INVESTIGATING STATISTICAL UNUSUALNESS IN THE CONTEXT OF A RESAMPLING ACTIVITY: STUDENTS EXPLORING CONNECTIONS BETWEEN SAMPLING DISTRIBUTIONS AND STATISTICAL INFERENCE*

Luis A. Saldanha and Patrick W. Thompson
Portland State University Arizona State University, U.S.A.

Reasoning proportionally about collections of a sample statistic's values is central to developing a coherent understanding of statistical inference. This paper discusses key developments that unfolded in a classroom teaching experiment designed to support students constructing such understanding. Instruction engaged students in activities that focused their attention on the variability among outcomes of randomly drawn samples. There occurred a critical shift in students' attention and discourse away from individual sample outcomes and toward the distribution of a collection of sample outcomes. This shift supported further developments concerning how to compare entire distributions of sample outcomes as a basis for conceptualizing a notion of statistical unusualness. We characterize aspects of these developments in relation to students' classroom engagement.

BACKGROUND

There is substantial evidence in the research literature that students, when asked to make judgments about outcomes of random sampling, tend to focus on individual samples and statistical summaries of them instead of how collections of sample statistics are distributed. For instance, Kahneman and Tversky (1972) produced empirical evidence to support their hypothesis that people often base judgments of the probability that a sample will occur on the degree to which they think the sample "(i) is similar in essential characteristics to its parent population; and (ii) reflects the salient features of the process by which it is generated" (ibid., p. 430). In later research, Kahneman and Tversky (1982) conjectured that people tend to take a singular rather than a distributional perspective when making judgments under uncertainty. A singular perspective is characterized by a focus on the causal system that produced the particular outcome and by an assessment of likelihood based on "the propensities of the particular case at hand". In contrast, a distributional perspective relates the case at hand to a sampling schema and views an individual case as "an instance of a class of similar cases, for which relative frequencies of outcomes are known or can be estimated" (ibid, p. 518).

Konold (1989) found strong empirical support for Kahneman and Tversky's (1982) conjecture. He presented compelling evidence that people, when asked questions that are ostensibly about probability, interpret such questions as asking to predict with certainty the outcome of an individual trial of an experiment. The participants in Konold's study often based their predictions of random sampling outcomes on causal explanations instead of information obtained from repeating an experiment. Konold (ibid.) referred to these combined orientations as the outcome approach. Moreover, he noted that his participants adhered strongly to this approach, even in the face of evidence designed to impel them to change their perspective. Decades earlier, Piaget & Inhelder (1951) documented the same robust orientations among young children who participated in their experiments.

Sedlmeier & Gigerenzer (1997) conducted an extensive analysis of decades of research on the effects of sample size on statistical prediction. They concluded that participants across a diverse spectrum of studies who incorrectly answered tasks involving a distribution of sample statistics probably interpreted task situations and questions as being about individual samples.

Some notable studies have addressed students' understanding of sampling distributions and related ideas in instructional settings (Well, Pollatsek, & Boyce, 1990; delMas, Garfield, & Chance, 1999; Sedlmeier, 1999). These studies suggest that engagement in carefully designed instructional activities using computer simulations of drawing many samples can help orient

* Research reported in this paper was supported by National Science Foundation Grant No. REC-9811879. The first author gratefully acknowledges the support provided by the College of Liberal Arts and Sciences of Portland State University for the production of this report.

students' attention to collections of sample statistics when making judgments involving samples. Analyses in these studies did not focus on characterizing students' evolving conceptions and imagery of ideas in relation to their engagement in instruction. Consequently, these studies offer limited insight into interactions between engagement, learning, and instruction.

The ubiquity of variability in random processes and their outcomes is a central idea in statistics (Cobb & Moore, 1997). Despite its centrality, students' understanding of variability and our comprehension of variability's role as an organizing idea in statistics instruction have received little research attention (Shaughnessy, Watson, Moritz, & Reading, 1999). Rubin, Bruce, & Tenney (1991) elaborated a conceptual analysis in which they proposed that the integration of two seemingly contrasting ideas underlay a coherent understanding of sampling and inference: 1) *sampling representativeness*—the expectation that a sample taken from a population will often have characteristics similar to that population's, and 2) *sampling variability*—the expectation that different samples selected from a common population will differ among each other and from the sampled population. Rubin et al.'s (ibid.) investigation of statistically untrained high school students' reasoning on sampling and inference tasks showed that students did not integrate these two ideas to reason about distributions of sample outcomes. Instead, one or the other expectation seemed more salient in students' minds, depending on the task. Notably, ideas of re-sampling were neither at the foreground of the authors' conceptual analysis nor part of the student tasks.

Schwartz, Goldman, Vye, & Barron (1998) and Watson & Moritz (2000) both characterized sampling as a method of indirectly obtaining information about a larger population by directly obtaining information from only a relatively small and representative subset of the population. Neither characterization, however, entailed images of the repeatability of the sampling process nor of the variability that we can expect among sampling outcomes.

Saldanha and Thompson (2002) argued that the singular interpretation of likelihood (Kahneman & Tversky, 1972, 1982; Konold, 1989) and conceptions of sampling that do not foreground ideas of variability and the repeatability of the sampling process are problematic for learning statistical inference because they disable one from considering the relative unusualness of a sampling process' outcome. Drawing on the results discussed here and on data from a teaching experiment, Saldanha and Thompson (2002) characterized a conception of sampling that entails images of the repetitive sampling process, the bounded variability among sampling outcomes, and the fuzzy similarity between sample and population. In their characterization these images are linked schematically to form an organized system of ideas that, they claim, supports building deep connections between sampling and inference. They thus propose this scheme-based conception of sampling as a powerful and enabling instructional endpoint.

PURPOSE AND METHODS

This report is part of a study investigating students' abilities to conceive the ideas of variability, samples, and sampling distributions as an interrelated scheme. The study is the second in a sequence of two investigations that employed constructivist teaching experiments (Steffe & Thompson, 2000) as the method of inquiry. Our aim was to produce epistemological analyses of these ideas—ways of thinking about them that are schematic, imagistic, and dynamic—and hypotheses about their development in relation to students' engagement in classroom instruction (von Glasersfeld, 1995; Thompson & Saldanha, 2000).

Eight students—one 10th-grader, three 11th-graders, and four 12th-graders—enrolled in a year-long non-AP statistics course at a suburban high school in the Southeastern United States participated in a 17-session classroom teaching experiment during their fall semester. All students had completed a standard Algebra II course that included a short unit on statistics and probability. This was their only known prior formal instruction in statistics.

Students' understandings and emerging conceptions were investigated in three ways: 1) by tracing their participation in classroom discussions (all instruction was videotaped); 2) by examining their written work; 3) by conducting individual clinical interviews. Three research team members were present in the classroom during all lessons: one author designed and conducted the instruction; the other observed the instructional sessions and generated field notes; a third member operated the video camera(s).

Two overarching and related themes permeated instruction throughout the experiment's

various phases: 1) the process of randomly selecting samples from a population can be repeated under similar conditions, and 2) judgments about sampling outcomes can be made on the basis of relative frequency patterns that emerge in collections of outcomes of similar samples[1]. Instructional activities were anchored around these themes, which were intended to support students' developing a distributional perspective of sampling and likelihood (Kahneman & Tversky, 1982). With these themes in mind to guide the experiment's progress, activities and lessons were revised daily according to what the research team perceived as important issues that arose for students in each instructional session. Indeed, the instructional methodology was flexible and responsive to local interactions, allowing for extemporaneous and in situ diversions from the planned instructional activities whenever the instructor deemed that such would advance the overarching instructional agenda. This feature was made possible by the fact that the designer and instructor were one and the same person.

The instructional activities were designed and conducted as discussion-based inquiry-oriented investigations. In accordance with our research and instructional agenda, the mathematical content of the teaching experiment was light on calculations and symbol use, but heavy on explication, description, and connection of ideas[2]. This agenda was enacted by the team members in their on-going interactions with students; we moved to negotiate a culture of sense-making in the classroom by placing a high premium on and promoting pro-active participation as listening, reflecting, questioning, conjecturing, and explaining and describing one's own and others' thinking about mathematical ideas under discussion.

Our report focuses on broad developments in students' participation that point to their thinking as they engaged in part of a sequence of classroom activities occurring in an early phase of the teaching experiment. The activities were designed specifically to support students' abilities to reason proportionally about collections of values of a sample statistic, and to conceive these collections as distributions. We characterize key attentional and discursive shifts that occurred in the classroom and the co-evolution of students' ideas and engagement with instruction.

The instructional sequence began with a concrete sampling activity that focused students' attention on the variability among outcomes of samples drawn randomly from populations having a parameter of interest whose value was unknown. Students then used computer simulations to investigate what it means to think that an outcome of a stochastic experiment is unusual. The sequence concluded by having students compare multiple collections of sampling outcomes with the aim of deciding whether a given collection had an unusual distribution. The unfolding of this sequence and student ideas that emerged as they engaged in it is described in the next section of the paper.

RESULTS

In the initial sampling activity small groups of students hand-drew random samples from different dichotomous populations of objects, the compositions of which were unknown to them[3]. They recorded the samples' outcomes (i.e., the number of a particular kind of object in each sample) and investigated patterns in them. Discussions centered on what those patterns suggested about the proportions in the sampled populations, what individual sample outcomes might be if the experiment were repeated, and how to decide whether two sets of outcomes were similar or dissimilar. In these discussions, we first focused students' attention on the variability among sample outcomes, and there soon emerged a consensus that this variability makes problematic any claim about a population's composition that is based on an individual sample's result. This led to the idea of looking at collections of sample outcomes, instead. Each group drew 10 small samples of equal size from a relatively large population of objects and the class investigated how these collections, as a whole, were distributed.

---

[1] Similar samples share a common size, selection method, and parent population. Furthermore, they are selected to obtain information about a common population characteristic.

[2] The most sophisticated calculations used in the course were proportions.

[3] Just prior to this activity, the instructor led a whole-class discussion centered on a simulated sampling demonstration that broached the ideas of random sample, population, and making an inference from the former to the latter.

At this point students' discourse began shifting away from individual samples and toward collections of samples. The ensuing discussions focused on how to look at a collection of sample outcomes in order to claim something about the composition of the underlying population. One pair of students had selected 10 samples of 5 candies each from an opaque sack of red and white candies, the proportion of which was unknown to everyone. After some discussion of their results and noticing that they had many samples having more whites than reds (see Figure 1.), students concluded that there were more white candies than red ones in the sack. They based this conclusion on their observation that 80% of the samples were "heavy on the white"—that is, contained three or more white candies. We henceforth refer to this collection of outcomes (see Figure 1) as Result 1.

Number of red candies		0	1	2	3	4	5
Number of samples	*(Result 1)*	0	5	3	1	0	1

Figure 1. The outcome of drawing ten samples of 5 candies each from a population having 50% red and 50% white candies.

Thus, students seemed to reason that if a majority of the samples in the collection each contained a majority of white candies, then this was sufficiently strong evidence to infer that the population also contained a majority of white candies. This line of reasoning suggests that students were able to coordinate two levels of thinking: one level involves individual samples and their composition; another level involves partitioning the collection of sample compositions in order to ascertain what proportion of them are composed mostly of white candies. This was the first instance in which students displayed an ability to operationalize the criterion for deciding what to infer about the sampled population. It is worth noting that at this point, their inference seemed to be quasi-quantitative—it still lacked a precise measure. They claimed only that the population, like a majority of the samples, had a majority of white candies.

When the population proportion was finally revealed to them, students were surprised to learn that there were actually equal numbers of red and white candies in the sack. A sustained discussion then ensued about two questions: first, "might these results be unusual?", and then "how might we investigate this question?". It was in this context that one student suggested repeating the process of selecting 10 samples of 5 candies several times with the intent of comparing this sequence of results to Result 1: "test it over and over", she said, without specifying the nature of the proposed comparison.

At this juncture we introduced Prob Sim (Konold & Miller, 1996), a sampling simulation program, as a tool to efficiently simulate repeating the experiment of drawing 10 samples of 5 candies from a large population of 50% red and 50% white candies. Each time that we collected 10 samples, we displayed a histogram of the number of samples having various numbers of red and white candies (see Figure 2).

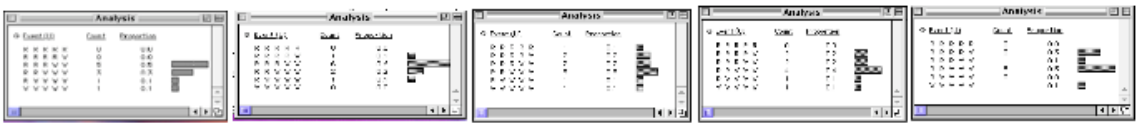

Figure 2. The sequence (from left to right) of outcomes of the simulated candy-sampling experiments.

Discussion of the results of each simulated experiment focused on making sense of and interpreting the information displayed in the Analysis window of the program.[4] For each collection of 10 sample results, the discussion concluded with a class vote taken to decide whether that collection was similar or dissimilar to Result 1. At first, students were uncertain about how to decide; some expressed opinions that they were unable to justify. We suggested that they think of a collection in terms of the relative weight of parts of its histogram. It emerged through the discussion that the characterization "heavy toward the white" was a criterion for deciding that a collection of simulated results resembled Result 1. The class applied this criterion over 5 simulations (see Figure 2), and the decisions were recorded and displayed in a table as

[4] Each Prob Sim window shows results of selecting 10 samples of 5 candies each from a large population evenly-split with red ("R") and white ("W") candies. Each row of the table appearing in a window displays an (unordered) outcome of the sample space. The number of samples having that outcome is expressed in absolute terms, as a proportion, and as a length of a histogram bar.

shown in Figure 3. In the end students agreed that Result 1 was not unusual once they saw that three of the five simulations produced results "similar" to it.

Distribution #	1	2	3	4	5
Similar ?	No	No	Yes	Yes	Yes

Figure 3. Record of decisions of whether the simulated process of selecting 10 samples of 5 candies produced a distribution of outcomes similar to Result 1.

DISCUSSION AND CONCLUSION

We find it useful to consider the instructional episodes described here as a sequence of phases unfolding out of cycles of interaction between instruction and student thinking and engagement. From our perspective, a first critical development occurred with students' realization—prompted by an orienting cue from instruction—that the variability among sample outcomes necessitated a consideration of how collections of outcomes were distributed in order to infer the underlying population's composition. This led to a second phase of engagement marked by a focus on multiple sample outcomes and re-sampling from various populations to accumulate collections of such outcomes.

In the case of the samples drawn from the population of candies, students seemed to reason proportionally about this collection; their inferences were based on the relative number of samples in the collection having a majority of white candies. We hypothesize that this reasoning entails thinking of a collection of sampling outcomes as having a two-tiered structure: on a first level one focuses on individual sample compositions and develops a sense of their accumulation; on a second level, one objectifies the entire collection and partitions it to determine a part's weight relative to the entire collection. This entails quantifying two different attributes—the composition of a collection of samples and the composition of individual samples—and coordinating these quantities so as to not confound them. These conceptual operations and their coordination can be taken to characterize critical aspects of imagining a collection of sampling outcomes as comprising a distribution. In retrospect, this line of reasoning was crucial to students' continued productive engagement in the instructional activities. We propose that it underlay their eventual ability to agree on a method for comparing entire distributions of sample proportions and to use this method as the basis of a rule for deciding when two such distributions are similar.

These developments, which marked the emergence of a third phase of the episode, were driven by interactions between two critical events: 1) the tension that students experienced when perceiving a discrepancy between their inference and the actual population proportion, and 2) instruction that capitalized on this tension by promoting a culture of inquiry around its resolution (e.g., by asking students to investigate whether their surprise was warranted). In this third phase students employed the earlier idea of collecting multiple samples and re- applied it to entire distributions of samples. That is, they were able to structure the simulated re- sampling results in collections of 10 and compare the distribution of each 10-sample collection to Result 1—a development facilitated by Prob Sim's presentation format and by the instructor's suggestion to record each result. In this phase of engagement students' focus was thus on entire distributions of 10-sample collections. As decisions about the similarity of each simulated distribution to Result 1 accumulated, they were recorded and organized (see Figure 3) to facilitate thinking about their collection itself as a distribution. Indeed, students were able to arrive at a judgment about the relative unusualness of distributions like Result 1 by considering what proportion of the collection of 5 distributions (of 10 sample outcomes) were "similar" to it, or, equivalently, what proportion of the distribution of 5 similarity decisions were "Yes". Students' reasoning here can thus be described as entailing conceptual operations similar to those described above, but applied to a more complex object: a collection of distributions of sample outcomes.

In sum, as students participated in these directed activities and discussions their thinking appeared to progress in complexity from focusing on single sampling outcomes, to reasoning proportionally about a collection of sampling outcomes, and finally to reasoning proportionally about a collection of such collections.

REFERENCES

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly (104)*, 801-823.

delMas, R. C., Garfield, J., & Chance, B. L. (1999). *Exploring the role of computer simulations in developing understanding of sampling distributions*. Paper presented at the American Educational Research Association, Montreal.

Glasersfeld, E. v. (1995). *Radical constructivism: A way of knowing and learning*. London: Falmer Press.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 509-521). New York: Cambridge University press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology (3)*, 430-454.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*(1), 59-98.

Konold, C., & Miller, C. (1996). *Prob Sim*. Computer Program. Amherst, MA.

Konold, C. & Pollatsek, A. (2002). Data Analysis as the Search for Signals in Noisy Processes. *Journal for Research in Mathematics Education, 33* (4), 259-289.

Piaget, J., & Inhelder, B. (1951). *La genèse de l'idée de hasard chez l'enfant*. Paris: Presses Universitaires de France.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed). *Proceedings of the Third International Conference on Teaching Statistics (Vol. 1, pp. 314-319)*. Dunedin, New Zealand: ISI Publications in Statistical Education.

Saldanha, L. A. & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257-270.

Schwartz, D. L., Goldman, S. R., Vye, N. J., & Barron, B. J. (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on statistics: learning, teaching, and assessment in grades K-12 (pp. 233-273)*. Mahwah, NJ: Lawrence Erlbaum.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Lawrence Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*, 33-51.

Shaugnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999). *School students' acknowledgment of statistical variation*. Paper presented at the Research Pre-session Symposium of the 77th Annual NCTM Conference, San Francisco, CA.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education (pp. 267-306)*. Mahwah, NJ: Lawrence Erlbaum.

Thompson, P. W., Saldanha, L. A., & Liu. Y. (2004). *Why statistical inference is hard to understand*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, April 2004.

Thompson, P. W., & Saldanha, L. A. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick, G. Martin, & D. Schifter (Eds.), *Research Companion to the Principles and Standards for School Mathematics (pp. 95-113)*. Reston, VA: NCTM.

Thompson, P. W., & Saldanha, L. A. (2000). Epistemological analyses of mathematical ideas: A research methodology. In M. L. Fernandez (Ed.), *Proceedings of the Twenty Second Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (Vol. 2, pp. 403-408)*, Tucson, AZ. Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31* (1), 44-70.

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