Int J of Sci and Math Educ (2017) 15:1039–1056
DOI 10.1007/s10763-016-9737-7

Post-secondary Science Students' Explanations of *Randomness* and *Variation* and Implications for Science Learning

Rebekka Darner Gougis[1] · Janet F. Stomberg[1] · Alicia T. O'Hare[1] · Catherine M. O'Reilly[1] · Nicholas E. Bader[2] · Thomas Meixner[3] · Cayelan C. Carey[4]

Received: 2 February 2016 / Accepted: 1 March 2016 / Published online: 23 March 2016

Abstract The concepts of randomness and variation are pervasive in science. The purpose of this study was to document how post-secondary life science students explain randomness and variation, infer relationships between their explanations, and ability to describe and identify appropriate and inappropriate variation, and determine if students can identify sources of variation. An instrument designed to test statistical concepts was administered to 282 college students from three universities, ranging from introductory non-science majors to science graduate students. Students readily distinguished between causes of variation. A naïve *no-pattern concept of randomness* persisted from first-year non-science majors to senior-level science majors, contributing to incorrect responses on the variation instrument. Students' expressions of randomness were better predictors of performance on the variation instrument than their expressions of variation. It is argued that inclusion of everyday language uses of randomness in instruction can bridge the gap between vernacular and scientific uses of this term.

Keywords Conceptual development · Language use in science · Nature of science · Randomness · Variation

✉ Rebekka Darner Gougis
rldarne@ilstu.edu

1 Illinois State University, Normal, IL, USA

2 Department of Geology, Whitman College, Walla Walla, WA 99362, USA

3 Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721-0011, USA

4 Department of Biological Sciences, Virginia Tech, Mail Code 0406, Blacksburg, VA 24061, USA

Introduction

Science learning requires synthesis of complex ideas across broad disciplines, such as variation and randomness, which require sophisticated understanding in order to conduct and understand science. Recognizing and understanding variation and randomness is necessary to sample and analyze data and perform statistical analyses, as well as construct sophisticated conceptions of scientific principles, such as evolution, climate change, and water quality. Recent Next Generation Science Standards (NGSS Lead States, 2013) in the USA have stressed patterns of inquiry that span all sciences and convey the nature of science as a collection of common practices (Scientific and Engineering Practices, SEPs) and overarching ideas (Crosscutting Concepts, CCCs). The concepts of randomness and variation are important but implicit components of SEPs and CCCs, namely the practices of *planning and carrying out investigations* and *analyzing and interpreting data*, and the crosscutting concepts of *patterns* and *scale*, *proportion*, *and quantity.* Adeptly planning and carrying out investigations and analyzing and interpreting data requires the scientist to reason about variation observed across data and decipher patterns in data, given the frequency of some observations in proportion to others, which allow the scientist to form a conclusion. Both randomness and variation have frequent vernacular uses, indicating alternative non-scientific meanings of these terms. For example, a millennial student may use random to refer to an unexpected event; similarly, the common adage, everything happens for a reason, indicates a belief that nothing is genuinely random. While the NGSS movement applies directly to pre-K-12 levels, the SEPs and CCCs apply to all levels of inquiry and thus still need to be emphasized in post-secondary science education. Little attention is paid to how the concepts of randomness and variation are addressed in the post-secondary science classroom, if at all, and there is even less understanding of how these concepts are brought to bear when engaging in SEPs and CCCs. This manuscript initiates a new line of inquiry to examine these relationships by documenting aspects of graduate and undergraduate students' uses of randomness and variation, demonstrating how these ideas are brought to bear when solving probability tasks, and discussing the relevance of these language uses for teaching post-secondary science.

Background

Learning how to adeptly perform investigations and analyze data requires construction of many scientific concepts that also have everyday meanings. The current line of inquiry examines randomness and variation; therefore, contextualized examples are provided across different fields of science to demonstrate how many scientific concepts rely on good working knowledge of randomness and variation. Additionally, a framework for conceptual development, relationships between everyday language and scientific language, and a literature review of the development of sophisticated concepts of randomness and variation are presented.

'Random' and 'Variation' Contextualized

Decontextualized definitions of "randomness" and "variation" contribute to their abstractness and difficulty to understand. Therefore, examples of climate change, soil science, water quality, and evolution are provided to demonstrate the pervasive need for scientific understanding of these concepts. Interpreting current climate conditions while taking into account long-term trends requires understanding patterns of weather variation appropriate to a particular climate. When students lack understanding of climate outliers (an anomalous wet or dry year in a climate record), they are likely to conclude that a very cold winter disproves current warming trends. Complicating this issue, the Intergovernmental Panel on Climate Change (2013) predicts increased variability in precipitation, air temperature, and storm events for many geographic regions as climates change. Thus, the extreme degree of variability is counterintuitive evidence in support of climate change—a very cold year is a datum that supports warming trends. Understandably, students may have difficulty identifying and quantifying increased variability over time in comparison to linear rates of change, as well as interpreting long-term trends in climate. Visualizing the variation inherent in long-term time series and summary metrics of variability, such as the standard deviation of daily precipitation within a year, may help students conceptualize how both long-term trends and variability change over time.

A basic soil ecology experiment demonstrates pervasiveness of variation and randomness concepts in experimental design. A soil ecologist examines differences in the amount of organic carbon between grazed and ungrazed grasslands and measures organic carbon concentrations in both treatments. The goal is to describe differences that exist between treatments, requiring an analysis of variance. Many students may reason that to accomplish this goal, they need to examine two samples: one from a grazed and one from an ungrazed soil. The student may naïvely conclude that grazing caused the observed difference between samples. However, scientists know that sampling only once from each treatment does not allow the variance to be analyzed, which furthermore prevents a scientist from determining what variation is due to the treatment (ungrazed vs. grazed) and what variation is due to the random nature of sampling and processes not specifically measured in the experiment (Bennington & Thayne, 1994). Soil, like many natural systems, is spatially variable (Conant & Paustian, 2002), and this natural variation occurs along a background of random differences. When studying systems that have variability due to both natural factors and variation due to random sampling, learners are called upon to determine reasonably confident attributions of variation. Unfortunately, they are often asked to do so without a thorough exploration of the different causes of variation in data.

Examining water quality is a unique context in which the concept of variability can be particularly challenging. Difficulty arises because it is necessary to use observations of stream chemical composition over time and space to understand causes of pollutant levels from several factors, such as land management and point source discharges. Variation may appear random when it is influenced by many environmental factors. The matters are complicated since average variation in stream contaminant concentrations cannot be determined by a simple mean because the average concentration of a contaminant depends on the concentration and the volumetric streamflow occurring over a time period. Students are typically accustomed to using an average (mean) to convey central tendency of a varied data set, but a mean contaminant concentration does not take into account volumetric flow, an important factor contributing to variation observed in water quality. Thus, developing understanding of water quality in flowing streams is reliant on a learner's ability to reason logically about variation in more than one factor at a time.

Finally, concepts of variability and randomness are also integrated with an understanding of evolutionary theory. Basic tenets of evolutionary theory are that all genetic variation initially arises from mutation, and mutations' resultant phenotypes are random regarding the advantage or disadvantage they impart in the particular environment in which the organism lives. Mutations are not teleological; they do not happen in order to provide an advantage. They occur due to random errors in DNA replication, and their effects are random with respect to the environment. Tetrapod ancestors, for example, could not have experienced mutations *in order to* disperse onto land; this would have been nonrandom. Rather, mutations happened in tetrapod ancestors and some were disadvantageous while others imparted phenotypes that increased fitness in terrestrial environments. Thus, those individuals were more suited for the terrestrial environment and thus selected for by in the terrestrial environment, resulting in their genes being more prevalent in subsequent generations.

When students do not have a scientific concept of random, they may be more likely to reason that genetic variation arises to provide a particular advantage in a given environment. Similar to mutation, understanding of genetic drift requires a scientific concept of randomness. Some biologists argue drift has an equal or greater impact on evolutionary trajectories than natural selection, because in numerous cases, it has been identified as a primary cause of divergence (Knowles & Richards, 2005). When drift occurs in a large population, variation in a gene pool seldom changes from one generation to the next because all members of the population have an equal chance of being impacted, but random chance in a small population can yield drastic effects on alleles represented in the gene pool. Thus, when students make this distinction between drift's effects on large vs. small populations, they are called upon to reason about randomness according to scale.

The examples of climate change, soil science, water quality, and evolution are provided to demonstrate how sound knowledge of randomness and variation is needed to learn a range of scientific concepts. While there are many everyday uses of the term "random," the concept of random referring to events having equal probability is most applicable to scientific contexts, such as those explored. When understanding random in this way, a student can reason how random mutations provide new genetic variations that serve as raw material for evolution. Likewise, she can justify the assumption that a randomly selected sample is likely to be representative of a population. Similarly, there are many concepts of variation, but defining "variation" as a mathematical term to describe data points' distances from a mean is most advantageous when encountering scientific contexts. A student understanding variation in this way can explain how an analysis of variance allows us to determine if data are likely to have arisen from distinct populations or the same population. Due to the pervasiveness of these terms in science, educators have good reason to develop students' scientific concepts of randomness and variation, which necessitates considering how these terms are used in everyday vs. scientific contexts.

Scientific and Everyday Language

Defined communities, such as the scientific community, use signs and symbols as resources to enable communication among its members, and language is such a system

(Lemke, 1990). Terms such as "random" and "variation" hold different meanings in scientific communities, compared to the everyday use of the terms. Therefore, communicators need to aware of the intended and un-intended audiences when crafting and sending communications, since terms are conceptualized in relation to other words or concepts and likely over-generalized from a more familiar context. Here, languages-for-specific-purposes, or groups of terms that hold different meanings or implications in various communities, is useful for understanding how individuals within sciences or from different communities may use the same term to carry different meanings (Pecman, 2014). For example, when an evolutionary biologist speaks of variation, what come to mind is likely genetic variation and its central role in evolutionary processes, whereas magnetic variation may be invoked by a physicist. Likewise, random may mean equal probability of an occurrence to a statistician, but an unexpected occurrence to a millennial student.

Difficulty in instruction regarding randomness and variation arises due to language, with some terms having distinct scientific meaning used in various ways in everyday contexts, even in textbooks (Anakkar, 2014). Brown and Kloser (2009a) argue that students' everyday language is a potential tool that educators can use to support conceptual learning in science, because when students use their everyday language to discuss scientific ideas, educators gain authentic "windows into the students' scientific understandings of the world" (Yeo, 2009, p. 913). Students' use of vernacular to express their conceptions of scientific phenomena is preferred over students mimicking text and teachers' speech without understanding, which offers the educator little information about students' conceptions (Yeo, 2009).

Brown and Kloser (2009a, b) argue language cannot be clearly divided between everyday and scientific language. Rather, they posit a conceptual continuity, or the existence of varying levels of continuity between science ideas as expressed in multiple communities, both scientific and non-scientific. This perspective can serve as a framework for understanding how to craft learning opportunities that maximize the relationship between students' native ways of understanding and science instruction. Conceptual continuity appears to be complementary to language-for-specific-purposes model (Pecman, 2014), in which she claims that learners often overgeneralize from the context in which the concept was constructed. Drawing upon students' everyday language has been successful in teaching scientific concepts to English language learners (Ryoo, 2015). Regardless of students' English proficiency, students' everyday understandings and scientific understandings should not be separate from one another, and the language used to develop scientific understandings must begin with students' everyday language. Sophisticated understanding of any concept encompasses multiple cognitive representations of that concept that are both internal and external to science, which facilitate traversing the conceptual continuities between contexts. Students' prior knowledge and experiences plays a critical role in this conceptual development.

Conceptual Development

Development and understanding of scientific concepts like variation and randomness is not the accumulation or transfer of factual information from teacher to learner (Hemmerich, Van Voorhis, & Wiley, 2015; Scott, Asoko, & Leach, 2007). Rather, it involves the construction of conceptual structures that are consistent with the learner's everyday experiences (Hemmerich, et al., 2015; Vosniadou, 2002). The learners' mental models allow them to develop expectations within, make predictions, and pose tentative explanations about their lived experiences, and in doing so, prior knowledge substantially influences further learning (Scott et al., 2007). When these expectations, predictions, and explanations are confirmed by experiences, the conceptual framework is reinforced and enriched, but when they are not confirmed by experiences, the framework is restructured and revised. This view of learning specifies that in order for instruction to be successful, it must connect prior knowledge, or the conceptual structures built over a lifetime of lived experience, to the new material being taught (Limón, 2001). Furthermore, instructors must strategically create experiences for students that are inconsistent with unscientific aspects of the students' existing conceptual structures to elicit cognitive conflict; allowing students to make distinctions between contextualized uses of a given concept or term that differ across communities (Hemmerich et al., 2015; Limón, 2001). Instructors must understand that conceptual change is gradual and complete accommodation of conceptual structures is unlikely after a single instructional intervention. However, Yore and Treagust (2006) suggest that learners can improve understanding of a concept by using one representation to constrain interpretation of a second representation. For example, a student who understands random to mean *anything is possible* can be prompted to use that definition to specify what anything refers to when a random sample is used to make generalizations about a population in an experiment. This approach may also help the student appreciate how a single, coherent conception of a term is useful across contexts to explain a wide range of phenomena (Scott et al., 2007). This example demonstrates how learners' prior knowledge of many scientific terms, such as variation and randomness, originates in unscientific contexts, and development of the scientific concept must begin with their prior knowledge—scientific or unscientific.

Learning the Nature of Randomness and Variation

Mathematics and science educators have long known that students' concepts of statistical terms are quite different from scientists' concepts of those terms. Garfield and Ahlegren (1988) documented difficulties that students encounter in learning probability and statistics and found that statistics, as they are used in the scientific community, are seldom taught prior to college. This is apparent when students are accustomed to number crunching rather than reasoning about why the mathematical procedures are needed to understand scientific data. Students often believe that a mean is a computational item rather than one of several possible ways to represent the middle of a data distribution (central tendency), which is a logical outcome given the instructional emphasis on procedural knowledge rather than mathematical reasoning. We assume that typical methods for teaching statistics do not begin with students' everyday experiences and language uses of terms such as average and likelihood, let alone support students in building a robust conceptual continuity between everyday and scientific uses of these terms.

Randomness is a concept for which instruction begins at an early developmental level and is foundational for developing understanding of typical Gaussian statistical methods (Heinicke & Heering, 2013). Bryant, Nunes, Evans, Gotardis, and Terlektsi (2014) guided elementary students to comprehend three central ideas while constructing a sound concept of randomness: (1) defining and understanding randomness as unpredictable or lacking a pattern; (2) a defined sample space (i.e. within a given situation, what are possible results?); and (3) The quantification of probability (i.e. performing a proportional calculation of the probability of possible outcomes). The first of these criteria provides opportunity to build upon students' everyday language use of random. Bryant et al. (2014) demonstrated impressive success with elementary students using this strategy, and conceiving randomness as being unpredictable or lacking a pattern is developmentally appropriate for elementary students. However, post-secondary students who conceive of randomness solely as one of these three criteria, such as randomness being anything that lacks a pattern (i.e. *the no-pattern concept of randomness*), would be viewed as holding a misconception from a scientific perspective.

Misconceptions regarding randomness are important to identify. Smith, diSessa and Roschelle (1993) argued that a misconception should be viewed as a worthwhile conception, which is constructed from the learner's experience and should serve as the starting point for the development of a more sophisticated conception of the target idea. Batanero, Godino, and Vallecillos (1994) point out that a plausible reason for why so many students develop distaste for statistics is because it is taught too abstractly and does not build upon students' experiences, language, and concepts.

The no-pattern concept of randomness is not sufficient to understand numerous statistical and scientific principles that students encounter in high school and college. This is because most scientific analyses and many scientific principles require interpretation of probability and randomness is the starting point to understanding probability. Learners who do not have a scientific concept of randomness fail to reason about adequate sample size (Barragués, Guisasola, & Morais, 2006), this inadequacy can be demonstrated with a hypothetical case in which a learner holding this concept is asked to solve the following problem modified from Watson, Kelly, Callingham, and Shaughnessy's instrument (2003).

> A class of 27 students did an activity with a spinner that was half shaded, the other half unshaded. Each student spun the spinner 50 times and the results for the number of times it landed on the shaded part were recorded for each student. Out of 50 spins, what would you predict is the average number of times it landed on the shaded part? Describe what you would expect the variation around that average to look like.

Here, the learner is being asked to make a prediction for an experiment involving a random event. While it is true that there is no way to predict the outcome of each independent event (one spin of the spinner) because the shaded and unshaded halves of the spinner have equal probabilities (random), we *can* predict that approximately half of the spins will land on the shaded region. We can also predict that if we were to graph the distribution of results across all 27 students in the class, a normal distribution would be most likely. Learners who hold the no-pattern concept of randomness would struggle to make such predictions, particularly if the association between *random event* and *unpredictable* and *no pattern* is strong. We suggest that such a student would be likely to claim that a randomly scattered distribution would be most likely, rather than a normal distribution. This may be due to being unable to differentiate between the result of one event and a set of results (Barragués et al., 2006).

Related to but distinct from randomness is the concept of variation, which Wild and Pfannkuch (1999) identify as "the centerpiece of statistical thinking" (p. 235). Understanding variation involves recognizing the different faces of variability, including "overall spread in a data set, variability between two data sets, variability as measurement error, etc." (Garfield & Ben-Zvi, 2007, p. 386). Watson and Kelly (2004) claim that there are three aspects of statistical variation that are necessary for analyzing outcomes of repeated trials: differences from an expected theoretical value, between the experimental and expected distributions, and between plausible and unrealistic variation. While scientists employ all three of these aspects of variation when conducting science, a primary goal in the natural sciences is to understand natural variation or variation not due to deviance from a mean or prototype but rather the differences observed in nature, which relates to the third dimension of understanding variation.

Randomness and variation are central, albeit often implicit, ideas in science. These concepts have numerous facets of meaning across scientific disciplines and everyday contexts that cannot be overlooked when teaching. Students' everyday language is pertinent to the conceptual development of variation and randomness, becomes integrated into how they talk about scientific phenomena, and thus reveals their prior knowledge at which science instruction ought to begin. The contexts in which students' understandings of variation and randomness have been examined are largely mathematical, with little attention to how randomness and variation are conceptualized in science contexts. Furthermore, misconceptions of randomness and variation are common and can potentially lead to faulty reasoning about scientific sampling, experimentation, and conventional statistical applications.

Purpose

The developmental trajectory of randomness and variation between elementary school and post-secondary students has not been well documented, despite being potentially informative for science education research and teaching practices. Thus, we seek to examine post-secondary students' understanding of randomness and variation. First, we ask, what are post-secondary students' expressions of randomness and variation when they are asked to define them in a science course? Secondly, what, if any, relationships exist between students' conceptions of these terms and their performance on a basic probability tasks? Third, are students able to distinguish between different sources of variation in collected data?

Methods

This study is a large-scale, multiple-site, Web-based survey of university biology or geology students in the USA. Data were collected in geology or biology courses spanning the developmental continuum of post-secondary student levels.

Participants

Participating students were from three universities and were samples of convenience. Some students were at a large Midwestern public doctorate-granting university ranked very selective (Barron's College Division Staff, 2015) and enrolled in an introductory general education biology course for non-science majors (n = 145), an introductory biology course for science majors (n = 94), or a graduate biostatistics course (n = 9). A second group of participating students were enrolled in a mid-level science major hydrogeology course (n = 21) at a small public master's-granting university also ranked very selective (Barron's College Division Staff, 2015). A third group of students included an upper-level freshwater biology course (n = 13) at a very large, public, doctorate-granting university ranked highly selective (Barron's College Division Staff, 2015). IRB approval was granted by the first author's institution and then approved by the other institutions' IRBs.

Instructors showed students a 4-min video in class at the beginning of the semester that explained the study to recruit volunteers. Instructors then emailed their students a link to complete the questionnaire online; the first page of the online questionnaire reiterated the study explanation that they heard during the recruitment video. Willing participants then proceeded to complete the online questionnaire in which students typed their responses into comment boxes, which took about 20 min. Participants were offered minimal course credit (i.e. less than 2 % of final course grade) for questionnaire completion.

Instrument

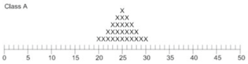
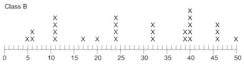
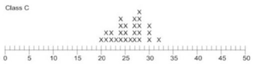
The 9-item instrument was composed of seven items from Watson and Kelly's (2004) instrument and two additional items that were created for this study; all items were open-ended (Table 1). The survey collected data on students' concepts of randomness, sampling, and variation. Part 1 of the instrument consisted of the 4-item Distributional Variation Subscale (DVS), which measures a student's ability to "describe appropriate variation … and identify appropriate and inappropriate variation in an established distribution" (Watson & Kelly, 2004, p. 125). This scale was chosen because it poses a basic probability task that is accessible by all levels of students being sampled, and thus serves as a performance indicator for distinguishing appropriate and unlikely variation in data. Watson et al. (2003) determined reasonable validity for the original DVS. The first three authors developed the scoring rubric to differentiate subtle differences between student responses while still capturing all responses in the data set.

An exploratory factor analysis was used to establish validity of this scale, using principal axis factoring as the extraction method, which is preferred when assumptions of normality are disrupted (Costello & Osborne, 2005). Bartlett's test of sphericity was significant (χ^2 (36) = 172.0, $p < 0.0001$), indicating significant correlated factors among the DVS items. Initial eigenvalues showed that the first factor explained 23.5 % of the variance, the second factor 16.2 % of the variance. The second, third, and fourth DVS items, each asking students to determine if a different data distribution is likely to be real, contributed to the factor structure. The third and fourth DVS items loaded onto the first factor, with 0.63 and 0.52 loadings, respectively. The second DVS item loaded on the second factor with a loading of 0.78. The first item did not contribute to the factor structure and variance explained by items 2–4, so it was not included in analyses.

This analysis indicates that the DVS measures two aspects of a student's ability to distinguish likely and unlikely variation. The first aspect (items #3 and #4) corresponds to the ability to recognize realistic range or spread of a data set. The second aspect (item

Table 1 Questionnaire items and scoring rubric

Questionnaire Items and Scoring Rubric

Part I	Rubric
<u>Distributional Variation Subscale (Watson & Kelly, 2004)</u> Three classes of 27 students each did an activity with this spinner (see image); each student spun the spinner 50 times and the results for the number of times it landed on the shaded part were recorded for each student.  1. Out of 50 spins, what would you predict is the average number of times it landed on the shaded part? Describe what you would expect the variation around that average to look like.	i. A whole-number score (0-4) was determined by giving one point for responses that included each of the following: • The average would be 25. • Distribution would likely be bell-shaped. • There would be about the same number of spins above and below the mean (i.e., equal frequency). • There would be about an equal spread of data on either side of the mean (i.e., equal range).
2. In some cases, the class just made up the data without actually doing the experiment. Class A's data is shown below, which is a symmetrical distribution with an average of 25 spins landing on the shaded part of the spinner. State whether or not you think the data are real or made up. Then explain why you think this. 	ii. A whole-number score (0-3) was determined by giving one point for responses that included each of the following: • These data are likely made-up. • Real data are not likely to be exactly symmetrical. • Perfect symmetry is not likely due to the randomness of the spinner, which is likely to produce some deviation from what is expected.
3. Class B's data is shown below, which also has an average of 25 spins landing on the shaded part of the spinner. State whether or not you think the data are real or made up. Then explain why you think this. 	iii. A whole-number score (0-3) was determined by giving one point for responses that included each of the following: • These data are likely made-up. • The data are too varied; the distribution is too broad; the range is too large. • The data contain too many unlikely data points, such as the data point that indicates a student landed on the shaded part of the spinner all 50 times.
4. Class C's data is shown below, which also has an average of 25 spins landing on the shaded part of the spinner. State whether or not you think the data are real or made up. Then explain why you think this. 	iv. A whole-number score (0-3) was determined by giving one point for responses that included each of the following: • These data are likely real. • These data have a realistic range. • These data have a realistic distribution that is roughly bell-shaped but not perfectly symmetrical.
Part II	**Rubric**
<u>1. Sample</u> What does "sample" mean? Please give an example of a "sample."	A whole-number score (0-4) was determined by giving one point for responses that included the following: • A plausible example • A statement that explained how a sample is representative of a larger whole. • A sample is a piece of something. • A sample is used to predict or generalize about the larger whole.
<u>2. Random</u> What does "random" mean? Please give an example of something that happens in a "random" way.	A whole-number score (0-2) was determined by giving one point for responses that included the following: • A plausible example • A statement that explained that all the possible outcomes have an equal chance, probability, or likelihood.
<u>3. Variation</u> What does "variation" mean? Please give an example of something that varies.	A whole-number score (1-3) was assigned according to the following: 1. Different forms or varieties of something 2. Changing or different outcomes/results across time and/or space 3. Distance or difference from a mean measurement
Part III	**Rubric**
<u>1. Rocks</u> A group of geology students on a field survey are instructed to weigh various sedimentary rocks. Each student takes turns weighing each rock. Below are a group of 10 students' measurements of the same piece of sandstone. 30.0g 31.2g 30.5g 23.2g 31.0g 31.1g 29.6g 30.3g 30.0g 30.8g The mean of these values is 29.8g. What do you think accounts for the variation seen in this data?	A whole-number score (0-2) was determined by giving one point for responses that included the following: • A plausible example • Human or measurement error
<u>2. Squirrels</u> A group of mammalogy students on a field trip collect five 13-lined ground squirrel adult males. One student weighs all the squirrels. Below are the weights for each squirrel. 110.5g 124.8g 236.8g 189.5g 144.9g The average weight of these squirrels is 161.8g. What do you think accounts for the variation seen in this data?	A whole-number score (0-2) was determined by giving one point for responses that included the following: • A plausible example of the cause of the variation with an explanation • Natural variation among squirrels

#2) corresponds to the ability to recognize the unlikelihood of a perfectly Gaussian distribution. Due to two aspects of variability understanding being measured by the DVS, aggregated items #3 and #4 (Cronbach's α = 0.51) and item #2 will be used separately as performance indicators for variability understanding, referred to hereafter as variation factors 1 and 2.

Part II of the instrument consisted of three separate items, which asked students to define "variation," "random," and "sample," and then to give an example of each. The first two of these terms is the focus of this study, but students' expressions of sample potentially elucidates their thinking about randomness and variation, because a sample must represent the variability in a whole, which is often accomplished through random selection. While construct validity of these items had previously been established (Watson et al., 2003), the first three authors developed a scoring rubric (Table 1) to account for more than one meaning of a term if it was included in a student's response.

Part III of the instrument consisted of two items created for this study and asked students to distinguish between variations in data due to methodological error vs. naturally occurring variation in the phenomenon studied. To establish construct validity, the first author created drafts of these items, presented them to the research team, and edits were made collaboratively until all members judged them to be reasonable indicators of the target features. Items in parts II and III of the instrument were treated as single-item factors and were analyzed separately.

Scoring Procedures

Initial qualitative analysis of student responses began with category construction for each item (Merriam, 2009), using a subset of about 30 responses. Three researchers (first three authors) independently open-coded responses by placing them in emergent categories that represented scientifically sound but qualitatively different responses. They then collaboratively collapsed open codes into the categories of responses in the scoring rubric. For most items, a significant number of students conveyed more than one category in their responses. For example, in response to part I, item 1 (Table 1), some students simply stated "25," while others stated "25" *and* that the distribution would be bell-shaped. Thus, the decision was made to assign scores to responses, based on number of correct response categories satisfied. The one exception to this process was with part II, item 3, for which students did not provide more than one meaning for variation; rather, responses were placed into one of three mutually exclusive categories, ranked 1–3 to indicate more preferred responses from a scientific standpoint.

After this scoring rubric was established, the researchers independently recoded the original subset of data using the rubric. Inter-rater reliability for each item was measured using Fleiss's kappa (κ), which calculates the degree of agreement in classification among three or more coders over that which would be expected by chance (Geertzen, 2012). Using this measure, $\kappa \geq 0.61$ indicates substantial agreement among coders (Viera & Garrett, 2005); so, for items in which initial coding attempts yielded $\kappa \leq 0.61$, the three researchers revisited the scoring rubric and discussed discrepancies in interpretation, after which rescoring occurred. These processes were iterated until $\kappa \geq 0.61$ was consistently achieved for all items. Once this was achieved, researchers scored the remaining data independently or in pairs. In cases when responses were scored by pairs, researchers came to a consensus on the appropriate score through discussion with the third researcher.

Generally, the scoring rubric accounted for scientifically acceptable responses and did not identify misconceptions; wholly incorrect responses from a scientific perspective were assigned a 0 score. However, the no-pattern concept of randomness occurred so frequently in part II, item 1 responses (Table 1) that researchers saw the need to document its frequency; the first author enumerated instances of the no-pattern concept across all the data.

Statistical Analyses

RQ#1: How do post-secondary students express randomness and variation? The scored responses of post-secondary students' expressions of random, sample and variation were examined using two methods. First, three Kruskal-Wallis one-way analyses of variance (ANOVA) were performed on all part II items separately using educational level (introductory non-science majors, introductory science majors, mid-level science majors, upper-level science majors, science graduate students) as the predictor variable and individual item score as the response variable. The nonparametric ANOVA was chosen due to the ordinal nature of the data and large differences in sample size. Second, frequencies of the no-pattern concept across education level were compared using a parametric one-way ANOVA; the parametric analysis was chosen in this case because data are continuous and did not violate the normality assumption.

RQ#2: Do students' conceptions of randomness and variation predict performance on a basic probability task? Six separate Kruskal-Wallis ANOVAs were performed using part II items as separate predictors of variation factors 1 and 2 from part I. Additionally, a correlation was performed to see if those students who expressed the no-pattern concept in part II, item 1 were more likely to state that the scattered distribution was the real distribution in part I, item 3.

RQ#3: Are students able to distinguish between different sources of variation? Students' abilities to distinguish between different sources of variation (measurement error vs. natural variation in the phenomenon being studied) were explored using Kruskal-Wallis ANOVAs. Two ANOVAs were performed for both part III items separately using educational level as the predictor variable and item score as the response variable.

Results

RQ #1: How do post-secondary students express randomness and variation?

Kruskal-Wallis one-way ANOVAs revealed statistically significant differences across educational level in how students defined "random" (H (4) = 29.6, $p < 0.0001$); sample (H (4) = 12.4, $p = 0.014$); and variation (H (4) = 30.6, $p < 0.0001$). Post hoc pairwise Mann-Whitney comparisons with a Bonferroni correction revealed that when asked to define "random," graduate students scored significantly higher than non-science majors ($U = 956$, $p = 0.001$) and introductory science majors ($U = 654$, $p = 0.007$). Post hoc comparisons indicated that when asked to define "variation," graduate students scored significantly higher than mid-level science majors ($U = 109$, $p = 0.001$); introductory science majors ($U = 137$, $p < 0.001$); and introductory non-science majors ($U = 121$, $p < 0.001$). Post hoc comparisons indicated that when asked to define "sample," upper-level science majors scored significantly better than introductory non-science majors ($U = 1220$, $p = 0.003$) and graduate students ($U = 22.5$, $p = 0.009$).

Students at all levels exhibited the no-pattern concept of randomness in their definitions of "random" and in item 3 of part I. Twenty-nine of the 281 participants (10.3 %) across all student levels provided a definition that matched the no-pattern concept. Among introductory non-science majors, 8 % of responses conveyed the no-pattern concept. Among introductory science majors, 13.8 % conveyed this concept. Fifteen percent of mid-level science major students provided the no-pattern concept, and one of each of the upper-level student and graduate students revealed this naïve concept. There were no significant differences in the frequency of the no-pattern concept across level of student. Students at all levels exhibited this concept, exemplified by the following student responses:

> This data looks more realistic to me because there are outliers, and there is no specific pattern. There shouldn't be a specific pattern in a random spinner experiment.
>
> - Introductory non-science major
>
> Real, because the data is spread out and random. There is no perfect pattern created.
>
> - Introductory science major
>
> This data set seems more realistic due to the random outcomes which are plotted on the graph. This type of data is more common than a perfectly symmetrical one.
>
> - Mid-level science major
>
> This data seems real because the values are spread out at random.
>
> - Upper-level science major
>
> Real. Data is more widely distributed and looks close to a real scenario.
>
> - Science graduate student

Students' definitions of "variation" fell into one of three distinct categories, corresponding to progressively higher scores on this item: Different forms or varieties of something (scored 1), changing or different outcomes/results across time and/or space (scored 2), and distance or difference from a mean measurement (scored 3). Forty-four percent of students across all levels provided a definition of variation that was given a score of 1. Thirty-nine percent of students provided a definition that was scored as a 2. Only 17 % of students provided a definition that was scored 3, which is the definition most useful for understanding statistical variability, and this was the most common definition provided by science graduate students.

RQ#2: Do students' conceptions of randomness and variation predict performance on a basic probability tasks?

Part II items were tested as predictors of variation factors 1 and 2 from part I, which measured ability to recognize a realistic range or spread in a real data set and ability to recognize that a perfect bell curve is not likely in a real data set, respectively. Students' responses to part II, item 1, asking students to define random, successfully predicted variation factor 2 (H (2) = 14.1, p < 0.001) but not variation factor 1. Part II, item 2 responses were not good predictors of variation factor 2, but they were good predictors of variation factor 1 (H (3) = 9.35, p = 0.025). Scores on part II, item 3, asking students to define variation, did not predict scores on either factors of variability understanding.

Students who demonstrated the no-pattern concept in part II, item 1 were more likely to say that the scattered distribution was the more realistic distribution in part I, item 3 (r [279] = 0.15, p = 0.01). There were students at all levels that claimed the more scattered distribution was more likely (51.7 % of introductory non-science majors, 50 % of introductory science majors, 40 % of mid-level majors, 23 % of upper-level science majors, and 56 % of graduate students), and this claim was made even when the no-pattern concept was not revealed in their definitions of random.

RQ#3: Are students able to distinguish between different sources of variation?

Part III items asked students to differentiate among causes of variation in data. Part III, item 1, requires students to recognize naturally occurring variation, no statistically significant differences across level of students were detected. On part III, item 2, asking students to recognize variation due to measurement error, many students misunderstood the question to read it as asking about several different rocks, rather than a single rock being measured by many students; these responses were eliminated from the analysis. Among remaining responses (n = 187), a significant difference was detected across student levels (H (4) = 16.70, p = 0.002), but post hoc comparisons could not detect the source of this significance.

Discussion

Several limitations of this study prevent far-reaching generalizations about how post-secondary science students understand randomness and variation. For example, internal reliability of variation factor 1 remained low, indicating there are aspects of variability understanding that are not captured by the measure. When variation factor 2 is included, only 39.7 % of the variability on the DVS was explained. Furthermore, in this study, it was assumed that when students were asked to define random, sample, and variation in a science course, the most salient context would be that of science. This is a risky assumption, given students' tendencies to extend understanding from non-scientific contexts into their scientific understanding. These less-than-satisfactory circumstances highlight a need for a comprehensive measurement of statistical understanding, as well as qualitative studies that deeply observe the emergence of these terms in students' practices, paying particular attention to how concepts of variation, sampling, and randomness are brought to bear while doing science. Only such a mixed-method approach would be able to elucidate causal mechanisms that give rise to the understandings of randomness and variation observed in this study.

Despite these limitations, this study demonstrates that generally, expressions of randomness and variation become more sophisticated with higher educational levels. This finding is not surprising, since we would generally expect students' reasoning to become more sophisticated with advanced studies. However, large proportions of students sampled in this study revealed conceptions of randomness and variation that are either not helpful or are potentially detrimental in understanding statistical variability and randomness's role in scientific practices. A disturbingly common and naïve definition of random provided by students at all levels revealed a no-pattern concept of randomness, or when students understand random to mean anything that lacks a pattern. Use of this naïve concept was revealed when students claimed that randomly scattered distributions are more likely than bell-shaped distributions in a basic probability experiment. Unlike definitions of the terms "random," "sample," and "variation," frequency of the no-pattern concept did not change with educational level, indicating a significant inadequacy of instruction to help novice scientists develop a useful concept of randomness as it applies to science. Konold and Higgins (2003) argue that students overcome a conceptual milestone when they begin to see a data set as a whole with its own emergent properties, rather than a group of individuals with their own distinct characteristics. It appears as if students who thought the scattered distribution was most likely focused on individual spins of the spinner, rather than the distribution of aggregated data. Opportunities to analyze distributions of data, rather than constructing graphs using individual data points, would be a positive step in helping students to see data distributions as tools useful toward answering scientific questions, rather than a collection of individual data points (Garfield & Ben-Zvi, 2007).

Randomness and variation traverse all scientific contexts, influencing a multitude of scientific practices, including experimental design, sampling, and data analysis. The natural events that constitute scientists' investigational foci inherently contain randomness and natural variation, thus requiring scientists to grapple with uncertainty. Such skills are brought to bear when identifying measurement errors, defining the boundaries of the problem space to be examined, and identifying variations of the occurrences of objects or events of interest across contexts. While uncertainty has been studied in the statistics education community (Williams 2012), exploration of scientific uncertainty, development of the skills scientists have to deal with uncertainty, and potentially fruitful instructional practices have yet been examined by science educators (Ruggeri, 2012). Epistemological views of measurement influence how students estimate uncertainty in measurements and calculations (Caussarieu & Tiberghien, 2016). Yet, this is an under-emphasized problem space within the nature of science, demonstrating that the nature of science is not an exact body of knowledge. Because of the lack of precise boundaries, epistemological views of measurement and uncertainty are important areas for further research.

Teaching Implications

When students have multiple experiences that develop appreciation for variation, such as exploring a single variable across contexts by examining probability sampling, representing changes in the variable, engaging in inference, discussing the relationship between sample and population, and describing variation in multiple ways, a foundation of statistical understanding is established (Watson, Callingham, & Kelly, 2007). Garfield and Ben-Zvi (2007) point out that allowing students to explore different ways to represent the same data set becomes increasingly easier for instructors as technology provides tools to create data representations in a timely fashion. However, non-critical selection of a graph or display option in a software package without consideration of the type of data and purpose could reinforce misconceptions about randomness, variation, and samples. As discussed earlier, prompting students to use one representation of a concept, such as the no-pattern concept, to constrain interpretation of a more scientific representation, is likely to promote conceptual development, and such instructional strategies can be supported through use of technology.

These findings also demonstrate that even subtle differences in students' language use surrounding the concepts of randomness, variation, and sampling lead to important differences in how these conceptions are brought to bear when solving basic probability tasks. The teaching of randomness and variation, particularly in the context of science, is a prime example of how developing educators' understanding of students' everyday language use could be helpful in supporting students' construction of more scientific conceptions of these terms (Brown & Kloser, 2009b). Without understanding of students' conceptual continuities between their everyday use of the terms "random" and "variation," educators are unable to take advantage of students' conceptual and linguistic resources. Tang (2011) acknowledges that when two or more discourses about natural phenomena are directly juxtaposed in the classroom, conflict can arise. This conflict can lead to cognitive dissonance that creates opportunity for learners to construct new conceptual knowledge by linking it to their prior knowledge and everyday language.

Brown and Spang (2008) put forth a pedagogical method for facilitating construction of new conceptual knowledge, based on students' everyday language. Disaggregate instruction involves introducing scientific ideas in everyday language, which requires the teacher to have facility with students' everyday language, followed later by specific science language instruction. One strategy employed in disaggregate instruction is double-talk, in which parenthetical speech patterns allow the teacher to offer both vernacular and scientific terms and phrases in the same idea (Brown, 2011). While we do not claim that the use of double-talk when teaching about randomness in the context of science will fix the naïve no-pattern concept, such instructional strategies would reveal early on in instruction the inconsistencies between the no-pattern concept and the scientific concept of random, thereby providing an opportunity for students to differentiate the uses of the term across different genres. For example, when students provide the no-pattern concept when asked to explain randomness, instructors might simply ask students to specify what exactly lacks a pattern, followed by prompts to explicitly distinguish between a single random event, such as a spin of spinner, and a distribution that summarizes multiple random events. In short, instructors who are unaware of the everyday language use of random by their students will ultimately allow conceptions like the naïve no-pattern concept to continue throughout university science coursework (Brown, 2011).

Acknowledgments This research was supported by the National Science Foundation (1245707). We thank the instructors of courses who allowed us to collect data from their students, and we are grateful to the students for participating in this study. We acknowledge the value of the reviewers' suggestions and thank Professor Larry Yore's support and guidance during the development of this article.

References

Anakkar, A. (2014). Semantic difficulties in science and their implications for education. *Journal of Baltic Science Education, 13*(4), 444–447.

Barragués, J. I., Guisasola, J. & Morais, A. (2006). Chance and probability: What do they mean to university engineering students? *International Journal of Mathematical Education in Science and Technology, 37*(8), 883–900.

Barron's College Division Staff (2015). *Profiles of American colleges 2015*. Hauppauge, NY: Barron's Educational Series.

Batanero, C., Godino, J. D. & Vallecillos, A. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology, 25*(4), 527–547.

Bennington, C. C. & Thayne, W. V. (1994). Use and misuse of mixed model analysis of variance in ecological studies. *Ecology, 75*(3), 717–722.

Brown, B. A. (2011). Isn't that just good teaching? Disaggregate instruction and the language identity dilemma. *Journal of Science Teacher Education, 22*(8), 679–704.

Brown, B. A. & Kloser, M. (2009a). Conceptual continuity and the science of baseball: Using informal science literacy to promote science learning. *Cultural Studies of Science Education, 4*(4), 875–897.

Brown, B. A. & Kloser, M. (2009b). A view of the tip of the iceberg: Revisiting conceptual continuities and their implications for science learning. *Cultural Studies of Science Education, 4*(4), 921–928.

Brown, B. A. & Spang, E. (2008). Double talk: Synthesizing everyday and scientific language in the classroom. *Science Education, 92*(4), 708–732.

Bryant, P., Nunes, T., Evans, D., Gotardis, L. & Terlektsi, E. (2014). Learning that randomness is fair but loaded dice are not. *Mathematics Teaching, 242*, 4–8.

Caussarieu, A. & Tiberghien. A. (2016). When and why are the values of physical quantities expressed with uncertainties: A case study of a physics undergraduate laboratory course. *International Journal of Science and Mathematics Education*. doi:10.1007/s10763-016-9734-x.

Conant, R. T. & Paustian, K. (2002). Spatial variability of soil organic carbon in grasslands: Implications for detecting change at different scales. *Environmental Pollution, 116*(Suppl. 1), S127–S135.

Costello, A. B. & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, & Evaluation, 10*(7), 1–9. Retrieved from http://pareonline.net/getvn.asp?v=10&n=7.

Garfield, J. & Ahlegren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44–63.

Garfield, J. & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching learning statistics. *International Statistical Review, 75*(3), 372–396.

Geertzen, J. (2012). *Inter-rater agreement with multiple raters and variables*. Retrieved from https://mlnl.net/jg/software/ira/.

Heinicke, S. & Heering, P. (2013). Discovering randomness, recovering expertise: The different approaches to the quality in measurement of Coulomb and Gauss and of today's students. *Science & Education, 22*(3), 483–503.

Hemmerich, J. A., Van Voorhis, K. & Wiley, J. (2015). Anomalous evidence, confidence change, and theory change. *Cognitive Science*. doi:10.1111/cogs.12289.

Intergovernmental Panel on Climate Change (2013). *Climate change 2013: The physical science basis. (Fifth Assessment Report of the Intergovernmental Panel on Climate Change)*. New York, NY: Cambridge University Press.

Knowles, L. L. & Richards, C. L. (2005). Importance of genetic drift during Pleistocene divergence as revealed by analyses of genomic variation. *Molecular Ecology, 14*(13), 4023–4032.

Konold, C. & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.

Lead States, N. G. S. S. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Lemke, J. L. (1990). *Talking science: Language, learning and values*. Norwood, NJ: Ablex.

Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and Instruction, 11*, 357–380.

Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. San Francisco, CA: Wiley & Sons.

Pecman, M. (2014). Variation as a cognitive device: How scientists construct knowledge through term formation. *Terminology, 20*(1), 1–24.

Ruggeri, N. L. (2012). Scientific uncertainty and its relevance to science education. *Dissertation Abstracts International, 73*, 3-A.

Ryoo, K. (2015). Teaching science through language of students in technology-enhanced instruction. *Journal of Science Education and Technology, 24*(1), 29–42.

Scott, P., Asoko, H. & Leach, J. (2007). Student conceptions and conceptual learning in science. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 31–56). New York, NY: Routledge.

Smith, J. P., diSessa, A. A. & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115–163.

Tang, K. S. (2011). *Hybridizing cultural understandings of the natural world to foster critical science literacy* (Doctoral dissertation). Retrieved from University of Michigan Library. (2011-09-15T17:17:00Z).

Viera, A. J. & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine, 37*(5), 360–363.

Vosniadou, S. (2002). On the nature of naïve physics. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 61–76). Berlin, Germany: Springer.

Watson, J. M. & Kelly, B. A. (2004). Statistical variation in a chance setting: A two-year study. *Educational Studies in Mathematics, 57*(1), 121–144.

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.

Watson, J. M., Callingham, R. A. & Kelly, B. A. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning, 9*(2), 83–130.

Wild, C. J. & Pfannkuch, M. (1999). Statistical thinking and empirical enquiry. *International Statistical Review, 67*(3), 223–248.

Williams, A. S. (2012). Worry, intolerance of uncertainty, and statistics anxiety. *Statistics Education Research Journal, 12*(1), 48–59.

Yeo, J. (2009). Finding science in students' talk. *Cultural Studies of Science Education, 4*(4), 913–919.

Yore, L. D. & Treagust, D. F. (2006). Current realities and future possibilities: Language and science literacy—empowering research and informing instruction. *International Journal of Science Education, 28*(2–3), 291–314.