Yes (forced)	378

Table 2 Interviews

Pseudonym	Major	Prior statistical experience	Interview date (duration, in minutes)		
			First	Second	Third
Ocean	Biology, Pre-dentistry	Took an AP Statistics course but not the exam	10/06 (62)	10/10 (70)	-
Britney	Exercise and Sport Science	Took an AP Statistics course	10/13 (62)	10/20 (58)	-
Karen	International Affairs	Took an AP Statistics course	10/14 (65)	10/21 (50)	10/24 (44)
Chloe	Political Science	No experience	10/20 (55)	10/21 (56)	-

Data Analysis

Analysis of Homework Questions

Of the five homework questions, two asked students to compare the variability of two data sets without first asking the students to choose the more variable distribution (Q3: test scores and Q5: raw data set). All of these responses were read and coded into one of 4 categories: a clear indication of the distribution that was more variable (1 or 2/A or B), clear indication that the variability is the same for both groups, or no clear choice made. For the questions in which students were first asked to choose the most or least variable distribution, Q7: test scores provided only a forced choice between the two distributions, but the other questions, Q2 simple dot plots, Q6 shoes, allowed students to indicate the variability was the same or differences could not be determined.

Once all responses were coded to indicate which distribution the subject had selected as most or least variable, the reasoning strategies in student responses were analyzed. The initial coding followed Arnold's (2013) framework, which presented two characteristics, spread and density, for variability. The specific descriptors for spread were range, IQR, range as an interval, interval for high and/or low values, and interval for groups. The features for density were clustering density, majority (mostly, many), and relative frequency. It should be noted that there are many developmental (and in general more detailed) frameworks for variability (for a comprehensive review, see Langrall et al., 2017, p. 494). Arnold's framework appeared to be more appropriate in analyzing student responses to homework questions in this study as they were open-ended but relatively brief data.

Independent random samples of 100 responses were selected for each question. Two coders (the authors) independently coded the responses using the features listed in the Arnold framework. After the coding of the 100 randomly selected responses was completed observed that we needed additional categories to code responses. Overall, Arnold's framework served as a first step in the analysis of student responses, but its influence was less prevalent in the subsequent iterations of

coding. The final coding scheme contained eight possible codes for students' reasoning (Table 3). For the student responses that were too ambiguous to give any of the eight categories, the code "None" is given such as a student response to Q6: "Because the females have more shoes that they own giving us more data to work with and it being more variable."

Random samples of responses were recoded independently by the authors of this paper, who then met to discuss differences in code and to refine the codebook. All responses to each question were then coded by one author. Once one author had coded all of the responses to a question, the coding was exchanged between the authors for review. All differences were discussed and in some cases, a second or third round of coding and review occurred until both authors agreed on all codes for all responses.

Analysis of Interviews

The analysis of the interviews followed the seven-step analytical model proposed by Powell et al. (2003). The phases are (a) viewing the video attentively, (b) describing the video, (c) identifying critical instances, (d) transcribing (selectively), (e) coding, (f) building a storyline of the issue under investigation, and finally (g) composing narrative.

As a first step to the analysis, we viewed the videos to become familiar with the approaches the participants used when working on the interview tasks. For the second step of the analysis, we wrote our description of how the interview participant addressed each task and the follow-up questions. In this step, we created a running summary of each participant's ways of reasoning about variability. The next step was identifying critical instances of students' reasoning. The focus was on the type of knowledge resources on which the interviewees depended when reasoning about variability. Any tension (similar to the use of "cognitive conflict" in statistics education literature (Biehler et al., 2018; Reading & Reid, 2007; Watson & Kelly, 2007)) that arose because an interviewee's ways of informal reasoning conflicted with each other might be regarded as a critical instance.

Detecting some of the patterns and inconsistencies of reasoning and determining how participants reframed their ways of reasoning as they worked on different datasets and distributions was crucial. Also, we attempted to focus on participants' repeated patterns of argument and interpretations that were not consistent with normative statistical reasoning. We were particularly interested in the relationships between participants' approaches to tasks and the characteristics of the tasks. This step also included writing summaries that captured each participant's main ways of reasoning. We used these summaries to identify the reasoning strategies that participants used in the subsequent interviews.

The interview participants' particular ways of reasoning and critical instances were noted, and related parts of the video recordings were transcribed. Next, the coding of the video segments took place based on the features given in Arnold's framework (Arnold, 2013) and the eight categories we employed in analyzing student responses to homework questions. The next two steps of the analysis, building a storyline and composing a narrative, took place after the coding procedure. Some of the transcribed

Table 3 The coding scheme

Category	Description	Example (HW question)
Distance to center	Specifically mentioning the relative position of the values to a measure of center or using standard deviation (when the term standard deviation was not referenced in the stem of the question)	The second set because there are more numbers that are farther away from my estimated mean than in the first set (Q5)
Clustered or spread out	Responses that argue a distribution is less variable because the observations are grouped or spread out or spread apart. The word spread by itself is not sufficient for this category; it must be modified by an adjective indicating more spread out than the other distribution	They are all spaced across the board without concentrated amounts in one particular area (Q2)
Range or IQ	The response had to mention either that the range for the distributions was the same or make a comparison of the IQR. Responses that used the word range differently, for example, the data cover a "broad range of scores" or "larger range of variables," were not included in this category	Class A has a larger IQR, a smaller minimum, and a larger maximum than class B (Q3)
Consistent or diverse	Responses that argue that a distribution is more varied because there are different values or the numerical values in a distribution are diverse or the distribution is less varied because the observations in the dataset are very similar to each other. When a response specifically mentioned modes with frequencies it was included in this category	I chose class B because there are more students with different scores, whereas in class A there is a clear mode, so most students received that score, creating less variability (Q7)
Even distribution or spread	Responses that indicated an even distribution or that the data were evenly spread out	The 2nd dataset has more variability because the numbers are more evenly spread out (Q5)
Shape	A response was coded in the shape category if it contained information mainly about the shape of the distribution such as indicating a distribution to be skewed or normal	Class A has more variability compared to class B since the first histogram is a skewed to right graph and the second one is more bell-shaped (Q3)
Extreme values	The response had to make explicit use of extreme values, outliers, or very large or very small data values in the argument	There are fewer responses that would be considered outliers (Q6)
Frequencies	An argument that specifically mentions frequencies. for example, the number of observations in of bins in histograms or that mentioned repetition of data values, especially when given raw data	I know the scores from the class I chose are more variable because there are far more responses in each category (Q7)

events were kept as episodes to exemplify students' reasoning in the "Findings" of this paper. In addition, we discussed part of the preliminary results with a capable colleague (who was also studying students' statistical reasoning) and asked him to suggest counterarguments to or alternative interpretations of our findings.

Findings

In this section, first, we will present the results for each homework question. Then, we will summarize the homework results and interview findings according to each research question.

Homework Results

The analysis of student responses to homework data suggested that students employed various ways to reason about variability in each question. Table 4 presents the frequencies of the eight reasoning categories in each homework question, and the additional category of "None" for the inconclusive responses.

As Table 4 shows, the students' approaches to variability differed substantively as the data feature and presentation changed from one homework question to another, and more than one code was needed in most of the student responses. In addition, some of the student responses were ambiguous or simply missing: these responses were coded in the category "None." In total, 8.3% of the responses to Q2, 12.7% of the responses for Q3, 4.8% of the responses for Q5, 16.7% of the responses for Q6, and 16.4% of the responses for Q7 were assigned no reasoning codes. In the next paragraphs, we will present the results for each homework question.

In Q2, three dot plots with few values were given and asked students to choose the graph with the least variability. The distribution of answer options and also the frequency of each coding category are given in Fig. 2. The option Y ($n=448$; 41.7% of the responses) and the option that claimed Z ($n=415$; 38.7% of the responses) to be the least variable distribution constituted almost 80% of the responses. When we look at the reasoning mechanisms used by students to justify option Y, out of 597 codes, clustered/spread out (261) and distance to center/SD (128) were the most frequent. In brief, most of the students who answered this question were able to focus on clustering of data values or their average distance to center when reasoning about variability.

The reasoning mechanisms used by students to justify the data set Z as the least variable indicated that approximately half of the students who chose this answer option indicated a primitive notion that we coded in the category even. The students employed expressions such as equally spaced, evenly distributed (or spread), uniformly distributed, consistent (constant, even), and spread. These justifications suggested that they hold a primitive variability notion that has not been reported in the literature before, focusing on whether the overall spread was haphazard (as in X or Y) or followed a pattern (as in Z).

Q3 asked to compare the test scores between two classes, class A and class B, represented in histograms, and it was designed not to have any identifiable extreme values. More students chose class A to be more variable (37%), followed by those who

Table 4 The results for the coding of homework questions

		Categories									
Question	Option	Shape	Frequencies	Distance to center	Even distribution or spread	Clustered or spread out	Range or IQR	Consistent or diverse	Extreme values	None	Total codes
2	X	3	0	7	9	40	3	18	5	12	97
	Y	18	2	128	20	261	26	35	75	36	597
	Z	13	1	16	281	97	9	82	13	31	541
	Can't	1	5	1	8	15	46	12	3	10	101
	Total	35	8	152	318	413	84	147	96	89	1336
3	A	56	27	16	4	17	14	11	6	19	170
	B	35	13	11	12	31	7	3	3	10	125
	Same	8	1	0	2	1	9	0	0	2	23
	Undecided	90	3	1	5	2	3	2	0	9	115
	Total	189	44	28	23	51	33	16	9	40	433
5	1	1	2	28	12	86	63	40	17	20	269
	2	4	13	25	31	172	164	644	7	21	1081
	Same	0	0	0	0	3	43	3	0	9	58
	Undecided	0	0	2	0	1	0	2	0	1	6
	Total	5	15	55	43	262	270	689	24	51	1414
6	Female	101	12	16	37	425	35	76	15	111	828
	Male	37	7	9	23	121	7	36	16	37	293
	Same	3	2	1	1	7	39	2	0	8	63
	Can't	0	0	0	0	2	3	2	0	8	15
	Total	141	21	26	61	555	84	116	31	164	1199
7	0	25	55	24	43	59	9	27	2	48	292
	1	24	26	52	19	38	3	21	4	14	201
	Total	49	81	76	62	97	12	48	6	62	493

concluded a comparison between the classes could not be determined (32%) and then that class B was more variable (26%). Q7 was similar to Q3 except for the fact that the data were represented using dot plots. The two dot plots were designed not to have any identifiable extreme values. Approximately 39% of the students did not mention variability in their answers. In addition, approximately 41% of the students chose class A as more variable, and 16% of the students chose class B to be more variable.

Student responses to Q3 and Q7 collectively suggest that a remarkable proportion of students arrived at no conclusion in comparing the two distributions' variability or did not discuss variability at all. Besides, the reasoning mechanisms provided by students for variability also differed remarkably between these two very similar questions as Fig. 3 shows the distribution of coding categories for Q3 and Q7. According to the figure, the most common reasoning category in Q3 was related to the shape of the distribution (44%) irrespective of the choice students made (Table 4). Very few student justifications included the normative statistical definition of variability, the average distance to the center (a total of 28 codes out of 433). The three highest categories in Q7, on the other hand, were clustered or spread out (20%), frequencies (16%), and distance to center (15%) in decreasing order.

In Q5, students were asked to compare the variability of the two datasets that were given in the raw form. Seventy-five percent of the students chose the second dataset to be more variable, and almost half of the students employed the category consistent or diverse as Fig. 4 shows. Student responses included expressions such as more different data values, diverse, and consistent. This result strongly suggested that students tend to think variability in terms of the variety of data values especially when the datasets are small and represented in raw form.

In Q6, which had two dot plots to show the number of pairs of shoes owned by a large group of females and males, 66% of the students concluded that the distribution for females and 26% of students concluded that the distribution for males was more variable. As Fig. 5 shows, the clustered category was the most common in student responses to this homework question. Only 154 students mentioned extreme values and, when they did, the responses tended to be vague.

Overall, the category clustered or spread out was observed in differing weights in all of the homework questions and was the most common category in Q2, Q6, and Q7 (Table 4). Other categories were the highest or among the most preferred reasoning mechanisms in homework questions; even distribution or spread was the second most common category in Q2, shape was the most common in Q3, consistent or diverse was the most common in Q5, and frequencies was the second most common in Q7. In conclusion, the analysis of student responses suggested inconsistent and diverse reasoning mechanisms by students. The results collectively suggest a lack of focus and regularity by students when reasoning about variability.

Results According to the Research Questions

In the previous section, we presented students' reasoning about variability for each homework question. In this section, we will summarize the results according to each research question and explicate them with interview findings.

All of the homework questions asked students to compare variability between data sets with the same range, so the responses should all contribute to answering

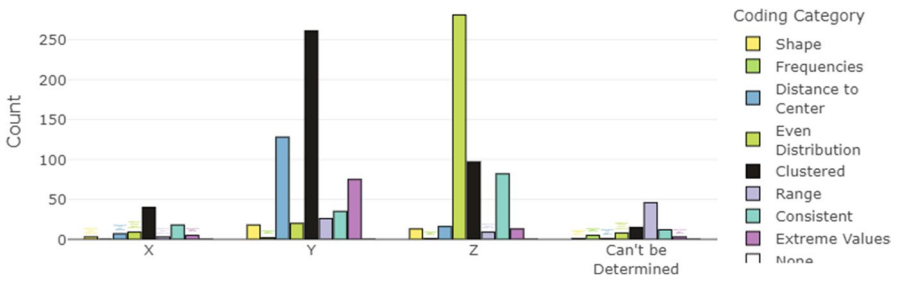


Fig. 2 The distribution of coding categories for Q2 options

research question a. How do undergraduate student's reason about variability when the datasets or distributions to be compared have the same range. For each question in which students were allowed to indicate the variability was the same, however, 5 to 6% of the respondents chose that option (Table 5). The most common identifiable justification was that the data sets or distributions had the same range. There were some students who when responding to the same question indicated that the variability was the same because there exist the same number of data values, the same sum of data values, or the same mean. These justifications did not occur with any regularity in the rest of the data set. As the relatively low percentages for *same* and *cannot be determined* options in Table 5 shows, most of the students were directed to other ways to compare the variability.

Overall, analysis of homework data suggested three informal approaches to variability. Primitive notions suggested by students were as follows: (1) variability as the extent to which consecutive data points are approximately equally distant from each other (Q2), (2) variability as the extent to which data values are different from each other (Q5), and (3) variability as the extent to which frequencies are different from each other (Q3 and Q7). Note that findings of these homework questions contribute also to our investigation of the other two research questions.

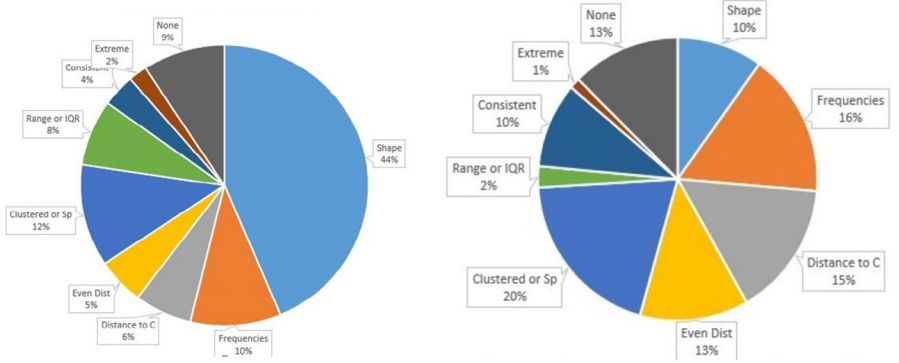


Fig. 3 The distribution of coding categories for Q3 and Q7

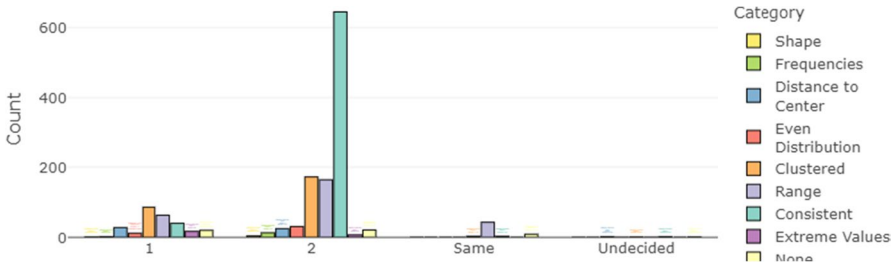


Fig. 4 The distribution of options and coding categories for Q5

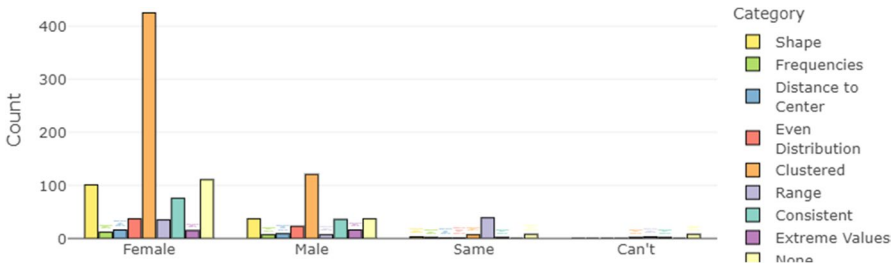


Fig. 5 The distribution of options and coding categories for Q6

The first two informal approaches to variability were remarkable because they are also closely related to the second research question. The students treated variability as the extent to which consecutive data points are different from each other, which we called consistent or diverse in coding student responses to homework questions. The use of consistent or diverse category was at 11%, 4%, 49%, 10%, and 10%, respectively, for questions 2, 3, 5, 6, and 7 (Table 4). The remarkable high percentage in Q5 allows us to conclude that the students who lack the normative meaning of variability tend to think variability in terms of the variety of data values especially when the datasets are small and represented in raw form.

The interview findings helped decipher the thinking mechanism for the above informal reasonings. When Chloe, Britney, and Karen observed that range did not provide distinguishing information about variability because the compared groups had the same range, they suggested the distribution with more distinct values should be more variable. Chloe's response to a task of the first interview illustrates this notion:

> Maybe range does not have in terms of variability because this range is a lot bigger, but if you got repeat so… you have more *varied responses* [emphasis added]. These are the very similar responses [refers to the first dataset], whereas this one has very similar responses—three people here. Yeah, sometimes range doesn't matter

The third informal notion, regarding variability as the extent to which frequencies are different from each other, was also evident in interviews. Chloe, for example,

decided that uniform distributions (see Q7 in Fig. 1) were *less* variable because the presence of different frequencies in dot plots or histograms meant larger variability. Accordingly, she confirmed that normal distributions would be more variable than uniform distributions, only because the frequencies were different from each other.

Questions 3 and 7 were designed not to have extreme values so the responses to these questions could be used to address research question b: how do undergraduate students reason about variability when the datasets or distributions to be compared have no extreme observations? The shape category was the most common in Q3, clustered or spread out (20%), frequencies (16%), and distance to center (15%) were the most common in Q7. These results for Q3 and Q7 indicate that most of the students were able to employ another method of reasoning to compare variability when extreme values were not available in datasets. The justifications using extreme values were only 2% and 1% of the responses, at Q3 and Q7, respectively. However, responses indicated a misconception commonly reported in the literature: If the "frequencies vary," then, the distribution should be more variable. According to the students, Class A was more variable because "There is a larger frequency range for Class A than Class B, so … concluded that Class A is more variable than Class B."

As discussed in "Methodology," students might have considered values in the data sets for questions 2, 5, and 6 to be extreme so we hypothesized that responses to those questions might include extreme values in the justifications. The use of extreme values was at 7%, 2%, and 3% of responses, respectively, in questions 2, 5, and 6 (Table 4). In brief, students rarely used extreme values in their reasoning about variability.

The analysis of interview data confirmed that extreme values did not have a primary role in the participants' reasoning. Britney claimed that without extreme values, "[The distribution] would be less variable because it would have a less range that it falls in between because you would shorten the range…" When asked why the presence of extreme values results in a more variable distribution, Chloe said, "Outliers increase *range of values* [emphasis added], and … in my own words, *hitting more numbers* [emphasis added]" (Fig. 6). As clear from the figure, with extreme values, an increase in range follows, so the probability of more different numbers in a distribution increases. Note that the first three themes in the figure are also listed as primitive notions in this study.

Table 5 Students' use of *same* and *cannot be determined* options

Question	Response		Justification	
	Same	Cannot be determined	Same range	None
2	NA	126 (12%)	46 (37%)	55 (44%)
3	15 (5%)	NA	9 (60%)	-
5	59 (6%)	NA	43 (72%)	-
6	60 (6%)	15 (2%)	43 (57%)	3 (4%)

NA stands for not applicable. The percentages under justification were calculated from the totals in the response column, not from the total number of responses

The second research question we attempted to answer was how undergraduate students reason about variability when the datasets or distributions to be compared have no extreme observations. Overall, interviews suggested that students often preserved their preexisting and more dominant notions such as shape, variety of values, clustered/spread out, and frequencies. The limited use of the *extreme values* terminology supports the conclusion that extreme values were peripheral in undergraduate students' reasoning about variability.

Only question 5 has a different number of values in the data sets so we assumed responses to the other homework questions could be used to address research question c. The analysis of student responses to the homework questions except Q5 suggested that the category consistent or diverse was less applicable in these questions. According to Table 4, the use of the category consistent or diverse was at 11%, 4%, 10%, and 10%, respectively, in questions 2, 3, 6, and 7. As the reasoning mechanisms in these homework questions were extensively discussed, we now report the findings of the interview tasks to better understand how students employed this reasoning mechanism.

The treatment of variability as to *how different the values are from each other* was first observed when each interview participant described variability at the beginning of the interviews. Ocean realized that describing variability based on the notion of difference was inconclusive in her investigation of variability in various tasks throughout the interviews. She reorganized her approach when she started to work on the tasks in which the datasets to be compared in terms of variability were given in the form of raw data (homework question 5). The following excerpt illustrates the reasoning mechanism she employed for variability:

> Initially, I was thinking of variability as…just like…the different data points [emphasis added] so… something is more variable if different data points were recorded with all different but here you can see that they were, like, three of them were all the same but still constituted for the higher variability…it does not always have to be. Each individual data point does not have to be different; it just depends on how far each varies from the mean [emphasis added].

The other three interview participants maintained the same primitive notion—possibly even more strongly as we explained when reporting the findings for the first two research questions. Variability meant merely different or change for the participants, and the observation of various numbers in datasets seemed to bolster interview participants' treatment of variability in this way. When the notion could not be employed because of the design of the question or other features of the tasks that overshadowed this property, the participants employed fragmented and inconsistent strategies. The following excerpt illustrates how Karen dealt with the situation:

> Well, both have repeating values in them. This one [the first dataset] has forty, forty-two. This one [the second dataset] has seventy and thirteen repeating itself …yeah, I think they have the same variability because they are not drastically different and they both have similar characteristics.

For the same situation, Britney started to use the notion of clustering of values to justify her decision. The following excerpt illustrates how she started to establish her

approach of "clustering" and whether or not the presence of values in a distribution "grouped together" implied less variability:

> Maybe, I would say the second dataset because, I guess, the eighteen is more like the median. The repeating values are more outside. Seventy is further away from the eighteen and, oh well, ten to forty is kind of a big jump. These values are kind of less likely in relation to each other…I mean more spread apart. Like it goes, if it was in a dot diagram [might mean dot plot], all these numbers will be like on this end it kind of jumps like from eighteen to seventy either or side and this one lay down here [inaudible]. Well, actually maybe the first dataset would be more variable since it has more spread out. It has some datasets from the end, forty kind of in the middle, and seventy-seven up here rather than the second one just having more like clustered on the lower end and the higher end.

In this section, we first reported student responses to each homework question. Next, we summarized these results and interview findings to answer the research questions.

Discussion

In this study, we examined student responses to homework questions and interview tasks to frame students' particular ways of reasoning about variability in three restricted situations reflecting known primitive notions of variability. We will elaborate on the findings of the study by discussing how our students performed in terms of *distance to the center* notion, present the limitations, and suggest implications for instruction and research.

The more comprehensive conclusion of the study could be outlined in three points. The findings collectively suggest that (i) students' reasoning about variability is often fragmented, incomplete, or even contradictory, (ii) students tend to take variability in terms of its colloquial meaning, *change*, and (iii) characteristics of the distribution and the choice of "external representations" (Langrall et al., 2017, p. 509) disproportionally influence students' reasoning.

Results showed that some of the students could not discuss variability further. The rest of the students who proceeded with the discussion suggested different intuitive ways to compare variability. The informal ways we extracted from homework data and explained further with interviews were (i) interpreting variability as the extent to which consecutive data points were approximately equally distant from each other, (ii) variability as the extent to which data points are different

Fig. 6 Chloe's link between extreme values and variability

from each other, and (iii) interpreting variability as the extent to which frequencies of bins (in histograms) were different.

The findings that emerge from the present study suggest students at the undergraduate level maintain thinking about variability as "overall spread and differences in data values (e.g. not all values are the same)" (Garfield et al., 2007, p. 142). Ways of thinking, such as (a) smaller range values mean less variability because the values will be similar and (b) when values are close together they are more similar hence less variable, were commonly observed. English and Watson (2015) and Garfield and Ben-Zvi (2005) suggested students' recognition that *observations vary from one to another* is an essential understanding of variability, especially in early school years. The expectation for students, however, is to make the transition from the colloquial meaning of variability—things that change—to the normative understandings over time (Ciancetta, 2007; Kaplan et al., 2010; Watson & Kelly, 2008). In our case, this progress was missing among many students, and the primitive conceptions that are reported for K-12 students are still held by the undergraduate students.

Regarding variability as a measure for *how different the values are from each other* suggests that students regard variability in a way that is more applicable to categorical data because, for categorical data, measuring variability is based on how frequently "the observations *differ from one another* [emphasis added]" (Kader & Perry, 2007, p. 2). Although this understanding is necessary (English & Watson, 2015; Garfield & Ben-Zvi, 2005), students should start to focus on how far observations are, on average, from the center for univariate quantitative data. In the next figure, we illustrate the distinct levels of understandings that we extracted from our students' ways of reasoning about variability.

Figure 7 illustrates students' approaches from vague to more deliberate consideration of the measure of center, with examples for each theme. The distinct themes in the figure share some commonalities with other researchers' categorizations (see, e.g. Makar, 2014; Makar & Confrey, 2005; Reid & Reading, 2008). Many of our students' consideration of variation could be described as "weak" according to Reid and Reading (2008) because they (i) "incorrectly describe variation," (ii) "poorly express description of variation," and (iii) "refer to irrelevant factors to explain variation" (p. 51). The themes also align with delMas and Liu's (2005) learning trajectory for standard deviation: "…understandings of the standard deviation that did not consider variation about the mean to more mean-centered conceptualizations that coordinated the effects of frequency (density) and deviation from the mean" (p. 55). In brief, our findings reveal some of the distinct phases that students could present while developing the normative meaning of variability. Both instruction and future research studies could benefit from these steps in creating learning trajectories for variability.

Limitations and Implications of the Study

Two important limitations of the study are the lack of detail in students' consideration of variability in their responses to homework questions and the narrow focus of the study. Researchers who want to collect large-scale data should be careful in interpreting students' reasoning in answering constructed-response questions that

are not subject to grading. We employed homework questions that were limited in terms of representation, context, and prevalent characteristics of the given distributions. Narrowing the focus might restrict the generalizability of findings to broader situations. In addition, designing these homework questions to elicit students' specific ways of reasoning might have triggered some of the informal notions more than some others that we could not observe in this study. As Biehler et al. (2018) and Çatman Aksoy and Işıksal Bostan (2021) suggest, enriching the quality, the variety, and the context of the task(s) in instruction and research settings could substantially support students' ability to reason about variability. Future research should investigate students' aforementioned primitive notions in distributions using more observations and a more diverse set of representations such as boxplots, stem and leaf plots, density curves, or for a more diverse set of shapes of distributions.

The study showed that undergraduate students' reasoning about variability is centered around some primitive notions such as change, consistency, predictability, and "how different the values are from each other." These notions, however, are more appropriate when addressing variability for categorical data. Students should be well aware of the difference between categorical and univariate numerical data, and then investigate variability's meaning for these types of data. Our study showed that people are more inclined to think variability in a way that is more appropriate for categorical variables. Therefore, instructors should provide more experience with data recognition, students' intuitions for statistical concepts, and their effect on statistical learning and investigations.

Although students may hold several primitive statistical notions, these notions tend to activate in appropriate circumstances (diSessa, 1993) such as when a particular characteristic is highlighted in a given distribution. Similarly, Langrall et al. (2017) suggest external representations as one of the three overarching areas of research in supporting student learning in statistics. External representations behave as thinking tools for students, and prompting students with multiple representations such as data in raw form, dot plots, or histograms may help students conceptualize variability more thoroughly (Gougis et al., 2017). For example, when students are given unordered data with some numerical values more than once, they found it convenient to interpret variability as difference and compared variability across

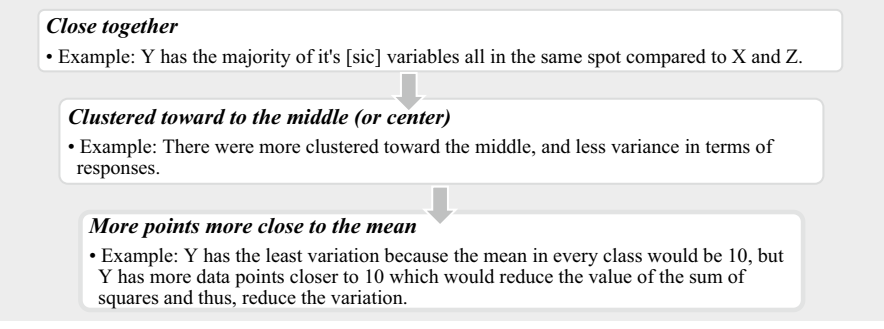


Fig. 7 Student explanations from less clear to clearer consideration of center

distributions according to this notion. The students, however, might regard range to be more convenient if the observations in datasets to be compared are ranked in order, as ordering data values foreground the difference between ranges.

Focusing on students' inconsistent reasoning mechanisms, which are generally called "cognitive conflicts" by statistics educators (see, e.g. Reading & Reid, 2007; Watson & Kelly, 2007), is also suggested to be effective in teaching statistics (Biehler et al., 2018, p. 150). When and how do more standard meanings and measures of variability become adoptable by students is still a question to investigate. In our case, we had students at different levels in terms of their attainments to the formal meaning of variability. Future research and teaching should reveal more on how students could coordinate their primitive notions and standard statistical meanings through different external representations and tasks (Biehler et al., 2018, p. 149; Kaplan et al., 2018; Langrall et al., 2017). To summarize, there is a need to illustrate how students could develop a more robust understanding of statistical concepts based on their primitive reasoning mechanisms.

Acknowledgements The work described in the article was based on the first author's doctoral dissertation entitled "Undergraduate Students' Informal Notions of Variability" under the guidance of the second author at the University of Georgia, Athens, GA.

Declarations

Conflict of Interest The authors declare no competing interests.

References

American Statistical Association (ASA). (2016). Guidelines for assessment and instruction in statistics education: College report 2016. Alexandria, VA: ASA.

Arnold, P. (2013). *Statistical investigative questions: An enquiry into posing and answering investigative questions from existing data* [Doctoral dissertation]. The University of Auckland.

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) report II*. American Statistical Association (ASA) and National Council of Teachers of Mathematics (NCTM).

Biehler, R., Frischemeier, D., Reading, C., & Shaughnessy, J. M. (2018). Reasoning about data. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 139–192). Springer. https://doi.org/10.1007/978-3-319-66195-7_5

Bock, D. E., Velleman, P. F., & De Veaux, R. D. (2012). *Stats: Modeling the world – AP edition* (4th ed.). Pearson.

Çatman Aksoy, E. & Işıksal Bostan, M. (2021). Seventh graders' statistical literacy: An investigation on bar and line graphs. *International Journal of Science and Mathematics* Education, *19*(2), 397–418. https://doi.org/10.1007/s10763-020-10052-2

Ciancetta, M. A. (2007). *Statistics students' reasoning when comparing distributions of data* (Publication No. 3294660) [Doctoral dissertation, Portland State University]. ProQuest Dissertations Publishing.

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55–82.

diSessa, A. (1993). Toward an epistemology of physics. *Cognition & Instruction, 10*(2–3), 105–225. https://doi.org/10.1080/07370008.1985.9649008

English, L. D., & Watson, J. M. (2015). Statistical literacy in the elementary school: Opportunities for problem posing. In F. M. Singer, N. F. Ellerton, & J. Cai (Eds.), *Mathematical problem posing:*

From research to effective practice (pp. 241–256). Springer. https://doi.org/10.1007/978-1-4614-6258-3_11

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4*(1), 92–99.

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 117–147). Erlbaum.

Gougis, R. D., Stomberg, J. F., O'Hare, A. T., O'Reilly, C. M., Bader, N. E., Meixner, T., & Carey, C. C. (2017). Post-secondary science students' explanations of randomness and variation and implications for science learning. *International Journal of Science and Mathematics Education, 15*(6), 1039–1056. https://doi.org/10.1007/s10763-016-9737-7

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

Kader, G., & Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education, 15*(2), 1–16. https://doi.org/10.1080/10691898.2007.11889465

Kaplan, J. 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

Kaplan, J. J., Lyford, A., & Jennings, J. K. (2018). Effects of question stem on student descriptions of histograms. *Statistics Education Research Journal*, *17*(1), 85–102.

Langrall, C. W., Makar, K., Nilsson, P., & Shaughnessy, J. M. (2017). Teaching and learning probability and statistics: An integrated perspective. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 490–525). National Council of Teachers of Mathematics.

Lann, A., & Falk, R. (2003). What are the clues for intuitive assessment of variability? In C. Lee (Ed.), *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy* (pp. 1–24). Central Michigan University.

Loosen, F., Lioen, M., & Lacante, M. (1985). The standard deviation: Some drawbacks of an intuitive approach. *Teaching Statistics, 7*(1), 2–5.

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. (2016). Developing young children's emergent inferential practices in statistics. *Mathematical Thinking and Learning, 18*(1), 1–24. https://doi.org/10.1080/10986065.2016.1107820

Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal, 4*(1), 27–54.

National Governors Association Center for Best Practices [NGA Center] & Council of Chief State School Officers [CCSSO]. (2010). Common core state standards for mathematics. Authors.

Pingel, L. A. (1993). Variability – Does the standard deviation always measure it adequately? *Teaching Statistics, 15*(3), 70–71. https://doi.org/10.1111/j.1467-9639.1993.tb00659.x

Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *Journal of Mathematical Behavior, 22*(4), 405–435. https://doi.org/10.1016/j.jmathb.2003.09.002

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5*(2), 46–68.

Reading, C., & Reid, J. (2007). Reasoning about variation: Student voice. *International Electronic Journal of Mathematics Education, 2*(3), 110–124.

Reid, J., & Reading, C. (2008). Measuring the development of students' consideration of variation. *Statistics Education Research Journal, 7*(1), 40–59.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Information Age.

Watson, J. M. (2007). The role of cognitive conflict in developing students' understanding of average. *Educational Studies in Mathematics, 65*(1), 21–47. https://doi.org/10.1007/s10649-006-9043-3

Watson, J. M., & Kelly, B. (2007). Assessment of students' understanding of variation. *Teaching Statistics, 29*(3), 80–88. https://doi.org/10.1111/j.1467-9639.2007.00295.x

Watson, J. M., & Kelly, B. (2008). Sample, random and variation: The vocabulary of statistical literacy. *International Journal of Science and Mathematics Education, 6*(4), 741–767. https://doi.org/10.1007/s10763-007-9083-x