

A Learning Progression for Variability

ETS RR–20-05

James H. Fife
Kofi James
Stephanie Peters

December 2020



ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

John Mazzeo
Distinguished Presidential Appointee

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Senior Research Scientist

Tim Davey
Research Director

John Davis
Research Scientist

Marna Golub-Smith
Principal Psychometrician

Priya Kannan
Managing Research Scientist

Sooyeon Kim
Principal Psychometrician

Anastassia Loukina
Senior Research Scientist

Gautam Puhan
Psychometric Director

Jonathan Schmidgall
Research Scientist

Jesse Sparks
Research Scientist

Michael Walker
Distinguished Presidential Appointee

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

A Learning Progression for Variability

James H. Fife, Kofi James, & Stephanie Peters

Educational Testing Service, Princeton, NJ

The concept of variability is central to statistics. In this research report, we review mathematics education research on variability and, based on that review and on feedback from an expert panel, propose a learning progression (LP) for variability. The structure of the proposed LP consists of 5 levels of sophistication in understanding variability, ranging from the learning that occurs prior to Grade 6 through an expert understanding of variability. Following our analysis of the variability research, the full LP is presented along with example tasks designed to elicit evidence of understanding at each of the proposed levels. The LP described in this report constitutes a new theoretical structure that must be independently validated vis-à-vis empirical recovery of the proposed levels by analyzing student responses to tasks designed to target different levels of the progression.

Keywords Distribution; extreme values; learning progression; mathematics; measures of spread; proportional reasoning; random; range; sample; statistics; variability

doi:10.1002/ets2.12286

The notion of variability is central to statistics (Cobb & Moore, 1997; Garfield & Ben-Zvi, 2005; Reading & Shaughnessy, 2000, 2004; Torok & Watson, 2000; Watson et al., 2003; Watson & Kelly, 2002). According to Watson and Kelly (2002), “Variation is at the heart of all statistical investigation. If there were no variation in data sets, there would be no need for statistics” (p. 1). Cobb and Moore (1997) marked the relationship between mathematics and statistics with the observation that the need for statistics “arises from *the omnipresence of variability*” (p. 801). In the Common Core State Standards for Mathematics (CCSSM), the first standard in the statistics and probability domain, the Grade 6 standard 6.SP.1, requires that students “recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers” (National Governors Association Center for Best Practices & Council of Chief State School officers [NGA/CCSSO], 2010, p. 45). According to the American Statistical Association’s *Guidelines for Assessment and Instruction in Statistics Education* (GAISE), “It is this focus on *variability in data* that sets apart statistics from mathematics” (Franklin et al., 2007, p. 6). An understanding of variability is essential for an understanding of statistics and for understanding the distinction between statistics and other areas of mathematics.

Nonetheless, prior to the seminal work of Shaughnessy et al. (1999), there was little research on students’ understanding of variability (Shaughnessy, 1997, 2007). Shaughnessy (1997) suggested that this lack of research could have been due to the lack of importance attributed to variability and measures of spread in school curricula, in which most of the emphasis was on measures of central tendency. Shaughnessy further suggested that the hesitancy of curriculum developers and teachers to teach measures of spread was due in part to the belief that teaching measures of spread, and therefore teaching variability, meant teaching standard deviation, whose definition is difficult to motivate and whose computation is cumbersome. Although the CCSSM begins its statistics standards in Grade 6 with the concept of variability, standard deviation is not introduced in the standards until high school. Prior to high school, spread and variability are measured using interquartile range and mean absolute deviation (NGA/CCSSO, 2010). Similarly, the GAISE propose a framework for K–12 statistics education in which spread is measured at the lowest level (Level A) using range and at the second level (Level B) using mean absolute deviation. Standard deviation is not introduced until Level C (Franklin et al., 2007). The CCSSM and GAISE each provide a framework for introducing the concepts of variability and measures of spread without dwelling on the technical definition and detailed computation of standard deviation.

In this report, we analyze the concept of variability and, based on our review of the relevant literature together with existing sets of standards and guidelines and feedback from an expert panel, present a proposed learning progression (LP) for variability. A *learning progression* is “a sequence of successively more complex ways of thinking about an idea that might

Corresponding author: J. H. Fife, E-mail: jfife@ets.org

reasonably follow one another in a student’s learning” (Smith et al., 2004, p. 5). In the context of the Educational Testing Service (ETS) *CBAL*[®] learning and assessment tool (Bennett, 2010; Bennett & Gitomer, 2009), an LP was described as

a description of qualitative change in a student’s level of sophistication for a key concept, process, strategy, practice, or habit of mind. Change may occur due to a variety of factors, including maturation and instruction, and each progression is presumed to hold for most, but not all, students. As with all scientific research, the progressions are open to empirical verification and theoretical challenge. (ETS, n.d., List Item 2)

An LP identifies what Clements and Sarama (2004) have called “developmental progression[s] of levels of thinking” (p. 83). Each level of an LP “characterizes a phase of student thinking en route to target performance” (Graf et al., 2019, p. 1). The rationale behind LP development is to provide a map that can guide assessment design, instructional practice, and student learning (Graf et al., 2019). While LPs in related domains will have points of connection between them, thus creating networks of learning paths, the developmental levels are best understood by recognizing individual LPs belonging to individual domains (Daro et al., 2011).

The current work constitutes the first step toward creating an LP for variability, based on our analysis of existing research in the domain of variability in K–12 mathematics education (Mislevy et al., 2003; Riconscente et al., 2016) and feedback from subject matter experts. The proposed LP must be independently validated, a procedure that involves a series of iterative steps (Graf & van Rijn, 2016, 2019). These steps include analyzing student responses to tasks designed to target different levels of the LP; this analysis can involve cognitive interviews with a small sample of students or a more formal psychometric analysis of responses using an appropriate psychometric model. This process is iterative; after each stage of the analysis, it may be appropriate to revise the LP or the tasks targeted at different levels. If the analysis of the LP confirms the specification and the ordering of the levels of the LP, then the LP can be examined for its efficacy as a teaching tool; this examination can include an analysis of teachers’ interpretations of the LP and the effectiveness of their implementation of the LP. The framework we propose here will serve as the foundation for subsequent validation steps that may inform revisions to the LP and/or to tasks designed to assess a student’s performance within the LP. This LP for variability and LPs for probability (Mejia Colindres & Peters, 2019) and data display (Kim & Oláh, 2019) together provide a road map that can guide learning in three key areas that contribute to statistical literacy.

To obtain a list of articles for our literature review, we conducted a search of the literature using EBSCOhost, JSTOR, and Google Scholar. For a search term, “variability” was too general, producing far too many irrelevant references. The search term “reasoning about variability,” however, was more targeted and returned approximately 20 high-quality articles. We followed up with searches on the authors of these articles, articles in the reference lists of these articles, and articles that referenced these articles. What emerged from these searches was a sequence of articles, beginning with Shaughnessy et al. (1999), that traced the progress of the research on students’ understanding of variability.

One of these papers (Reading, 2004) appeared in a special issue of *Statistics Education Research Journal* devoted to papers dealing with research on reasoning about variability. The articles in this issue and in a subsequent issue on the same topic—10 articles altogether—presented a picture of the state of the art, at that time, of research on reasoning about variability. Several of these articles covered research that is relevant to the development of a LP about variability.

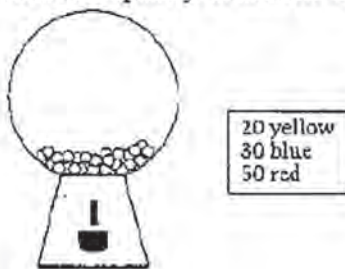
We also reviewed two documents that set forth general sets of standards for statistics education: the GAISE report (Franklin et al., 2007) and the statistics portion of the CCSSM (NGA/CCSSO, 2010). Finally, the work of Peters (2011) on the meaning of a robust understanding of variation provided the basis for Level 5 of our LP.

Initial Research on Students’ Understanding of Variability

Gumballs and Lollies

The impetus to undertake the line of research that began with Shaughnessy et al. (1999) was an analysis of student responses to an item on the 1996 National Assessment of Educational Progress (NAEP; see Shaughnessy, 2007); the item is shown in Figure 1. By asking for the most likely number of red gumballs, the item reflected the then prevalent curricular emphasis on central tendency over variability. In an analysis of all the statistics items from the 1996 NAEP mathematics assessment, Zawojewski and Shaughnessy (2000) observed that in a sample of 232 students, only one student gave a response to this item that was a range of possible numbers of red gumballs instead of a single number.

Think carefully about the following question. Write a complete answer. You may use drawings, words, and numbers to explain your answer. Be sure to show all of your work.



The gumball machine has 100 gumballs; 20 are yellow, 30 are blue, and 50 are red. The gumballs are well mixed inside the machine.

Jenny gets 10 gumballs from this machine.

What is your best prediction of the number that will be red?

Answer: _____ gumballs

Explain why you chose this number.

Figure 1 The gumball item on the 1996 National Assessment of Educational Progress mathematics assessment. From *National Assessment of Educational Progress, 1996, National Center for Education Statistics* (<https://nces.ed.gov/nationsreportcard/nqt>). In the public domain.

Task Design

This observation led Shaughnessy et al. (1999) to conduct an experiment in which the gumball task was revised so as to encourage students to think about a range of outcomes instead of a single outcome and then was administered to 324 students in Grades 4–12 in the United States and Australia. In the revision, gumballs were changed to hard candies for the US students and to lollies for the Australian students. (In Australia, a hard candy is called a “lollie.”) This revised task has come to be known as the lollies task. An abridged version of the revised task (with lollies) is shown in Figure 2. The number of lollies and the proportions of colors are the same as those in the NAEP item, and as with the NAEP item, the revised task asks for the number of red lollies in a sample of 10 lollies randomly drawn from the bowl of 100 lollies. The revised task, however, includes additional questions about what will happen if the sample of 10 lollies is taken six times (with the lollies being returned to the bowl after each sample is taken). Three forms of the response are requested: (a) a list of the likely number of red lollies in each sample, (b) a selection of what the student thinks is the most likely list from a series of choices, and (c) an indication of the probable range of the number of red lollies. For each of these responses, the student is asked why the student’s response is the most likely. The task then moves to two more difficult questions: one dealing with a larger sample size and the other dealing with a larger number of draws. These questions in the revised task were designed to force students to confront the fact that, while the average number of red lollies in a large number of samples may equal five, the actual number of red lollies will vary from sample to sample.

Student Responses

Shaughnessy et al. (1999) found that students gave a variety of responses to the tasks. Some gave responses that indicated they expected numerous possibilities for the number of reds that could occur in a sample of 10 lollies, giving responses such as 0, 1, 4, 7, 9, and 10 reds. Shaughnessy et al. interpreted these responses to mean that the students perhaps thought that each possible number of reds was equally likely or that “anything can happen” in a chance experiment. Other students gave responses, such as 6, 7, 8, 8, 7, and 9, that overpredicted the number of reds; these students tended to give explanations that focused on the large number of reds in the bowl (50 reds) instead of the proportion of reds (one half reds). Students who focused on the proportion of reds were more likely to make appropriate predictions about both the center of the distribution of reds and the spread. Finally Shaughnessy et al. noticed that Grade 12 students who had had a probability course were more likely to respond 5, 5, 5, 5, 5, 5; Shaughnessy and his colleagues conjectured that these students were

Student Response Form

1A) Suppose we have a bowl with 100 lollies in it. 20 are yellow, 50 are red, and 30 are blue. Suppose you pick out 10 lollies.
How many reds do you expect to get? ____
Would this happen every time? Why?

1B) Altogether six of you do this experiment.
What do you think is likely to occur for the numbers of red lollies that are written down?
Please write them here.
____, _____, _____, _____, _____, _____
Why are these likely numbers for the reds?

1C) Look at these possibilities that some students have written down for the numbers they thought likely. Which one of these do you think best describes what might happen?
a) 5, 9, 7, 6, 8, 7
b) 3, 7, 5, 8, 5, 4
c) 5, 5, 5, 5, 5, 5
d) 2, 3, 4, 3, 4, 4
e) 7, 7, 7, 7, 7, 7
f) 3, 0, 9, 2, 8, 5
g) 10, 10, 10, 10, 10, 10
Why do you think the list you chose best describes what might happen?

1D) Suppose that 6 students did the experiment—pulled out ten lollies from this bowl, wrote down the number of reds, put them back, mixed them up.
What do you think the numbers will most likely go from? From ____ (low) to ____ (high) number of reds.
Why do you think this?

1E) Suppose that 6 students each pulled out 50 lollies from this bowl, wrote down the number of reds, put them back, mixed them up.
What do you think the numbers will most likely go from this time?
From ____ (low) to ____ (high) number of reds.
Why do you think this?

1F) Suppose that 40 students pulled out 10 lollies from the bowl, wrote down the number of reds, put them back, mixed them up. Can you describe what the numbers would be, what they'd look like?
Why do you think this?

Figure 2 Abridged lollies task. From “Student Perceptions of Variation in a Sampling Situation,” by C. Reading and M. Shaughnessy, in T. Nakahara and M. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, p. 4 - 90), 2000, International Group for the Psychology of Mathematics. Copyright 2000 by C. Reading and M. Shaughnessy.

accustomed to answering questions—such as, “What is the probability that ... ?”—that require single-number answers rather than a range of possible outcomes.

Theoretical Frames: The Work of Torok and Watson

In their 1999 study, Shaughnessy and his colleagues had evaluated written student responses. Expanding on this work, Torok and Watson (2000) explored student understanding of variability in more depth by conducting in-person interviews, based in part on the lollies task, with students in Australia. They interviewed 16 students: two boys and two girls from each of Grades 4, 6, 8, and 10. The students were given two versions of the lollies task—one with 50 red, 20 yellow, and 30 green lollies and one with 70 red, 20 yellow, and 10 green lollies—and they were given two additional tasks that

focused on realistic situations in which there were nonrandom sources of variation. For the lollies task, after the students responded to the tasks, they were asked to draw six handfuls of lollies from a bowl of lollies and to record the number of reds in each handful. They were then given the opportunity to modify their answers to the earlier questions.

On the basis of the students' responses to these tasks, Torok and Watson (2000) identified four "levels of developing concepts of variation" (p. 153). Their descriptors of these levels are shown in Figure 3. The names given to the levels themselves are not particularly enlightening, but the descriptors show a gradual progression in understanding of variability. The authors observed that students whose understanding was at the lowest level (Level A) tended to focus on individual outcomes and predicted too much variation, similar to Shaughnessy et al.'s (1999) suggestion that some students may have thought that each possible number of reds was equally likely. Students with an understanding at this level tended to view the average as the mode; for example, when a fourth-grade student was asked to explain the statement that the average height of a group of students was 130 cm, he said that meant that most of the students were 130 cm tall. These students were also unduly influenced by experimental outcomes and often reasoned from unrelated factors instead of the proportion of lollies of each color. Finally, these students "never volunteered expressions relating to variation and showed poor knowledge of terms associated with variation when specifically asked" (Torok & Watson, 2000, p. 157). For example, one student could not explain the meaning of "average maximum temperature" or how it could be calculated.

Torok and Watson (2000) found that students with a Level B understanding "readily acknowledged variation" (p. 157) and usually provided ranges of numbers instead of specific values in response to the lollies task. They were more likely to use proportional reasoning to some extent but were likely to produce responses with too much or too little variation. According to Torok and Watson (2000), "This appeared to stem from a conflict between proportional ideas (e.g., half red or 50% red) and alternative ideas (e.g., most reds or more reds)" (p. 157). This is consistent with Shaughnessy et al.'s (1999) observation that some students seemed more focused on the number of reds than on the proportion of reds. Finally, students with an understanding at this level "seldom referred to variation explicitly without prompting but had reasonable knowledge of terms associated with variation when asked" (Torok & Watson, 2000, p. 158).

Students with a Level C understanding had stronger proportional reasoning skills but were sometimes led by those skills to produce ranges that were too narrowly clustered about the mean. "Their use of language ... was more sophisticated ... [and they] had a strong knowledge of terms associated with variation" (Torok & Watson, 2000, pp. 159–160). For example, students whose understanding was at this level were able to make proper use of terms like *random* and *average*. Only two of the 16 students exhibited Level D thinking; both of these students gave appropriate responses to all of the questions, displaying an appropriate balance between variation and clustering. Like the students whose understanding was at Level C, the students who exhibited Level D thinking "had a strong knowledge of terms associated with variation" (Torok & Watson, 2000, p. 160).

Continuing the Research

Expanding on the work of Shaughnessy et al. (1999) and Torok and Watson (2000), Watson et al. (2003) prepared an assessment instrument with 16 multipart items that covered chance variation and data variation as well as sampling variation. They administered their assessment to 746 students in Grades 3, 5, 6, and 9 in 10 public schools in Australia. Based on the student responses to these 16 items, Watson et al. identified four "levels of increasing understanding" (p. 11) and gave the following names and descriptors for these levels:

Level 1: Prerequisites for Variation (p. 11)

- Students are likely to justify responses with stories or personal experiences.
- Students recognize variation only in the simple context of "not looking the same every day" (p. 11) or in describing a surprising outcome.
- Students cannot interpret graphs and tables.
- Students give numerically inappropriate responses to questions involving chance.

Level 2: Partial Recognition of Variation (p. 12)

- Students generally use unquantified statements, such as "anything can happen" (p. 12), to justify chance outcomes.
- Students are likely to make flawed interpretations of graphs.
- The terms *sample*, *random*, and *variation* are "likely to be familiar but students have difficulty expressing the concepts in words" (p. 12).

<p><i>Level A: Weak appreciation of variation</i></p> <ul style="list-style-type: none"> • Acknowledge variation • Provide responses that suggest a very weak understanding of proportional ideas • Focus on individual outcomes without consideration of the set • May refer to the average as the most common individual value • Provide answers with inconsistent degrees of variation and clustering • Are easily swayed by experimental results • Do not produce meaningful summary graphs (for 40 draws) • Never refer to variation explicitly, show poor knowledge of variation terminology • Have poor general knowledge of real world situations <p><i>Level B: Isolated appreciation of aspects of variation and clustering</i></p> <ul style="list-style-type: none"> • Readily acknowledge variation • Provide responses that suggest a very weak understanding of proportional ideas • May provide answers in terms of sub-ranges or specific values • May refer to the average as a value within a range of common values • May provide answers with consistently too much or too little variation and clustering • Are moderately swayed by experimental results • Generally attempt summary graphs but do not produce meaningful ones (for 40 draws) • Never refer to variation explicitly, have reasonable knowledge of variation terminology • Have variable knowledge of real world situations <p><i>Level C: Inconsistent appreciation of variation and clustering</i></p> <ul style="list-style-type: none"> • Readily acknowledge variation • Exhibit strong proportional thinking and may provide responses that imply representativeness, such as the “perfect” sample of 5 red, 2 yellow, and 3 green • Provide answers in terms of specific outcomes in the context of a set of outcomes • May provide answers with consistently too much or too little variation and clustering • Are only slightly influenced by experimental results • Produce equivalent of time series graphs to summarise data • Explicitly refer to variation, may have strong knowledge of variation terminology • Have basic general knowledge of real world situations <p><i>Level D: Good, consistent appreciation of variation and clustering</i></p> <ul style="list-style-type: none"> • Readily acknowledge variation • Provide responses that suggest a conflict between proportional ideas; or exhibit strong proportional thinking and provide responses that imply representativeness • Provide answers as specific outcomes in the context of a set of outcomes • Consistently provide answers with an appropriate level of clustering • Are only slightly influenced by experimental results • Produce frequency or time series graph to summarise data • Explicitly refer to variation, usually have strong knowledge of variation terminology • Have good general knowledge of real world situations

Figure 3 Descriptors of student performance for the four levels of developing concepts of variation. From Torok and Watson (2000), figure 2. Copyright 2000 by Springer Nature. Reprinted with permission.

Level 3: Applications of Variation (p. 12)

- In explaining situations involving variation and sampling, students will “focus on some appropriate aspects of the concepts while ignoring or being misled by others” (p. 12). Examples include the following:
 - Find the mean of a data set but not appreciate the importance of the variation in the data.
 - Provide partial analysis of graphs while missing overall trends.
 - In analyzing chance contexts, demonstrate variation but not the appropriate degree of variation (either too much or too little).
 - When selecting samples, focus on representative sampling methods (sampling methods that produce outcomes that might be considered typical) or chance sampling methods (sampling methods that produce outcomes purely by chance), but not both simultaneously.

- Students attempt to give definitions of *sample*, *random*, and *variation*, but their definitions “do not achieve a high level of sophistication” (p. 13), perhaps relying on examples to explain the terms.

Level 4: Critical Aspects of Variation (p. 13)

- Students understand the importance of variability as well as central tendency in analyzing data.
- Students are likely to make statistically appropriate analyses of graphs.
- In analyzing chance contexts, students will demonstrate appropriate variation.
- In sampling situations, students are likely to detect sources of bias, such as nonrepresentativeness, “as well as make appropriate suggestions on their own” (p. 13).
- Students are likely to give sophisticated definitions of the terms *sample*, *random*, and *variation* without relying on examples to explain the terms.

Concurrent with the work of Torok and Watson (2000) and Watson et al. (2003), Reading and Shaughnessy (2000, 2004) conducted in-person interviews with students to gain deeper understanding about students’ responses to the original lollies task (Figure 2). They conducted interviews with 12 students from Australia (six from primary school—Grades 4, 5, and 6—and six from secondary school, in Grades 9 and 12). The students were asked to respond to the questions in the lollies task, though only the one Grade 12 student was asked to respond to Questions E and F. Based on their analysis of the student responses, Reading and Shaughnessy (2004) concluded that most of the responses could be placed into one of two groups—those that attempted to describe the variation and those that attempted to explain the cause of the variation. (A few responses fell into both groups.) This led Reading and Shaughnessy (2004) to develop two hierarchies of development for variability, one for description and one for causation.

Description Hierarchy

- Level 1: *Concern with either middle values or extreme values.* Students describe the distribution either by focusing on the unlikelihood of obtaining a lot of extreme values or the high likelihood of obtaining a lot of middle values, but not both (p. 214).
- Level 2: *Concern with both middle values and extreme values.* Students describe the distribution by mentioning the likely extreme values (e.g., the maximum and the minimum) and by describing what is happening in between (p. 215).
- Level 3: *Discuss deviations from an anchor (not necessarily central).* Students describe the distribution in terms of deviations from an anchor value that is not a central value (p. 216).
- Level 4: *Discuss deviations from a central anchor.* Students describe the distribution in terms of deviations from a central anchor (p. 216).

Causation Hierarchy

- Level 1: *Identify extraneous causes of variation.* Students gave causes of variation that focused on physical properties of the sampling, such as the visibility of red lollies versus the other colors or the location in the bowl of the various colors (p. 217).
- Level 2: *Discuss frequencies of colors as a cause of variation.* Students gave causes of variation that focused on the number of reds rather than the proportion of reds, for example, a student might think that 50 is a lot of red lollies and so there must be a lot of red lollies in the *sample* without considering that 50 lollies is just one half of the lollies in the bowl (p. 218).
- Level 3: *Discuss proportions of colors as a cause of variation.* Students gave causes of variation that focused on the proportion of reds and other colors, for example, students might use the fact that one half of the lollies were red to assert that they would expect one half of the lollies in the sample to be red (p. 218).
- Level 4: *Discuss likelihoods based on proportions.* Students gave causes of variation that attempted to infer the likelihood of obtaining a certain color from the proportion of that color in the bowl, for example, a student might say that obtaining five lollies in the sample was a more likely outcome than obtaining three or eight lollies (p. 219).

Reasoning about Variability: Two Special Issues of *Statistics Education Research Journal*

The early 2000s initiated a wave of further research on students' developing understanding of variability. In July 2003, for example, reasoning about variability was the theme of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy. The presentations and discussions at that conference led to the 2004 publication of a special issue of *Statistics Education Research Journal*, devoted entirely to six articles that grew out of that conference (Bakker, 2004; Ben-Zvi, 2004; Ben-Zvi & Garfield, 2004; Gould, 2004; Hammerman & Rubin, 2004; Reading, 2004). Four additional papers from the conference (delMas & Liu, 2005; Garfield & Ben-Zvi, 2005; Makar & Confrey, 2005; Pfannkuch, 2005) were published in the May 2005 issue of the journal. Several of these authors discussed research relevant to the development of an LP for variability.

Ben-Zvi: "Reasoning About Variability in Comparing Distributions"

For example, Ben-Zvi (2004) observed how two seventh-grade students developed an understanding of variability in the course of comparing two distributions—in this case, the number of letters in the surnames of students in a US class and in an Israeli class. Ben-Zvi identified seven stages through which the students progressed:

Stage 1: *On what to focus: Beginning from irrelevant and local information.* Students focus on irrelevant aspects of the data (e.g., three of the American names begin with "Mc"; p. 48).

Stage 2: *How to informally describe the variability in raw data.* Students make informal statements that do not take into account the relevant variability in the data (e.g., US surnames are longer than Israeli surnames; p. 49).

Stage 3: *How to formulate a statistical hypothesis that accounts for the variability.* Students modify informal statements to account for the variability (e.g., US surnames are usually longer than Israeli surnames, but not always; p. 50).

Stage 4: *How to account for variability when comparing groups using frequency tables.* Students support their claims with references to specific features of the distributions but sometimes struggle with variability in the data (e.g., US names are longer than Israeli names because in Hebrew words are usually written without vowels; p. 51).

Stage 5: *How to use center and spread measures to compare groups.* Students begin to use measures of center (mean, median, mode) and spread (range) to compare the two distributions, but their responses may seem procedural. They do not seem to understand what the measures mean or what the distinction is between measures of center and measures of spread (p. 53).

Stage 6: *How to model variability informally through handling outlying values.* Students do not understand the meaning of *outlier*, thinking that "outlier" means one of the least frequent values (p. 54).

Stage 7: *How to notice and distinguish the variability within and between the distributions in a graph.* Students generate graphical displays of the data and use them to compare the distributions. Students sometimes have difficulty interpreting a graph displaying both distributions, being uncertain which features of the graph are relevant (p. 55).

Owing to the small sample size and the limited nature of the study, Ben-Zvi (2004) warned against broad generalizations of its results. But he did suggest three "learning phenomena" (p. 60) that students may experience:

- Students' prior knowledge [may] be engaged in interesting and surprising ways, possibly hindering progress in some instances but making the basis for construction of new knowledge in others,
- many questions [may] make little sense to the students, or, alternatively, will be reinterpreted and answered in different ways than intended, and
- students' work [may] inevitably be based on partial understandings, which will grow and evolve. (p. 60)

Reading: "Student Description of Data While Working With Weather Data"

Reading (2004) investigated the extent to which the Reading and Shaughnessy (2004) description hierarchy, developed in the context of sampling variability, could be applied to student learning in the context of data variability and inference making. A total of approximately 65 students in Grades 7, 9, and 11 were presented with a month's worth of rainfall and temperature data for their hometown, with each student receiving data for a different month. The students were asked to describe the weather in their town for the month for which they had data.¹ Reading then classified

the student responses according to the Reading and Shaughnessy hierarchy. Most of the responses were classified as Level 1 or Level 2, describing variation using extreme values, middle values, or both. Reading observed, however, that while some of the responses described variation numerically, other responses described variation exclusively in words. She labeled responses with no numeric descriptions as “qualitative” and responses containing numeric descriptions as “quantitative.”

Reading (2004) concluded that the qualitative/quantitative distinction in the responses was more sophisticated than the Reading and Shaughnessy (2004) Level 1 or Level 2 distinction. As a result, she replaced Levels 1 and 2 of the Reading and Shaughnessy hierarchy with two new levels, qualitative and quantitative. Then she applied the structure of the observed learning outcomes (SOLO) taxonomy framework (Biggs & Collis, 1982) to each level and used the framework to classify the responses into three sublevels—unstructural responses, multistructural responses, and relational responses. Unstructural responses gave one description (qualitative or quantitative) to summarize the variability in the data, multistructural responses gave more than one description, and relational responses provided a link between multiple descriptions. Quantitative relational responses usually linked extreme and middle values; qualitative relational responses were uncommon. According to Reading (2004), the responses at the qualitative level were structurally similar to responses in Levels 1 and 2 of the Reading and Shaughnessy hierarchy but were “expressed in the less statistically mature qualitative form” (p. 97).

Reading (2004) also compared her hierarchy with the Watson *et al.* (2003) levels of understanding. She equated her qualitative level with Watson *et al.*'s Level 1 and her quantitative level with Watson *et al.*'s Level 2. She also suggested that Levels 3 and 4 of the Reading and Shaughnessy (2004) hierarchy correspond to Levels 3 and 4 of the Watson *et al.* levels of understanding.

Makar and Confrey: “Articulating Meaning in Statistics”

Research on teachers and on instructional interventions has provided additional insight into how the concept of variability is acquired. Makar and Confrey (2005) discussed how preservice teachers use nonstandard language to talk about variability. The authors observed 17 preservice secondary mathematics and science teachers enrolled in a one-semester course on assessment. The subjects were asked “to compare the relative improvement of test scores between two groups of students” (Makar & Confrey, 2005, p. 47). The preservice teachers could use standard statistical terms to describe the distributions of test scores, but they also used nonstandard terminology when the standard terms did not express the thoughts they were trying to convey:

The diversity and richness of their descriptions of variation and distribution demonstrated that the prospective teachers found many ways to discuss these concepts, and that through their nonstandard language they were able to articulate keen awareness of variation in the data. (Makar & Confrey, 2005, p. 47).

delMas and Liu: “Exploring Students’ Conceptions of the Standard Deviation”

delMas and Liu (2005) conducted a study of college students in which a computer program was used to help students develop an understanding of standard deviation. As noted earlier, the concept of standard deviation is difficult to motivate and messy to compute, and if standard deviation is taught in elementary statistics courses, students likely develop only a procedural understanding. According to delMas and Liu (2005),

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. A conceptual model of the standard deviation is needed to develop instruction that promotes the concept. (p. 56).

For delMas and Liu (2005), a conceptual understanding of standard deviation requires, among other things, an understanding of three concepts—distribution, mean, and deviation from the mean. This is reasonable, since the standard deviation is defined to be the (square root of) the mean of the (squared) deviations from the mean of the individual data points. Based on the results of their study, they concluded that students had a range of conceptual understandings of standard deviation. delMas and Liu (2005) noted that some students had ideas “inconsistent

with a coherent conception of the standard deviation” (p. 79), whereas others had ideas that “capture some relevant aspects of variation and the standard deviation, but may represent a cursory and fragmented level of understanding” (p. 79). Still others had ideas that “represent much closer approximations to an integrated understanding [of standard deviation]” (delMas & Liu, 2005, p. 79). Finally, some students “demonstrated an ability to coordinate the effects of several operations on the value of the standard deviation, an indication of a more integrated conception” (delMas & Liu, 2005, p. 79).

Garfield and Ben-Zvi: “A Framework for Teaching and Assessing Reasoning About Variability”

Finally, as a conclusion to the collection of articles dealing with reasoning about variability, Garfield and Ben-Zvi (2005) listed what they believed to be the seven key areas of conceptual understanding of variability²:

1. *Developing intuitive ideas of variability.* Recognize that variability is everywhere. Some things vary more than others, but we can try to understand the causes of the variability.
2. *Describing and representing variability.* Graphs of data can show the variability in the data, with different graphs sometimes showing different aspects of the variability. Different numerical summaries (e.g., range, standard deviation, interquartile range) can tell us different things about the variability, with different summary statistics being more useful for different types of variation.
3. *Using variability to make comparisons.* When comparing two or more data sets, it is important to distinguish between the variability within each group and the variability between groups.
4. *Recognizing variability in special types of distributions.* Understand the properties of normal distributions.
5. *Identifying patterns of variability in fitting models.* Understand how to determine how well a model fits data by looking at the variability of the deviations of the data from the model.
6. *Using variability to predict random samples or outcomes.* Understand properties of sample variability. Large samples vary more than small samples, but the sample statistics from large samples vary less than sample statistics from small samples.
7. *Considering variability as part of statistical thinking.* Any statistical investigation must always consider the variability of the data.

Garfield and Ben-Zvi (2005) stated that while the order of the ideas is

increasingly sophisticated, progress in students’ construction of meanings is not linear but rather complex and is better captured by the image of spiral progression. . . . Ideas related to variability must be constantly revisited along the statistics curriculum from different points of view, context and levels of abstraction, to create a complex web of interconnections among them. (p. 95)

The GAISE Report

In 2007, the GAISE report (Franklin et al., 2007) was released with the endorsement of the American Statistical Association. As mentioned earlier, the report established a framework for K–12 statistics education. The framework identifies four steps in statistical problem solving and indicated the role of variability in each step: (a) formulate questions—*anticipate variability*; (b) collect data—*acknowledge variability*; (c) analyze data—*account for variability*; and (d) interpret results—*allow for variability*. The framework then describes how each of these four steps can develop over three levels. For example, as mentioned previously, students with a Level A understanding use the range of data to measure the spread; at Level B, students may use mean absolute deviation; and at Level C, they may use standard deviation.

The three levels in the GAISE framework correspond roughly to grade levels, but specific alignments are not made. While Level A might correspond to elementary school, Level B to middle school, and Level C to high school, the authors made the point that if a middle school student has not been exposed to statistical concepts in elementary school, then the middle school student’s understanding will begin at Level A. Similarly, if a high school student has not had Level A and Level B exposure before high school, then he or she will not immediately begin with a Level C understanding.

Table 1 Elements and Reasoning Indicative of Robust Understanding of Variation

Element	Perspective		
	Design perspective	Data-centric perspective	Modeling perspective
Variational disposition	Acknowledging the existence of variability and the need for study design	Anticipating reasonable variability in data	Anticipating and allowing for reasonable variability in data when using models
Variability in data for contextual variables	Using context to consider sources and types of variability to inform study design or to critique study design	Describing and measuring variability in data for contextual variables as part of exploratory data analysis	Identifying the pattern of variability in data or the expected pattern of variability for contextual variables
Variability and relationships among data and variables	Controlling variability when designing studies or critiquing the extent to which variability was controlled in studies	Exploring controlled and random variability to infer relationships among data and variables	Modeling controlled or random variability in data, transformed data, or sample statistics
Effects of sample size on variability	Anticipating the effects of sample size when designing a study or critiquing a study design	Examining the effects of sample size through the creation, use, or interpretation of data-based graphical or numerical representations	Anticipating the effects of sample size on the variability of a sampling distribution

Note. Adapted from Peters (2011), figure 13. Copyright 2011 by International Association for Statistical Education.

A Robust Understanding of Statistical Variation

Subsequent work by Peters (2011) considered additional aspects of variation whose understanding is required for what she called a robust understanding of variation. She interviewed 16 statistics teachers as they solved three variation tasks. On the basis of the data she collected, she identified indicators of a robust understanding of variability and classified these indicators according to four aspects of variability and three perspectives of reasoning about variability. The four aspects of variability are (a) variational disposition, (b) variability in data for contextual variables, (c) variability in relationships among data and variables, and (d) effects of sample size. *Variational disposition* includes creating design strategies for collecting data that acknowledge or anticipate variability. It also includes recognizing unreasonable variation, perhaps due to a data entry error. *Variability in data for contextual variables* includes interpreting summary measures of variability (e.g., standard deviation for univariate data, correlation coefficient for bivariate data) and fitting models to data. *Variability in relationships among data and variables* involves strategies to control variability when designing studies; it also involves the understanding of the distinction between controlled and random variability. Finally, *effects of sample size* involve understanding the effect that the size of a sample has on the variability in the sample and on the variability in the statistics used to characterize the sample.

For Peters (2011), a robust understanding of variability involves the integrated understanding of these four elements across three perspectives: a design perspective, a data-centric perspective, and a modeling perspective. The indicators of this understanding are summarized in Table 1.

Expert Panel Review

A preliminary version of this work was reviewed in Summer 2017 by a panel of experts in the field of statistics; two of these experts were professors of mathematics education at research universities, and one was an assessment development specialist at ETS. The panelists reviewed the LP and provided feedback based on a set of guiding questions aimed at assisting us in refining the LP. For example, we asked the members of the panel if the descriptions of the levels in the progression made meaningful cognitive and mathematical distinctions and if they made useful instructional distinctions, if there were gaps in the sequencing of the levels, and if the panelists had additional suggested modifications to the progression. We also asked the panelists to comment on the example tasks at each level, what grade range each level might represent, and how we might communicate with teachers regarding the progression.

After reviewing the panel members' written responses to our questions, we conducted a virtual meeting with the panel in Fall 2017. Following the panel's recommendations, we made some changes to the progression. We incorporated some additional research, especially English and Watson (2016), and we included a more detailed discussion of the GAISE (Franklin et al., 2007). We eliminated a detailed discussion of measures of central tendency, although we retained some references to misconceptions regarding these measures in the LP. We removed standard deviation from the lower levels of the LP, because in both CCSSM and GAISE, this concept is not introduced until late (high school in CCSSM and Level C in GAISE), and we revised the example tasks to better align with the levels of the LP.

A Learning Progression for Variability

While the authors we discussed developed their own descriptors of the levels many students pass through as they gain an understanding of variability, there is a good deal of commonality among these ideas. And while many of these authors are based in Australia and conducted their research with Australian students, the mathematics standards in the Australian Curriculum (Assessment and Reporting Authority, 2010) are sufficiently similar to the CCSSM that we can combine their research with the research of the US-based experts to develop our own LP for variability; see Table 2. This hypothetical LP has five levels; the first four are based on the work of Shaughnessy et al. (1999), Torok and Watson (2000), Watson et al. (2003), Reading and Shaughnessy (2000, 2004), and others. For three of these levels, we have taken the names given by Watson et al.; for Level 1, we thought "Naive Understanding of Variability" was a more descriptive title than Watson et al.'s "Prerequisites for Variation." Our Level 5 incorporates the work of Peters (2011) on robust understanding of variability. The full LP is presented in Table 2; an overview of the five levels follows:

- Level 5: Robust Understanding of Variability.
- Level 4: Critical Aspects of Variability.
- Level 3: Applications of Variability.
- Level 2: Partial Recognition of Variability.
- Level 1: Naive Understanding of Variability.

When aligned with the CCSSM, Levels 1–4 correspond to an understanding of variability appropriate for middle or high school students. Research by English and Watson (2016), however, has suggested that students as early as fourth grade can gain an understanding of variability through the administration of carefully planned tasks, and the GAISE report (Franklin et al., 2007) suggests that elementary school students can use the range of a set of data as a measure of its spread. Lehrer and Kim (2009) found that students in Grades 5 and 6, when engaged in modeling data, can recognize and devise measures of variability. Hence, in the proposed LP, we have aligned Level 1 with grades earlier than Grade 6. Level 2 aligns with Grade 6, Level 3 aligns with Grade 7, and Level 4 aligns with high school. Level 5 corresponds to an understanding that might be approached by an advanced university statistics student or in-service statistics teacher. These alignments, however, are only intended to demonstrate how the levels of our LP correspond with standards in the CCSSM. One must also keep in mind the admonition in the GAISE report. A Grade 6 student who has had no previous instruction in statistics will begin at Level 1 before progressing to Level 2, and high school students will not be able to begin Level 4 until they have mastered Level 3 (Franklin et al., 2007, p. 13).

Students with a Level 1 understanding in the proposed LP often provide arguments that are based on an idiosyncratic understanding of variability or on their previous experiences, and not on an analysis of the data. This was observed by Torok and Watson (2000), who remarked that students at their Level A often reason from factors unrelated to the distribution; by Watson et al. (2003), who observed that Level 1 students are likely to justify responses with stories about their personal experiences; and by Ben-Zvi (2004), for whom students at his Stage 1 focus on irrelevant aspects of the data.

But not all naive understandings are due to the use of irrelevant information. While Shaughnessy et al. (1999) did not construct developmental levels, they observed several stages in student understanding (or misunderstanding). In particular, they observed that some students expected a wide range of possible numbers of red lollies in a sample of 10 lollies, perhaps because students with a low level of understanding believe that the various possible outcomes are equally likely, that anything can happen in a chance experiment. This is consistent with Torok and Watson's (2000) observation that students at their Level A often predict too much variation, and it accounts in part for Watson et al.'s (2003) observation that students at their Level 1 give responses that are numerically inappropriate. Watson et al. also found that students at their Level 2 may make unquantified statements, such as "anything can happen," to justify their conclusions.

Table 2 A Learning Progression for Variability

Level	What students can and cannot do	Tasks appropriate for this level
5 Robust understanding of variability <i>University students and in-service teachers</i>	<ul style="list-style-type: none"> • Variational disposition^a <ul style="list-style-type: none"> • Acknowledge the existence of variability and the need for study design. • Anticipate reasonable variability in data. • Anticipate and allow for reasonable variability in data when using models. • Variability in data for contextual variables <ul style="list-style-type: none"> • Use context to consider sources and types of variability to inform and critique study design. • Describe and measure variability in data for contextual variables as part of exploratory data analysis. • Identify the pattern of variability in data or the expected pattern of variability for contextual variables. • Variability in relationships among data and variables <ul style="list-style-type: none"> • Control variability when designing studies or critique the extent to which variability was controlled in studies. • Explore controlled and random variability to infer relationships among data and variables. • Model controlled or random variability in data, transformed data, or simple statistics. • Effects of sample size <ul style="list-style-type: none"> • Anticipate the effects of sample size when designing a study or critiquing a study design. • Examine the effects of sample size through the creation, use, or interpretation of data-based graphical or numerical representations. • Anticipate the effects of sample size on the variability of a sampling distribution. 	<p>The final exam for an abstract algebra class included a question asking for a proof; the question was worth 15 points. The professor for the class asked two of his graduate students to score the proofs, each scoring the proofs for half of the students. The graduate students each used the same rubric to score the proofs, but the professor was concerned that they were not interpreting the rubric in the same way. The average score of the proofs scored by one of the graduate students was 9.7, while the average score of the proofs scored by the other graduate student was 10.3. What can the professor conclude about the scores assigned by the two graduate students? What additional information does the professor need to draw a meaningful conclusion about the scoring of his graduate students? Additional information might include the following:</p> <ul style="list-style-type: none"> • The distribution of scores assigned by each graduate student • The standard deviation of the scores assigned by each graduate student • The range of scores assigned by each graduate student • Whether or not the papers were assigned at random to the graduate students for grading • The number of scores at each score point assigned by each graduate student

Table 2 Continued

Level	What students can and cannot do	Tasks appropriate for this level
4 Critical aspects of variability <i>High school students</i>	<ul style="list-style-type: none"> Students at this level exhibit proficiency in drafting arguments based on proportional reasoning. For example, when asked to predict the number of red candies in a sample drawn from a known population, they would focus their attention on the proportion of red candies instead of the number; as a result, they would be likely to make appropriate predictions about both the center and the spread of a distribution. Students will demonstrate in their responses an appropriate degree of variation with an appropriate balance between variation and clustering. Students understand the importance of variability as well as the importance of central tendency in characterizing a data set. Students are able to give sophisticated definitions of technical terms such as <i>sample</i>, <i>random</i>, and <i>variation</i> (Watson et al., 2003). Students can give a statistically appropriate analysis of graphs; they can generate graphical displays of data and use them to compare two or more distributions (Ben-Zvi, 2004; Watson et al., 2003). Students are likely to detect sources of bias in samples, such as nonrepresentativeness. Students can describe a distribution in terms of deviations from a central value (mean, median, or mode; Reading, 2004; Reading & Shaughnessy, 2000, 2004; Watson et al., 2003). Students have a conceptual understanding of standard deviation as a measure of variability. Students can translate between graphical, numeric, and symbolic representations of distributions. 	<p>The following histograms represent the age distributions of the two countries.^b</p> <p>Age Distribution in the United States</p> <p>Age Distribution in Kenya</p> <ol style="list-style-type: none"> How do the shapes of the two histograms differ? Approximately what percentage of people in Kenya in 2010 were between the ages of 0 and 10 years? Approximately what percentage of people in the United States in 2010 were between the ages of 0 and 10 years? Approximately what percentage of people in Kenya in 2010 were between the ages of 70 and 100 years? Approximately what percentage of people in the United States in 2010 were between 70 and 100 years? The population of Kenya in 2010 was approximately 41 million people. Approximately how many people in Kenya were between the ages of 0 and 10 years? Between 60 and 100 years? If you had visited a city in Kenya in 2010, do you think you would likely have seen many teenagers? Would you likely have seen many people over 70 years old? Explain your answers based on the histogram.^b



Table 2 Continued

Level	What students can and cannot do	Tasks appropriate for this level																																														
3	<ul style="list-style-type: none"> Students at this level will have some facility with proportional reasoning but will sometimes produce responses with too much or too little variation; the range may be insufficiently clustered around the mean or may be too tightly clustered around the mean (Torok & Watson, 2000; Watson et al., 2003). Students at this level can calculate the measures of central tendency (mean, median, and mode) and spread (range and mean absolute deviation) but may not appreciate the importance of variation. Their understanding of these measures is procedural and not conceptual; they do not understand what the measures mean or what the distinction is between measures of center and measures of spread (Ben-Zvi, 2004; delMas & Liu, 2005; Watson et al., 2003). Students at this level may attempt to define technical terms like <i>sample</i>, <i>random</i>, and <i>variation</i>, but they may rely on examples to explain the meaning of the terms (Watson et al., 2003). Students may be able to provide a partial analysis of graphs while missing overall trends (Watson et al., 2003). When selecting samples, students may focus on representativeness or randomness, but not both. A distribution may be described in terms of deviations from an anchor value that is not a central value (Reading, 2004; Reading & Shaughnessy, 2000, 2004; Watson et al., 2003). Students may not understand what an outlier is, thinking that an outlier is the least frequent value. 	<p>Below are the heights of the players on the University of Maryland women's basketball team for the 2012–2013 season and the heights of the players on the women's field hockey team for the 2012 season.</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Field Hockey Player Heights (inches)</th> <th>Basketball Player Heights (inches)</th> </tr> </thead> <tbody> <tr> <td>66</td> <td>64</td> </tr> <tr> <td>62</td> <td>66</td> </tr> <tr> <td>65</td> <td>63</td> </tr> <tr> <td>62</td> <td>64</td> </tr> <tr> <td>70</td> <td>64</td> </tr> <tr> <td></td> <td>65</td> </tr> <tr> <td></td> <td>68</td> </tr> <tr> <td></td> <td>62</td> </tr> <tr> <td></td> <td>67</td> </tr> <tr> <td></td> <td>65</td> </tr> <tr> <td></td> <td>62</td> </tr> <tr> <td></td> <td>64</td> </tr> <tr> <td></td> <td>75</td> </tr> <tr> <td></td> <td>65</td> </tr> <tr> <td></td> <td>76</td> </tr> <tr> <td></td> <td>74</td> </tr> <tr> <td></td> <td>68</td> </tr> <tr> <td></td> <td>72</td> </tr> <tr> <td></td> <td>67</td> </tr> <tr> <td></td> <td>69</td> </tr> <tr> <td></td> <td>74</td> </tr> <tr> <td></td> <td>79</td> </tr> </tbody> </table>	Field Hockey Player Heights (inches)	Basketball Player Heights (inches)	66	64	62	66	65	63	62	64	70	64		65		68		62		67		65		62		64		75		65		76		74		68		72		67		69		74		79
Field Hockey Player Heights (inches)	Basketball Player Heights (inches)																																															
66	64																																															
62	66																																															
65	63																																															
62	64																																															
70	64																																															
	65																																															
	68																																															
	62																																															
	67																																															
	65																																															
	62																																															
	64																																															
	75																																															
	65																																															
	76																																															
	74																																															
	68																																															
	72																																															
	67																																															
	69																																															
	74																																															
	79																																															
Grade 7	<ul style="list-style-type: none"> Students at this level will have some facility with proportional reasoning but will sometimes produce responses with too much or too little variation; the range may be insufficiently clustered around the mean or may be too tightly clustered around the mean (Torok & Watson, 2000; Watson et al., 2003). Students at this level can calculate the measures of central tendency (mean, median, and mode) and spread (range and mean absolute deviation) but may not appreciate the importance of variation. Their understanding of these measures is procedural and not conceptual; they do not understand what the measures mean or what the distinction is between measures of center and measures of spread (Ben-Zvi, 2004; delMas & Liu, 2005; Watson et al., 2003). Students at this level may attempt to define technical terms like <i>sample</i>, <i>random</i>, and <i>variation</i>, but they may rely on examples to explain the meaning of the terms (Watson et al., 2003). Students may be able to provide a partial analysis of graphs while missing overall trends (Watson et al., 2003). When selecting samples, students may focus on representativeness or randomness, but not both. A distribution may be described in terms of deviations from an anchor value that is not a central value (Reading, 2004; Reading & Shaughnessy, 2000, 2004; Watson et al., 2003). Students may not understand what an outlier is, thinking that an outlier is the least frequent value. 	<p>Based on visual inspection of the data, which group appears to have the larger average height? Which group appears to have the greater variability in the heights?</p> <p>Compute the mean and mean absolute deviation (MAD) for each group. Do these values support your answers in part (a)?</p> <p>How many of the 12 basketball players are shorter than the tallest field hockey player?</p> <p>Imagine that an athlete from one of the two teams told you she needs to go to practice. You estimate that she is about 65 inches tall. If you had to pick, would you think that she was a field hockey player or that she was a basketball player? Explain your reasoning.</p> <p>The women on the Maryland field hockey team are not a random sample of all female college field hockey players. Similarly, the women on the Maryland basketball team are not a random sample of all female college basketball players. However, for purposes of this task, suppose that these two groups can be regarded as random samples of all female college field hockey players and all female college basketball players, respectively. If these were random samples, would you think that female college basketball players are typically taller than female college field hockey players? Explain your decision using answers to the previous questions and/or additional analysis.</p>																																														

Table 2 Continued

Level	What students can and cannot do	Tasks appropriate for this level
2	<p>Partial recognition of variability Grade 6</p> <ul style="list-style-type: none"> Students at this level are likely to have difficulty with proportional reasoning. For example, they may focus on the number of red candies in a population instead of the proportion of red candies. As a result, they may overestimate the number of red candies (Torok & Watson, 2000). Students are likely to use unquantified chance statements to describe outcomes, such as “Anything can happen” (Watson et al., 2003). Students may interpret the pattern in a distribution but may not make a specific reference to the variation exhibited in the distribution. Terms like <i>sample</i>, <i>random</i>, and <i>variation</i> are likely to be familiar, but students may not understand the precise mathematical meanings of these terms or may have difficulty expressing the mathematically accurate meanings of these terms in words (Watson et al., 2003). Students may make statements that informally account for variability. They may acknowledge variation by giving a range of numbers in response to a question instead of a specific value (Torok & Watson, 2000). Students will describe features of a distribution numerically. They can support their claims with references to specific features in the data (Reading, 2004). At this level, students will have difficulty describing a particular distribution. Students’ reasons do not reflect understanding of chance or variation. Students’ explanation of variability as displayed in graphs is apt to be flawed (Watson et al., 2003, p. 23). Students at this level can calculate mean absolute deviation and understand that it measures the variability in a population (NGA/CCSSO, 2010). 	<p>1. Suppose we have a bowl with 100 candies in it. 20 are yellow, 50 are red, and 30 are blue. Suppose you pick out 10 candies. How many reds do you expect to get? ____ Would this happen every time? Why?</p> <p>2. Altogether six of you do this experiment. What do you think is likely to occur for the numbers of red candies that are written down? Please write them here. ____, _____, _____, _____, _____, _____ Why are these likely numbers for the reds?</p> <p>3. Look at these possibilities that some students have written down for the numbers they thought likely. Which one of these do you think best describes what might happen?</p> <p>a. 5, 9, 7, 6, 8, 7 b. 3, 7, 5, 8, 5, 4 c. 5, 5, 5, 5, 5, 5 d. 2, 3, 4, 3, 4, 4 e. 7, 7, 7, 7, 7 f. 3, 0, 9, 2, 8, 5 g. 10, 10, 10, 10, 10</p> <p>Why do you think the list you chose best describes what might happen?</p> <p>4. Suppose that 6 students did the experiment—pulled out 10 candies from this bowl, wrote down the number of reds, put them back, mixed them up. What do you think the numbers will most likely go from? From _____ (low) to _____ (high) number of reds. Why do you think this?</p>

Table 2 Continued

Level	What students can and cannot do	Tasks appropriate for this level
<p>1 Naïve understanding of variability</p> <p><i>Before Grade 6</i></p>	<ul style="list-style-type: none"> Students at this level may reason from factors unrelated to the distribution, they may attempt to justify their responses with stories about personal experiences, or they may focus on irrelevant aspects of the data (Ben-Zvi, 2004; Torok & Watson, 2000; Watson et al., 2003). Students may overpredict variability, perhaps because they think that all possible outcomes are equally likely; they have the idea that “anything can happen” in a chance experiment (Torok & Watson, 2000; Watson et al., 2003). Students recognize variation only in the simple context of “not looking the same every day” or in describing a surprising outcome (Watson et al., 2003, p. 11). Students at this level will have a poor knowledge of technical terms associated with variation (Torok & Watson, 2000). Students may make informal statements that do not take into account the variability in the data. Students may describe features of a distribution in words and not numerically (Reading, 2004). Students at this level may confuse different measures of central tendency, such as mean and mode (Torok & Watson, 2000). Students may be able to read information from a graph but have difficulty interpreting information obtained from a graph or integrating information obtained from several graphs (Watson et al., 2003). Responses to questions involving chance are apt to be numerically inappropriate. Students at this level will not have developed measures of variability. 	 <p>1. A class used this spinner. Out of 50 spins, how many times do you think the spinner will land on the shaded part? Why do you think this?</p> <p>2. A class of students recorded the number of years their families had lived in their town. Here are two graphs that students drew to tell their story (Watson et al., 2003).</p> 
		<p>a. What can you tell by looking at graph 1?</p> <p>b. What can you tell by looking at graph 2?</p> <p>c. Which graph tells the story better? Why?</p>

^aThe descriptors and the task at this level are adapted from Peters (2011), figure 13. Copyright 2011 by International Association for Statistical Education. ^bAdapted from New York State Education Department (2013). CC BY-NC-SS.

Watson et al. (2003) found that students at Level 1 recognize variation only in the context of “not looking the same every day” (p. 11), while Torok and Watson (2000) observed that students at their Level A had a poor knowledge of technical terms associated with variation. They said that students at higher levels had a reasonable knowledge of technical terms, but Watson et al. (2003) clarified this by observing that students with a Level 2 understanding are familiar with technical terms (in particular, *sample*, *random*, and *variation*) but do not understand the meanings of these terms, whereas students with a Level 3 understanding attempt to give definitions of these terms but may resort to giving examples to explain their meanings. It is not until Level 4 that students can give sophisticated definitions of these terms without relying on examples (Watson et al., 2003).

Reading (2004) extended the Reading and Shaughnessy (2000, 2004) description hierarchy to apply to data variability as well as sample variability. She discovered that Levels 1 and 2 of the Reading and Shaughnessy (2004) hierarchy could be reorganized into a qualitative level and a quantitative level; the first corresponds to Watson et al.’s (2003) Level 1, and the second corresponds to Watson et al.’s Level 2. Students with a Level 1 understanding describe features of a distribution in words and not numerically, whereas students with a Level 2 understanding describe features numerically. This is consistent with the general description of Level 1 in our LP as representing a naive understanding of variability.

Difficulties or misconceptions common with students at Level 1 of our LP include the tendency to confuse the mean with the mode (thinking of the mean as the most common value; Torok & Watson, 2000), difficulty interpreting tables and graphs (Watson et al., 2003), and a lack of conceptual understanding of standard deviation (delMas & Liu, 2005).

Shaughnessy et al. (1999) and Torok and Watson (2000) both observed that, in the lollies task, some students focused on the number of red lollies (50) rather than the proportion of lollies (one half). As a result, they tended to overpredict the number of red lollies. Torok and Watson placed these students in their Level B, corresponding to our Level 2. Ben-Zvi (2004) observed that students at his Stage 3 (out of seven stages) made informal attempts to account for variability. This corresponds to Torok and Watson’s observation that students at their Level B acknowledge variation by providing a range of responses instead of specific values.

Students at Level 3 of our LP have some facility with the application of proportional reasoning to problems of variation but may produce responses with too much or too little variation. This observation is supported by the findings of Torok and Watson (2000) and Watson et al. (2003). Watson et al. also observed that Level 3 students can calculate the mean but do not appreciate the importance of variation. This is consistent with Ben-Zvi’s (2004) Stage 5, in which students can calculate measures of center and spread but do not have a conceptual understanding of these measures. It is also consistent with delMas and Liu’s (2005) finding that students at this level have a cursory and fragmented understanding of standard deviation.

Watson et al. (2003) found that students at Level 2 are likely to make flawed interpretations of graphs, students at Level 3 might provide a partial analysis of a graph while missing overall trends, and students at Level 4 are likely to make statistically appropriate analyses of graphs. This last finding is consistent with Ben-Zvi’s (2004) final Stage 7, in which students generate graphical displays of data and use them to compare distributions.

Reading and Shaughnessy (2000, 2004) distinguished between students who describe a distribution in terms of deviations from an anchor value that is not a central value and students who describe a distribution in terms of deviations from a central value. Reading (2004) identified the first group of students with Watson et al.’s (2003) Level 3 and the second group with Level 4, as we have done in our LP.

Finally, our Level 5 is adapted from Peters (2011); see Table 1. An understanding at this level constitutes an understanding of multiple aspects of variability and multiple contexts in which variability occurs. It also includes an understanding of the interconnections between variability and related concepts (Peters, 2011). It represents the level of understanding of variability that one would expect of advanced college students and in-service teachers.

Progressive Learning about Variability

Finally, the progression of ideas in the LP that we have developed is consistent with the seven key areas of conceptual understanding identified by Garfield and Ben-Zvi (2005). These key areas were themselves extracted from the articles in the special issues of the *Statistics Education Research Journal* and other research. But, as Garfield and Ben-Zvi pointed out, learning is not always linear but is often helical, and topics must be continually revisited from different points of view, different contexts, and different levels of abstraction. While an LP can provide a path for a student to follow in the student’s journey from naivete to conceptual understanding, it is a path that the student may often need to retrace.

Notes

- 1 The two sets of data (rainfall and temperature) were presented to the students separately, with a teaching episode in between. While there were about 65 students receiving each set of data, it was not exactly the same 65 students due to variability in class attendance.
- 2 In Garfield and Ben-Zvi (2005, pp. 93–95), each of the seven key areas is followed by three to six bullet points providing more explanation. We have quoted the seven key areas and summarized the bullet points for each area.

References

- Assessment and Reporting Authority. (2010). *Mathematics*. Australian Curriculum. <https://www.australiancurriculum.edu.au/f-10-curriculum/mathematics/>
- Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, 3(2), 64–83.
- Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, 8(2–3), 70–91. <https://doi.org/10.1080/15366367.2010.508686>
- Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). Springer. https://doi.org/10.1007/978-1-4020-9964-9_3
- Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, 3(2), 42–63.
- Ben-Zvi, D., & Garfield, J. (2004). Research on reasoning about variability: A foreword. *Statistics Education Research Journal*, 3(2), 4–6.
- Biggs, J. B., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press.
- Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, 6, 81–89. https://doi.org/10.1207/s15327833mtl0602_1
- Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly*, 104, 801–823. <https://doi.org/10.1080/00029890.1997.11990723>
- Daro, P., Mosher, F. A., & Corcoran, T. B. (2011). Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction. Consortium for Policy Research in Education. <https://doi.org/10.12698/cpre.2011.rr68>
- delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, 4(1), 55–82.
- Educational Testing Service. (n.d.). *The CBAL® mathematics competency model and provisional learning progression*. <https://www.ets.org/cbal/mathematics/>
- English, L. D., & Watson, J. M. (2016). Development of probabilistic understanding in fourth grade. *Journal for Research in Mathematics Education*, 47(1), 28–62. <https://doi.org/10.5951/jresmetheduc.47.1.0028>
- Franklin, C. A., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. American Statistical Association. <http://www.amstat.org/education/gaise/>
- Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, 4(1), 92–99.
- Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal*, 3(2), 7–16.
- Graf, E. A., Peters, S., Fife, J. H., van Rijn, P. W., Arieli-Attali, M., & Marquez, E. (2019). *A preliminary validity evaluation of a learning progression for the concept of function* (Research Report No. RR-19-21). Educational Testing Service. <https://doi.org/10.1002/ets2.12257>
- Graf, E. A., & van Rijn, P. (2016). Learning progressions as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 165–189). Routledge.
- Graf, E. A., & van Rijn, P. W. (2019, April 5–9). *Cycle for validating a learning progression* [Paper presentation]. Annual Meeting of the American Educational Research Association, Toronto, ON, Canada.
- Hammerman, J. K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal*, 3(2), 17–41.
- Kim, E. M., & Oláh, L. N. (2019). *Elementary students' understanding of geometrical measurement in three dimensions* (Research Report No. RR-19-14). Educational Testing Service. <https://doi.org/10.1002/ets2.12250>
- Lehrer, R., & Kim, M.-J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal*, 21(2), 116–133. <https://doi.org/10.1007/BF03217548>
- Makar, K., & Confrey, J. (2005). “Variation-talk”: Articulating meaning in statistics. *Statistics Education Research Journal*, 4(1), 27–54.
- Mejia Colindres, C., & Peters, S. (2019). *A learning progression for probability* [unpublished manuscript].

- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02
- National Assessment of Educational Progress. (2006). *NAEP question tool*. ICES NCES. <https://nces.ed.gov/nationsreportcard/nqt>
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. http://www.corestandards.org/wp-content/uploads/Math_Standards1.pdf
- New York State Education Department. (2013). *Algebra 1 module 2 topic B lesson 8*. EngageNY. <https://www.engageny.org/resource/algebra-i-module-2-topic-b-lesson-8>
- Peters, S. A. (2011). Robust understanding of statistical variation. *Statistics Education Research Journal*, 10(1), 52–88.
- Pfannkuch, M. (2005). Thinking tools and variation. *Statistics Education Research Journal*, 4(1), 83–91.
- Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal*, 3(2), 84–105.
- 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). Kluwer. https://doi.org/10.1007/1-4020-2278-6_9
- Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Kyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89–96). Psychology of Mathematics Education. <http://www.igpme.org/publications/current-proceedings/>
- Riconscente, M. M., Mislevy, R. J., & Corrigan, S. (2016). Evidence-centered design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 40–63). Routledge.
- Shaughnessy, J. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph & K. Carr (Eds.), *People in mathematics education* (Vol. 1, pp. 205–237). Kluwer.
- Shaughnessy, J., Watson, J., Moritz, J., & Reading, C. (1999, April 22–24). *School mathematics students' acknowledgment of statistical variation* [Paper presentation]. 77th Annual Conference of the National Council of Teachers of Mathematics, San Francisco, CA, United States.
- Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics* (pp. 957–1010). National Council of Teachers of Mathematics.
- Smith, C., Wisner, M., Anderson, C. W., Krajcik, J., & Coppola, B. (2004). Implications of research on children's learning for assessment: Matter and atomic molecular theory. Center for Education, National Research Council.
- Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal*, 12(2), 147–169. <https://doi.org/10.1007/BF03217081>
- Watson, J. M., & Kelly, B. A. (2002). Can grade 3 students learn about variation? In B. Phillips (Ed.), *Proceedings of the sixth international conference on teaching statistics: Developing a statistically literate society* (pp. 1–6). International Association for Statistical Education. http://iase-web.org/documents/papers/icots6/2a1_wats.pdf?1402524960
- 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, 1–29. <https://doi.org/10.1080/0020739021000018791>
- Zawojewski, J. S., & Shaughnessy, J. M. (2000). Data and chance. In E. A. Silver & P. A. Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 235–268). National Council of Teachers of Mathematics.

Suggested citation:

Fife, J. H., James, K., & Peters, S. (2019). *A learning progression for variability* (Research Report No. RR-20-05). Educational Testing Service. <https://doi.org/10.1002/ets2.12286>

Action Editor: James Carlson

Reviewers: Edith Aurora Graf and Gabrielle Cayton-Hodges

CBAL, ETS, the ETS logo, and MEASURING THE POWER OF LEARNING. are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS RESEARCHER database at <http://search.ets.org/researcher/>