EFFECTS OF QUESTION STEM ON STUDENT DESCRIPTIONS OF HISTOGRAMS

Jennifer J. Kaplan
University of Georgia
jkaplan@uga.edu

Alexander Lyford
Middlebury College
alyford@middlebury.edu

Jeremy K. Jennings
University of Georgia
jkjennings@gmail.com

ABSTRACT

Assessment is necessary, but difficult, in statistics education. Multiple choice assessments are common, particularly for research purposes. Open-ended assessments may be more adept at revealing student understanding,but ensuring their validity can be difficult. The study presented here examines differences in student descriptions of histograms for different question prompts and scenarios in order to understand how best to ask such questions in research and teaching situations. The results show that different ways of asking students to describe histograms and different scenarios and shapes of histograms lead to systematic differences in student descriptions of histograms. The paper concludes with suggestions for phrasing questions, both for research and classroom assessment, and provides directions for future research based on our results.

Keywords: *Statistics education research; Undergraduate introductory statistics; Assessment*

1. INTRODUCTION

Two of the central recommendations of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report are to stress conceptual understanding, rather than mere knowledge of procedures, and to use assessments to improve and evaluate student learning (American Statistical Association, 2005; GAISE College Report ASA Revision Committee, 2016). In order to help students develop conceptual understanding it may be necessary to combine these recommendations. We can use meaningful assessments that reveal student thinking to provide formative assessment and show students through the assessment process that we as instructors value conceptual understanding (Pellegrino, Chudowsky, & Glaser, 2001). Assessing conceptual understanding, however, is a "challenge faced by all educators in statistics education" (Gal & Garfield, 1997, p. 7). Many STEM disciplines, including statistics, have developed concept inventories, which are typically multiple-choice assessments in which the distracters were derived from common student misconceptions, to assess student conceptual understanding (D'Avanzo, 2008; Knight, 2010; Libarkin, 2008). Multiple-choice concept inventories are efficient to administer to large groups of students. They may not, however, reveal students' understanding as well as constructed-response questions, also known as open-response or short answer questions, in which students must write an answer in their own words (Bennett & Ward, 1993; Birenbaum & Tatsouka, 1987; Bridgeman, 1992; Kuechler & Simkin, 2010). In addition, methods for ensuring the reliability and

validity of multiple choice assessments are better documented in the literature (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) than they are for the construction of open-ended assessment items.

In this paper we report the creation of constructed response items to assess student descriptions of histograms. Section 2 contains a description of the importance of histograms to student learning in statistics, a framework for assessing the descriptions written by students, and a brief description of the issues involved in the reliability and validity of constructed response items. We then describe a randomized study to explore differences in student descriptions of histograms under different conditions including shape, scenario, and prompt. The results presented not only reveal differences in student writing about histograms across the conditions, but also general results about student thinking about histograms. In particular, as is predicted by the literature, students are less likely to attend to the feature of variability in a distribution than to any of the other features (Ben-Zvi, 2004).

2. BACKGROUND

2.1. THE IMPORTANCE OF HISTOGRAMS

Histograms are commonly used graphical displays of the distribution of a quantitative variable (Arnold & Pfannkuch, 2014). Histograms are adept at revealing the distribution of data values, especially the shape of the distribution and any outlier values. Previous research, however, indicates that undergraduate students tend to hold various misconceptions about histograms (delMas, Garfield & Ooms, 2005; Kaplan, Gabrosek, Curtiss, & Malone 2014; Meletiou-Mavrotheris & Lee, 2010). For example, students may view all graphs with bars as bar graphs or case value graphs (see Kaplan et al., 2014 for discussion), not distinguishing the difference between these graphs and histograms (delMas et al., 2005; Meletiou-Mavrotheris & Lee, 2010). Confusing histograms and case value graphs, thinking that each bar represents one observation instead of the aggregate of all observations that take certain values, can lead to confusion between what is represented by the horizontal and vertical axes in a histogram (Bright & Friel, 1998; Friel, Curcio, & Bright, 2001). These misconceptions may inhibit students' abilities to read and describe histograms as a statistician would. Difficulties that students have in reading and reasoning about histograms are particularly problematic for two reasons: 1. being able to read histograms is an aspect of being a statistically literate citizen (American Statistical Association [ASA], 2005; Gal, 2002; Rumsey, 2002) and 2. being able to read histograms is a foundation for understanding inferential procedures (delMas et al., 2005; Noll & Hancock, 2015). Histograms are considered to be common graphical displays based on the prominence of their use in introductory statistics texts (see for example, Agresti & Franklin, 2012; Moore, 2009; Utts & Heckard, 2012), in the popular press (Manjoo, 2013; Quealy, Cox, & Katz, 2015), and, increasingly, in K–12 education (ASA, 2007; Common Core State Standards Initiative [CCSSI], 2010; National Council for Teachers of Mathematics [NCTM], 2000). The understanding of the connection between histograms and density functions or distributions is necessary to understand sampling distributions, which are, in turn, the foundation to understanding of statistical inference (Madden, 2011; Meletiou-Mavrotheris & Lee, 2002a). In formal inferential procedures, we compare the observation we have to what we would expect to happen in the long run in order to make a decision about a hypothesis. Without an understanding of distribution, which begins with an understanding of histograms, a student will have difficulty assessing the likelihood of the observation and making a judgement about the hypothesis (Noll & Hancock, 2015).

2.2. A FRAMEWORK FOR STUDENT DESCRIPTIONS OF HISTOGRAMS

A complete description of a distribution of a quantitative variable displayed in a histogram is one that addresses shape, center, and variability in the context of the data (delMas et al., 2005; Noll & Hancock, 2015). Watson (2005) writes, "A larger objective in terms of the goal of statistical literacy when students leave school is to be able to tell a story from a context with a graph that displays variation, clustering,

middles and surprises" (p. 189). There is, however, limited literature that explores how students develop language for describing distributions or what "describing data distributions encapsulates" (Arnold & Pfannkuch, 2014, p. 1) in the minds of students.

Arnold (2013) provides a comprehensive framework for analyzing student descriptions of distributions. The distributional description framework (DDF, Figure 1) combines and supplements two previously existing distribution-related frameworks (Bakker & Gravemeijer, 2004; Ben-Zvi, Gil, & Apel, 2007) and was developed through actual student descriptions of distributions (Arnold, 2013). The DDF contains twenty-eight specific features that are measures, depictions, or descriptors of distributions. These specific features are organized into five overarching statistical concepts based on the Informal Inferential Reasoning (IIR) theoretical framework of Ben-Zvi et al. (2007): contextual knowledge, distributional, graph comprehension, variability, and signal and noise. The second level of organization for the DDF is grounded in Bakker and Gravemeijer's (2004) characteristics of distribution combined with the ideal data-dialogue framework of Pfannkuch, Regan, Wild, and Horton (2010).

Overarching statistical concepts	Characteristics of distribution	Specific features measures/ depictions/descriptors
Contextual knowledge	Population	1. Target or other acceptable population
	Variable	2. Variable, 3. Units, 4. Values
	Interpretation	5. Statistical feature described in contextual setting
	Explanation	6. Possible reason for a feature
Distributional	Aggregate view	7. General shape sketched (correctly) 8. Hypothesis and prediction
	Symmetry	9. Overall shape
	Modality	10. Modality
	Skewness	11. Position of the majority of the data
	Individual Cases	12. Highest and lowest values
Graph comprehension	Decoding visual shape	9. Overall shape 13. Parts of the whole
		10. Modality
	Unusual features	14. Gaps, 15. Outliers
Variability	Spread	16. Range, 17. Interquartile range 18. Range as an interval 19. Interval for high and/or low values 20. Interval for groups
	Density	21. Clustering density 22. Majority (mostly, many) 23. Relative frequency
Signal and noise	Centre	24. Median, 25. Mean
	Modal clumps	26. Peak(s), 27. Local mode, 28. Modal group(s)

Figure 1. Distribution framework for curriculum level 5 (ages 13–15). Reprinted from "Describing distributions" by P. Arnold and M. Pfannkuch, *Proceedings of the Ninth International Conference on Teaching Statistics*, p. 2, 2014, International Statistical Institute.
Reprinted with permission from the authors.

The DDF suggests that when discussing the context of a distribution, students should attend to the population from which the data were collected as well as the variable measured and the units and possible values of the variable: characteristics 1, 2, and 3 in Figure 1. Sample responses A and B, below, illustrate the use of the DDF and its numbering system when reading student writing about distributions. In the

sample response from Student A, the population is New Zealand (1), the variable measured is household debt (2) and the units are given as NZ$ (3). Student B gives the population as Tokoeka kiwis (1), with the variable, heights (2), and the units in cm (3). With respect to shape, Arnold (2014) notes that students attend to two aspects of shape that are combined to provide a good description of the shape of a distribution. Students first tend to notice symmetry or lack thereof and next notice the number of peaks, or modality, of a distribution. Within the DDF (Figure 1) these ideas are noted as 9. overall shape and 10. modality. In addition, the lack of symmetry, or skewness, can be described by students as the location of the majority of the data (characteristic 11). Student A combines right skewed (overall shape) and unimodal (modality) to provide a good description of shape. Similarly, Student B combines approx. symmetrical (overall shape) and bimodal (modality). Arnold notes that the strongest descriptions of distributions are those that connect shape and context, including possible reasons for the features seen in the data. For example, when students are presented with a right skewed distribution of reaction times to a particular event, a strong description would include a mention that the skewness is due to the impossibility of negative reaction times. A second example is descriptions of bimodal distributions that indicate a mixture of two distinct populations in the sample as seen in the last statement provided by Student B (Arnold, 2013; Arnold & Pfannkuch, 2014).

> Sample Response: Student A
> The distribution of the NZ (1) household debt (2) is right skewed (9) and unimodal (10). The debts peak (26) at approximately $10,000 (3). The debts range from $0–$200,000 (18). Between $0–$80,000, the debts are approximately symmetrical, where they are tightly grouped (9, 13, 21). There is a short tail from $100,000 to $200,000 (19). The middle household debt (24) is approximately $50,000 (Arnold & Pfannkuch, 2014, p. 4).

> Sample Response: Student B
> The distribution of the heights (2) of Tokoeka kiwis (1) is approx symmetrical (9) and bimodal (10). The heights range from 35-43 cm (3, 18). The middle Tokoeka kiwi height is 39 cm (24). The heights peak at around 36.5 and 40 cm (26, 27). The heights are tightly grouped in two groups (13, 21) one between 36.1–39 cm (20) and another between 39–42 cm (20). These two groups might mean the two different genders (5) (Arnold, 2013, p. 225).

The characteristics enumerated in the DDF associated with centers of a distribution are majority (22), median (24), mean (25), peak (26), local mode (27) and modal group (28). Students A and B both quantify the median by giving a value of the middle value of the variable. In addition, Student B mentions two peaks, which may be considered local modes, at 36.5 cm and 40 cm. The characteristics associated with variability are highest and lowest values (12), gaps (14), outliers (15), range (16), interquartile range (17), and intervals for the range of the data set (18) or some subset of the data set (19, 20), and clustering density (21). Neither of the student example responses contains specific measures of variability, such as range (16) or interquartile range (17), but both contain examples of descriptions of intervals. Both students describe the interval of values of the entire data set (18). In addition, Student A describes the interval that contains the upper tail of the data (19) and Student B describes the interval for each of the two subsets (20). Finally, both students mention clustering density (21) using the phrase “tightly grouped.”

2.3. CONSTRUCTED RESPONSE ITEMS

Constructed-response scores are not only better than multiple choice items at revealing student understanding (Bennett & Ward, 1993; Birenbaum & Tatsouka, 1987; Bridgeman, 1992; Kuechler & Simkin, 2010), they can also have a greater correspondence with clinical interview scores than multiple-choice test scores (Nehm & Schonfeld, 2008), and the correspondence between clinical interviews and constructed-response scores persists when the constructed responses are scored by computers (Beggrow,

Ha, Nehm, Pearl, & Boone, 2013). They are not, however, exempt from the need to exhibit a high degree of validity and reliability as defined by educational measurement standards (see for example, American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). Validity is a measure of the appropriateness of an assessment for the intended purposes of the assessment (Miller, Linn, & Gronlund, 2013). At the item level, the validity of a constructed response question is measured by the degree to which the student responses elicited by the question match the intentions of the question writer (Pollitt, Ahmed, Baird, Tognolini, & Davidson, 2008). Reliability is a measure of the degree to which an assessment tool produces stable and consistent results. For example, student results should remain relatively constant under multiple administrations, different versions of an assessment tool should produce similar student results, and different questions designed to assess the same concept should function similarly with respect to student responses (Crocker & Algina, 1986). There are many factors that can affect item validity or reliability including the format, features, context, or semantic structures of the item (Chi, Feltovich, & Glaser, 1981; diSessa, Gillespie, & Esterly, 2004; Gentner & Toupin, 1986; Greeno, 2009; Sabella & Redish, 2007; Silver, 1979). For example, a question based on a detailed context may cue students to answer using common knowledge about the context rather than the disciplinary knowledge the question was designed to elicit (Sweiry, 2013). This can have an impact on the validity of the question because the question is actually measuring a student's knowledge of the context instead of the content knowledge associated with the question. Context can also have an impact on the reliability of an instrument. When multiple questions addressing the same content are written using different contexts, students may not provide the same quality of answers because of an interaction with their knowledge of the context. Another factor that may impact the reliability and validity of a constructed response question is the question stem itself, which may provide cues to students about what constitutes a complete response (Crisp, Sweiry, Ahmed, & Pollitt, 2008). These cues impact the validity of an item because student responses may not reflect the actual knowledge of the content being assessed. Instead, the responses may be a parroting back of the stem of the question. In fact, studies in the sciences have shown that even small alterations in context and question wording influence the content of student responses to constructed-response questions (Nehm, & Ha, 2011; Schurmeier, Atwood, Shepler, & Lautenschlager, 2010; Weston, Haudek, Prevost, Merrill, & Urban-Lurain, 2015).

For this study, we were interested in creating constructed-response questions in which students would write descriptions of data distributions shown in histograms. To investigate the conditions under which the items would have a high level of validity and reliability, several versions of the items were written and distributed at random to students. The items were of two different contexts and had three different wordings of the question stems. The research questions associated with the research study are:

With respect to the shape, center, and variability within the context of a given histogram,

1. To what extent do student descriptions of histograms differ when the question stems are modified?
2. Are the results consistent across two separate scenarios?

In answering these research questions, we intend to provide guidance to researchers in creating more items designed to assess students' abilities to give a complete description of the distribution of data, one that includes mention of shape, center, and variability in the context of the data, when the data are provided in the form of a histogram.

3. METHODOLOGY

In this section, we first discuss the academic settings from which the student samples were drawn and the demographics relating to the samples. We then describe the data collection procedures used. The section ends with a description of the analysis, which includes the rubric for coding student descriptions of histograms.

3.1. SETTING AND SAMPLES

The data were collected during spring and fall semesters in 2014 from students in an introductory statistics course offered by the Department of Statistics at a large research university in the Southeastern United States. The course meets the quantitative reasoning general education requirement for undergraduate students at the university. Topics covered include 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, and two-way tables and the chi square test of independence. Each fall and spring semester approximately 1,300 students take the course, separated into seven lecture classes taught by three or four different lecturers. The students also attend a weekly computer lab led by one of approximately ten teaching assistants who are graduate students in the Department of Statistics. The approved human subjects protocol included a waiver of documentation of consent, so all responses to the research items were included in the data set.

The course from which the data were collected was highly coordinated. The lecturers were given a schedule for the coverage of statistics content and most used a version of the same presentation slides and the same example problems. All students had access to the lecture notes of all of the lecturers and completed the same lab activities and homework assignments. Graphical representations of quantitative variables were covered before summary measures of center and variability. Students were first instructed to use dot plots and histograms to describe the shape of the distribution. Instruction was provided on symmetric, skewed, and bimodal distributions. Instruction on shape was followed by instruction on the use of the mean and standard deviation or median and IQR to describe the measures of center and variability of a distribution. Students were instructed on the use of point and click software to create graphical representations and calculate summary measures. Students were asked to describe the shape of histograms and estimate summary measures or relative frequencies when given data in the form of histograms on homework and lab assessments. In addition, students were instructed and assessed on the relative positions of the mean and median in symmetric and skewed data. While the common textbook for the course (Agresti & Franklin, 2012) clearly stated that the key features to describe a distribution were shape, center, and variability, the assessment items used in this study were the only opportunity for students to provide a description of data given in the form of a histogram.

3.2. DATA COLLECTION

All students in a given semester were shown the same histogram, but were randomly assigned to one of three different prompts, shown as A, B, and C, below. The right-skewed, annual income scenario (Figure 2a: Atlanta Income) was used in spring 2014. The unimodal and symmetric, Student Sleep scenario (Figure 2b: Student Sleep) was used in fall 2014. The horizontal axis was labeled but captions were not provided to the students. Prompt A contains only the instruction to describe the distribution, without mention of the variable. This prompt is referred to as the *Distribution Only* prompt. Prompt B does not include the word *distribution* and refers only to the variable name and is called the *Variable Only* prompt. Prompt C includes both the term *distribution* and the variable name so is referred to as the *Both* prompt. The final phrase in prompts B and C was changed to reflect the different contexts of the graph in each scenario (Atlanta Income or Student Sleep)

A. Describe as completely as possible the distribution shown in the histogram. (DISTRIBUTION ONLY)
B. Describe as completely as possible what the graph tells you about the (yearly income of adults in Atlanta)/(number of hours high school students in Georgia sleep on school nights). (VARIABLE ONLY)
C. Describe as completely as possible the distribution shown in the histogram, being sure to explain what the graph tells you about the (yearly income of adults in Atlanta)/(number of hours high school students in Georgia sleep on school nights). (BOTH)

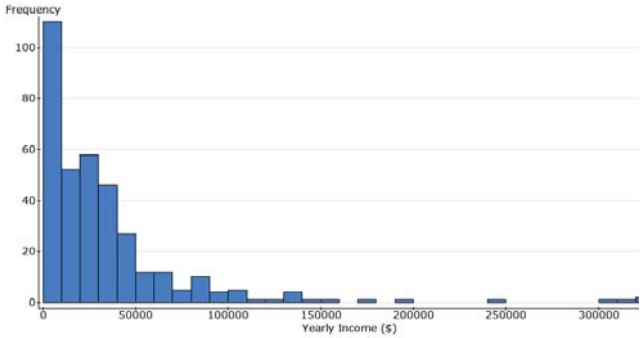

Figure 2a. Histogram of Atlanta income data

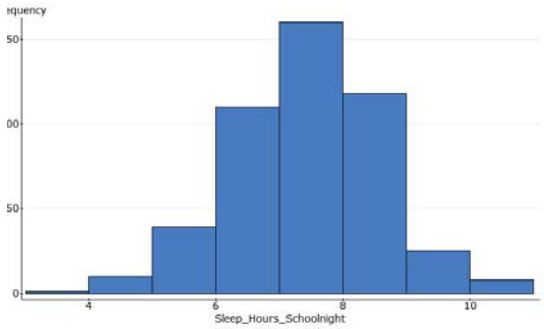

Figure 2b. Histogram of student sleep data

Student responses were collected as part of a homework assignment administered through WebAssign (http://www.webassign.net/), an online homework submission application. The homework assignments were made available for students at the end of the second week of class after learning about histograms in lectures. At the time of the administration of the homework, a relatively equal number of examples of both symmetric and skewed distributions had been used in the course. In both semesters, students had two weeks to complete the assignments. Although students could discuss homework questions with one another, a string checker applied to the data determined that no two student responses were identical. In addition, during the coding phase, any time an unusual phrase occurred, for example something that read as though it had been copied from a dictionary,all responses were searched for that phrase. Through this process, no examples of copied text were found and there was no evidence of strong, unnatural similarities between any two student responses in terms of word usage and sentence structure.

In total, there were 1155 responses given in spring 2014 and 1178 in fall 2014, distributed across the prompts as specified in Table 1. Electronic files of the responses were obtained from the course coordinator. The analysis was done using the electronic files.

Table 1. Number of responses by semester, scenario, and prompt

		Prompt			
Semester	Scenario	Distribution Only	Variable Only	Both	Total
Spring 2014	Income	331	425	399	1155
Fall 2014	Sleep	369	407	402	1178
	Total	700	832	801	2333

3.3. ANALYSIS

The analysis of differences between student responses to the prompts was based on the differences in the proportion of responses in which the four categories—shape, center, variability, and context—were mentioned. The decision to code specifically for these four categories was made to align with the DDF Framework described in Section 2. The rubric used to code the responses can be seen in Table 2. To be coded as addressing shape, a response had to include a description of at least one component of shape specified by the DDF Framework: overall shape or modality. For the Student Sleep data, the students tended to use the phrases *symmetric*, *bell-shaped*, or *normal*, although *unimodal* was sufficient to be coded as containing reference to shape. Similarly, responses to Atlanta Income tended to include the phrase *right skew* without mention of modality, but *unimodal* would have been sufficient to be coded as containing reference to shape.

To be coded as addressing center, responses to both prompts needed to include a correct value for a measure of center. Aligned with the DDF Framework, this could include a discussion of the location of the mean, median, mode, or the modal clump. For our purposes, the modal clump was defined to be the

section of the graph containing a sufficiently large proportion of the data. For the Student Sleep data, this was roughly students getting between 7 to 9 hours of sleep, and for the Atlanta Income data this equated to adults making between $0 and $50,000 annually. To be coded as addressing variability, the DDF Framework necessitates a discussion of the range of the data or a subset of the data as either a number or an interval. We, however, expanded this definition to include a discussion of both maximum and minimum values, meaning that a response describing both the lowest and highest results and their relative distance apart, along with other potential outlying data values in each dataset, constituted a response being coded as discussing variability. A student response that exemplified this coding was "Students tended to sleep lots of different hours, there are some as low as 2 hours and others as many as 12 hours, so there were lots of different hours they slept." Whereas this response does not address the range directly, the student clearly recognizes the inherent variability in the data as evidenced by several outlying values. Thus, the response was coded positively for variability.

To be coded as addressing context, responses had to contain at least two of the three aspects mentioned in the DDF Framework: variable, units, and population. For the Student Sleep data, if a student wrote "Students tend to sleep 8 hours," it was coded positive for context as sleep is the variable, hours the units, and students the population. A similar mechanism was used for the Atlanta Income data, where a response such as "People tend to make less than $50,000" was coded as correct for context, with people being the population, income (here described via "make $50,000") being the variable, and dollars, as indicated by the dollar sign, being the units. A response of "People tend to make less than 50,000" was not coded as correct for context due to the lack of units and lack of description of the variable of interest, income (i.e., these people could have been "making" 50,000 paper airplanes). Responses such as "People tend to earn less than 50,000" were coded as correct for context because the use of the word "earn" allowed us to assume that the units were dollars. The next two sections describe the process by which the coding rubric was refined.

Table 2. Coding rubric

Category	Requirements for belonging in each category
Shape	Must correctly discuss the shape of the histogram by describing Student Sleep as symmetric, unimodal, bell-shaped, or approximately normal (unimodal with right skew for Atlanta Income).
Center	Must give a valid measure of center (e.g. mean, median, mode, average) and correctly state its location.
Variability	Must discuss either the range of the data, highlight potential outliers, locate the maximum and minimum values, or give an approximation of the variability directly.
Context	Must answer the question within the context of the problem by using the appropriate variable with the appropriate units (e.g., 8 hours, 10,000 dollars) and identifying the subject of each unit (e.g., students, Atlanta adults). At least two of the three aspects mentioned in the DDF Framework: variable, units, and population, must be present for a response to be coded for context.

Two separate raters each read the first 100 student descriptions of the Student Sleep histogram and determined independently whether each response described the shape of the histogram, gave an accurate measure of the center, discussed its variability, and/or responded within the context of the problem. After coding the first 100 submissions for the Student Sleep histogram in this manner, the two raters checked their coding for agreement, and it was found that 88 of these responses were coded in an identical manner. Of the remaining 12 responses, 10 of the disagreements in coding between raters were due to two different interpretations of the *variability* category, which was changed slightly to be less ambiguous, and two of the disagreements were due to semantic interpretations. These differences were discussed and a

consensus was reached for each response. Based on the revised rubric (Table 2), the raters then independently coded the remaining 1,078 student responses. The Atlanta Income data were scored following the same procedure. The initial inter-rater agreement values for independent coding of both datasets (Tables 3 and 4) indicate reasonable levels of agreement for the rubric categories. Overall agreement measures the proportion of responses that were coded identically by each rater (e.g., both raters independently agreed that a particular response was given in context). Cohen's Kappa (Cohen, 1960), a more robust measure of agreement, takes into account the probability that each successful agreement occurred by chance. The remaining disagreements in both the Atlanta Income and Student Sleep data sets were discussed amongst the raters, and a consensus was reached, ultimately leading to categorization of all 1,155 and 1,178 student responses, respectively.

Table 3. Inter-rater agreement for Atlanta income data

Category	Cohen's Kappa	Overall Agreement
Shape	.941	.977
Center	.907	.938
Variability	.803	.925
Context	.969	.980

Table 4. Inter-rater agreement for student sleep data

Category	Cohen's Kappa	Overall Agreement
Shape	.960	.985
Center	.946	.976
Variability	.811	.922
Context	.934	.972

Once the responses were coded, summary statistics (relative frequencies and bar graphs) were examined to identify similarities and differences in responses across prompts and scenarios. Four separate logistic regression models were then created, one for each of the four categories: shape, center, variability, and context. The response variable for each model was whether the description contained the statistical category of interest (e.g., shape, center, variability, or context) and the predictor variables were scenario and question prompt. The interaction term between prompt and scenario was included in all models. A significant interaction term demonstrates evidence that the effects of the individual predictor variables on the response variable are not additive. For example, an interaction term would be significant if students are more likely to respond in context when given the *Distribution Only* prompt or when given the Student Sleep scenario, but the same students are less likely to respond in context when given both the *Distribution Only* prompt and Student Sleep scenario. The results of these analyses presented in the next section help determine whether the proportion of student responses containing each of the four statistical categories remained constant across different prompts and scenarios.

4. RESULTS

To assess the extent of the differences in student responses across prompts (research question 1) and scenarios (research question 2) with respect to the shape, center, variability, and context, the percentage of responses in each of the four categories (Table 5) under each condition was calculated. Across all conditions except the *Variable Only* prompt, students more frequently discussed the shape of the histogram than any other feature. Overall, approximately 73% of the responses mentioned shape, but only 52% of the students who received the *Variable Only* prompt mentioned shape. In contrast, the students who received the *Variable Only* prompt had the highest rate of mentioning center (73%) and context (81%) compared to the other conditions. Students discussed the center and context in varying amounts

depending upon which prompt and scenario combination the student received. Only 27% of students discussed the center and context in the *Distribution Only* prompt, whereas 73% and 81% of students discussed the center and context in the *Variable Only* prompt, respectively. Student responses across all conditions were least likely to contain information about variability regardless of prompt or scenario. Overall, only 28% of the responses contained a mention of variability.

Table 5. Percentage of responses in each category

	Shape	Center	Variability	Context	Total
By Scenario					
Atlanta Income	71.1%	44.4%	24.2%	57.7%	1155
Student Sleep	75.0%	67.4%	32.1%	69.9%	1178
By Prompt					
Distribution Only	90.3%	26.9%	15.3%	27.1%	700
Variable Only	52.2%	73.3%	36.4%	81.4%	832
Both	79.7%	63.5%	30.8%	77.8%	801
All Responses	72.7%	55.8%	28.0%	63.6%	2343

Although student responses to both scenarios tended to involve a description of the center and the correct context of the histogram, a larger proportion of students discussed center and context with the Student Sleep scenario than with the Atlanta Income scenario, an absolute increase of 23 and 12 percentage points and a proportional increase of 52% and 21%, respectively. Student responses to the Student Sleep scenario were 4 and 8 percentage points more likely to contain a discussion of the shape and variability, respectively, compared to the Atlanta Income scenario. Despite the differences in these proportions, the effects of each prompt were similar across both scenarios.

Figure 3 further addresses the research questions by illustrating the similarities in student responses across the three prompts for both scenarios: Student Sleep and Atlanta Income. In both scenarios, students receiving the *Distribution Only* prompt nearly always (approximately 90% of the time) described the overall shape of the histogram. Students receiving the *Distribution Only* prompt, however, were unlikely

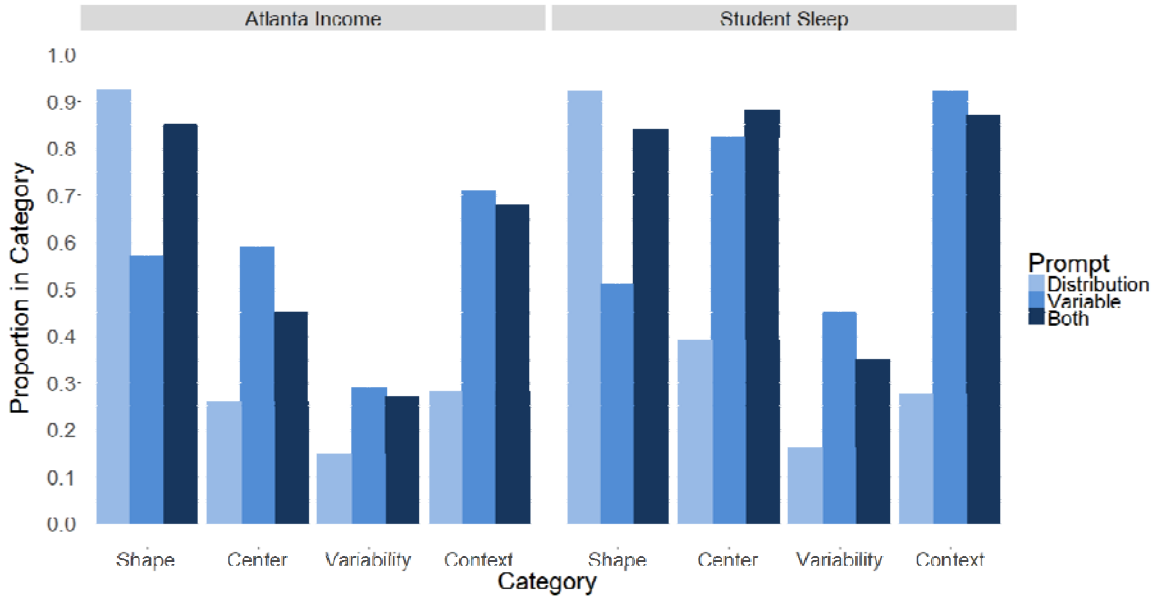

Figure 3. Descriptions by prompt and scenario

to discuss any other aspect of the histogram; they discussed the center, variability, and context in fewer than 30% of the responses across both scenarios. Students receiving the *Both* or *Variable Only* prompts were very likely to formulate a response in the context of the scenario.

Logistic regression models were created to assess the significance of the similarities and differences in student responses across the prompts and scenarios. Table 6 shows the results of the logistic regression analyses. Statistically significant (p-value < 0.05), positive coefficients are represented by plusses (+), statistically significant, negative coefficients are represented by minuses (—), and non-significant terms are represented by dots (·). The baseline response for the logistic regression models was a student response to the *Both* prompt of the Atlanta Income scenario. Thus, all relationships displayed in Table 6 are relative to this baseline so, for example, the significant positive coefficient, for the Context category and Scenario: Sleep factor, means that a student was significantly more likely to give a response in context when responding to the Student Sleep scenario than when responding to the Atlanta Income scenario for students who were presented with the *Both* prompt.

Table 6. Logistic regression analysis results

	Predicted Category			
Factor	Shape	Center	Variability	Context
Scenario: Sleep	+	+	+	+
Prompt: Distribution Only	+	—	—	—
Prompt: Variable Only	—	+	·	·
Scenario: Sleep/Prompt:Distribution	·	—	·	—
Scenario: Sleep/Prompt:Variable	—	·	·	·

+ significant positive relationship, — significant negative relationship, · non-significant relationship

For all four response categories—shape, center, variability, and context—both the Prompt and the Scenario led to significant differences in the response patterns. For students who were presented with the *Both* prompt (baseline), those who were presented with the Student Sleep scenario were significantly more likely to mention all four aspects of the distribution than students who were presented with the Atlanta Income scenario. For students who were presented with the Atlanta Income scenario (baseline), those who were presented with the *Distribution Only* prompt were significantly more likely to mention Shape, but significantly less likely to mention Center, Variability, and Context than students who were presented with the *Both* prompt. Students with the Atlanta Income scenario who were presented with the *Variable Only* prompt were significantly more likely to mention Center and significantly less likely to mention Shape than those with the *Both prompt*. There were, however, three significant interaction terms, each with a large negative coefficient. These coefficients were larger in magnitude than those of the corresponding main effects. Thus, students receiving the Student Sleep scenario and *Variable* prompt were significantly less likely to mention the shape of the distribution compared to students receiving the Atlanta Income scenario and *Both* prompt. Likewise, students receiving the Student Sleep scenario and *Distribution* prompt were significantly less likely to mention either the center or the context of the histogram when compared to the baseline.

5. DISCUSSION

5.1. SUMMARY OF RESULTS

Across all conditions, students were most likely to mention the shape of the distribution, followed by context and then center. Students were least likely to discuss variability across all conditions. Approximately 72% of students across all prompts and both scenarios, Student Sleep and Atlanta Income,

failed to discuss potential outlying points, maximum and minimum values, the spread or range of the data, or a formal approximation of the variability in the data. The tendency of students to neglect mention of variability is well documented in the literature (see for example, Ben-Zvi, 2004; Leavy & Middleton, 2011; Meletiou-Mavrotheris & Lee, 2002b, Reading & Shaughnessy, 2004). In order for a student to present a formal discussion of variability, he or she must first recognize the need to mention variability and also have located the center of the distribution, as formal measures of variation are a function of the distance of the points from the center (Jones & Scariano, 2014; Kader & Jacobbe, 2013). In a formal discussion of variability, a student must also note something about the shape of the distribution (Konold & Pollatsek, 2002), as the shape determines how points systematically fall far from the center (skewed distribution) or close to the center (bell-shaped distribution). Thus, a formal discussion of variability almost always implies a discussion of shape and center, although the converse may be generally untrue. Although responses in this study did not require a formal discussion of variability to be coded as containing a discussion of variability, the results are congruent with the literature: nearly 90% of students who discussed variability also discussed shape and center, but only 30% of students who discussed shape and center also discussed variability.

The wording of each prompt given to the students in both the Atlanta Income scenario and the Student Sleep scenario evoked different patterns of student responses. When the word *distribution* was included in the prompt (*Distribution Only* and *Both*), students were significantly more likely to mention the shape of the histogram when compared to the students who received the Variable Only prompt, which did not contain the word *distribution*. There was an absolute increase of 27.5 and 38.1 percentage points and a proportional increase of 52% and 71%, for *Both* and *Distribution Only,* respectively. This may indicate that students equate the word *distribution* with the word *shape*. In other words, students may think that a synonym for the statistical definition of the word *distribution* is *shape.*

Students who responded to the Student Sleep scenario were significantly more likely to give correct descriptions for each of the four aspects of the distribution as compared to students who responded to the Atlanta Income scenario. With respect to center and variability, the added difficulty seemed to be related to the difficulty of identifying the mean or median value from a right-skewed distribution as compared to a unimodal and symmetric distribution. Students who responded to the Student Sleep scenario were more likely to mention the median (or mean) and the modal clump, writing, "the median number of hours of sleep was 7, but most students get between 6 and 9 hours." These responses were coded as having center (7 hours) and variability (between 6 and 9 hours), in other words, belonging to both categories, Center and Variability. In contrast, students who provided a description of the center of the right-skewed Atlanta Income scenario tended to report only the modal clump without a specific measure of center—for example, "most people make less than 50,000." Such a response would be coded as containing a measure of center, but not variability. If the student had included the phrase, "but some people make over 300,000," the response would have also been coded as addressing variability through mention of the upper value. The relative ease of identifying the center of the symmetric distribution may have prompted the students to provide more information about the distribution, leading to a discussion of variability.

The example responses given in the previous paragraph also provide insight into the difference in inclusion of context across the two scenarios. The examples for the Atlanta Income scenario would not have been coded as having context because only one of the three context characteristics were mentioned. In each example, the values are mentioned (50,000 and 300,000, respectively), but there are no units so it is not clear what the people are making (i.e., income is at best implied). The examples for the Student Sleep scenario do have complete context: students, sleep, and hours. Note that it is relatively easy to write a sentence about the Atlanta Income scenario in which the context is implied but not specified. In contrast, it is more difficult to construct a response to the Student Sleep scenario that has the ambiguity associated with the context. This may explain why students who responded to the Student Sleep scenario were more likely to include a complete description of the context of the scenario.

Student responses in the Student Sleep scenario often included a description of the histogram's shape and center in tandem. Although we have not included a formal analysis of the conditional probabilities across categories, we noted that a typical student response to the Student Sleep scenario was: "The graph

is bell-shaped with a mean of about 8 hours." Unlike responses from the Student Sleep scenario, responses to the Atlanta Income scenario rarely exhibited this coupling of shape and center. A common response to this scenario was: "A majority of the adults make very little. The mean would be very low." The differences in the overall shapes of the two histograms, bell-shaped versus right-skewed, likely contributed to the large discrepancy in the number of responses containing a description of shape between the two scenarios—students seemed more likely to identify and explain that a histogram was bell-shaped than that it was right-skewed. In summary, the *Both* prompt seemed to cue students to provide the most complete description of the histogram for both scenarios, and the responses to the Student Sleep scenario seemed to evoke more complete responses than responses in the Atlanta Income scenario.

5.2. IMPLICATIONS FOR RESEARCH

The results indicate that question prompt and scenario both impact the response patterns of undergraduate students' written descriptions of histograms and, therefore, the validity and reliability of the assessment items. With respect to question prompt, researchers should consider the purpose of the assessment when choosing from among the prompts described. For example, researchers with a particular interest in investigating what students recognize as key aspects of a distribution of univariate quantitative data may want to use a *Distribution Only* prompt. Those with an interest in investigating what students recognize as important features of a display of univariate quantitative data may choose to use a *Variable Only* prompt. The *Both* prompt seems to cue more complete descriptions, lessening the validity of the item for those specified purposes. The authors undertook this study to develop assessment items for a project designed to provide feedback on student learning to instructors of large lecture introductory statistics classes. Whereas typical introductory statistics textbooks use a mix of *Distribution Only* and *Both* prompts in their exercises (see for example, Agresti & Franklin, 2012; Moore 2009; Utts & Heckard, 2012), we believe the *Both* prompt will provide more authentic feedback for instructors about how well their students have mastered the student learning outcome associated with describing distributions of univariate quantitative variables in context. Using a *Distribution only* prompt, may limit student responses to discussions of shape and provide little information on whether they have learned how statisticians conceptualize distributional features.

There are two limitations to the research that provide opportunities for future research: 1) the results for Scenario are confounded with the shape of the distributions shown in the histogram and 2) the research was completed at one institution using students in the same course. The first limitation is not trivial to address as the shape and context of the distribution are generally linked. In fact, according to Arnold (2013), the most complete descriptions of distributions are those that include an explanation of the contextual reason for the shape exhibited by the data. In the case of the Atlanta Income scenario, a symmetric distribution would not be a reasonable model so the use of a symmetric distribution to model Income data might confuse students who are aware of the typical shape of the distribution of incomes, leading to an assessment item that has lower levels of validity and reliability. In contrast, a second, right skewed distribution could be created for the Student Sleep scenario. It is plausible that on weekends the distribution for Student Sleep is right skewed, meaning most students sleep between 5 and 9 hours a night on the weekend with a few getting more sleep.

Although we hypothesize that the difference in response patterns for center and variability across the scenarios, Student Sleep and Atlanta Income, is due to the different shape of the data and not the scenario of the data, it is possible that it is easier for students to conceptualize and quantify a measure of center for the number of hours students sleep, a scenario with which they would be more familiar, rather than the annual income for residents of a city for which they might have less contextual knowledge. This is particularly true because all of the participants were students who have slept, but fewer of the participants might have lived in Atlanta and received an income. Replication studies with other scenarios and shapes will aid in understanding the relationship between scenario, shape, and quality of student descriptions of distributions of univariate data.

The limitation associated with the use of students from a single, highly-coordinated course at a single institution may mean that the responses are not generalizable beyond our particular setting. That said, we believe the treatment of descriptions of univariate distributions in the course is typical of such courses and included examples of distributions of various shapes and from a variety of contexts. In addition, this limitation can be addressed through replication studies at other institutions. In fact, we have created a trained ensemble of machine learning algorithms to categorize student descriptions of the two histograms, Student Sleep and Atlanta Income. New data can therefore be collected and analyzed in a short time to replicate the results presented and ascertain whether they are an artifact of the university or generalizable to a larger population. We plan to validate the models using data collected from other institutions. In addition, we plan to create categorization models for student responses to more histograms in order to provide the statistics education community with a set of assessments that will provide formative feedback in real time for instructors of statistics. This is one of the future directions associated with the project. A second future direction is the investigation of the interaction between misconceptions about histograms held by a student and that student's ability to describe a histogram, particularly given the literature on student misconceptions of histograms (see, for example, delMas et al., 2005; Kaplan et al., 2014; Meletiou-Mavrotheris & Lee, 2010). The research team is currently analyzing data that included both items to assess student misconceptions about histograms and student descriptions of histograms in an effort to shed light on this second direction for research.

One interesting finding of this study was the positive relationship between the use of the word *distribution* in the prompt and mention of shape in the student descriptions. This finding may indicate that students use the word *shape* as a synonym for *distribution*. This finding is perhaps not unexpected given that *distribution* has been identified in the literature as a potentially lexically ambiguous word in statistics (Kaplan, Fisher, & Rogness, 2009; Richardson, Dunn, & Hutchins, 2013) and may provide a direction for study by researchers interested in lexical ambiguity in statistics. Given that distribution underlies a breadth of content in statistics, addressing students' misunderstanding that distribution is a synonym for shape could have a large impact on student learning in statistics. More globally, the findings of this study serve as a reminder for researchers, instructors, and test developers of the potential effects of wording on the reliability and validity of assessment items. The development of assessment items should include careful examination and piloting of initial versions to ensure the final item has a high level of validity and reliability aligned with the intended content and purpose of the assessment item.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation under Grant No. 1322962. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank the personnel of the Automated Analysis of Constructed Response (AACR) research project. In particular, the Question Development Cycle (QDC) team provided background and theory that helped in the design of the research. In addition, the authors are grateful for the comments made by the two anonymous reviewers and the associate editor. These comments were invaluable to improving the organization of the paper. Any errors or issues, however, are solely the responsibility of the authors and not their helpers.

REFERENCES

Agresti, A., & Franklin, C. (2012). *Statistics: The art and science of learning from data* (3nd ed.). Upper Saddle River, NJ: Pearson Education, Inc.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing.* Washington DC: AERA.

American Statistical Association (2005). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: Author.
[Online: http://www.amstat.org/education/gaise/GAISECollege.htm]

American Statistical Association (2007). *Guidelines for assessment and instruction in statistics education: Pre-K-12 framework*. Alexandria, VA: Author.
[Online: http://www.amstat.org/asa/files/pdfs/GAISE/GAISEPreK-12_Full.pdf]

Arnold, P. (2013). *Statistical investigative questions: An enquiry into posing and answering investigative questions from existing data* (Unpublichsed doctoral thesis). The University of Auckland, New Zealand.
[Online: https://researchspace.auckland.ac.nz/handle/2292/21305]

Arnold, P., & Pfannkuch, M. (2014). Describing distributions. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9),* Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8G1_ARNOLD.pdf]

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168). Dordrecht, The Netherlands: Kluwer.

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2013). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? *Journal of Science and Educational Technology, 23*(1), 160–182.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: L. Erlbaum Associates.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal, 3*(2), 42–63.
[Online: https://iase-web.org/documents/SERJ/SERJ3(2)_BenZvi.pdf]

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? Helping young students to reason and argue about some wider universe. In D. Pratt & J. Ainley (Eds.), *Reasoning about statistical inference: Innovative ways of connecting chance and data. Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5, August, 2007), University of Warwick, England* (pp. 1–26). Warwick, England: University of Warwick.

Birenbaum, M., & Tatsouka, K. K. (1987). Open-ended versus multiple-choice response formats—It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 329–341.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253–271.

Bright, G. W., & Friel, S. N. (1998). Graphical representations: Helping students interpret data. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching and assessment in grades K-12*. Mahwah, NJ: Lawrence Erlbaum.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Common Core State Standard Initiative (CCSSI) (2010). *Common core state standards for mathematics*. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Crisp, V., Sweiry, E., Ahmed, A., & Pollitt, A. (2008). Tales of the expected: The influence of students' expectations on question validity and implications for writing exam questions. *Educational Researcher, 50*(1), 95–115.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

D'Avanzo, C. (2008). Biology concept inventories: Overview, status, and next steps. *Bioscience, 58*(11), 1079–1085.

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty with reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Reasoning about distribution: A collection of research studies: Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy*, Auckland, New Zealand. [CD-ROM]. Auckland: University of Auckland.
[Online: https://www.causeweb.org/cause/archive/artist/articles/SRTL4_ARTIST.pdf]

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science, 28*(6), 843–900.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124–158.

GAISE College Report ASA Revision Committee (2016). *Guidelines for assessment and instruction in statistics education college report 2016*. Alexandria, VA: American Statistical Association.
[Online: http://www.amstat.org/education/gaise]

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities, *International Statistical Review*, *70*(1), 1–51.

Gal, I., & Garfield, J. B. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. B. Garfield (Eds.). *The assessment challenge in statistics education*. Amsterdam, The Netherlands: IOS Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*(3), 277–300.

Greeno, J. G. (2009). A theory bite on contextualizing, framing, and positioning: A companion to son and goldstone. *Cognition and Instruction, 27*(3), 269–275.

Jones, D. L., & Scariano, S. M. (2014). Measuring the variability of data from other values in the set. *Teaching Statistics, 36*(3), 93–96.

Kader, G., & Jacobbe, T. (2013). *Developing essential understanding of statistics for teaching mathematics in grades 6–8* (P. S. Wilson, Vol. Ed.) In R. M Zbiek (Series Ed.), *Essential understanding series*. Reston, VA: National Council of Teachers of Mathematics.

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2009). Lexical ambiguity in statistics: What do students know about the words association, average, confidence, random and spread? *Journal of Statistics Education*, *17*(3), 1–19.
[Online: http://ww2.amstat.org/publications/jse/v17n3/kaplan.pdf]

Kaplan, J. J., Gabrosek, J. G., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education*, *22*(2).
[Online: http://www.amstat.org/publications/jse/v22n2/kaplan.pdf]

Knight, J. K. (2010). Biology concept assessment tools: Design and use. *Microbiology Australia, 31*, 5–8.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, 33, 259–289.

Kuechler, W. L. & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *The Decision Sciences Journal of Innovative Education, 8*(1), 55–73.

Leavy, A. M., & Middleton, J. A. (2011). Elementary and middle grade students' constructions of typicality. *Journal of Mathematical Behavior, 30*(3), 235–254. doi:10.1016/j.jmathb.2011.03.001

Libarkin, J. C. (2008, Oct 13-14). *Concept inventories in higher education science.* Manuscript for the National Research Council Promising Practices in Undergraduate STEM Education Workshop 2, Washington, DC.

Madden, S. R. (2011). Statistically, technologically, and contextually provocative tasks: Supporting teachers' informal inferential reasoning. *Mathematical Thinking and Learning*, *13*(1-2), 109–131.

Manjoo, F. (2013, June 6). You won't finish this article: Why people online don't read to the end. *Slate Online*.
[Online: http://www.slate.com/articles/technology/technology/2013/06/how_people_read_online_why_you_won_t_finish_this_article.html]

Meletiou-Mavrotheris, M., & Lee, C. (2002a). Student understanding of histograms: A stumbling stone to the development of intuitions about variation, In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics,* [CD-ROM] Cape Town, SA. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://iase-web.org/documents/papers/icots6/10_19_me.pdf]

Meletiou-Mavrotheris, M., & Lee, C. (2002b). Teaching students the stochastic nature of statistical concepts in an introductory statistics course, *Statistics Education Research Journal, 1*(2), 22–37.
[Online: https://iase-web.org/documents/SERJ/SERJ1(2).pdf]

Meletiou-Mavrotheris, M., & Lee, C. (2010). Investigating college-level introductory statistics students' prior knowledge of graphing, *Canadian Journal of Science, Mathematics, and Technology Education*, *10*(4), 339–355.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th edition), Boston, MA: Pearson.

Moore, D. S. (2009), *The basic practice of statistics* (5th ed.). New York: W.H. Freeman and Company.

National Council of Teachers of Mathematics (NCTM) (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*(3), 237–256.

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching, 45*(10), 1131–1160.

Noll, J., & Hancock, S. (2015). Proper and paradigmatic metonymy as a lens for characterizing student conceptions of distributions and sampling. *Educational Studies in Mathematics, 88*(3), 361–383.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press

Pfannkuch, M., Regan, M., Wild, C. J., & Horton, N. (2010). Telling data stories: Essential dialogues for comparative reasoning. *Journal of Statistics Education*, *18*(1), 1–38.
[Online: https://www.tandfonline.com/doi/abs/10.1080/10691898.2010.11889479]

Pollitt, A., Ahmed, A., Baird, J., Tognolini, J., & Davidson, M. (2008). *Improving the quality of GCSE assessment.* Carrickfergus, UK: Qualifications and Curriculum Authority.

Quealy, K., Cox, A., & Katz, J. (2015, Feb. 17). At Chipotle, how many calories do people really eat? *The New York Times Online.* Retrivied from: https://www.nytimes.com/interactive/2015/02/17/upshot/what-do-people-actually-order-at-chipotle.html?_r=1]

Reading C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201–226). Dordrecht, The Netherlands: Kluwer.

Richardson, A. M., Dunn, P. K., & Hutchins, R. (2013). Identification and definition of lexically ambiguous words in statistics by tutors and students. *International Journal of Mathematical Education in Science and Technology*, *44*(7), 1007–1019.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/rumsey2.html]

Sabella, M. S. & Redish, E. F. (2007). Knowledge organization and activation in physics problem solving. *American Journal of Physics, 75*(11), 1017–1029.

Schurmeier, K. D., Atwood, C. H., Shepler, C. G., & Lautenschlager, G. J. (2010). Using item response theory to assess changes in student performance based on changes in question wording. *Journal of Chemistry Education, 87*(11), 1268–1272.

Silver, E. A. (1979). Student perceptions of relatedness among mathematical verbal problems. *Journal for Research in Mathematics Education, 10*(3), 195–210.

Sweiry, E. (2013, October). *A framework for the qualitative analysis of examinee responses to improve marking reliability and item and mark scheme validity*. Paper presented at the 39th Annual Conference of the International Association for Educational Assessment, Tel Aviv, Israel.

Utts, J. M., & Heckard, R. F. (2012), *Mind on statistics* (4th ed.). Boston, MA: Brooks/Cole Cengage Learning.

Watson, J. (2005). Developing an awareness of distribution. In K. Makar (Ed.), *Reasoning about distribution: A collection of research studies: Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy*, Auckland, New Zealand. [CD-ROM]. Auckland: University of Auckland.

Weston, M., Haudek, K. C., Prevost, L. B., Merrill, J., & Urban-Lurain, M. (2015). Examining the impact of question surface features on students' answers to constructed response questions on photosynthesis. *CBE Life Sciences Education, 14*(2). doi:10.1187/cbe.14-07-0110

JENNIFER J. KAPLAN
Department of Statistics
401 Brooks Hall
310 Herty Drive
Athens, GA 30601