5-2011

Statistical Analysis When the Data is an Image: Eliciting Student Thinking About Sampling and Variability.

Margaret A. Hjalmarson

Tamara J. Moore
Purdue University, tamara@purdue.edu

Robert Delmas

STATISTICAL ANALYSIS WHEN THE DATA IS AN IMAGE: ELICITING STUDENT THINKING ABOUT SAMPLING AND VARIABILITY

MARGRET A. HJALMARSON
George Mason University
mhjalmar@gmu.edu

TAMARA J. MOORE
University of Minnesota
tamara@umn.edu

ROBERT DELMAS
University of Minnesota
delma001@umn.edu

ABSTRACT

1. INTRODUCTION

Within statistics education, there is a growing interest in understanding students' application of understanding about variability and sampling given the relative lack of research in either area (Shaughnessy, 2007). The task examined in this paper elicited students' knowledge of these concepts within a small-group problem solving task completed by teams of first-year engineering students. In the Nanoroughness task, teams of students designed a procedure for quantifying the roughness of a material surface using digital images generated by atomic force microscopy. The procedure required students to apply statistical methods in order to aggregate the data. The focus of this article is the subsequent analysis of the responses to the task and the questions raised by that analysis.

The Nanoroughness task is unique but critical as a statistical modeling task for two reasons. First, the students needed to use statistical measures to develop a measure that would describe a qualitative characteristic (roughness) without any prompting as to what statistical procedures were relevant. There are different ways to conceptualize roughness of a surface. Sandpaper's roughness depends on the grain size of the sand. A road may be rough if it has randomly occurring large holes but smoother if the bumps are evenly distributed. The challenge in developing quantitative measures to define qualitative characteristics is that different quantitative analyses emphasize different variables and the students needed to both analyze and apply statistical procedures relevant to the context. For instance, determining which member of a set is the "most rough" or the "least rough" will depend on what measurements were selected, and how those measures were analyzed. The second unique characteristic of the task is that the students also needed to define a sampling procedure for an image that would facilitate quantifying the variability in the surface portrayed in the digital image. Usually when students need to take measurements of a population, the population is a discrete set of objects. In this case, the data set was a continuous surface. From the data set, the students need to determine the relevant population (e.g., every point on the surface, every peak on the surface, peaks and valleys). Such continuous populations are not unique within engineering and the sciences and occur in a variety of contexts where characteristics need to be measured and operationally defined.

The task was implemented in a first-year engineering course that served as an introduction to basic tools of engineering with an emphasis on MatLab® and Excel® as technological tools. The Nanoroughness task was used in the course to introduce students to the real work of engineers who must not only calculate statistics but also analyze and interpret the results. Our research asked a two-part question. First, *what is the quality of student responses to the Nanoroughness task?* To answer this we looked at the viability of the model they had created and how well they had explained their procedure for comparing the roughness of images. Second, *what statistical models were elicited by the task?* We specifically looked at the sampling methods students used and then how the students analyzed the data set they had created. In this paper, we describe the quantitative and qualitative analyses we completed of a sample of student responses.

2. LITERATURE REVIEW

The relevant literature related to students' understanding of data analysis falls into two broad categories: measures of variance or distribution and data sampling. Watson, Kelly, Callingham, and Shaughnessy (2003) defined variation as "the underlying change from expectation that occurs when measurements are made or events occur" (p. 1). "Roughness", almost by definition, is physical variation in a surface. For measures of central tendency and distribution, students need to apply statistics to describe characteristics of a population. In this case, the population is an image representing a surface. They need to determine which statistics to select in order to describe characteristics of the surface and compare different surfaces to each other.

Some studies have looked at how students use the mean as a statistical measure. Pollatsek and colleagues (Konold, Pollatsek, Well, & Gagnon, 1997; Pollatsek, Lima, & Well, 1981; Well, Pollatsek, & Boyce, 1990) have examined students' understanding of the mean as a measure of central tendency. Their studies confirm that undergraduate students can compute the mean, but they don't necessarily know how to interpret what it indicates about a data set. In a study of pre-service elementary teachers' understanding of the mean, median and mode, Groth and Bergner (2005) found the students often had algorithmic conceptions of the terms and limited understanding about when to apply the statistical measures (i.e., how to select appropriate measures of central tendency). In the case of Nanoroughness, the mean was one option among many from which students could choose to describe the data set, and other measures of central tendency (e.g., median, mode) could be computed with different results. The students also could incorporate a measure of variability into their procedure (e.g., standard deviation, range) as a measure of roughness.

An additional layer of complexity beyond determining what statistical measure to use was that the population was not well-defined. One interpretation of the population could be all the peaks and valleys on the surface so the students needed to select a subset of peaks and valleys as a sample. The population could also be defined as every point on the image so the students needed a method for either sampling from the points on the image or analyzing the pixel color value of every point on the image. Students' first decision was to determine what points they would use for their population and then what (if any) subset of those points to use. For instance, just the peak points? The entire image? If using the peak values, how many peak values?

Often statistics tasks presented to students require them to analyze numerical data sets where the values are given or where they can take physical measurements. In this case, the students needed to determine what to measure given an image that represented the universe of the data set. Konold and Pollatsek (2002) have described this as a process of

separating signal from noise where students need to determine what to attend to in a data set and then determine critical variables. In other studies of students completing similar modeling tasks, the sample set was clearly defined and students needed only to determine if possible values needed to be eliminated (Hjalmarson, 2007). In order to define a sample set in the case of a continuous surface, students need to determine what and where to measure.

Very little research had been conducted on students' understanding of variability prior to 1999 (Reading & Shaughnessy, 2004; Shaughnessy, 1997), despite the central role the concept plays in statistics (Hoerl & Snee, 2001; Moore, 1990; Snee, 1990) and an apparent conceptual gap in students' understanding of variability (Reading & Shaughnessy, 2004; Shaughnessy, 1997). A few investigations have been conducted into students' understanding of sampling variability and instructional approaches that affect this understanding. Reading and Shaughnessy (2004) presented evidence of different levels of sophistication in elementary and secondary students' reasoning about sample variation. Meletiou-Mavrotheris and Lee (2002) found that an instructional design that emphasized statistical variation and statistical process produced a better understanding of the standard deviation, among other concepts, in a group of undergraduates. Students in the study saw the standard deviation as a measure of spread that represented a type of average deviation from the mean. They were also better at taking both center and spread into account when reasoning about sampling variation in comparison to findings from earlier studies (e.g., Shaughnessy, Watson, Moritz, & Reading, 1999).

Shaughnessy (1997; Reading & Shaughnessy, 2004) noted that the standard deviation is both computationally complex and difficult to motivate as a measure of variability. Part of this difficulty may stem from a lack of accessible models and metaphors for students' conceptions of the standard deviation (Reading & Shaughnessy, 2004). Most instruction on the standard deviation tends to emphasize teaching a formula, practice with performing calculations, and tying the standard deviation to the empirical rule of the normal distribution. This emphasis on calculations and procedures does not necessarily promote a conceptual understanding of standard deviation. Part of the difficulty may also stem from students' misunderstanding of how variability can be represented graphically. For example, when presented with a histogram, some students judged the variability of the distribution on the basis of variation in the heights of bars, or the perceived "bumpiness" of the graph, rather than the relative density of the data around the mean (Garfield, delMas, & Chance, 1999). DelMas and Liu (2005) provided evidence that experience with a computer environment designed to promote students' ability to coordinate characteristics of variation of values about the mean moved from simple, one-dimensional understandings of the standard deviation toward more mean-centered conceptualizations that coordinated the effects of frequency (density) and deviation from the mean.

Shaughnessy (2007) identified three types of reasoning about variability that can be addressed by statistics instruction: variability within data; variability between samples; and variability within sampling distributions. The current study highlights a problem context where students need to consider variability within data to produce a solution. A factor that may impede students understanding of variability in data is the lack of problems that naturally motivate the need to measure variability. Some examples are provided in the literature (e.g., Konold & Kazak, 2008; Lehrer & Schauble, 2003), but there are few. Many studies have focused on students understanding of sampling variability and sampling distributions (e.g., Chance, delMas, & Garfield, 2004; delMas, Garfield, & Chance, 1999; Kelly & Watson, 2002; Reading & Shaughnessy, 2004; Rubin, Bruce, & Tenney, 1991; Shaughnessy et al., 1999; Torok & Watson, 2000) but not on variability within data, per se, and the contexts often do not depict real world problems.

The study reported here illustrates a situation that naturally elicits the use of measures of variability within data and the design of a sampling method to solve a real world problem in an undergraduate engineering course.

2.1 MODEL- ELICITING ACTIVITIES INCLUDING STATISTICAL ANALYSIS

The Nanoroughness task is an example of a model-eliciting activity. The design process used for the task has been described elsewhere (Hjalmarson, Diefes-Dux, & Moore, 2008; Moore & Diefes-Dux, 2004) and for the purposes of this paper, we focus on the types of statistical models revealed in the students' work. Model-eliciting activities for students require the development of models or procedures rather in addition the production of answers (Lesh, Hoover, Hole, Kelly, & Post, 2000; Zawojewski, Hjalmarson, Bowman, & Lesh, 2008). For instance, in Nanoroughness, the students needed to define a procedure for defining roughness of a surface and then select the roughest sample by implementing the procedure. The central product for the task is the procedure and not just the computed values and subsequent ranking of samples by roughness. Requiring students to describe their procedure as the product of their problem solving process naturally elicits students' thinking about the statistical concepts and makes misconceptions more evident. Hjalmarson (2008) and Doerr and English (2003) have described students' thinking in other model-eliciting activities requiring data analysis. Moore (2008) examined teachers' solutions to the Nanoroughness task. A common feature of all of these tasks is that the task statement doesn't specifically ask students to use statistical analysis. However, common types of statistical measures are elicited by each task (e.g., mean is often selected when students need to analyze a table of data).

Because model-eliciting tasks require students to describe a procedure as a central activity in the problem-solving process, the tasks can serve as a launching point for discussions about the meaning of central concepts inherent in the problem and it's context. For instance, in the Nanoroughness task, the students have to find a way to quantify variation in the surface. By making their assumptions explicit via the procedure they design, the possibility for discussion of the constraints and affordances of their procedures is possible. The same design questions occur in any context where qualitative characteristics are quantified. Every statistical measure provides different information about the data set. For example, the mean can be used to describe the central tendency of a data set, but it can obscure the range of the values. Standard deviation is a measure of variation but only relative to the mean of the data set.

Another salient feature of model-eliciting activities is that students worked in groups. The procedures needed to be explicit both to the client for the product and to their team members. Making the procedures explicit opened them to questioning by the team members, prompting students needed to consider other operational definitions of roughness. The combination of group work and the requirement to design a procedure that was tested on a given data set can cause students to go through cycles of refinement of their solutions. For instance, many groups may start by computing the mean for their data set because it is a familiar statistic. Some groups may move on to other measures after seeing the results of their analysis using the mean (particularly if the mean didn't differentiate images).

Assessment of students' responses to model-eliciting activities has often first focused on describing the characteristics of students' models (Carmona-Dominguez, 2004; Hjalmarson, 2007). Since the models are a procedure including multiple considerations, the assessment of these models typically includes finding patterns or common themes in the models that can be sorted into types or categories (e.g., there are common methods

students used for finding a sample from the image). In addition, students' models emerge at different degrees of quality typically because a model is incomplete. For instance, students may leave out critical information necessary for someone else to successfully implement their model. In the Nanoroughness task, for example, students may have described the need for generating a sample data set, but not described how to generate a sample (e.g., by finding random points on the surface, drawing random lines). The quality assurance guide described by Lesh and Clarke (2000) is one example of an assessment tool used to categorize students' models by how well they meet the needs of the client and how well the procedure can be generalized to similar situations.

3. METHOD AND MATERIALS

3.1 COURSE INFORMATION

The student work analyzed for this study was drawn from a first-year engineering course in fall 2003. 1478 students were enrolled in a first-year engineering problem solving and computer tools course at a large, public Midwestern university. The students included 1203 males and 275 females. The students were divided into laboratory sections of approximately 25 students per section in order to complete the nanoroughness model-eliciting activity in class. The activity was their fourth and final model-eliciting activity during the semester. Within the laboratory sections, the students were divided by the graduate teaching assistant into long-term teams of three to four. The students were in two different teams for the course, one in the first half of the semester and one in the second. Since they were working with their second team for the semester, the teaching assistants could use information gathered about students' prior experiences to generate teams. The only fixed rules (set by the department) were that a team should have 3 or 4 students total and no fewer than two females or fewer than two international students. We selected 35 responses from 35 teams in different sections of the course for this analysis.

3.2 NANOROUGHNESS LABORATORY ACTIVITY

The Nanoroughness Laboratory Activity is broken into two distinct parts: an individual task and a team modeling task. The individual task consisted of the students reading a short description of the company that supposedly hired the team, and then answering questions designed to elicit their initial interpretations of roughness (Figure 1).

The students posted their individual responses online using a format generated by the department. Once they had completed the individual responses, the teaching assistant released the responses to the rest of the team members. The team then compared and contrasted the individual responses in order to negotiate team definitions for roughness. Once they came to common definitions, the teaching assistant provided the team with the modeling task. Prior to working on the modeling task, the students were provided with a description of Atomic Force Microscopes (AFM) and procedures for taking digital images of materials at the molecular level and a sample of images generated with an AFM. The teams had about an hour and a half to develop their procedure and write the memo to the client. Since this was the students' first draft of a procedure in the context, we expected some level of incompleteness in their procedure descriptions and that there were aspects of the situation they might miss. However, the task did elicit students' initial thinking about sampling and the application of statistical measures.

The second part of the activity required student teams to create a procedure for measuring roughness at the nanoscale level given AFM images of gold. Here, the AFM

images were like topographical maps with a height bar that represents the third dimension. Sample A (see Figure 2) represents one of the three different samples of gold with different scales that were provided to the teams to create their procedure for measuring roughness. The teams were asked to respond to the client in a memo that would allow the client to measure the roughness of any surface using an AFM image. The questions the teams responded to in their memos are in Figure 2. See Moore (2008) or Moore and Diefes-Dux (2004) for more information on the Nanoroughness Task.

1) How do you define roughness?
2) What procedure might you use to measure the roughness of the pavement on a road?
3) Give an example of something for which degree of roughness matters.
4) For your example, why does the degree of roughness matter?
5) How might you measure the roughness (or lack of roughness) of this object?

Figure 1. Individual thinking questions on concepts of roughness.

Please reply in a memo with your answers to the following:

1. The series of steps that can be used to measure roughness of the nanoscale material using the AFM images.
2. A description of how the procedure would work by applying it to gold samples A, B, and C that are attached to this memo.
3. A description of what information your team would need in order to imporve your procedure to quantify the roughness of the gold.

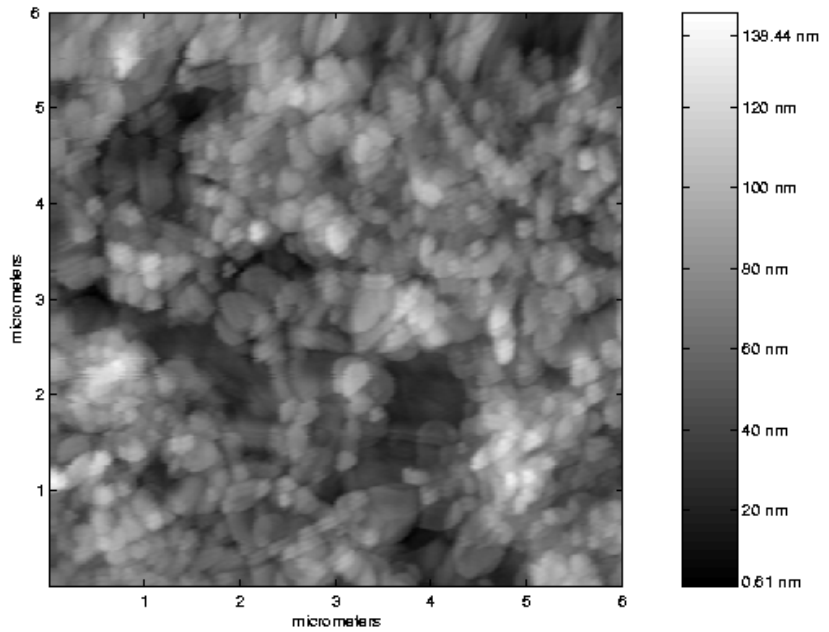

Figure 2. Nanoroughness task problem statement with an example of an AFM image of gold.

4. DATA ANALYSIS

Data analysis focused on the students' responses to the task shown in Figure 2. The team responses were coded in two stages: the assessment of team responses and the description of typical response characteristics. The coding for the assessment of team responses used the Quality Assurance Guide (Lesh & Clarke, 2000) described in the next subsection to assign a numeric score to each response. The Quality Assurance Guide has also been used in other studies including the analysis of student work (Carmona-Dominguez, 2004; Chamberlin, 2002; Moore, 2006). In order to describe the characteristics of different solutions, a qualitative analysis was used to first develop descriptors based on the coders' reading of the student work and then assigning those descriptors to student work. The qualitative analysis resulted in descriptions of the types of student responses in terms of the types of statistical measures the students used and how those measures were implemented in the roughness procedure.

A team of three researchers assessed and categorized the student responses to the model-eliciting task. All three had worked with designing modeling tasks for this course. One was an interdisciplinary mathematics/engineering educator who had been the principal designer for the Nanoroughness sequence. One was a mathematics educator who led the task design group of engineering and education graduate students and faculty. She also had experience scoring tasks using the quality assurance guide. The third was a materials engineering graduate student who had worked with the task design team. The research team was intentionally combined with a blend of experience from engineering, engineering education, and mathematics education in order to provide a variety of perspectives on the student responses.

4.1 QUANTITATIVE ANALYSIS WITH THE QUALITY ASSURANCE GUIDE

The Quality Assurance Guide (see Table 1) described in Lesh and Clarke (2000) was selected to quantitatively assess the tasks. The levels are designed to categorize how well the students' procedure fulfilled the needs of the client and how well they explained their procedure in a generalizable fashion. The range of responses goes from level 5, where the response met the needs of the client for the present situation and for other similar situations as well, to level 1, where the response was going in the wrong direction and the team would need to rethink the procedure completely. The levels in between include varying levels of detail and description. The number of responses in each level is also shown in Table 1. Few of the responses were expected to receive a score of 5 on the quality assurance guide due to the fact that it was the first iteration of the teams' solutions and was turned in after 1.5 hours of working on the problem in a laboratory setting. However, the teams continued to work on this problem through follow-up activities that lead to a project. The research team coded the responses by first reading and scoring a team's procedure individually and then coming to consensus on a final rating. A student team sample (Figure 3) is provided to illustrate the use of the Quality Assurance Guide.

Table 1. The Quality Assurance Guide and Total Number of Responses in the Team Samples for each category.

Score	Performance Level	How useful is the product?	Total Number of Responses
1	Requires redirection	The product is on the wrong track.	6

		Working longer or harder won't work. The students may require some additional feedback from the teacher	
2	Requires major extensions or revisions	The product is a good start toward meeting the client's needs, but a lot more work is needed to respond to all of the issues.	16
3	Requires editing and revisions	The product is on a good track to be used. It still needs modifications, additions or refinements.	9
4	Useful for this specific data given, but not shareable and reusable OR Almost shareable and reusable but requires minor revisions	No changes will be needed to meet the immediate needs of the client for this set of data, but not generalized OR Small changes needed to meet the generalized needs of the client.	2
5	Sharable and reusable	The tool not only works for the immediate situation, but it also would be easy for others to modify and use it in similar situations.	2

To determine the roughness of each sample, we would first draw a number of lines across the sample. Obviously, the more lines drawn would result in a more accurate approximation, but also take more time. The lines should be in a ratio to the scale of each sample. For example, if we draw a 1 micrometer line on a 2 micrometer by 2 micrometer sample, we would then draw lines of 3 micrometers on a 6 micrometer by 6 micrometer sample. After we had drawn a number of random lines, we would take 10 evenly spaced readings of the height from each. From this recorded data, we could then calculate the mean height across the line. Having taken the measurements for several different lines, we could assume that to be the mean height of the entire sample. Once we calculated the mean, we could then figure the standard deviation using the data points we had recorded. The smoother substance would have a lower standard deviation. Furthermore, if the peaks and valleys that the lines intercept are graphed using straight lines to connect the peaks and valleys, we could then calculate the area. This allows for correction of samples that have fewer peaks but the peaks cover a larger area thereby making the sample rougher. To apply this to the samples we are given, we suggest that five lines be drawn across each sample. In sample A, we would draw lines of 3 micrometers. In sample B, we would draw lines of .5 micrometers, and finally in sample C, we would draw lines of 1 micrometer. We would then take approximate height measurements at 10 evenly spaced points on each line, and record the data. We would then calculate the mean height of each sample, and then the standard deviation using the data we recorded. The data could then be plotted using the distance along the line as the x-axis and the height as the y-axis. If the points are connected, we could then calculate the area under the graph. By comparing these two values we could come up with the smoother substance. To better obtain the values, samples with the same scale would have been more useful, along with a more scientific way to determine the height than judging against a color scale.

Figure 3. Sample student response (Team Alpha) to the Nanoroughness task.

On the Quality Assurance Guide, Team Alpha (the response shown in Figure 3) received a score of four indicating that the solution was almost shareable and reusable but required minor revisions. The team's description of graphing the peaks and valleys was not clear to the reader. It was not clear how the data would be plotted on a graph (e.g., Is each line plotted on a different graph or is all the data plotted on one graph?) They also didn't explain how they would aggregate the information from the graphs. Depending on the values, the area under the curve for a plot with more difference between peaks and valleys could have the same area as a curve with fewer, more uniform peaks. It is not clear how the standard deviation would be used to differentiate these two scenarios. Team Alpha used typical statistical measures (mean and standard deviation) as will be described in the qualitative analysis section which follows. They also employed a random line method for generating their sample set of data points which other teams used as well.

4.2 QUALITATIVE ANALYSIS

The qualitative analysis focused on describing the students' models for measuring roughness. Model-eliciting activities such as the Nanoroughness task ask students to develop a model for quantifying a qualitative characteristic. A model includes objects,

operations on the objects and relationships between the operations (Lesh & Clarke, 2000). From a statistical perspective, the objects are the data points and the operations are measures such as the mean and standard deviation. In the present task, the students needed to carry out two statistical processes: defining a sample and quantifying variability. As discussed previously, roughness in and of itself is a way of describing variability in a surface. The challenge was to create a data set and measures that could be used to compare the variability of different surfaces. For instance, the task requires students to develop a method for sampling. In order to develop codes for the teams' sampling methods, several questions were considered. How do students think about random samples in this context? How do they go about generating a random sample from the population as they have defined it? For variability, what measures do students calculate to quantify variability? What do they see as the relationships between those measures (e.g., what do students infer from the mean or the standard deviation?).

Two groups of codes emerged for classifying the responses to the Nanoroughness MEA: sampling method and statistical measures. Descriptors were generated after the responses had been assigned scores and the descriptors were categorized by type. The sampling method codes describe the teams' methods for selecting data from the images (i.e., selecting points to include in their data set or subsets of the data). The teams were given just the images without numerical information about individual pixels. They had a scale that showed the height in relation to the pixel color, but they did not receive data about individual pixels. The images intentionally were provided with different scales so that the teams would have to quantify the information in order to determine a ranking of the images by roughness rather than just selecting an image visually. Not all of the teams noted or accounted for the difference in scale of the images in their method. The differences in sampling method are also important because they reflect differences in how the teams defined the sample set (e.g., the whole image or the peaks in the image) they needed to measure. The following sections will discuss each group of qualitative codes: sampling, procedures, and statistical measures.

4.2.1 SAMPLING

Sampling is an important aspect of defining a quantitative measure of roughness into a usable measure for this problem. As in any statistics problem, how the sample is generated can impact the resulting analysis. Since the data for this problem are images with infinitely many quantifiable points, the teams needed to design a method for generating a data sample. At this point, they did not have specific quantitative data about the pixel values on the images nor did they have a statistical procedure for analyzing the sample. They had a scale showing the correspondence between the gray scale shade of the point and the height at that point (see Figure 2). In a later follow-up task, they were provided with a data file containing the pixel values and the corresponding height of the surface at each pixel. For the task analyzed in this paper, they had to estimate the shade of each pixel from the color bar. Table 2 provides an overview and description of the codes used to describe the teams' sampling methods, as well as the number of teams whose response to the task received each code. Most teams recognized the need for randomness in the sample in order to avoid skewed data. However, there are multiple methods for randomly generating a sample. Many teams used randomly selected points or lines to generate a structure on the image. The sampling context used in this task is unique in the sense that students are not usually asked to generate a sampling method before defining measures to compute based on the sample. Additionally, the data points in statistics tasks

are often obviously discrete objects (e.g., people, trials, measurements) rather than continuous surfaces.

Table 2. Sampling codes describing students' method for sampling data and number of teams (N=35) whose response to the task received each code.

Code	Description	Number of Responses
Adjust the scale	Making an adjustment in the data set (e.g., by only using a portion of the image) for the difference in scale between images or convert the scale.	12
Random points	Selecting some number of points on the image randomly as data points	10
Drawing a grid over the image	Draw a grid on the image either to create subsets of data within the cells or along the gridlines	8
Random lines	Drawing random line(s) on the image	7
Eyeball method	Just "looking" at the picture (e.g., finding the peaks that look the biggest) to pick data points	7
Whole picture	Using every point of data on the image	4
Note the scale	Noting that the scale is different, but no adjustments within the procedure to account for the difference	4
Cross-section	Taking a slice of the image and using height data only from the particular slice.	3
Random area	Drawing a random area	1

Adjust the scale

Adjust the scale represented the teams' recognition that the images provided were not the same size, and therefore, represented different size populations. Students tended to deal with this in one of two ways: taking subsets of the data in equivalent pieces (e.g., Team B below) or asking the company to only give them samples taken on the same scale. For the method of taking subsets, many teams did not give explicit directions on how to do this, only stating that it needed to be done, while others told the company how to do it (e.g., Team A below). Relevent excerpts from team responses are as follows:

> Team A: "Since the different graphs are in different units, we must convert the scales to the same units before commencing calculations…In order to improve the procedure, we need all of the pictures to have the same scale so we can get rid of the conversion process."

> Team B: "Put the heights of a certain area into data points. This area should be one that could be universally used. For the sake of simplicity, we are using the area of 1 [square] micrometer."

Random Points

Many teams indicated the need for selecting points from the image. If teams had a method for selecting points, they were coded in a manner that indicated the method. Teams that received the random points code either did not indicate how to select the points or indicated that a computer would select the points.

> Team C: "Use a computer imaging program and the AFM images to determine the height of the surface using a randomly selected statistical sampling of no fewer than 100 data points."

> Team D: "We would approximate the height of 30 different random points on each graph."

Draw a Grid over the Image

Some teams chose to draw grid lines on the images as a method to either sample data points at the intersections of the grid lines or to subdivide the images into cells. If a team used the grid lines to divide the image into cells, they often employed another method of sampling within each cell. As will be discussed in the statistical measures section, teams often aggregated results from individual cells.

> Team E: "You would start by breaking the AFM image into a grid. Then find the height at each of the grid intersection points and store them in a data file."

> Team F: "The first step is to divide each of the samples into tiny, individual sections that measure 0.25 micrometers by 0.25 micrometers." (Note – this team went on to perform calculations on a sample of points inside each square of the grid.)

Random Lines

The *random line* code was given to teams who chose to lay lines on the image as a method to collect a sample. Several teams indicated that this was a way to account for different size images. It is also a method used by engineers when calculating similar kinds of measures. As with the random points code, the students attended to the need for randomness in a sample.

> Team G: "To determine the roughness of a sample, we would first draw a number of lines across the sample. Obviously, the more lines drawn would result in a more accurate approximation, but also take more time. The lines should be drawn in a ratio to the scale for each sample. For example, if we draw 1 micrometer lines on a 2 micrometer by 2 micrometer sample, we would then draw lines of 3 micrometers on a 6 micrometer by 6 micrometer sample."

> Team H: "Randomly drawn lines of random lengths are placed on the sample. The length of each line is measured. An interval for how often a measurement of height [is taken] is determined by the scale of the axis divided by 10."

Eyeball Method

The *eyeball method* was a code assigned to procedures that required "looking at the graph" to determine how to proceed. This was in ineffective method since it wasn't clear what to look at on the graph and there was no quantifiable procedure to provide a clear method for differentiation between samples (i.e., teams who employed it scored a 1 or a 2 on the Quality Assurance Guide). For instance, Team I used the idea of consistent colors in the image but did not provide a definition of consistency that could be used repeatedly by different users.

> Team I: "Our team looked at the contrast and consistency of the color in each image. The roughness of the gold was found by using the contrast of each sample. The higher colors were whiter and the darker colors were lower. We decided to use this method because if the colors were consistent in the image then the sample was generally smooth. If there were drastic color differences in the sample, it showed that the sample is rough."

Less Common Codes

There are four other codes for sampling that emerged from our data, but were not as common as the codes listed above. The code *whole picture* was given when a team provided a method for using what they perceived as all of the data in the image. Some teams indicated that they would use the grayscale pixel information (even though they were not told that this was an option) as the data points for the image. This was an effective strategy. Whereas, others just noted that it was necessary to collect all of the heights. This was significantly less effective. The code *note the scale* was given to teams whose procedure recognized the fact that the scales were different, but did not make adjustments for these differences. *Cross-section* was a code given to teams whose procedure took a "slice" of the image and used the height data just from that particular slice. The code *random area* represented a procedure that found a random area size within the image and then used the data within that area to continue with their measure of roughness.

4.2.2 MEASURES OF CENTRAL TENDENCY

In defining roughness, the teams tended to provide a single numerical representation of the height of their samples of gold. The teams were seeking to represent the typical height of the surface and explain how the typical height related to the roughness of the surface. This resulted in the use of measures of central tendency. As with other measures, the task did not indicate what measures the students needed to compute and any measures students found were elicited by the task. As noted in Table 3, a large majority of teams computed the mean.

Table 3. Measures computed as part of the procedure for data analysis and the number of teams (N=35) whose response to the task received each code.

Code	Number of Responses
Mean	22
Standard Deviation	23
Maximum/minimum/range	9
Histograms	6
Median	3
Informal measures for spread (e.g., modified standard deviation)	2
Mode	2

That students were drawn to the mean is not surprising. The same phenomenon occurred in other statistical model-eliciting tasks in engineering (Hjalmarson, 2007). This is consistent with the literature indicating students do know how to compute the mean, (e.g., Pollatsek, Lima, & Well, 1981) even if they don't understand the meaning or, in this case, haven't defined roughness in a way that calculating the mean would make sense. Computing the mean for this task is a complex endeavor that starts with asking: the mean of what (e.g., peaks, valleys)? Students may also compute means of multiple subsets of the data and then need to aggregate the values in some fashion. They would then need to decide if having a high or low mean is an indicator of greater roughness. This, of course, depends on the answer to the first question.

The complexity in the task is not in the computation but in determining what to compute and how to interpret the results particularly in light of other measures (e.g., standard deviation). Samples with the same mean could have very different appearance (e.g., a sample with consistent heights would have the same mean as a surface with an approximately equal number of high and low heights). As shown in the first coding schema, the students used various sampling schemes to create a data set. Interpreting the results usually meant they looked for low values of the mean of the heights in their sample. The student work in Table 4 shows a variety of ways that teams employed the mean in their procedures, starting with effective solutions and moving toward ineffective solutions as measured by the Quality Assurance Guide (QAG). None of the teams that scored a one on the QAG used any measure of central tendency of heights in their solutions. There were five teams that used median or mode to represent the central tendency of the heights of the gold samples. All three teams that used the median also used the mean in their solutions.

Table 4. Examples of teams' sampling methods and how they employed the mean.

Team	Sampling Method	Use of the Mean
Team H (QAG Score = 5)	This team laid random lines and devised a method for how often to measure height along each line	"The average (mean) height of the [sampled] points for each line is determined and the average height of the lines for each sample is determined. "
Team G (QAG Score = 4)	This team laid random lines and devised a method for how often to measure height along each line.	From this recorded data, we could then calculate the mean height across the line. Having taken the measures for several different lines, we could assume that to be the mean height of the entire sample."
Team J (QAG Score = 3)	This team laid a grid and collected their sample data from the intersections of the grid lines.	"Calculate the mean of the samples."

Team K (QAG Score = 2)	This team laid a grid and then used the eyeball method within each grid to measure "bumps."	"Then [we] estimated the average height of the bumps on the surface per square nanometer, by using the height scale provided. Then we used our data to figure out the average height of the bumps on the picture."

It is worth stating that use of the mean in and of itself did not indicate whether or not a procedure for measuring roughness was effective or not. The manner in which the teams sampled and how they interpreted the mean were better indicators of the quality of their solution. In order to generate a measure of roughness, the students needed to move beyond central tendency (since in isolation the results are ambiguous) and toward other measures. All but two of the teams that used the mean also had some measure of variability. The two teams that did not use a traditional measure of variability in their solutions were Team H (QAG Score: 5) and Team K (QAG Score: 2).

4.2.3 MEASURES OF VARIABILITY

Shaughnessy (2007) distinguishes between variability (likelihood of change) and variation (measurement of change). For instance, students could be analyzing variability between samples or the variation in a data set. The task served as an introduction to thinking about variation and variability of a data set as well considerations for determining a quantitative measure of the variation or variability. "Roughness" is, in and of itself, variation in a surface. Engineers have multiple, context-dependent methods for quantifying that variation or roughness. The task also required students to analyze variability between images. Most of the groups calculated a standard deviation as part of their data analysis. However, finding maximum and minimum values was another method students used to describe the variation in the surface. For example, one student team wrote, "Know the maximum and minimum heights of each image. Measure the height of each peak and valley of each line and find the average of those heights." What is important to note is that the task elicited these constructs from the students. Nowhere in the problem statement was it prescribed that students calculate any particular measure or that they should use statistical methods at all.

Most groups moved beyond measures of central tendency to measures of variability in the surface. Eighteen of the groups computed both the standard deviation and the mean. Their use of the statistics varied. It was not necessarily the case that a team created a sample, computed the mean for the values in the sample and then computed the standard deviation (though many groups did). As discussed previously, they had different methods for determining the sample data to use, and there were subtle but important distinctions in how the groups first determined a sample and then calculated statistics. Their ways of thinking about how to quantify variability interacted with their sampling methods. For example, some groups used a local-global approach to the data. They first found the standard deviation for a subset of the image and then aggregated across subsets to determine a value that represented the whole image. This may have been accomplished with an area model (i.e., subdividing the image into regions) or with a line model (i.e., drawing random lines on the image) in order to find subsets of the data as discussed previously. For example, one group wrote, "The standard deviation of the height of the material of each line would then be determined. Using these standard deviations of heights, the average of all the standard deviations of heights could be used to determine a total average standard deviation of the whole surface given."

Variability in a data set is a measure of how different values in the data are from each other. Some groups interpreted this variability by finding the range, maximum, or minimum values. This interpretation focuses on the extreme values in the data set. Some groups calculated the standard deviation for the sample. They interpreted a larger standard deviation to mean that the surface was rougher. As an example of this, one team wrote, "Once we have all our data points we would take the standard deviation of the heights. So that gives us how far the data points are away from the mean, therefore if a surface has a high standard deviation then it has a high roughness because there is a greater change in surface height." It was not always clear if the students understood what the standard deviation indicated about a data set or whether they were calculating it because it was a natural choice after the mean was computed. However, knowing that the standard deviation should be larger for rougher samples is one indication that the students understood that the standard deviation measures the spread of the data set relative to the mean. A higher mean would indicate taller bumps in the surface. A higher standard deviation would indicate greater variation in the bumps.

When considering variability in the nanoroughness context, it is important to ask "variability of what?" There are at least two interpretations of variability in the context. The first looks at how tall the peaks are or how low the valleys are and attempts to quantify the spread between them. The second interpretation examines variation in peak height. One requires quantifying a range from maximum and minimum values. The other requires quantifying the consistency in peak height. For example, one group wrote "…look at how many 'bumps' there are and their size. We can compare the colors of the pictures obtained with the color bars, if the images have surfaces that are of mostly similar colors, then we can conclude that they are mostly of similar heights (since similar colors represents similar heights) and when all of the particles are of similar heights, they should make up a nice even surface."

5. DISCUSSION

Without prompting, the task elicited students' conceptions of sampling and variability within a context where these two concepts were naturally intertwined. Students needed to consider how to measure variability by first considering what population was varying, how to generate a sample of that population and then how to quantify the variability. The students generated different statistics (e.g., mean, standard deviation) and then created procedures for aggregating and interpreting the outcomes. "Variability of what?" was a foundational question. The students had to both generate a procedure and interpret the results of their model. The two components of the procedure were sampling and quantifying the variability in the surface. Both of these tasks are somewhat unique in that statistics typically emphasizes discrete populations (e.g., people, objects) rather than measuring the characteristics or properties of a material (a fairly common engineering task). We have divided our discussion of these results into implications for teaching and research to describe how the task could be used in the classroom and areas for further investigation into students' understanding of sampling and variability.

5.1 IMPLICATIONS FOR TEACHING

Part of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) project, funded by the American Statistical Association (ASA), was the development of six recommendations for the teaching of introductory college statistics courses (see Franklin & Garfield, 2006). The six GAISE recommendations are: (1) Emphasize

statistical literacy and develop statistical thinking; (2) Use real data; (3) Stress conceptual understanding rather than mere knowledge of procedures; (4) Foster active learning in the classroom; (5) Use technology for developing conceptual understanding and analyzing data; and (6) Integrate assessments that are aligned with course goals to improve as well as evaluate student learning. The Nanoroughness task provides statistics instructors with an activity that meets all of these guidelines. The entire Nanoroughness Task can be found at http://modelsandmodeling.net. In this task, students are immersed in a meaningful, real-world problem based on actual images of gold surfaces. Students are engaged in a task that is similar to problems confronted by professional statisticians (Wild & Pfannkuch, 1999). The task requires students to use or construct appropriate measures of center and variability, and to build a model from these measures to address a problem. Sampling schemes need to be devised in order to deal with the large amount of information in each image. The task assesses students' understanding of statistical concepts such as center and variability in that correct conceptions are needed to produce viable models. And the task, itself, provides a window into students' understanding of these statistical concepts, as well as information that can be used to remedy misunderstandings and misconceptions.

Another instructional feature of the Nanoroughness task is that it naturally elicits the use of statistical measures and the need for taking a sample. The samples of gold do not differ with respect to the average height of the pixels. Yet, the three samples can be distinguished visually. This requires students to come up with a measure to estimate roughness, and a measure of variability appears to be a natural choice. The Nanoroughness task could follow instructional sessions on the standard deviation, providing a natural extension of the concept and measure to a natural setting.

The results indicate that this activity can be used to identify misunderstandings that students have about measures of center and variability. About half of the teams did not identify appropriate units of analysis, measures of center, or measures of variability, and their interpretation of what the standard deviation represents and how it related to the concept of smoothness was not well-reasoned. These students' methods could provide starting points for helping them to develop a better understanding through activities that require them to operationalize their methods and to test if their methods actually identify the smoothness of each sample in a reasonable way. For example, one team proposed calculating the mean deviation of each point in the sample, and summing the mean deviations as a measure or roughness. The sample with the lowest sum would be the smoothest. However, the sum of mean deviations is always zero, so this measure would not distinguish the three samples. Testing the method could provide a springboard for exploring what a mean deviation represents and guided discussion could be used to develop a deeper understanding of the mean and the standard deviation.

The task also provides students with an opportunity to apply sampling schemes if they are covered prior to the task. The teams came up with different methods for sampling from the populations. An extension to the activity would be to have students discuss and compare the different sampling methods. This can be used to develop students' understanding of bias in sampling, the issue of representativeness, and how large a sample needs to be to provide an accurate assessment of a model. Students could analyze the results from different sampling methods under the same operational definition of nanoroughness. Questions that could be addressed are: Do the different sampling methods produce similar results? Are some sampling methods better than others (and under what criteria)? If so, what makes them better? Addressing these questions could lead to discussions of randomness, when a method uses randomness and when it does not, and whether random sampling produces a more representative sample than other methods.

Issues of sample size could also be explored (e.g., How many points do you need to provide an accurate estimate of the nanoroughness for a piece of material? Is there a minimum size? Is there a sample size above which accuracy does not improve appreciably?)

Another set of activities that can follow naturally after the Nanoroughness task are sessions that explore sampling distributions, or distributions of measures of nanoroughness from different samples. These follow-up activities could be designed to help students explore whether different samples from the same population produce the same or similar estimates of nanoroughness and if sample size is related to the variability in estimates of nanoroughness from different samples. In the same spirit as an MEA, students could be asked to design methods for answering these questions and to evaluate the effectiveness of the various methods. These tasks would provide additional practice with applying concepts such as sampling, random selection, and the distribution of a variable, extending this concepts from a single sample to multiple samples.

5.2 IMPLICATIONS FOR RESEARCH

This study provides evidence that the Nanoroughness MEA naturally elicits application of concepts such as measures of center, variability, and sampling to a modeling task. The evidence indicates that students take several different approaches in applying these statistical concepts, and that the task produces artifacts that provide a window into students' conceptual understanding. These findings raise several questions that should be addressed in future research.

Most teams used some type of measure of variability, and many used the standard deviation. The activity asks teams to evaluate their own models and the models of other teams. This could lead to a better understanding and appreciation of variability in data. What we do not know is the nature of students' understanding of variability, and more specifically of the standard deviation, both before and after completing the Nanoroughness task. Similarly, students who participate in the Nanoroughness task can be expected to develop a better understanding and appreciation for sampling and sampling methods as a result of critiquing the sampling methods used by different teams. Items and tasks from research studies on students' understanding of variability and the standard deviation (e.g., Chance et al., 2004; delMas & Liu, 2005; C Reading & Shaughnessy, 2004; Shaughnessy et al., 1999) and of sampling methods (Watson & Kelly, 2005, 2006) could be administered prior to and after students participate in the MEA to determine if changes in their understanding and thinking do occur.

It would also be informative to conduct a comparative study where all students receive the same initial instruction on measures of center of variability, but are then randomly assigned to either receive additional instruction on these topics or to participate in the Nanoroughness MEA. The additional instruction could cover the same number of class sessions as the MEA, and engage the students conceptually (e.g., applications of the concepts in a variety of contexts to promote a deeper understanding and transfer). Comparison of assessment results would address the question of whether or not the MEA is more effective in developing students conceptual understanding of these topics.

The design principles used to develop an MEA imply that participation in an MEA should increase the likelihood of transfer. MEAs include many of the conditions that have been shown to increase retention and transfer of knowledge and problem-solving to new contexts: solve carefully designed problems; develop familiarity with each context; confront students' misconceptions and intuitions; help students see similarities and differences; guide students to find the general principle behind the example; emphasize

deep (relational or structural) features over surface features; promote the development of mental frameworks for connecting information (Schwartz, 2004; Schwartz, Varma, & Martin, 2008). Questions of whether or not students who participate in the Nanoroughness MEA have better retention and are more likely to develop effective solutions to analogous problems need to be addressed.

6. REFERENCES

Carmona-Dominguez, G. (2004). *Designing an Assessment Tool to Describe Students' Mathematical Knowledge*. Purdue University.

Chamberlin, S. A. (2002). *Analysis of interest during and after model eliciting activities: a comparison of gifted and general population students*. Purdue University.

Chance, B., delMas, R. C., & Garfield, J. B. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking.* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

delMas, R. C., Garfield, J. B., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. . *Journal of Statistics Education*, *7*(3). Retrieved from http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm .

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

Doerr, H. M., & English, L. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for Research in Mathematics Education*, *34*(2), 110-136.

Franklin, C. A., & Garfield, J. B. (2006). The GAISE project: Developing statistics education guidelines for grades pre-K-12 and college courses. In G. Burrill & P. Elliott (Eds.), *Thinking and reasoning with data and chance: 2006 NCTM yearbook* (pp. 345-376). Reston, VA: National Council of Teachers of Mathematics.

Garfield, J. B., delMas, R. C., & Chance, B. (1999). The role of assessment in research on teaching and learning statistics.

Groth, R. E. (2005). An investigation of statistical thinking in two different contexts: Detecting a signal in a noisy process and determining a typical value. *The Journal of Mathematical Behavior*, *24*(2), 109-124.

Hjalmarson, M., Diefes-Dux, H., & Moore, T. (2008). Designing model development sequences for engineering. In J. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and modeling in engineering education: Designing experiences for all students* (pp. 37-54). Rotterdam, The Netherlands: Sense Publishers.

Hjalmarson, M. A. (2007). Engineering students designing a statistical procedure for quantifying variability. *The Journal of Mathematical Behavior*, *26*(2), 178-188.

Hoerl, R., & Snee, R. (2001). *Statistical Thinking: Improving Business Performance*. Pacific Grove, CA: Duxbury Press.

Kelly, B. A., & Watson, J. M. (2002). Variation in a chance sampling setting: The lollies task. In B. Barton, K. Irvin, M. Pfannkuch, & M. J. Thomas (Eds.), *Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia: Mathematics education in the South Pacific*. Sydney, Austrlaia: MERGA.

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, *2*(1), Article 1.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289.

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 151-167). Voorburg, The Netherlands: International Statistical Institute.

Lehrer, R., & Schauble, L. (2003). Origins and evolution of model-based reasoning in mathematics and science. In R. A. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: A models & modeling perspective on mathematics problem solving, learning & teaching* (pp. 59-70). Mahwah, NJ: Erlbaum.

Lesh, R., & Clarke, D. (2000). Formulating operational definitions of desired outcomes of instruction in mathematics and science education. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp. 113-149). Mahwah, NJ: Lawrence Erlbaum.

Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for developing thought-revealing activities for students and teachers. In A. E. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 591-646). Mahwah, New Jersey: Lawrence Erlbaum.

Meletiou-Mavrotheris, M., & Lee, C. (2002). Teaching students the stochastic nature of statistical concepts in an introductory statistics course. *Statistics Education Research Journal [Online]*, *1*(2). Retrieved from http://fehps.une.edu.au/serj.

Moore, D. S. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of Giants* (pp. 95-138). Washington D.C.: National Academies Press.

Moore, T. (2006). *Student team functioning and the effect on mathematical problem solving in a first-year engineering course*. Purdue University.

Moore, T.J. (2008). Model-Eliciting Activities: A case-based approach for getting students interested in material science and engineering. *Journal of Materials Education*, 30(5-6), 295 - 310.

Moore, T., & Diefes-Dux, H. (2004). Developing model-eliciting activities for undergraduate students based on advanced engineering content. In *34th ASEE/IEEE Frontiers in Education Conference*. Savannah, Georgia.

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12*, 191-204.

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

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 314-319). Voorburg, The Netherlands: International Statistical Institute.

Schwartz, D. L. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129-184.

Schwartz, D. L., Varma, S., & Martin, L. (2008). Dynamic transfer and innovation. In S. Vosniadou (Ed.), *Handbook of conceptual change* (pp. 479-506). Mahwah, NJ: Erlbaum.

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In *Proceedings of the Twentieth Annual Conference of the*

Mathematics Education Resarch Group of Australasia (pp. 6-22). Rotorua, NZ: University of Waikata.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (2nd ed., Vols. 1-2, Vol. 2, pp. 957-1009). Charlotte, NC: Information Age Publishing.

Shaughnessy, J. M., Watson, J. M., Moritz, J. B., & Reading, C. (1999). School mathematics students' acknowledgment of statistical variation.

Snee, R. (1990). Statistical thinking and its contribution to quality. *The American Statistician*, *44*, 116-121.

Torok, R., & Watson, J. M. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal*, *9*, 60-82.

Watson, J. M., & Kelly, B. A. (2005). Cognition and instruction: Reasoning about bias in sampling. *Mathematics Education Research Journal*, *17*(1), 24-57.

Watson, J. M., & Kelly, B. A. (2006). Expectation versus variation: Students' decision making in a sampling environment. *Canadian Journal of Science, Mathematics and Technology Education*, *6*, 145-166.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, *34*(1), 1-29.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, *47*(2), 289-312.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, *67*, 223-265.

Zawojewski, J., Hjalmarson, M., Bowman, K., & Lesh, R. (2008). A modeling perspective on learning and teaching in engineering education. In J. S. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and modeling in engineering education: Designing experiences for all students* (pp. 1-16). Rotterdam, The Netherlands: Sense Publishers.