"IS THIS SAMPLE UNUSUAL?": AN INVESTIGATION OF STUDENTS EXPLORING CONNECTIONS BETWEEN SAMPLING DISTRIBUTIONS AND STATISTICAL INFERENCE

By

Luis A. Saldanha

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOHPY
in

Education and Human Development

August, 2004

Nashville, Tennessee

Approved:

Patrick W. Thompson
Paul Cobb
Clifford Konold
Kay McClain
Philip S. Crooke

To Alberta and Tomás.

In recognition of their sacrifices and their spirit of life-long learning.

ACKNOWLEDGEMENTS

The completion of this dissertation entailed the support of a number of remarkable individuals and entities. Foremost, I thank my research advisor, Patrick W. Thompson, for providing an intellectually stimulating doctoral apprenticeship of the kind I can only aspire to emulate with my own future graduate students. From our first virtual encounter in 1995, when he detected some fledgling potential in me, through our close collaboration on this research project from 1999 till 2002, Pat provided conditions propitious for fostering my intellectual growth and development from novice researcher to collaborator and co-author. Thank you, Pat: I cannot imagine a richer and more intellectually enabling and generative research apprenticeship.

I thank the other members of my committee for their availability, their patience, and their insightful feedback: Paul Cobb, Cliff Konold, Kay McClain, and Phil Crooke. In particular, I wish to acknowledge Paul Cobb's role in helping to create a research-intensive environment within the Department of Teaching and Learning that is providing a fertile training ground for a next generation of mathematics education researchers.

In addition I must express my gratitude to several members of the support staff in the Department of Teaching and Learning for their frequent assistance throughout my time there: Diane Nelson, Angie Saylor, Dana Thomas, and Sandra Uti.

The research reported in this study was funded by National Science Foundation Grant No. REC-9811879. Much thanks is owed to the foundation as well as to the students who participated in this study for making this research possible. My involvement in this research project was also enabled, in part, by FCAR doctoral fellowship number 972688, awarded by the Government of Quebec for the years 1996-1999.

The writing of this dissertation took place out-of-residence, in the quiet town of Hanover, New Hampshire and in the daily company of my partner, Kathryn J. Lively—an opportunity made possible through her generous financial and emotional support. I am indebted to Kathryn for helping ensure my timely completion of this document.

Finally, I wish to acknowledge the support, in forms too numerous to list, of my immediate family during the tenure of my studies: Alberta, Tomás, Paulo, and Pedro.

TABLE OF CONTENTS

PART II

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

This dissertation explores the emergence of ideas among eight high school students as they participated in a classroom teaching experiment addressing statistical concepts. Instruction aimed to move students toward embedding statistical inference within the foundational idea of *sampling distributions*—the distributional structure of a collection of a sample statistic's values that one conceives as emerging in the long run, under repeated random sampling. This embedding is deep and the connections are important, yet both are rarely a subject of instruction in statistics. In contrast, instruction in this teaching experiment engaged students with the inner logic of statistical inference, pushing them to explore its deep structure.

Given the study's exploratory nature, the central aim of this dissertation is to gain insight into what students understood of the ideas addressed in instruction in relation to their engagement with that instruction. The method employed to develop such insight is the *retrospective analysis* (Cobb, 2000; Steffe & Thompson, 2000). That is an after-the-fact systematic and coordinated examination of the data generated in the teaching experiment.[1] In this particular case, my examination is directed at characterizing students' understandings in terms of plausible underlying imagery and conceptual operations.

My examination of these data takes a multi-pronged approach: I characterize instructional activities, instructional interactions and student engagements, and students' emergent and stable understandings. Moreover, my analyses attempt to capture the emergent and dynamic interplay among these components. This results in a characterization of the teaching experiment as a sequence of interrelated instructional activities unfolding in synergy with the emergence of students' ideas.

The dissertation is divided into two broad parts distinguished by their content and structure. Chapters I through IV constitute the first part. Chapter I develops a directed analysis of previous research, highlighting issues of relevance for this study. Chapter II provides the background and context for the teaching experiment, detailing its important aspects and situating it within the prior relevant research. Chapter III elaborates the theoretical lenses employed in my analyses, taking into consideration the perspectives that underlay the design and implementation of the

[1] My role as research assistant in conducting this teaching experiment (NSF Grant No. REC-9811879; P.I. Patrick Thompson) enabled me to position myself as a co-investigator in this analysis. The narrative in this dissertation is thus spun from the perspective of a member of the research team attempting to communicate with nonmembers.

teaching experiment. Chapter IV details the procedures and methods employed in my analyses of the data.

In the dissertation's second part analyses and results are distributed across Chapters V through IX. Instructional interactions are characterized as unfolding in a sequence of four interrelated phases and students' conceptions that emerged within them are detailed. Each of Chapters V through VIII addresses one distinct phase of instruction. Chapter IX gives a summary overview of the teaching experiment and elaborates conclusions.

Looking ahead, much of this study focuses on students' experiences and thinking as they worked with collections of a sample statistic's values, organizing them in ways that could provide a basis for making statistical inferences. The study indicates that in exploring the deep structure of statistical inference, students *grappled* with that structure. Conceiving a collection of a sample statistic's values as a sampling distribution entails the coordination of multiple objects and actions in a hierarchical structure. Students experienced significant difficulties developing this coordination, even when individual objects and actions seemed unproblematic for them to envision. This suggests that developing a coherent understanding of sampling distributions is non-trivial and may be rooted in abilities to navigate with facility among hierarchically structured objects and processes.

PART I

CHAPTER I

LITERATURE ANALYSIS

In this chapter I develop a directed analysis of prior research relevant to this study. The studies considered here are drawn from a broad base of research in stochastic reasoning, specifically research on understanding ideas related to sampling and statistical inference. Some of these research studies were purely psychological investigations, while others explored students' performance and learning in contexts involving designed instruction. My analysis of this research highlights and develops particular issues that are relevant for understanding the rationale for the design of the teaching experiment and this study. Explicit connections between these issues and this study will be elaborated throughout the second chapter.

§

In American cognitive psychology Kahneman and Tversky (1972) spearheaded research on understanding sampling when, on the basis of empirical evidence, they hypothesized 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). This hypothesis suggests that Kahneman and Tversky's subjects' focused their attention on *individual* samples.

In later research, Kahneman and Tversky (1982) conjectured that people, indeed, tend to take a *singular* rather than a *distributional* perspective when making judgments under uncertainty. In the former, one focuses on the causal system that produced the particular outcome and assesses probabilities "by the propensities of the particular case at hand" (ibid., p. 517). In contrast, the *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 analyses instead of information obtained from repeating an experiment. Konold (ibid.) referred to these combined orientations as the *outcome approach*. Moreover, he noted that this approach was quite robust among participants[1] in his experiments; they were not easily swayed to abandon causal analyses and to consider patterns in the outcomes of a repeated experiment as a basis for prediction, even in the face of evidence designed to impel them to do so. Decades earlier, Piaget and Inhelder (1951) documented similar robust orientations among young children who participated in their experiments involving prediction under uncertainty.

The distinction between singular and distributional perspectives of probability was discussed by the mathematician Richard von Mises (1957). Von Mises' defined probability as the relative frequency of a *repetitive* event in a reference class (collective) of such events. In his, strong frequentist, view it is nonsensical to pose probability questions about single events if all one really has in mind is a single event unrelated to a repetitive scheme and a collection of such events. The historian of probability Ian Hacking (1975) traced this duality between frequentist and singular interpretations of probability back to the time of Pascal and Fermat. On the one hand, at that time probability had cultural connotations having to do with degrees of belief or opinion warranted by authority. On the other hand probability had ties to observed frequencies, such as the co-occurrences between fever and disease, the number of comets, and the deaths of kings. By Hacking's account the development of a mathematically coherent formulation of the quantification of uncertainty is a story of the transition from the former to the latter as a prevalent perspective among key players in its development. This transition was rife with struggles to shed the former, deeply culturally entrenched, perspective.

Gigerenzer and his colleagues (Gigerenzer, 1998; Gigerenzer, 1994; Hertwig & Gigerenzer, 1999) used the distinction between frequentist and singular interpretations of probability as the basis of a framework for understanding people's decision-making strategies under uncertainty. Arguing from a position in evolutionary psychology which claims that the human perceptual systems and mind are hard-wired to attend to "natural frequencies" in the environment, Gigerenzer (1998) proposed that in many cases people's non-normative responses to probability

[1] Participants were psychology undergraduates at a major comprehensive university in the U.S.

questions can be attributed to their non-frequentist interpretations of the questions, especially in questions not couched in the language of frequencies. Some of Gigerenzer's work relates to sampling. In an important meta-study, Sedlmeier and Gigerenzer (1997) analyzed thirty years of research on understanding the effects of sample size in people's statistical predictions. They argued compellingly that subjects in a diverse spectrum of studies who incorrectly answered tasks involving a distribution of sample statistics may have interpreted task situations and questions as being about individual samples.

§

Analyses of the meanings and interpretations of probability have been reiterated often in the psychological, historical, and educational research literature. Analyses of what it can mean to understand sampling and inference, however, are more rare.

Rubin, Bruce and Tenney (1991) proposed that the central idea of statistical inference is that a sample provides *some* information about its parent population. This idea implies that we should place bounds on what can be inferred from a sample to the underlying population. In their view, this line of reasoning entails balancing two ideas which, from a deterministic perspective, may appear contradictory: *sampling representativeness*—"the idea that a sample taken from a population will often have characteristics similar to those of its parent population", and *sample variability*—"the contrasting idea that samples from a single population are not all the same and thus do not all match the population" (ibid., p. 314). Rubin et al. (ibid.) asserted that the integration of these two ideas is key to developing a coherent understanding of statistical inference. Moreover, their investigation of statistically naïve high school students' responses on sampling and inference tasks suggested that students did not integrate these ideas to reason about distributions of sample outcomes. Instead, students tended to focus on one or the other idea, depending on the task; in some situations ideas of sample representativeness were more salient in their mind, in others they focused on ideas of sampling variability.

Two recent investigations of school students' understandings of sampling (Schwartz et al., 1998; Watson and Moritz, 2000) drew on conceptual analyses of sampling and statistical inference. The authors of both studies characterized sampling essentially as a method of indirectly obtaining information about a larger population by directly obtaining information from only a relatively small subset of the population.

Schwartz et al. (ibid., pp. 240-242) described the structure of a statistical inference as having the following components: a population, a random procedure for selecting a sample from the population, a resulting sample, and an inference from the sample back to an estimate of the population. Schwartz et al. (1998) proposed that a coherent understanding of sampling and inference is difficult for students because it entails thinking about collections of cases instead of individual cases, and because interpreting certain everyday contexts (e.g., opinion survey situations) in terms of sampling and inference entails navigating tensions between causal and random perspectives. Schwartz et al.'s (ibid.) analysis and research with elementary school students focused on the problematicity of integrating two images or schemas that may appear as incongruent to statistically naïve people; that is, selecting only a *part* of a population, yet obtaining reliable information about the *entire* population.[2]

Watson and Moritz (2000) studied a large cross section of Australian school students' ideas about samples and sampling. By characterizing students' ideas and evaluating their relative sophistication, they constructed a developmental model of conceptions of sampling. The authors cited the following characterization as encompassing aspects of sampling that were most important for their study: " […] a subset of the population called a sample is selected. Although data are then collected only from or about the sample, conclusions are drawn (generalized) to the larger population as well … What is the essential nature of a sample? In a word, a sample should be 'representative'. This means that, effectively, a sample should be a small-scale replica of the population from which it is selected, in all respects that might affect the study conclusions" (Orr, in Watson and Moritz, 2000, p. 48).

Together, Schwartz et al.'s and Watson and Moritz's analyses of sampling and inference highlight the following important aspects:

1) The goal of sampling: to reliably obtain information about a population, whose entirety is inaccessible. Thus, such information must be obtained indirectly;
2) The method for attaining this goal: selecting a subset of the population in such a manner (called "random") that information obtained directly from it will provide similar reliable information about the larger population;

[2] Stigler (1986, pp. 161-169) gave an insightful discussion of Laplace's and Quetelet's intellectual struggles with the idea of using only a subset of the population to conduct a national census. His discussion suggests that there is a historical basis for thinking that the integration of these two schemas is conceptually non-trivial.

3) The desired relationship between the sample and population: the subset of the population should have the "representativeness" property—it should be a small-scale replica of the population;
4) The inference: generalizing from the sample to the larger population.

This characterization describes aspects and images presumably entailed in a coherent understanding of sampling and inference. A significant feature of this characterization is that one could arguably be mindful of its aspects and still be unable to make judgments about a sampling outcome's relative unusualness. A thought experiment will help drive this point home: assume the perspective of a statistically naïve person and suppose one selects a random sample from a very large population having unknown composition. In the absence of any other information, there is no reason to doubt the "representativeness" of the sample and an inference about the unknown population is made on the basis of the sample's composition. This lack of doubt can only be based on faith that the random selection process will produce a sample that reflects the sampled population. Now suppose that the random sampling process is repeated: one draws another sample of the same size from the same population and observes that it has a different composition than the first sample. On the basis of *this* sample alone one will reach a different conclusion about the population's composition. How does one reconcile these differences? How does one ascertain which sample is more representative? Is one or the other sample unusual? One can imagine repeating this thought experiment many times and asking similar questions about the resulting collection of samples we end up with.

The point here is that the expected variability among outcomes of randomly drawn samples necessarily makes sample "representativeness" a problematic notion. Without recourse to images of what is expected to occur in the long run when a sampling process is repeated many times, students will have little hope of building a coherent image of what it means for a sample to be "representative". Schwartz et al.'s (ibid.) and Watson and Moritz's (ibid.) conceptual analyses do not problematize the notion of sample representativeness, nor do they entail images of the repeatability of the sampling process and of the expected variability among sample outcomes. Consequently, students who develop this characterization as their encompassing image of sampling and inference are arguably disabled from judging whether a sample outcome is unusual and from understanding the deep connections between statistical inference and distributions of sample statistics. In this sense, the characterization of sampling and inference elaborated by

Schwartz et al. (ibid.) and Watson and Moritz (ibid.) cannot help students move beyond a singular view of sampling. Indeed, a literal reading of the characterizations given by these authors suggests that they were referring to individual samples. Put differently, Schwartz et al's. and Watson and Moritz's conceptual analyses leave ideas that are foundational to sampling and inference unpacked. As such, I argue that their analyses do not characterize a sufficiently rich conception of sampling to target for instruction.

Rubin et al. (1991) hinted at the problematicity of sample representativeness when they pointed out that the variability among sample statistics forces us to place bounds on what we can infer from a sample to its underlying population. They suggested that integrating sample representativeness and sampling variability into a coherent conception of inference rests on having a relative frequency interpretation of "likelihood". Thus, from this frame a representative sample is one whose statistic (say, a proportion) is *likely* to be close to that of the underlying population, in the sense that we expect a relatively large proportion of statistics calculated from numerous repetitions of the sampling process to lie within some "close" distance of the population parameter. Similarly, sampling variability leads us to believe that some samples are *likely* to differ from the population parameter, in the sense that we expect some proportion of statistics calculated from numerous repetitions of the sampling process to lie beyond some "close" distance of the population parameter.

§

As the above discussion highlights, sampling variability is a central idea in statistics. Despite its centrality, however, students' understandings of sampling variability and our comprehension of variability's role as a central organizing idea in statistics instruction has received little research attention (Shaughnessy et al., 1999).

Shaughnessy et al. (ibid.) investigated school students' ideas that might be related to sampling variability. After students had observed a repeated sampling experiment, these researchers focused on what students predicted as the most likely *range* of outcomes to occur in a small number of randomly selected samples. The range of a set of values gives the distance (i.e., the absolute value of the difference) between extreme values of the data set (sample proportions, in this case).

Thompson (personal communications, 2000-2002) proposed that it is not productive to think of sampling variability without having an underlying mental image of the repeatable process that produces sample outcomes. He argues that one must have such a dynamic process in mind about which it makes sense to ask "what is varying?" Without such imagery, he argues, sampling variability is essentially reduced to "differences in sample outcomes". In Thompson's view, a powerful conception of statistical/sampling variability should be tied to developing a sense of sampling distribution; it should enable students to move beyond mere differences in sample statistics and toward structuring a collection of statistics in terms of the proportion of them that lie within certain sub-ranges of the entire collection's range. Thompson's view of variability, thus, appears to resonate with Rubin et a.l's (1991) elaboration. The difference is that Thompson emphasizes having in mind the dynamic imagery of the repeatable sampling *process*, whereas Rubin et al. (ibid.) did not.

From Thompson's perspective, variability and range are not one and the same, as Shaughnessy et al. (1999) seemed to suggest. Because an infinite number of collections of a sample statistic's values can have the same range, the range tells us nothing more about such a collection's distributional structure. As such, the range of a collection of sample statistics is arguably the crudest measure of its variability. Although a student's predicted range may indicate her sense of expected extreme outcomes, it need not suggest anything about her ideas of variability and distribution in the sense elaborated by Thompson. While range is sometimes used as a measure of a distribution's spread, as the above analysis indicates this use pre-supposes that one has already conceptualized distribution in the sense elaborated by Thompson.

As this discussion makes clear, it is possible to have different kinds of statistical variability in mind. This points to a need to problematize statistical and sampling variability in the research literature. For instance, it would be productive to research

1) what students may have in mind by "variability"; 2) possible meanings of statistical variability and relationships between them; 3) conceptualizations of variability that may support the development of coherent reasoning about distributions.

§

In the research literature, most investigations of people's understanding of ideas related to sampling have been purely psychological. Some studies already discussed here, however,

involved an instructional component whose aim was to promote the development of students' understanding of particular ideas related to sampling. The effects of instruction on students' thinking were then assessed by comparing their performance on pre and post-instruction test questions.

Shaughnessy et al.'s (1999) instructional intervention consisted of having school students observe a repeated sampling experiment in a classroom: a demonstrator drew around 5 samples of 10 candies from a thoroughly-mixed jar containing red, blue, and yellow candies in known proportions. The classroom teacher drew students' attention to the number of red candies in each sample as she recorded the results on the board for public viewing. Shaughnessy et al. (ibid.) were interested in knowing how this demonstration would influence students' predictions about the range of sample outcomes (i.e., number of red candies in a sample) they expected in a collection of 5 such candy samples. Although their research report said little about the specific intent of the activity, the authors presumably hoped that it would impel students to revise their pre-instructional predictions of the most likely range of red candies in a collection of candy samples and to re-align their predictions with normative ones. The authors' intent was also implied in their concluding hypothesis that the sampling activity had allowed students to "see the variation" (ibid., p. 20).[3] Given the previous discussion about meanings of variability, without more information, however, it is unclear what Shaughnessy et al. (ibid.) meant by this statement.

Shaughnessy et al.'s report did not focus on aspects of student engagement in instructional interactions. For instance, we do not know whether the classroom teacher highlighted particular aspects of the sampling experiment and its results for students, nor what kinds of classroom discussions ensued about aspects of the sampling experiment. In addition, the authors seemed to have assumed an unproblematic interpretation of the experiment and the sampling outcomes in particular. Consequently, it remains unclear *how* instruction in Shaughnessy et al.'s study (ibid.) may have helped student development.

Schwartz et al. (1998) compared their 5th-grade students' pre and post-instructional test responses on questions about sampling methods (in both chance contexts and everyday contexts) in order to see how students' understanding evolved as a result of their engagement in

[3] In both pre and posttest Shaugnessy et al. (1999) asked 4th-grade students to predict the most likely range of possible outcomes (i.e., number of red candies in a sample) in the sampling experiments. They noted that a significant percentage of students changed their minds and predicted a more normative range of outcomes on the posttest.

instruction. Schwartz et al.'s (ibid.) instructional intervention was quite sophisticated: it was based on a well articulated and integrated design and assessment framework and it employed an interactive video environment that structured student activity around iterative cycles of generating and revising their solutions to an "anchor problem". The anchor problem—*The Big Splash*—was designed to create a shared context in which students could make their ideas about sampling public through discussions in a classroom setting. A central assumption of the instructional design was that the processes of objectifying thought and of negotiating meaning with one's peers and more knowledgeable others are fundamental to conceptual development.

Briefly, *The Big Splash* activity engaged students in the task of designing a sample survey, in an everyday context, in order to estimate an income.[4] The aim of the activity was to provide a context to help students align their various everyday ideas that resemble the part-whole relationship between sample and population into a coherent view of sampling and statistical inference.[5] The interactive video environment included demonstrations of sampling intended to promote discussion and reflection on relationships between sample and population (part-whole and proportional mini replica), on sample representativeness and sampling bias.

Schwartz et al. (ibid.) coded students' responses to the pre and post-tests to indicate their ideas about what constitutes a good sample, whether they understood that a sample was different from a population, and whether they understood that a sample provides information about a population. The authors reported their interpretations of students' understandings of the purpose of surveys, sampling bias, randomization and stratification. Analyses of these data led Schwartz et al. (ibid.) to conclude that students' understanding of the part-whole relationship between sample and population improved after instruction.[6] The authors hypothesized that the real-world

---

[4] *The Big Splash* is part of *The Adventures of Jasper Woodbury*—a series of video-based complex problem scenarios developed by the Cognition and Technology Group at Vanderbilt [CTGV] (1992). Details about *The Big Splash* and its underlying instructional design and assessment framework are also described in Schwartz et al.(1998, p. 250, pp. 262-264).

[5] In a previous study, Schwartz et al. (1998) concluded that children do not possess an abstract schema that they can use to understand all statistical situations. Instead, their intuitive statistical understanding is comprised of a collection of overlapping, and even incongruent, schemas that are differentially evoked depending on the particular problem context. Moreover, the authors claimed that these are "everyday" schemas that approximate isolated relationships within the structure of a statistical inference but that share only some features with the overall structure. The Big Splash activity built on these findings; it was designed to help student integrate the disparate everyday schemas that they bring into sampling contexts into a more unified and stable conception of sampling.

[6] Comparison of pre and post-test responses showed relatively large increases in the percentage of students who viewed a sample as different from the total population and in the percentage who viewed a sample as providing information about the total population. These results were accompanied by relatively large decreases in percentages

context of *The Big Splash* had helped students interpret the world in terms of the mathematical relationship that allows extrapolation from the sample to the whole population (ibid., p.265). The authors did not elaborate further on how this may have occurred.

Schwartz et al. (ibid.) reported that their results were less conclusive as to whether students' understanding of sampling bias improved. The data reportedly suggested that students' understanding of the part-whole relationship between sample and population was implicated in their understanding of sampling bias. However, their understanding of the part-whole relationship did not appear to be sufficient for understanding bias (ibid., pp. 265-266). Schwartz et al. (ibid.) also concluded that *The Big Splash* activity helped improve students' understanding of stratified random sampling and sampling using a randomizing device.[7]

In a study of college students' understanding of the variability of the arithmetic mean, Well, Pollatsek, and Boyce (1990) used an individualized instructional session in which students observed computer simulations of random sampling and graphical displays of the distribution of sample means. Instruction aimed to help students visualize the distribution of sample means and to distinguish situations involving sample means from those involving the scores within a single sample.

The instructional session progressed in two phases; first, students used the computer program to simulate selecting a large number of small samples from a population, and they observed the distribution of the sample means on screen. The interviewer drew students' attention to the difference between this distribution and that of the population (also displayed on screen), highlighting the difference in their spreads. Students' intuitions and understandings were then probed; they were asked to anticipate how variable was the distribution of sample means for samples ten times as large, both by estimating what proportion of such sample means they expected would fall beyond a certain cutoff and by sketching what they expected the sampling distribution to look like. In the second phase, this procedure was repeated with samples ten times as large. The interviewer drew subjects' attention to the difference in variability (spread) between the sampling distributions for the two sample sizes, pointing out that the latter was

---

of students who did not have these views. Thus, after instruction a much larger fraction of the students seemed able hold in mind two conceptions about samples that they might previously have considered to be incompatible.

[7] The authors based their conclusion on the relatively large increases, from pre to posttest, in the percentage of students that preferred either of these sampling methods over a biased method, and on the relatively large decreases in the percentage of students that preferred biased methods.

dramatically smaller than the former. Students were then asked why they thought the sampling distributions were so different. The session concluded by having students attempt several transfer problems (ibid., p. 307).

Well et al. (ibid.) reported that instruction helped students learn to distinguish between distributions of sample means and distributions of scores within a sample. However, after instruction many students apparently still did not realize that the distribution of means for large samples is less variable than that for small samples. The authors thus concluded that students did not come to understand how sample size influences the variability of the mean.[8]

Two recent studies (delMas, Chance, & Garfield, 1999; Sedlmeier, 1999) also reported improvements in college students' understanding of situations involving sampling distributions and probability as a result of their engagement in sustained instruction. Both studies extended the instructional approach used by Well et al. (1990) into a sequence of instructional activities embedded within technology-intensive environments. Computer simulations of drawing many samples were employed to help students shift their focus from individual samples to collections of samples when making judgments involving sample outcomes.

delMas et al.'s (1999) study investigated the role of computer simulations in the development of college students' understanding of sampling distributions. Their study progressed in 3 research cycles that were driven by reflexively related developments in two components: 1) instructional and software refinements, and 2) student performance and understanding. Student learning in each phase was typically assessed by comparing pre and posttest performance. The research was conducted over a period of 18 months in technology-intensive statistics classes at the college level.[9]

In the first phase of delMas et al.'s study students used the *Sampling Distributions* micro world (ibid.) to simulate drawing hundreds of samples of a given size from a population.[10] Students recorded the sample means for different sample sizes and they described the shape and

---

[8] Some of the issues raised in the last section with regard to unarticulated meanings of sampling variability also apply to Well et al.'s study. However, I will not raise these issues again here.

[9] A total of 283 non-mathematics majors enrolled in introductory statistics courses across three universities participated in the study.

[10] The microworld enabled the user to simulate drawing random samples from predefined populations and to control various parameters such as sample size, number of samples, and shape of a population distribution. The program provided graphical displays of a population's distribution and histograms showing the distribution of sample statistics. The design was intended to facilitate guided exploration and discovery of sampling distributions.

spread of the resulting sampling distributions.[11] Then they answered questions designed to have them reflect on ideas related to the central limit theorem: "what is the effect of sample size on the shape, center and, spread of sampling distributions?". Students repeated these activities with normal, skewed, and "unusually shaped" populations. In the first phase posttest, students were shown a histogram depicting a population's distribution. Their task was to decide which of a series of other histograms shown to them best represented a distribution of sample means for samples of size 500 randomly drawn from that population. They then answered similar questions for larger sample sizes. Despite a reported significant improvement from pre to posttest,

delMas et al. (ibid.) found that a significant number of students did not seem to understand basic implications of the central limit theorem. For instance, some believed that as sample size increases, a sampling distribution approximates the population distribution with respect to both shape and variability. Accordingly, the researchers redesigned the activity for a second phase of the study.[12]

The revised activity placed more emphasis on comparisons of shape and spread than on recording of parameters and statistics. Students were asked to make direct comparisons of their pretest "predictions" with the sampling distributions produced by the program. The design rationale was that providing students with opportunities to test their own predictions and confront possible inconsistencies between their expectations and observed outcomes would promote conceptual change (Posner et al., 1982). The data reportedly suggested that the second version group outperformed the initial version group with respect to choosing a correct or good response for each item (delMas et al., 1999, p. 10). Nevertheless, there were still students who believed that as sample size increases, the sampling distribution becomes more like the population with respect to shape and variability. The authors conjectured that these students were using their knowledge of the distribution of elements within samples and (mis)applying it to the behavior of sampling distributions.

A third revision of the activity was designed to address this misconception and to help students distinguish between individual samples and sampling distributions. In this phase of the research the software was modified to include a "samples" window that allowed students to

---

[11] For ease of narration, I use the term *sampling distribution* in place of *distribution of sample means* when describing delMas et al.'s study.

[12] A different sample of 141 college students enrolled in an introductory statistics class participated in this phase of the study.

easily compare the distribution of the elements within individual samples and the statistics for each sample. Other parts of the program were also modified to support thinking proportionally about sampling distributions.[13] The actual activity was also embedded within a situational context: a story problem was created in which students had to make a likelihood decision about the outcome from either of two different sized samples. The activity took students through a series of steps culminating in the comparison of two sampling distributions for samples of different sizes to answer a probability question. delMas et al. (ibid.) reported that overall student performance in the third phase was compatible with results from the second phase.

delMas et al.'s (ibid.) general conclusions were that neither a straightforward presentation of information nor the use of technology and activities grounded in learning theory (Posner et al., 1982) necessarily lead to a sound conceptual understanding of sampling distributions. Indeed, the researchers found students who, after having participated in the third refinement of the activity, still believed tenaciously that larger samples produce a sampling distribution that is similar in shape to the population. This particular finding might point to students' difficulties in conceptualizing sampling distributions and their underlying re-sampling scenarios in ways that entail making distinctions among various interrelated objects and actions (i.e., a population and its composition, selecting a sample from that population and recording a statistic's value, repeating the last action to accumulate a collection of sample statistic's values, structuring this collection as a distribution, etc.).

Sedlmeier (1999, pp. 128-139) explored the effects of a "training program" on college students' performance on sampling distribution tasks.[14] Instruction used a virtual urn-model to run repeated simulations of drawing colored (marked) balls. Students were shown how sampling distribution tasks could be represented by an urn model and solved by analogy to that model. The design principle of the training was to have students make connections between the urn simulations and the corresponding sampling distributions generated by the computer. Sedlmeier conjectured that presenting information to students in frequency formats would facilitate their making the desired connections.

---

[13]For instance, a moveable slider was added to the horizontal axis of the sampling distribution histogram in order to help students determine what percent of samples' statistics fall above or below an arbitrary value.

[14] Twenty one students from various departments within a German university participated in the study. In the first session of the study students took a pretest, received training, and took a first posttest. Students were tested a third and fourth time 1 week and 5 weeks after the first session.

The training began by having students observe sampling with replacement from an urn containing 5 red and 5 blue balls. They were then asked to imagine that two samples were drawn from the urn, one of size 10 and the other of size 40. They were to decide in which sample the proportion of blue balls would be more likely to deviate by more than 10% from the mean proportion of 50% blue balls. Students then followed a sampling demonstration: the computer drew repeated random samples of 10 balls and 40 balls and calculated the proportion of blue balls in each sample. The computer stacked these proportions in slots placed along a horizontal axis and students could watch a histogram of each sampling distribution emerge.[15] In a text window, the software "explained" that the sampling distribution for the smaller sample size was flatter around the mean and had a higher variance than that for the larger sample size. The text presented the conclusion that the smaller sample was more likely to yield proportions that deviate by 10% or more from the true population proportion (ibid., pp. 130-132). The computer repeated this demonstration with an extreme case; it compared sampling distributions for samples of size 10 and 1000.

After the demonstrations, students were invited to freely choose two sample sizes and let the program create corresponding sampling distributions of the proportions of blue balls. After they had repeated this several times, the program text window emphasized that with a larger sample size there would be more sample proportions (or means) near the true value than with a smaller sample size.

In the next phase of the training, students were shown how to model the Maternity Ward problem (see Appendix A) with the software by replacing red and blue balls with the labels "girl" and "boy". The computer then created sampling distributions with samples of size $n = 15$ (smaller hospital) and $n = 45$ (larger hospital). Students compared the two histograms and were asked to decide in which hospital there would be more days on which the proportion of boys would fall between 40% and 60%.

In the last phase of the training, students used the software to construct sampling distributions for a variant of the Post Office Problem (see appendix). The training concluded with a discussion of issues that should be considered when making judgments involving the impact of sample size. Two points were stressed in the discussions: first, samples can come from

[15] The software displayed histograms side-by-side or superimposed them for easy comparison of distributions.

different urns (e.g., heights of mean and women) and might therefore not be comparable. Second, one should always check whether the sampling process can be considered random or whether it might be biased or even deterministic (ibid., p. 135).

The data in Sedlmeier's study (ibid.) consisted of students' aggregated response rates on sampling distribution tasks that appeared on the four tests. Sedlmeier compared correct response rates across the four tests to measure immediate and long term effects of training, and he compared results across old and new tasks to measure the transfer effect.[16] On the basis of observed increases on these measures, Sedlmeier concluded that students' performance improved as a result of the training they had received.

I close this section by discussing two significant issues with regard to the instructional interventions and studies described in it. First, instruction was, in general, carefully and purposefully designed. Much of it was sustained and employed sophisticated environments and tools. In particular, the use of computer simulations to help students build imagery of sampling as a repeatable process and to facilitate and structure their exploration of the behavior of collections of sample statistics is arguably instructionally productive. As suggested by the publication dates of delMas et al.'s (1999) and Sedlmeier's (1999) studies, this use of instructional micro worlds seems to be on the cutting edge of research in statistics instruction. Although most studies reported improvements in students' understanding of sampling ideas after their engagement in instruction—as indicated by the performance measures described—there is also evidence to suggest that helping students change their perspectives and understandings of sampling (e.g., singular to distributional reasoning, sampling bias) will continue to pose a challenge for designers of statistics instruction.

The second significant issue revolves around the methodology employed by these studies to assess the effectiveness of instructional interventions. Most studies relied heavily, if not exclusively, on a common approach; comparison of pre and posttest performance measures as the main indicator of student understanding and development. This research methodology is canonical. Its merits and drawbacks are well understood and accepted by the research

[16] "Old tasks" were those, such as the Maternity Ward problem (see Appendix A), that students had already encountered in the training program. "New tasks" were similar in structure to the training tasks but used different cover stories (ibid. p. 129).

community, and I will not engage in a debate about its pros and cons here. However, I put that its predominant use in these studies reflects issues that go beyond practical and efficient ways of conducting research and touch on what these studies were really interested in capturing.

Consider Sedlmeier's (1999) training study as a case in point. He emphasized students' observable performance on particular tasks, but did not fold back from the data to characterize students' ways of thinking and understandings that might express themselves in ways consistent with the observable data. In another sense, also, Sedlmeier (ibid.) de-emphasized the role and agency of the learner in his study. His description of the instructional activities stressed showing students a desired connection (e.g., showing them how to map a virtual urn model onto a situation involving sampling distributions) and having them observe demonstrations and text produced by the computer environment, but it said little about students' own mental activity and reflections as they participated in these activities. In fact, the study generated little information about students' experiences; their engagement, interpretations, and understandings, and plausible interactions among these. These aspects of Sedlmeier's study suggest that his central aim was to provide "hard" data on the effectiveness of the training program on performance, not to generate insight into student understanding and development in relation to their engagement with the instruction.

Some of these issues also apply to Schwartz et al.'s (1998) and delMas et al.'s (1999) studies, though to a much lesser degree. Although concern with students' engagement, reasoning, and development appears to have been much greater to those researchers, test performance was still used as the overriding indicator of student understanding and progress. This necessarily restricted the kind of insights that they were able to generate.

In sum, the issue is that heavy reliance on performance measures is of limited value if unaccompanied by interpretations that suggest students' ways of thinking—their interpretations, their understandings and imagery, and their conceptual difficulties—that ostensibly express themselves in observable performance. As much as the instructional studies described here might add to our comprehension of statistics learning, their usefulness is constrained by their almost exclusive reliance on this research methodology. The information produced by this methodology provides little insight into what students' understood as they engaged in instructional activities and into aspects of their engagement that helped move their thinking in productive directions.

§

To summarize, ample evidence from research on understanding samples and sampling suggests that students tend to focus on *individual* samples and statistical summaries of them instead of how *collections* of samples are distributed. There is also evidence that students tend to base predictions about a sample's outcome on causal analyses instead of statistical patterns in a collection of sample outcomes. These orientations are problematic for learning statistical inference because they arguably disable students from considering the relative unusualness of a sampling process' outcome.

The instructional studies discussed here generally suggest that engagement in purposefully designed instructional activities embedded within technology-intensive environments can influence students' understanding of sampling ideas and their performance on tasks in productive ways. These studies assessed the effects of instruction on student thinking and development largely by comparing students' pre and post-instruction test response rates. This methodology constrained the insights gained into the relationship between the development of students' thinking and their engagement in instructional activities; analyses' "grain size" were relatively course and students' understanding and development were not characterized in terms of conceptual schemes that might underlie their observable behavior and performance patterns.

Finally, while various studies have elaborated components of understanding sampling and inference, these components have not been synthesized to characterize a conception of sampling as a scheme of interrelated ideas entailing repeated random selection, variability, distribution, and representativeness.

CHAPTER II

BACKGROUND AND EXPERIMENTAL CONTEXT

This study is based on the second in a sequence of two whole-class teaching experiments that addressed the same mathematical content matter. Both experiments investigated high school students' thinking as they participated in classroom instruction designed to support their conceiving sampling and inference as a scheme of interrelated ideas including repeated random selection, variability among sample statistics, and distribution.[1] These teaching experiments are related to the prior research by their linkages to and disconnections from the issues highlighted in my analysis of that research. These relations will be explicated as I summarize features of the experiments.

I begin by summarizing the first experiment and its findings (Saldanha & Thompson, 2002), as they bear on the second experiment and this study.

The First Teaching Experiment

Twenty-seven 11th- and 12th-grade students participated in a whole class teaching experiment conducted within in a non-AP semester-long statistics course offered during winter 1999 at a suburban high school in the Southeastern U.S. (ibid.). The experiment addressed ideas of sample, inference, sampling distributions, margins of error, and interrelations among them. The aim was to develop epistemological analyses of these ideas (Glasersfeld, 1995; Steffe & Thompson, 2000; Thompson & Saldanha, 2000)—ways of thinking about them that are schematic, imagistic, and dynamic—and hypotheses about their development in relation to students' engagement in classroom instruction. Toward this end, students' thinking about ideas addressed in instruction was investigated in three ways: by tracing their participation in classroom discussions (all instruction was videotaped), by examining their written work, and by conducting post-experiment individual interviews.

---

[1] These teaching experiments were designed and conducted by Patrick Thompson, principle investigator and director of the research project *"Investigating the role of multiplicative reasoning in the learning and teaching of stochastic reasoning"* (NSF Grant No. REC-9811879). This project entailed 5 teaching experiments conducted over a 40-month period, and involving three different groups of participants. My involvement as research assistant on this project entailed a variety of activities: assisting with the design of instruction and assessment; assisting with teaching; interacting with and interviewing participants; collecting, organizing, and analyzing data.

Instruction stressed two overarching and related themes: 1) the random selection process 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.[i] These themes were intended to support students' developing a distributional interpretation of sampling and likelihood (Kahneman & Tversky, 1982; Konold, 1989, von Mises, 1957). Though an a priori outline of the intended teaching and learning trajectories (Simon, 1995) guided the progress of the teaching experiment, adjustments were made to instruction according to the research team's interpretation of issues that arose for students in each lesson.

The first teaching experiment progressed over 9-consecutive lessons and unfolded in three interrelated phases. It began with directed discussions centered on news reports that mentioned data about sampled populations and news reports about populations per se (raising the issue of sampling variability). The experiment then progressed to activities that led to questions of "what fraction of the time would you expect results like these?". This entailed having students employ, describe the operation of, and explain the results of computer simulations of taking large numbers of samples from various populations with known parameter values. The experiment ended by examining simulation results systematically, with the aim that students see that distributions of sample proportions are relatively unaffected by underlying population proportions[2], but are affected in important ways by sample size.

A preliminary report of the teaching experiment (Saldanha & Thompson, 2002) elaborated a conception of sample and sampling that emerged from analyses of student data. A small number of student participants, generally those whose performance on instructional tasks was strong and who were able to hold coherent discourse about ideas highlighted in instruction, had developed a stable scheme of images centering on repeatedly sampling from a population, recording the value of a statistic, and tracking the accumulation of these values as they dispersed themselves in an interval around the sampled population parameter's value. These students seemed to have a *multiplicative conception of sample* (MCS), in which an encompassing image is of a sample as a quasi-proportional mini-version of the sampled population. Moreover, this conception entails a salient image of the repeatability of the sampling process and an anticipation of the bounded

[2] In Chapter VIII I qualify this assertion and I elaborate the design rationale for promoting it as an instructional endpoint.

variability among sampling outcomes that supports reasoning about distributions of outcomes. Figure 2.1 attempts to capture this characterization pictorially:

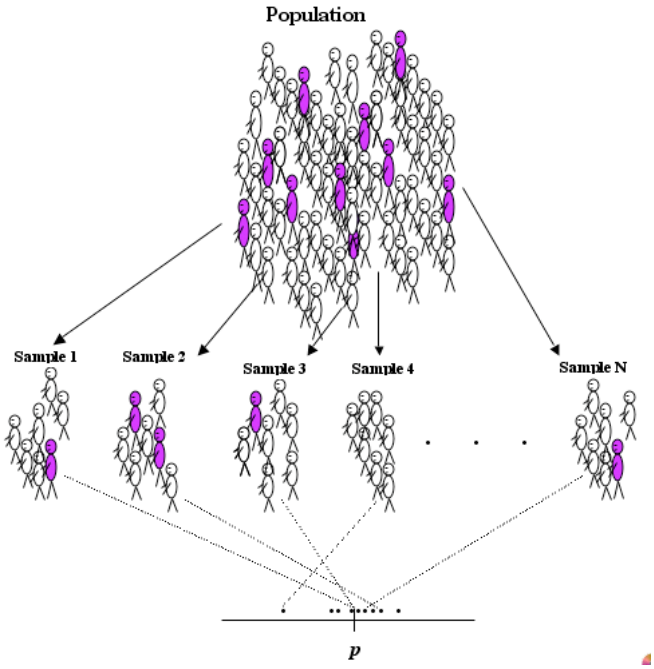


Figure 2.1. A multiplicative conception of sample entails a quasi-proportionality relationship between sample and population. Multiple samples are seen as multiple, scaled quasi mini-versions of the population.

There are two central reasons for highlighting MCS here. The first is to point out that MCS entails ideas already elaborated in the sampling research literature (Schwartz et al., 1998; Watson & Moritz, 2000; Rubin et al., 1991; Shaughnessy et al., 1999). However, it entails the composition of those ideas into an encompassing image that is arguably more empowering than the characterizations of samples and sampling elaborated in the literature. Because MCS entails an orientation to variability and the anticipated long-run behavior of a collection of a sample statistic's values, it provides a basis for understanding that judgments about particular sampling outcomes (e.g., confidence in their representativeness) must be made by appealing to distributions of such outcomes. MCS is therefore a powerful conception to target for instruction because it potentially enables those who develop it to understand the basis of statistical inference. Indeed, this speaks to the second reason for highlighting this conception here: it is that MCS constituted the research team's attendant vision of a desirable instructional endpoint in the second teaching experiment. As such, it helped guide the design of instructional activities in that experiment.

The Second Teaching Experiment

Purpose and perspectives

The second teaching experiment—the one forming the basis of this study—was conducted in fall 1999 within a year-long non-AP statistics and probability course given at the same high school in which the first experiment was conducted. This experiment addressed the same ideas and used a similar instructional approach as the first one (Saldanha & Thompson, 2002). In addition, the point of departure for this experiment was shaped by conjectures that the research team formulated about students' difficulties in the first experiment. Those conjectures will be elaborated in Chapter V, in Part II of the dissertation.

The experiment's aims were consistent with those of the first. The idea was to explore students' understandings of interconnections among ideas of sampling, statistical inference, and sampling distributions. This was done by engaging students in discussion-based classroom activities designed to support their making particular connections among these ideas that might, in turn, lead to their developing powerful understandings of them.

I use the word "support", in this last sentence, in a double sense: on the one hand it refers to an *end* of instruction, on the other hand it refers to a *means*. Both meanings did not carry equal weight in this teaching experiment. While it was important that instruction attempt to help students succeed in making particular targeted connections among ideas—indeed, those attempts are fully elaborated in Part II of this dissertation—achieving such success was of secondary concern to the research team. The primary concern was in having students engage substantively with instructional activities, so that interactions that flowed out of that engagement would support the emergence of significant mathematical behaviors which might constitute rich pointers to students' understandings, reasonings, interpretations, and general ways of knowing with respect to the ideas addressed in instruction.[3]

I hasten to add that this perspective is distinctly different from those of the prior relevant instructional studies. One apparent difference is that it places a high premium on gaining insight

---

[3] This perspective on research in learning and instruction draws on a metaphor from bio-chemical design experimentation. In that research paradigm, media are designed and "employed" to support the emergence (and perhaps development) of particular phenomena among agents interacting within those media. But having the phenomena of interest emerge is not, per se, the researchers' overriding interest. Rather it is studying the conditions and interactions that might or might not give rise to the phenomena's emergence. The first is a secondary goal, akin to a means, whereas the second is a primary one, akin to an end.

into what students understood as they engaged with instruction. As such, it takes both the local and global unfolding of instructional interactions as a primary data source. A related difference is that this perspective entails an encompassing orientation to look beyond students' observable behaviors, actions, and performance and interpret them as potential expressions of their underlying understandings, conceptions, and ways of knowing.

Participants and setting

Eight liberal-arts-bound students in Grades 10 through 12 participated in the teaching experiment. These eight students constituted the entirety of the class, the majority of whom enrolled in the statistics course without prior knowledge that it would entail experimental instruction. Students were notified on the first day of class what the experiment would involve; those students who remained agreed to participate in the study.[4] All students had completed a standard Algebra II course which included a short unit on statistics and probability—this was students' only known prior formal instruction in statistics.

The high school was located in an upper middle class suburb. The student population was fairly homogeneous in its racial and socio-economic make-up, consisting largely of white English-speaking adolescents from upper middle class backgrounds.[5] Like all academic courses offered at the school, assessment in the form of written in-class examinations and periodic progress reports in the form of quantitative and letter grades were mandatory and were expected by students, the administration, and parents. However, the research team was not bound to a particular curriculum and was relatively free to experiment with content and instruction.

Classes at the high school ran on an alternating weekly block schedule. The course usually met four times per week for periods of approximately 40 or 52 minutes each. The scheduling was such that on most days another class occupied our classroom during the periods immediately before and after ours. This constraint, together with the brief time period in which the 3-person research team had to set up and take down the audiovisual equipment needed to conduct and

[4] The research team was present and assumed responsibility for the course on the first day of classes. A few students trickled in and out of the course during the change period, in the first two weeks. 8 students participated beyond this period.

[5] One of the 8 student participants in this course was non-American, non-Caucasian, and a non-native English speaker. The rest were Caucasian Americans.

record each lesson, made it impossible to change the traditional row-seating arrangement of the classroom. Figure 2.2 is a schematic of the typical seating arrangement used in the classroom.

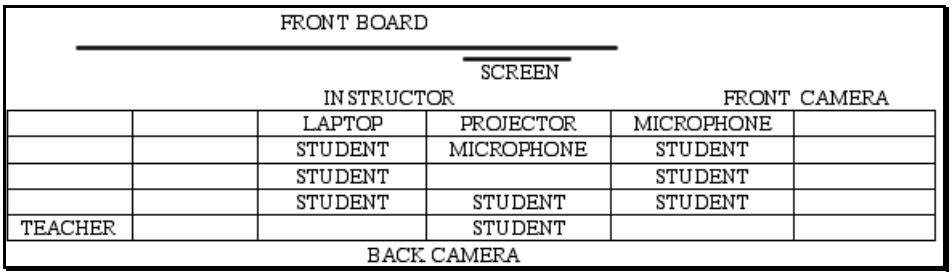


Figure 2.2. The typical class arrangement during lessons.

An experienced full-time mathematics teacher, then new at the school, was assigned as "formal" teacher of the course and was present during all lessons. The teacher had never taught statistics. He and the research team enjoyed an amicable rapport; he was cooperative, un-intrusive, and relieved to relinquish responsibility for teaching the beginning of the course to the research team. The teacher occasionally participated in the classroom discussions; sometimes as an advanced learner of the mathematics, other times as instructional facilitator.

Classroom culture and instruction

Instructional activities were typically designed and conducted as discussion-based inquiry-oriented investigations in a spirit consistent with that of *Principles and Standards for School Mathematics* (NCTM, 2000). In accordance with the research team's agenda, the mathematical content of the teaching experiment was designed to be light on calculations and symbol use and heavy on explication, description, and connection of ideas.[6] This instructional agenda was enacted by the research team members in their on-going interactions with students; they 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, and explaining and describing one's own and others' thinking about mathematical ideas under discussion.

Students' own accounts suggested that this milieu differed markedly from that in their other mathematics courses. As they indicated, an especially salient distinction for them was the discussion-based nature of activities and a style of engagement that demanded they attempt to

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

carefully describe, explain, and connect ideas. Another reportedly salient distinction for students was the instructor's frequent use of computing technology in classroom discussions. These impressions are consistent with those formed by members of the research team; their interactions with the mathematics faculty at the school indicated that a traditional lecture-based approach to teaching mathematics, typically adhering closely to chapter sequences in standard textbooks, was highly valued.[7] With one notable exception, this approach seemed to be the norm among the mathematics faculty in the school. Given these considerations, students likely experienced tensions between expectations for engagement in this teaching experiment and their history of engagement in school mathematics.

As I already mentioned, instruction in this teaching experiment was employed both as a means for supporting the development of students' understandings and as a tool for advancing the research agenda. Because the research goal was to explore students' thinking relative to the ideas addressed in instruction, instruction was necessarily flexible and allowed for considerable latitude in classroom interactions. Put simply, instruction followed students to the extent that the instructor, who was also the team leader, deemed it productive to do so. The flexibility of instruction is of central importance in this teaching experiment, imbuing it and its products with emergent features. These features will be elaborated and developed throughout the chapters in Part II of the dissertation.

Despite instruction's emergent features, two overarching themes were stressed in the content of instruction throughout the teaching experiment: 1) the process of randomly selecting samples from a population can be repeated under similar conditions, and 2) judgments about a sample's outcome (i.e., a statistic's value) can be made on the basis of relative frequency patterns that emerge in collections of outcomes of similar samples. The rationale for these themes drew on prior relevant research that indicates that students tend to base judgments about sampling outcomes on causal analyses of *individual* sample outcomes instead of statistical patterns in a *collection* of such outcomes. Further, as my analysis of prior research highlights, sampling and statistical inference are not typically treated as part of a system of interrelated ideas entailing repeated random selection, variability among sample statistics, and representativeness. Instruction in this experiment emphasized building connections among these ideas by anchoring

[7] This is based on anecdotal evidence obtained, in part, from the team's experience in supervising mathematics student teachers at this school.

them on the foundational notion of *sampling distributions*—that is, patterns of dispersion that emerge when a sample statistic's values aggregate over the long run.

Data corpus

The teaching experiment unfolded in a sequence of 17 consecutive classroom lessons over a period of 28 days. Individualized interviews were conducted with students at mid and post-experiment. All classroom interactions and interviews were audio-videotaped. Records of students' written work on activity-based assignments and on various assessment items given at different points were also collected. Other artifacts generated or employed in the teaching experiment include lesson plans, field notes, instructional activities, and computer programs. This set of materials constitutes the complete data corpus of this study.

CHAPTER III

THEORETICAL PERSPECTIVES

My analysis of the teaching experiment's unfolding is guided by a view of a synergistic cycle between three principle types of activity: the design of instruction, students' engagement and emergence of their ideas, and the research team's interpretation of student behaviors and conceptions (see Figure 3.1).

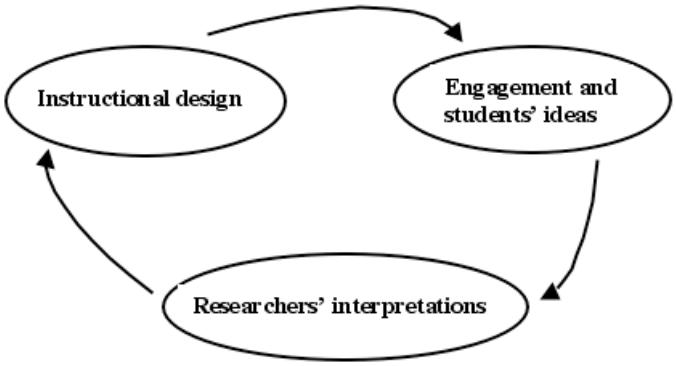


Figure 3.1. A cycle between 3 types of activity.

This was the typical cycle of activity entailed in the teaching experiment. The research team would design an instructional activity to address a particular mathematical idea or issue and with an envisioned mode of student engagement in mind. The instructor would then engage students with the activity, typically in a whole-class discussion format orchestrated so as to address the intended ideas and issues and to bring out students' interpretations and understandings related to them. Such engagements often led discussions in unanticipated directions, as part of the instructional agenda was to follow students' thinking whenever it seemed productive to do so. The instructor and research assistant would formulate impressions and interpretations of students' ideas and understandings that emerged during this engagement. These interpretations typically formed the bases of the design of more instruction, in conjunction with the overarching instructional agenda already elaborated in the last chapter.

The re-iteration of this cycle over time constituted a motive force that drove the evolution of instruction in tandem with the emergence of students' ideas as they engaged with instruction.

This image is much like Gravemeijer's (1994) and colleagues' (Gravemeijer et al., 2000) characterization of iterative mini-cycles of research and development that propelled their design experiments.

The retrospective analysis in this study retraced this cycle of activity as it unfolded over the course of the entire teaching experiment. Drawing on the data corpus generated in the teaching experiment, I analyze and describe instructional activities, I characterize students' ideas and understandings that emerged as they engaged with activities, and I describe how the research team's interpretations of the students' emerging understandings and conceptual difficulties fed back into the design of subsequent instruction.

This chapter elaborates three theoretical perspectives that came into play in my descriptions and analyses of these components:

1. Radical constructivism
2. Quantitative reasoning
3. Didactic objects and didactic models

These perspectives were employed by the research team as interrelated components of a framework that guided the design and of the teaching experiment. The team's aim was to generate conceptual analyses, of ideas addressed in instruction, that took into account students' experiences and engagement within instructional interactions. Since the aim of this study is consonant with that goal, I draw extensively on these perspectives in my analyses and therefore describe them in the coming pages.[1] Though these perspectives are seen as interrelated, I separate them here for the purpose of explication.

Radical Constructivism and Conceptual Analysis

Glasersfeld (1995) elaborated Piaget's genetic epistemology (1971, 1977) into a psychological theory of knowing that is commonly referred to as radical constructivism. A central image in radical constructivism is of individual cognizing agents engaging in purposive goal-directed activity, in which they strive to makes sense of their personal experiences in environments and social settings. Glasersfeld used the term "radical" in his constructivist theory

[1] A necessary first part of such a conceptual analysis—perhaps more aptly termed *epistemological analysis* (Thompson & Saldanha, 2000)—is a study of this kind that documents the emergence of students' ideas in relation to engagement with instruction.

to highlight its non-representational view of mind, knowledge, and knowing. Two basic tenets of this epistemological position are:

1) knowledge is not passively received but is built up by the cognizing subject;

2) cognition serves an adaptive function; to organize one's experiential reality, not to discover ontological reality (Glasersfeld, 1995, p. 51).

The "experiential reality" that Glasersfeld distinguished from ontological reality is defined essentially as the *structure* that a conscious cognitive agent creates in the stream of its otherwise amorphous experience. The activity of creating this structure—the building up of personal knowledge by organizing one's experiential world—is characterized as "operating". Drawing on the work of Ceccato (1949) and Piaget (1971, 1977) radical constructivism maintains that these operations can be explored. The lineage connecting these strands of research is worth elaborating briefly, as it provides a context for understanding Glasersfeld's approach to the analysis of conceptual operations.

Piaget's *genetic epistemology* (Piaget, 1971, 1977) was an interdisciplinary approach to understanding human intellectual, moral, and social development. Genetic epistemology made deep connections among biology, philosophy, psychology, and logic, and used both structural and functional approaches to understanding what might constitute human knowledge. The ideas that knowing is always a dynamic process, always involving mental operations, and that mental operations are always part of a larger system of operating, were central to Piaget's work.[2] On the other hand, while Piaget described mental structures as being organizations of mental operations, he emphasized the structural aspect of knowledge over the operational aspect of knowing. But he always grounded his notion of knowledge firmly in the idea that knowledge is not a copy of reality, but rather is built from and within a person's total neural activity (Thompson & Saldanha, 2000).

Working independently of the Piagetian school, Silvio Ceccato outlined what he called *tecnica operativa*, or operational technique, in which one must "consider any mental content (percepts, images, concepts, thoughts, words, etc.) as a result of operations" (Cecatto, as cited in Bettoni, 1998). That is, one must describe *consapevolezza operativa*, or conceptual operations (translated literally as "operating knowledge")[3] in order to answer the question "which mental

---

[2] Later in the chapter I characterize what Piaget meant by "mental operation".

[3] Quotations translated by M. Bettoni. Phrases translated by GO Translations, http://translator.go.com./.

operations do we perform in order to conceive a situation in the way we conceive it?" (Bettoni, 1998; Thompson & Saldanha, 2000).

Glasersfeld combined aspects of Ceccato's operational analysis and Piaget's genetic epistemology to develop an analytic method that he called *conceptual analysis*. The method's aim is to create models of concept construction that describe hypothetical conceptual operations by which one might come to know something in the way one apparently does. Glasersfeld's conceptual analysis frames reasoning and communicating as imagistic processes and knowledge as an emergent aspect of them (Glasersfeld, 1978).

I drew on radical constructivism for my analyses of individual students' understandings and ways of knowing in two central ways. First, it functioned as a background epistemological stance that served to constrain and orient the kinds of descriptions and explanations I proposed of students' thinking (Thompson, 2000). For instance, given the first tenet of radical constructivism, I tended not to presume that because students engaged in instruction that prompted them to attend to a certain way of thinking about a particular mathematical idea that they in fact developed a disposition to think about it in the intended way. This constrained me from taking for granted that students' understandings were in line with intended instructional endpoints, and at the same time it oriented me to look for evidence different from instructional actions as a basis for describing and explaining students' plausible understandings.

In a similar vein, given radical constructivism's second tenet, I tended not to presume that particular objects and inscriptions designed for and employed in instructional activities were transparent to students. Instead, I was oriented to take them as potentially problematic for students and perhaps having a variety of possible interpretations.

The second prominent role that radical constructivism played was in orienting me to characterize students' thinking in operational terms. Describing and explaining mathematical behaviors in terms of plausible underlying conceptual operations—the kind that are "near the surface of consciousness" and therefore "learnable"—provides a useful starting point for theorizing about ways of thinking and conceiving ideas that are more or less powerful. Moreover, this starting point can serve as a useful basis for designing instructional activities and

engagements intended to foster the development of particular ways of thinking about targeted ideas.[4]

While radical constructivism functioned as a background theory that served to constrain and orient the kinds of descriptions and explanations I proposed of students' thinking, it did not provide their content such as mathematical conceptions per se. Nor could it provide descriptions of instructional activities. For this I drew on two foreground perspectives: a theory of quantitative reasoning (Thompson, 1994) and an instructional design framework (Thompson, 2002).

Quantitative Reasoning

Thompson's (1994) theory of quantitative reasoning is about people conceiving situations in terms of quantities (i.e., things having measures or measurable attributes) and relationships among quantities. A powerful aspect of this theory is its focus on the psychology of *conceptualizing* quantity and measurement. In Thompson's frame, a quantity is not an ontologically real entity that exists independently of people's conceptualizations. Rather, *quantity* is a conceptual entity that a person constructs when conceptualizing situations: one thinks of a quantity when conceiving an attribute of an object as measurable. Thompson described this schematically; "it [quantity] involves an object-image, a conceptualized attribute of the object, a tacit understanding of appropriate units of measure, and a quantification process—a process by which one directly or indirectly assigns numerical values to the attribute" (Cortina, Saldanha, & Thompson, 1999). Key to Thompson's characterization is the idea that a coherent conception of quantity entails conceptualizing situations in ways that support conceiving of attributes embedded within them as measurable.[5] For only then can the question of how to measure and attribute (e.g., to determine "how much of it there is") be sensibly addressed.

Thompson's theory of quantitative reasoning is germane to this study in two salient ways. First, the theory underlay the research team's vision of the kinds of reasoning it hoped students

---

[4] This idea is fully developed in Thompson (2002).

[5] Thompson had a ratio scale measure in mind when he defined "conceiving an attribute of an object as measurable" to mean that one conceives of the attribute as segmentable, and that segmentation is in comparison to some standard amount of that attribute (Thompson & Saldanha, 2003)

would develop with respect to the statistical and probabilistic ideas addressed in instruction. As such, the theory guided the design of particular instructional activities.

The research team often approached particular statistical ideas addressed in instruction from the perspective of measurement and quantity. For instance, numerical data were often treated as representing measures of sampling outcomes—that is, attributes of a collection of items selected from a population. Similarly, a population parameter was treated as a measure of an attribute of the population of sampled items. Instruction had students work with collections of such data values in ways that were intended to raise issues having to do with quantifying attributes of them: "What proportion of a collection of sample statistic's values are within such and such range of the sampled population parameter's value?", "How can we measure how densely packed around the population parameter value is a collection of a sample statistic's values?", "How can we quantify how unusual a particular kind of sampling outcome might be?", "How can we quantify our confidence or expectation that a particular sample statistic's value is representative of the sampled population parameter's actual value?"

Much attention was given, in instruction, to engaging students in activities that might provide them with an experiential basis for thinking about these issues and addressing such questions more generally. The specifics of such instructional activities and the particular issues discussed within them with them will become clear throughout Chapters V through VIII. My aim here is to orient the reader to the significance of Thompson's theory for the design and implementation of instruction in the teaching experiment.

Second, Thompson's theory is also directed toward explicating the conceptual operations by which people come to conceive situations quantitatively, both developmentally (over time) and in specific settings. In its application to the analysis of conceptualizing statistical ideas, the frame was a useful tool in the retrospective analysis in this study. It provided a point of reference for characterizing students' mathematical reasoning and understandings in comparison with those targeted in instruction. More specifically, it oriented my descriptions of what students did and what difficulties they experienced in thinking of the statistical concepts in terms of ideas related to measurement and quantity. Thompson's theory draws heavily on Piaget's ideas of image, scheme, mental operation, and reflective abstraction to describe the emergence of quantitative ideas in terms of schemes of mental operations. The next three subsections elaborate these ideas, as my analyses of students thinking and understandings draw on them.

Action scheme, assimilation and accommodation

As I mentioned earlier, the idea of goal-directed action was central to constructivism and hence to Piaget's theory of knowing and genetic epistemology. Action, as Piaget thought of it, was always purposive and undertaken by an agent in the context of attaining some goal. But an action is not the same as an observable behavior. To Piaget, actions could range in complexity from the most basic, and directly observable, sensorimotor actions (e.g., the rooting reflex) to the most sophisticated imaginative actions (e.g., the group transformations of the square) that need not be expressible in any observable behavior (Thompson, 1994). Furthermore, actions are tied, to varying degrees, to experience—whether those experiences be largely concrete and sensori-motor or mental and reflective.[6]

Piaget (1971, p. 42) defined a "scheme" as "whatever is repeatable and generalizable in action". However, this characterization is deceptively simple for a construct that was such a unifying component of his theory. Glasersfeld (1995, p. 65) elaborated this into *action scheme*—a framework that posits a global structure of goal-directed action that, together with the processes of assimilation and accommodation, explains how an agent might come to know *whatever is repeatable and generalizable in action.* An action scheme consists of a three-part pattern:

1) Recognition of a certain situation (i.e., an internal state that is necessary for the activation of actions composing it);

2) a specific activity associated with that situation (i.e., the actions themselves);

3) an expectation that the activity produces a certain previously experienced result (i.e., an imagistic anticipation of the result of acting).

Glasersfeld (ibid., p. 65) explained that the "recognition" in part 1 is always the result of an *assimilation*, by which he meant that an agent "always reduces a new experience to already existing sensorimotor or conceptual structures". The process of *accommodation* is characterized very nicely in Glasersfeld's description of the workings of actions schemes:

> "An experiential situation is recognized as a starting-point of a scheme if it satisfies the conditions that have characterized it in the past. From an observer's point of view, it may manifest all sorts of differences relative to past situations that functioned as

[6] Piaget's idea of action was broader than I cast it here, entailing ideas of emotion and affect as well (Piaget, 1967).

trigger, but the assimilating organism (e.g., the child) does not take these differences into account. If the experiential situation satisfies certain conditions, it triggers the associated activity […] The activity, part 2, then produces a result which the organism will attempt to assimilate to its expectation part 3. If the organism is unable to do this, there will be a perturbation (Piaget, 1974, p. 264). The perturbation, which may be either disappointment or surprise, may lead to all sorts of random reactions, but one among them seems likely: if the initial situation 1 is still retrievable, it may be reviewed, not as a compound triggering situation, but as a collection of sensory elements. This review may reveal characteristics that were disregarded by assimilation. If the unexpected outcome of the activity was disappointing, one or more of the newly noticed characteristics may effect a change in the recognition pattern and thus in the conditions that will trigger the activity in the future. Alternatively, if the unexpected outcome was pleasant or interesting, a new recognition pattern may be formed to include the new characteristic, and this will constitute a new scheme. In both cases there would be an act of learning and we would speak of an 'accommodation'. The same possibilities are opened, if the review reveals a difference in the performance of the activity, and this again could result in an accommodation." (Glasersfeld, 1995. pp. 65-66).

Thus, accommodation is the altering of one's existing conceptual structure(s) impelled by one's efforts to assimilate situations. In this sense, accommodation is an operation that enables an agent's continued assimilation of encountered situations. But as Glasersfeld's description implies, the relationship is not unidirectional. Instead, assimilative and accommodative actions are reflexively related—either type can underlie and lead to the other type. Moreover, when taken together as a system of reflexively related operations, they have a generalizing effect in that they drive an agent's ability to engage in repeated and progressive goal-directed action.[7]

[7] Piaget distinguished among three related types of assimilations. (1) Functional or reproductive assimilation consists of repeating an action and of consolidating the action by this repetition; (2) recognitive assimilation consists of discriminating the assimilable object in a given scheme, and (3) generalizing assimilation consists of extending the field of this scheme (Piaget, 1977, pp. 70-71)

Imagery

The parenthesized statements appearing next to Glasersfelds' elaboration of the parts of an action scheme are Thompson's (1994) restatement of them. His language emphasizes that an agent's recognitive capacities are due to internal states that are the result of its own activity and actions. Furthermore, Thompson's restatement of part 3 relates schemes to *imagery*. Indeed, Piaget's notion of imagery was broad and tightly bound up with action schemes, this is one reason why it isn't easily explicated or comprehended.

Consistent with his non-representational view of mind, to Piaget images were not like static mental pictures, nor were they data structures produced by perceptual processes (Kosslyn, 1980). Instead, Piaget thought of images as dynamic and reconstitutive of experiences, and as entailing vestiges of the mental operations that constituted them. "Piaget focused on images as the product of *acting* […] To Piaget, images are residues of coordinated actions, performed within a context with an intention" (Thompson, 1996, p. 270). Piaget distinguished among three types of images according to the level of abstraction and development of the image's constituent operations:

1. An "internalized act of imitation … the motor response required to bring action to bear on an object … a *schema* of action."
2. "In place of merely representing the object itself, independently of its transformations, this image expresses a phase or an outcome of the action performed on the object …[but] the image cannot keep pace with the actions because, unlike operations, such actions are not coordinated one with the other."
3. "An image that is dynamic and mobile in character … entirely concerned with the transformations of the object … the image is no longer a necessary aid to thought, for the actions which it represents are henceforth independent of their physical realization and consist only of transformations grouped in free, transitive, and reversible combination". (Piaget, quoted in Thompson, 1994, p. 181)

Abstractions of sorts

Internalization and *interiorization* refer to the reconstruction of actions, at different levels of abstraction, that can enable mental imagery of situations involving the actions (Piaget, 1977; Thompson, 1994; Vuyk, 1981). Actions are internalized if they can be carried out in thought—that is, if the actions' execution and its result can be imagined without having to be

physically carried out. But internalization is the result of an initial abstraction and the resulting imagistic reconstruction may still be heavily constrained by aspects of the concrete experience. For instance, the reconstruction may occur only in real-time and feel like a mental replay of an experience. Interiorization is a further abstraction—"a progressive reconstruction and organization of actions" (Thompson, 1994, p.181)—that further emancipates the imagined actions from the initial concrete experience. For instance, an interiorized sequence of actions and its result may be imagined in an accelerated time frame and understood without the need to mentally play out each step of the composition.[8]

As Piaget's comments on imagery, above, imply, he thought of a mental operation as an internalized or interiorized system of coordinated actions.[9] Furthermore, he implied that while mental operations are always implemented in images, an image need not be tied directly to the origins of the operation (Thompson, 1994).

Piaget distinguished two broad types of abstractions that drive the reconstructions and reorganizations in internalization and interiorization. He spoke of *simple* or *empirical* abstraction*s* as deriving directly from *objects*: "… it is one thing to extract a character, x, from a set of objects and to classify them together on this basis alone, a process which we shall refer to as 'simple' abstraction and generalization …" (Piaget, quoted in Glasersfeld, 1995, p. 103). However, this characterization can be misleading to a reader who takes a realist interpretation and doesn't keep in mind that Piaget thought of objects as constructions. Moreover, Piaget was speaking from the perspective of a cognizing agent still unaware of the role of his own actions on his experiential reality. With this in mind, Glasersefeld (1995) more aptly called an abstraction "empirical" if it abstracts properties directly from sensorimotor experiences. For example, a child may abstract from his or her basic visual and tactile experiences with apples the rule "all apples are green and smooth". Such a child might be surprised when presented with the possibility of considering russet apples as apples.

---

[8] An example of interiorization, suggested by Pat Thompson (personal communication, 1997), is when one is able to imagine running through values in the domain of the function $f(x) = 3x - 2$ and generating a subset of its range without having to mentally perform the indicated mathematical operation for each value one runs through. Rather, this is an imagistic anticipation of the target set corresponding to a subset of the function's domain.

[9] Another property of operation not stressed in this characterization is *reversibility*. That is, an internalized system of coordinated actions is a mental operation for someone if one can imagine its decomposition and a corresponding change in state of affairs (situation or object).

The second type of abstraction—*reflective* abstraction—is derived from a cognizing subject's own activity and coordination of actions. Piaget (1971) illustrated reflective abstraction with an account of how a young child discovered the commutativity of addition by placing ten pebbles in various configurations (in a line and then in a circle) and then noticing that he always got ten pebbles regardless of whether he counted them from right to left or from left to right in a line, or clockwise or counterclockwise in a circle: " It is true that the pebbles, as it were, let him arrange them in various ways; he could not have done the same with drops of water. So in this sense there was a physical aspect to his knowledge. But the order was not in the pebbles; it was he, the subject, who put the pebbles in a line and then in a circle. Moreover, the sum was not in the pebbles themselves; it was he who united them. The knowledge . . . was drawn from, then, not from the physical properties of the pebbles, but from the actions that he carried out on the pebbles" (Piaget, 1971, p. 17).[10]

In a broad sense, Piaget formulated a theory of the development of human cognitive actions from the most basic to the most sophisticated. The internal mechanisms that he posited as driving this development were themselves described as sophisticated high-level actions that cognizing subjects need not be conscious of. Indeed, Piaget's theory characterizes cognitive development as a recursive, self-referential, and self-modifying process that is necessarily diachronic; a cognitive agent's constructive efforts and constructions are always constrained by its current organizational state and by a history of its constructive efforts. This was clear in the following quote: "… no behavior, even if it is new to the individual, constitutes an absolute beginning. It is always grafted onto previous schemes and therefore amounts to assimilating new elements to already constructed structures (innate, as reflexes are, or previously acquired" (Piaget, quoted in Glasersfeld, 1995, p. 62). The emergence, from this process, of relatively stable constructions and organizations of images are what underlie people's understandings or ways of knowing in specific settings. This view resonates strongly with Johnson's (1987) characterization:

> "Grasping a meaning is an *event* of understanding. Meaning is not merely a fixed relation between sentences and objective reality, as Objectivism would have it. What we typically regard as fixed meanings are merely sedimented or stabilized structures that emerge as recurring patterns in our understanding" (Johnson, 1987, p. 174).

---

[10] Piaget further distinguished among three types of reflective abstractions that depended on the developmental level of the cognizing agent's actions (Glasersfeld, 1995). I do not describe these here.

The language of conceptual schemes, imagery, and mental operations can be a powerful tool to describe students' understandings of situations they encounter in instruction and to describe the development of rich, connected, coherent understandings of mathematical ideas. It is also useful in both the design and analysis of instruction aimed at having those understandings emerge from students' engagement and reflection.

Didactic Objects and Didactic Models

The design of instruction in the teaching experiment was organized around the ideas of didactic objects and didactic models (Thompson, 2002). These ideas are grounded in the research team's interest in Glaserfeld's (1995) style of conceptual analysis. In Thompson's (2002) view, conceptual analyses of mathematical ideas and understandings cannot be carried out abstractly, but rather must be given in terms grounded in people's conceptual experience. Doing conceptual analyses entails imagining students having *something* in mind in the context of *discussing* that something. (ibid., 2002). Toward this end, instructional activities were designed with two central aims in mind: 1) to create opportunities for students and the instructor to discuss particular things, objects, or ideas that needed to be understood and to discuss how to imagine such things, and 2) to create opportunities for the instructor-researcher to ensure that specific conceptual issues would arise for students as they engaged in discussions with him.

When goals 1 and 2 above are realized with regard to a particular idea, they can end up producing instructional conversations (interactions) around that idea. Put simply, then, the overarching design rationale for instructional activities in the teaching experiment was grounded in the research team's desire to engineer situations that would engage students in instructional conversations that might support building psychological models of their understandings.

At the core of this design rationale, then, is an image of students purposively participating in conversations engineered and choreographed to foster reflection on some mathematical *thing*—an object, an idea, or a way of thinking. The term *didactic object* refers to "a thing to talk about that is designed to support reflective mathematical discourse involving specific mathematical ideas or ways of thinking." (Thompson, 2002; p. 210). The instructional activities employed in the teaching experiment were typically designed to entail such things. For example, the experiment's opening instructional activity engaged students in a sampling activity in which

they generated collections of a sample statistic's values. These collections of data values were then represented in frequency tables by the instructor. These tables organized the collections in a way intended to support having reflective conversations about how to construe and structure these collections so as to be able to compare them, for instance. These conversations turned out to be productive in a number of ways. But the frequency tables themselves did not constitute a didactic object. Rather, the way in which they were used by the instructor in the total activity made them a didactic object. Indeed, a didactic object is akin to a tool. But just as a hammer is not a tool unless conceived as such, an object is not in and of itself didactic. Rather, an object can become didactic in the "hands" of someone who conceives using it in ways that enable student engagements and conversations that foster productive reflection on specific mathematical things.[11]

Didactic objects are useful tools for helping develop conceptual analyses; when employed in teaching experiments to produce environments that foster reflective mathematical discourse, they help generate observable information. As such, they facilitate researchers' formulation of hypotheses about students' understanding and development in relation to their engagement with instruction.

The term *didactic model* (Thompson, 2002) refers to a model that an instructor or instructional designer has of "what they intend students will understand and how that understanding might develop" (ibid., p.211). This is their "image of all that needs to be understood for someone to make sense of the didactic object in the way he or she intends" (ibid., p. 221)—that is, an image that guides the designer or user's decisions concerning how the didactic object will be used, such as what conversations to have around them and what issues to raise in those conversations.

In this study the ideas of didactic objects and didactic models are reflected in my descriptions and analyses of instructional activities, which typically entail elaborating their design rationales and issues they aimed to address.

To summarize, the interrelated perspectives of radical constructivism, quantitative reasoning, and didactic objects and didactic models formed a framework that guided the design and

[11] Several examples of didactic objects are discussed at length in Thompson (2002).

implementation of the teaching experiment analyzed in this study. The framework was rooted in the idea of extending Glasersfeld's (1995) style of conceptual analysis to statistical and probabilistic ideas in the context of instructional interactions.

These frames helped texture this study in a number of ways: radical constructivism played the role of background theory, serving to constrain and orient the *types* of descriptions and explanations I developed of students' understandings and ways of knowing. Thompson's (1994) frame of quantitative reasoning was applied to statistical and probabilistic notions, casting them in terms of ideas of measurement and quantity, and providing the mathematical content of my descriptions and explanations of students' understandings. The ideas of didactic objects and didactic models (Thompson, 2002) are reflected in my descriptions and analyses of the instructional activities with which students engaged in the teaching experiment.

CHAPTER IV

ANALYTICAL PROCEDURES

With the interrelated frames of radical constructivism, quantitative reasoning, and didactic objects and didactic models in mind, in this chapter I describe the methods I employed to analyze the data generated in the teaching experiment. I do so on two levels. First, I give an overview of my approach, characterizing it in terms of a basic procedure consistent with the grounded-theory approach to conducting qualitative analysis. I then describe the specific procedures enacted that were tailor-made for guiding my analysis of the teaching experiment's data.

Analytical Approach from a Global Perspective

From a global perspective, the method I employed to generate descriptions, hypotheses, and explanations is consistent with grounded theory's (Glaser & Strauss 1967; Strauss & Corbin, 1990) procedures of continual review, constant comparison, and regeneration, as elaborated by Cobb and Whitenack (1996). This method can be characterized operationally as iterating a basic three-step procedure:

1. Search available data with an eye toward conceptualizing episodes and sequences of engagement that are suggestive of students' conceptions and understandings related to ideas addressed in instruction. Formulate initial impressions (e.g., descriptions and hypotheses) about students' understandings in particular contexts.
2. Test viability of initial impressions by searching data for supporting or contradictory evidence, as in step 1.
3. Adjust initial impressions on the basis of evidence obtained in step 2; refine, elaborate, abandon, or reconstruct initial impressions.

Analyses were generated by reiterating this procedure with the aim of developing increasingly stable and viable hypotheses and models of students' conceptions in relation to their engagement in instruction. A particular heuristic that guided my analyses at all levels was to search for evidence of relative stability/robustness and instability/non-robustness in students' thinking, imagery, and ways of operating. This heuristic oriented me to comprehend the limits of students' thinking and understanding and the situations that taxed those limits. In addition I

relied on a two other guiding principles. One is akin to a law of momentum: a conjecture lives until substantial contradictory evidence emerges, and its viability is strengthened as evidence in support of it accrues from the various data sources. The other principle is akin to a law of parsimony: I resisted attributing more understanding to a student than was necessary to account for a behavior I was trying to explain. Together, these principles helped constrain and guide my interpretation of the data.

The videotaped classroom lessons were examined first to formulate initial impressions concerning students' engagement in instruction and their conceptions related to ideas of sampling and distributions addressed in instruction. In subsequent examinations, I triangulated with the other data sources—students' written work and individual interviews—searching for evidence to test which of my initial impressions were robust enough to become potentially viable conjectures and which were refuted or needed revision and refinement. In iterating this procedure, the information I created from increasingly extensive cross referencing of the data sources served to increasingly constrain and orient my hypotheses and theory-building. As this regenerative process unfolded, my theory-building activity became increasingly directed, evolving from a state of relatively unstable initial hunches to one of relatively stable hypotheses supported by evidence from several sources. This process unfolded in a sequence of identifiable levels of analytic activity described in the next section.

Data Analysis from a Local Perspective: Procedural and Organizational Details

In this section I describe the procedural details of the analytic method mechanistically and reiteratively to impart a sense of how my analyses developed from the most basic to the most sophisticated levels. The data corpus generated in the teaching experiment is listed below in decreasing order of abundance:

- 17 videotaped classroom lessons;
- Numerous instructional artifacts: lesson plans, activity sheets, computer softwares;
- 14 documents of students' written work: 12 in-class activities and/or take home assignments;
- 2 formal in-class evaluations;

- 2 individual student interviews (one conducted at mid-experiment, the other at post-experiment).[1]

The analytical procedure consisted of conducting repeated passes over the collection of lesson videotapes—that is, directed viewing and examinations of the videotaped classroom lessons in chronological order.[2] In these examinations, I did not consider students' behavior and ideas as disembodied from instruction. Rather, I took classroom instruction and instructional interactions (e.g., engagements and conversations) as a focal point of my examinations and anchored my analyses around them. In this way, descriptions and analyses of students' understandings were grounded in their participation in instruction.

Level I: Preliminary examination of videotaped lessons

Pass 1

Analyses began by making a chronological pass over the entire collection of videotaped classroom lessons to identify times during which a topic related to sampling and sampling distributions was discussed (either explicitly or implicitly). For each lesson, I created a 30-second time sample of when such discussions happened and I identified rough starting and ending times of individual instructional activities.[3] In addition, for each lesson I created two files: a "students" file in which I recorded notes (theoretical and other kinds) that occurred to me about students' reasoning and engagement; a "lesson" file for notes about instruction and activities. The notes constituted a form of loose coding during the preliminary examination of the classroom data. Figure 4.1 displays the computer environment in which I worked to create these documents.

---

[1] All of this information was organized according to the chronological order of its production prior to my formal analysis of it.

[2] All video was digitized and transferred to CD ROM. There were two video images for most lessons: one captured all participants and the front board and projection screen, another captured students and a sideboard.

[3] This was a relatively low-level analysis in which I simply made a check mark, on a running time chart, in every 30-second interval that the discussion addressed a seemingly related topic (e.g., sampling, distribution, (un)usualness, (un)likelihood, expectation, rare results, variability, etc.).

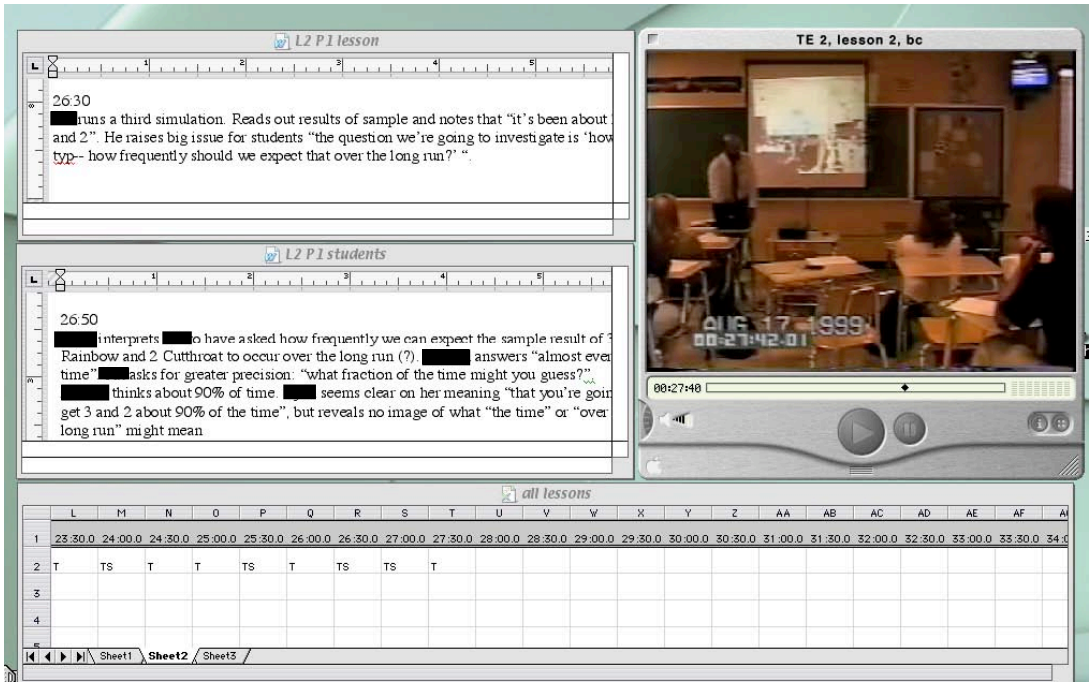


Figure 4.1. The computer environment in which I conducted preliminary analyses of the videotaped lessons.

Pass 2

The time sample files created in the first pass enabled me to make a second, and more directed, pass of the classroom lesson videos. In this pass I focused on filling out my initial rough identification of distinct instructional activities. This was done by building on the Lesson notes created in the first pass; I elaborated on what individual activities seemed to entail and what issues instruction raised, for instance. In addition, I began to triangulate with the other artifacts in the data corpus: I included in the Lesson file a copy of any relevant activity items such as guides and handouts. In the same pass, I also elaborated on the "Students" notes by adding whatever thoughts occurred to me concerning their utterances as they participated in the activities. The product of this second pass was a more elaborated set of notes about lessons and students.

By the end of the second pass I anticipated a need to move toward creating verbatim transcripts of segments of each lesson during which particular activities unfolded. This was necessitated by the nature of the classroom interactions; they were intricate, messy and protracted, and not easily analyzed without having captured their fine details. This decision precipitated a move to a second level of analysis, which I refer to as "transcription as analysis".

Level II: Transcription as analysis

Working in the same technological environment that I employed in the first level of analysis (see Figure 4.1), the second level focused on transcribing lesson segments that centered on the instructional activities identified earlier. This stage of the analysis was much more intense and protracted than the first; I did not merely transcribe to capture participants' utterances, I also used this activity as an opportunity to reflect more deeply on what was said and done among participants in the classroom. This stage consisted of two passes over the videotaped lessons. The first pass produced annotated lesson transcripts and copious theoretical notes about instruction and about students' ideas and engagement. In the second pass, I completed the lesson transcripts by including in them all textual information (e.g., activity guides and handouts, video screen shots, sketches, etc) employed in the classroom lessons.

This second level of activity produced a set of more elaborated notes about the lessons and about students' engagement and ideas. More importantly, it produced 17 lesson transcripts sufficiently detailed to code without need for further sustained video viewing. This precipitated the start of a third level of analysis, in which I worked largely with this new textual data.

Level III: Transcript coding

At this level of the analysis I conducted two passes over the lesson transcripts, each one entailing more directed and elaborate coding of instruction and of students' engagement and ideas. In the first pass, each lesson transcript was coded for: 1) characteristics of individual instructional activities, such as their purpose and underlying rationale, and the issues they sought to raise and engage students with; 2) characteristics of students' interpretations and understandings of these issues that emerged as they engaged in the individual instructional activities. The codes and descriptions developed for the instructional activities drew on the other forms of data such as lesson plans, activity handouts, and field notes generated during the experiment's progress.[4]

In addition, I coordinated the coding of students' ideas that emerged in classroom discussions in particular activities with analyses of their written work on tasks relevant to those activities (i.e., homework activities and in-class assessment questions).

[4] This activity occasionally entailed consultations with the team leader to clarify issues related to the intent of aspects of particular instructional activities.

Analyses of students' written work on such tasks entailed characterizing salient features of their responses and explanations. I also considered students' written responses and explanations in terms of their degree of consistency with "model" responses and explanations—those I imagined expressed interpretations and understandings targeted in instruction. Moreover, whenever a task was also explored in the individual interviews, those segments of the interviews were analyzed for additional insight into students' interpretations and understandings of the issue(s) and idea(s) at hand.

I should mention that analyses of students' relevant written work and interview segments were not conducted separately from the coding of the lessons transcripts. Instead they were conducted in tandem, whenever such data was produced in the chronological unfolding of the lessons. In this way, analyses of students' written work and interview items were sensitive to the unfolding of instruction and students' participation in it.

In the second pass of this stage of the analysis, I studied the codes created for individual activities and made connections across them. By coordinating my analyses of instructional activities with my developing sense of issues that arose for students as they engaged in the activities, I was able to aggregate instructional activities into sequences of them. This occurred with my identification of shifts in instruction (e.g., changes in approach or in the "big ideas" addressed) that I related to significant classroom developments and the research team's interpretation of them.

Level IV: Narrative construction

What emerged from my coding activities in the third stage of analysis was a rough outline for the structure of a descriptive narrative of the teaching experiment. The outline conceptualized four broad phases in which instruction unfolded. Using this outline as an organizational tool, I then constructed a narrative describing how instruction progressed from the start to the end of the teaching experiment in tandem with characterizations of students' understandings and interpretations that emerged as they engaged with instruction. Moreover, the narrative was designed to highlight the interdependence and co-evolution of instruction with students' engagement and salient understandings, thereby reflecting a central feature of the teaching experiment. The narrative construction occurred over a series of successive refinements and drew extensively on the notes and codes developed in the previous stages of analysis. At the

same time constructing the narrative constituted a form of analysis itself, providing a space for elaborating and refining descriptions and relations generated in previous stages that now served to refine the unfolding narrative itself.

This constructive process converged to the narrative that constitutes Chapters V through IX of the dissertation. The narrative intertwines descriptions and analyses of various aspects of the teaching experiment. These are elaborated in the overview of Part II of the dissertation.

PART II

OVERVIEW

In the introduction to this dissertation, I briefly described the teaching experiment as having unfolded in a sequence of interrelated instructional phases. In this part I describe these phases in detail and situate my analyses of students' conceptions and development within them.

I view the unfolding of instruction as giving rise to an *emergent instructional trajectory* because the evolution of instruction was driven by complex interactions between a priori designed instruction, the instructor's reflective interactions with students within particular instructional settings, and adjustments made to instruction on the basis of those reflective interactions.

As I mentioned in Chapter II, the instructional methodology used in the experiment was flexible to the point of allowing classroom interactions to "veer" in unanticipated ways from planned instructional agendas. The general aim in allowing relatively wide latitude in these interactions was to occasion their rich development, in particular to optimize the possibility of capitalizing on serendipitous events in the classroom. Thus, adjustments to instruction were made on a variety of scales relative to that of the overall experiment. This feature of the instructional design brings into question what, in my analysis, constitutes a unit called a *phase* of instruction. I addressed this problem by considering the "big mathematical ideas" that instructional activities and interactions centering on those activities aimed to broach and develop. Framing instructional activities in terms of their attendant design rationales enabled me to then relate activities in terms of these rationales and to thus aggregate activities into *sequences*. This approach also enabled me to consider critical *shifts* in the big ideas addressed within sequences of activities or critical shifts in the way instruction aimed to engage students with the big ideas. These shifts typically constitute the boundaries of what I call phases of instruction.

The first four chapters in this part describe instructional interactions as occurring in four phases and they elaborate shifts that suggest their evolution into an emergent instructional trajectory:

Phase 1: Orientation to statistical prediction and distributional reasoning

Phase 2: Move to conceptualize probabilistic situations and statistical unusualness

Phase 3: Move to conceptualize variability and distribution

Phase 4: Move to quantify variability and extend distribution

Figure IIa displays a time-line of the unfolding of this trajectory over the experiment's duration.

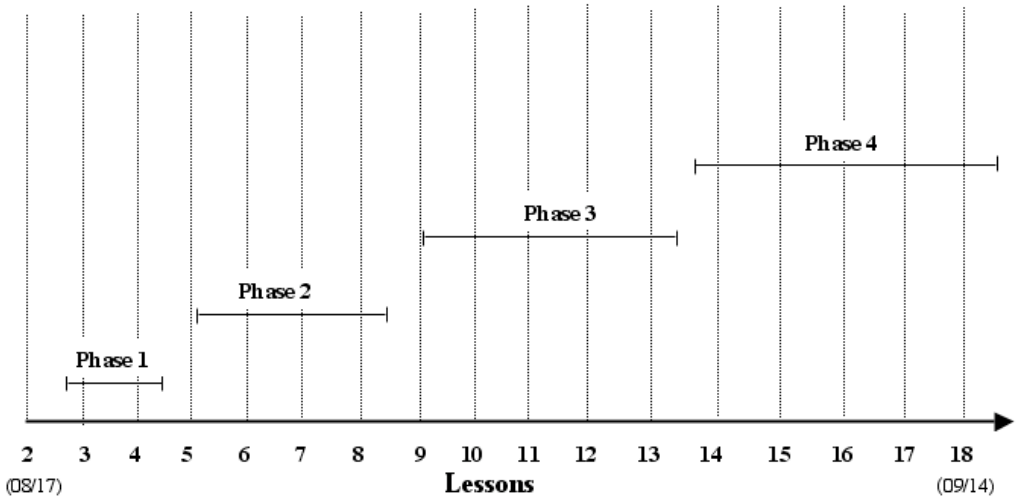


Figure IIa. Time-line of instructional phases across the duration of the teaching experiment.

Each phase is named with a thematic descriptor that suggests the central direction in which instruction aimed to move students. Each of Chapters V through VIII is devoted to events that unfolded in a single phase, and each elaborates four distinct yet interrelated levels of description and analyses; activities and sequences thereof, discussions and student conceptions embedded within them. Chapter IX gives a summary overview of the findings and elaborates conclusions.

Just as the entire experiment is characterized as unfolding in a sequence of phases, so, too, are individual phases seen as unfolding in a sequence of activities. Similarly, activities within phases are seen as unfolding in a sequence of instructional discussions, each of which often unfolded in phases. It is useful to keep this nested and fractal-like structure of the experiment in mind as an orienting perspective, as it can help the reader distinguish and coordinate these levels. The figure below depicts these levels and the bottom-up unfolding of interactions that drove the experiment's emergent structure to increasingly macro levels.

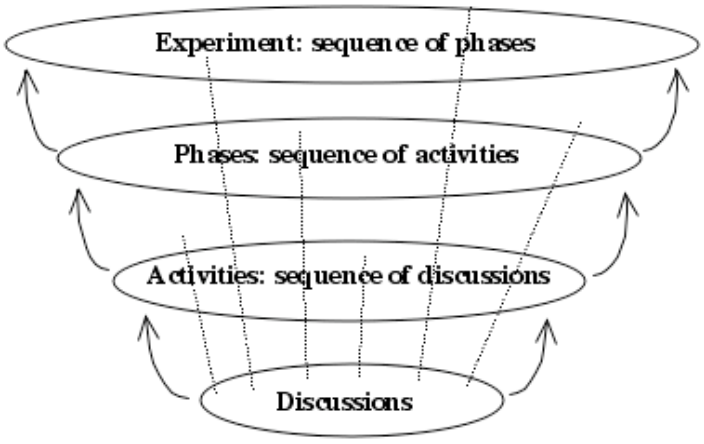


Figure IIb. A hierarchy of levels of description and analysis. The broken lines denote analyses of students' thinking extended to broader levels.

Although evidence of students' conceptions emerged largely in the "bottom" levels—within local classroom interactions centering around particular activities—analyses of these conceptions are not circumscribed within these levels. Instead, analyses are extended throughout higher levels. This feature enables conceptualizing the entire experiment as a structured whole emerging out of interrelations between the various levels of analysis.

The coming chapters culminate with rather compelling evidence, emerging near the conclusion of Phase 3 (Chapter VII), of students' profound difficulties in conceiving a distribution of sample statistics. That students experienced this difficulty is a surprise, especially after Chapters V and VII show their apparent ability to structure collections of sample statistics in ways that would seem to support their coming to think of them as distributions.

CHAPTER V

PHASE 1: ORIENTATION TO STATISTICAL PREDICTION AND DISTRIBUTIONAL REASONING

This chapter describes the first three instructional activities of the teaching experiment as they unfolded in a sequence over Lessons 2 through 4 (see Figure 5.1). Broadly speaking, these activities engaged students in exploring collections of sample statistics with the aim that they move toward making inferences about the composition of an underlying population on the basis of the make up of those collections.

Phase 1:	**Orientation to statistical prediction & distributional reasoning**	
Lesson	**Activity (A)**	**Duration**
2 (08/17)	A1: Sampling tangible objects, Part 1 and Part 2 (Discussion 1)	11 m.
3 (08/18)	A1, Results from Part 2 revisited (Discussion 2)	16 m.
3	A1, Part 3: Transitions to A2 (Discussion 3) a) Part 2 repeated and new results discussed b) Focus on result of 2nd trial of candy-sampling experiment	10 m.
3	A2: Investigating unusualness with *Prob Sim*	10 m.
4 (08/19)	A3: Assessment activity	18 m.

Figure 5.1. Chronological overview of activities in Phase 1.[1]

The chapter begins by elaborating the rationale for the design and implementation of Activity 1, drawing on conjectures that emerged from the research team's analyses of a previous teaching experiment. The chapter then characterizes discussions that unfolded around Activities 1 and 2, typically following their temporal order and highlighting instructional interactions and students' thinking that emerged within them. The final part of the chapter focuses on students' written work in Activity 3, supplemented with highlights from subsequent classroom discussions that serve to elaborate students' thinking.

Analyses highlight critical shifts in the class's foci of attention and discourse that delineate broad phases in instructional interactions, the whole of which gave rise to a hypothetical classroom developmental trajectory.

[1] The symbol "A1" in Figure 5.1 denotes Activity 1. In similar figures appearing at the beginning of each chapter, an analogous symbol denotes each of the other activities.

Prelude to Phase 1

The point of departure for instruction in this experiment was motivated by two conjectures that emerged from the research team's analyses of the first teaching experiment.[2] Instructional activities in that experiment centered on having students explore and interpret the results of computer simulations of drawing many samples from populations having known proportions (Saldanha & Thompson, 2002). Activities were intended to help students develop a sense for how a sample proportion's values get distributed around the underlying population proportion. My analyses of the experiment's data suggested that many students experienced two significant difficulties:

1) students could not reconcile the idea of sampling repeatedly from a population having a *known* proportion with the idea that we typically draw a single sample from a population to learn about its *unknown* proportion. The unresolved tension that students experienced between these two ideas seemed to have a lasting and significant impact on their ability to develop a coherent sense of sampling distributions.

2) students had no experiential basis for understanding what the sampling simulations were simulations of, and the activity of trying to make sense of the simulations lacked a concrete grounding for them. This, in turn, had a significant effect on their ability to engage with instruction.

In planning this—the second—experiment, the research team surmised that these difficulties might be circumvented were students to first engage in selecting samples of physical objects from populations, with the aim of inferring the underlying population proportion's unknown value. These considerations motivated the design of Activity 1 (see Figure 5.2), the experiment's opening instructional activity[3].

---

[2] A brief description of the first teaching experiment is provided in Chapter II.

[3] Activity 1, which was introduced in the first substantive lesson of the experiment, was immediately preceded by 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.

Activity 1: Rationale and Description

Sampling Activity

In this activity we will try to get a sense of populations of objects by investigating how the samples drawn from those populations are distributed. The populations consist of: red and yellow toothpicks, pennies and nickels, red and white candies.

1. Mix up the objects and, without looking, draw a sample of 5 objects from the bag. Record the number of yellow toothpicks, nickels, or red candies in your sample. Consider what *this* sample suggests about the population of objects in the bag.

2. Put the sample back in the bag and mix thoroughly.

3. Repeat this random sampling process 10 times.
Keep track of the number of yellow toothpicks, nickels, or red candies in each sample. Record your results in a table.

4. Interpret your results in terms of what they suggest about the contents of the bag.

Figure 5.2. Written guide for Activity 1.

The general aim of Activity 1 was to engage students in a concrete sampling experiment as a basis for having them draw conclusions about a population. The activity's first part (Question 1) asked students to make a prediction about the proportion of objects in a collection on the basis of the composition of a *single* sample randomly drawn from that collection. The style of reasoning that instruction intended to have emerge among students was consistent with the logic of statistical inference: "given that a sample is randomly selected from a population, we assume that it is a representative sample. Thus, under the assumption that the underlying population is *like* the sample, we can then hypothesize the population parameter's value". This line of reasoning had been explicated by the instructor in a pre-activity discussion (see footnote 1); he highlighted it again during the first part of this activity, which was conducted in a whole-class discussion format.

The activity's second part (Questions 3 and 4) had students repeat the sampling experiment 10 times with the aim that they keep track of and record the 10 sample outcomes of interest. Students were then asked to interpret their collection of outcomes in terms of what it suggests about the proportion of items in the sampled population. Thus, whereas students had previously made an inference on the basis of a single sampling outcome, this part of the activity oriented them toward making an inference on the basis of a *collection* of sampling outcomes. A central aim of this part was to provide students with an experiential basis for building imagery of the

repeatability of the sampling process, the variability among sampling outcomes, and the aggregation of individual sample outcomes into a collection of outcomes.

Students engaged in the activity in small groups, each of which randomly selected samples of 5 objects, without replacement, from one of the collections of objects.[4] Students knew that each collection was composed of two classes of objects—red and white candies, red and yellow toothpicks, and nickels and pennies, respectively. The proportions in each collection, however, were unknown to them.

Activity 1 unfolded in three parts over Lessons 2 and 3, each part entailing substantive discussions centering on a main idea of the part. The activity progressed through Question 1 during the last 11 minutes of Lesson 2, at which time a first discussion focusing on that question occurred. The activity also advanced into the second part (Questions 3) during Lesson 2, but no substantive discussions occurred then due to a lack of time.

The activity was revisited at the start of Lesson 3 (the next day), whereupon a second discussion focusing on students' sampling results in Questions 3 and 4, obtained during Lesson 2, occurred during the first 16 minutes. Immediately following this, the second part of the activity was repeated and students selected a second collection of samples from their respective populations. A third discussion focusing largely on one group of students' sampling results then unfolded over approximately 10 minutes. This third discussion served as a transition into Activity 2. The next three sections highlight instructional interactions and student conceptions that emerged within each of these discussions, respectively.

Discussion 1

The transition between the two parts of the sampling activity—Question 1 and Questions 3 and 4—developed within the context of a discussion in which the instructor sought to implicitly make the variability among sample outcomes a rationale for selecting more samples. He did this by raising the possibility that the any individual sampling outcome might be unusual. The instructional aim was to have students view the selection of multiple samples and the consideration of their outcomes as a natural strategy for investigating the possibility of unusualness and for obtaining more information about the underlying population proportion. The

[4] Each collection contained approximately 500 items in an opaque sack.

following discussion excerpt, lasting 2 minutes during the first part of the activity (Lesson 2), illustrates interactions that drove this transition.[5]

Episode 1, Lesson 2:

1. I: Now, based on what you just took if you had to make a guess about what's in the bag, besides toothpicks or coins or whatever, the relative amounts of things in the bag. Based on the amounts you took, what would you say is in the bag?
2. Peter: A lot more yellow ones.
3. I: A lot more yellow ones. Ok. If in fact that were representative of what's in the bag, what would you say is in the bag?
4. Peter: Toothpicks?
5. I: Ok, what did you get? What did you get?
6. Peter: I got four yellow ones…four yellow toothpicks and one red toothpick.
7. I: So if the population looked just like that sample, then four fifths of it is yellow. Right?
8. Cathy: Right.
9. I: So that's the idea. That if the population looked just like what you picked in terms of the relative amounts, then four fifths of the toothpicks in there would be yellow. All right?
10. Peter: I gotcha.
11. I: So now, if you were to judge what's in that bag according to what you picked, what would you say (points to Sarah)?
12. Sarah: Four-fifths pennies.
13. I: So you had four pennies and one nickel. So you would say four fifths of everything … if the contents of the bag looked just like what you picked. Four fifths of the coins in there are pennies. All right? And you got four whites and one pink, candy (points at Cathy). Boy, four and one is, is it. How likely do you suppose that is, that you got four and [one]?
14. Peter: I guess pretty good.
15. Nicole: It depends on what's in the bag.
16. I: That's right. It depends on what's in the bag. So, how would you uhh, how would you start getting more information about whether or not that in fact what you picked is an anomaly?
17. Peter & Kit: Take more samples.
18. I: Take more samples. All right. So that's the idea.
19. Nicole (inaudible)
20. Peter: Is that what we should do?

The excerpt illustrates the instructor's attempt to push students toward explicitly elaborating the inferential line of reasoning (lines 1-13), and to make that reasoning an object of classroom focus and discussion (Thompson, 2002; Cobb, 1998). The issue of unusualness arose out of a

[5] "I" denotes the instructor's utterances. All other utterances are students'.

serendipitous event: each group's sample, drawn from its respective population, turned out to have the same proportion of objects. The instructor used this event as a natural occasion to raise the question of the sampling outcome's likelihood (line 13). Nicole's response, in line 15, suggests that she had a strong sense that a sampling outcome depends on the sampled population's composition. The instructor, in turn, raised the issue of the outcome's unusualness and in doing so seemed to set up conditions that impelled some students to propose selecting multiple samples as a natural strategy for investigating this issue (line 17).

Thus, the transition from the first to the second part of the activity developed out of a pattern of interactions (Bowers & Nickerson, 2001) that seemed to impel members of the class to consider re-sampling as a natural solution to a problem. Moreover, these interactions were strongly orchestrated by the instructor who coordinated a priori and emergent instructional agendas. This approach of steering classroom interactions so as to facilitate and enable the emergence of key ideas among participants of the group was a hallmark of the instructional method employed in the experiment (Davis & Simmt, 2003). Though the particulars of such classroom instructional interactions were generally not pre-determined, having such interactions emerge was very much a *designed* feature of this experiment.

Episode 1 of Lesson 2 also provides glimpses of students' thinking with regard to inferential reasoning in the early part of the experiment. In particular, I draw the reader's attention to Peter's two highlighted utterances (lines 2 and 14).[6] I interpret the first utterance as suggesting his orientation to making gross quantitative inferences (Steffe, 1991), and I take the second utterance as suggesting a pre-quantitative conception of likelihood. That is, I hypothesize that at this early stage of engagement in the experiment, for Peter the relationship between a sample proportion and the sampled population's composition was not yet operationalized into a precise quantity. Instead, the relationship seemed focused on a rather gross comparison of the relative amounts of two colors of toothpicks in his sample. Similarly, I hypothesize that Peter's idea of likelihood was more like a good hunch ("I guess pretty good") that, while perhaps based on an unarticulated sense of frequency, did not entail a full-fledged quantification of expectation. Responses like Peter's, above, were not uncommon among students in the early phases of the experiment,

[6] Peter was somewhat of a leading figure in the class. In addition to having a prominent social profile within the class, Peter was vocal, he was engaged, and he readily shared his ways of reasoning with others. By virtue of his pro-active engagement, Peter was an important force in helping to crystallize the participatory norms and in driving instructional interactions like those exemplified in this excerpt.

especially in the absence of instructional scaffolding like that exemplified in lines 6-12 of this episode.

Concerning Nicole's sense (line 15) that the likelihood of an outcome of 4/5^{ths} depends on the sampled population proportion, her response does not suggest that ideas of variability among outcomes were at the foreground of her imagery. That is, there is no reason to believe that at that time Nicole was mindful of a *loose* dependence between sample and population composition; hers did not appear to be a schematized image entailing an anticipation of patterns in outcomes that might emerge over the long-run were one to re-sample frequently, and entailing a sense that the variability among outcomes is bounded and thus leads one to expect a *fuzzy* resemblance between sample and population.

Discussion 2

The transition from the first to the second part of the activity forced a shift in the classroom discourse and attention away from *individual* sample outcomes and toward *collections* of outcomes. The ensuing discussions in Lesson 3 then centered on how to look at collections of sampling outcomes—that is, how to structure such collections—in order to claim something about the composition of the underlying population. The instructor steered these discussions so as to occasion student reflections on interrelations among the ideas of re-sampling, variability among sampling outcomes, and aggregation of individual outcomes into a collection of outcomes. The instructional aim was to support students' construing these collections as distributions of sample proportions (amounts). This section draws on excerpts from the beginning of Lesson 3 to highlight ideas that emerged within these discussions.

The instructor began Lesson 3 by organizing students' collections of 10 sampling outcomes, obtained in the previous lesson, in frequency tables on the board (see Figure 5.3). The organizational structure of the tables was suggested by the instructor, who filled in their entries as each group of students called out its results. The tables' structure was intended to highlight two quantities: 1) all possible values of the sample statistic of interest (e.g., the number of red candies in a sample, listed in a table's top row) and, 2) the number of samples in a collection having each of those values (listed in a table's bottom row). This representational format was intended to facilitate partitioning the data collection in ways that support seeing it as a

distribution of outcomes and to thus suggest something about the underlying population proportion.

Discussions of the sampling outcomes typically centered on these tables and are suggestive of students' ways of interpreting them and what their contents might indicate about the sampled population. The following excerpt, comprised of two contiguous segments, lasted approximately 4 minutes during the beginning of Lesson 3.

Episode 2, Lesson 3:

Segment 1

1. I: Uhh, then I had you repeat taking the samples about ten times. Roughly ten times for each group. What did you notice about the samples? Were they all alike?
2. Peter: Nope!
3. I (points to Peter & Lesley): Ok, were all of yours $4/5^{ths}$ yellow?
4. Peter: No.
5. I: All right. So if you got one that was, say $3/5^{ths}$ yellow, then if you generalize from that one you would say "$3/5^{ths}$ of the toothpicks in this bag are yellow". Ok, clearly they're not both correct. Ok, so you take more samples. You can't, ok, so any time you took a sample could you predict with certainty what you were going to get?
6. Nicole: No.
7. Peter: No.
8. I: No. <u>But</u>, did the samples that you took start to follow a pattern?

(3-second silence)

9. Nicole: No.
10. I: Not at all? Ok, let's put your re—uhh results back up. Did any of you, ok do any of you have a different answer to that question—"did your samples start to follow a pattern?"
11. Cathy: Sort of.
12. I: Sort of, Cathy (motions to Cathy as though expecting elaboration)
13. Cathy: I mean, they were mainly like four to one but some of them are different
14. I: but some of them were off?
15. Cathy: Yeah.

In this segment, the instructor drew on students' experience in the second part of the sampling activity to raise the issue of uncertainty. The explicit issue was that sampling outcomes vary and therefore make predictions uncertain (lines 1-5). There seemed to be general consensus among students that this is true. The instructor's implicit issue was that variability among outcomes makes it problematic to infer a population proportion on the basis of *individual*

sampling outcomes. His idea was to use this problematicity to motivate students to consider *multiple* sampling outcomes in terms of suggestive patterns (lines 5 and 8).

At this point there is evidence of different orientations among students: Nicole (line 9) did not recollect there being any pattern in her outcomes. That is, her sense of pattern—whatever it may have been—did not admit seeing her sampling results as having a pattern. Cathy (lines 11-15), on the other hand, had a sense that most of her outcomes being 4/5^{ths} ("mainly four to one, but some are different") suggests a pattern. It would seem that Cathy was considering, albeit with trepidation, her collection of sampling outcomes in a way that entailed partitioning it into two classes of outcomes: those having a value of 4/5^{ths}, which constituted the majority, and those having other values. I put that her sense was, therefore, of a relatively *gross* pattern that entailed a blurring of the distinctions among individual outcomes and a re-consideration of these distinctions relative to the entire collection. There is evidence, in the next segment, that Nicole's perspective was qualitatively different from this.

The next segment is a continuation of the first and provides more information about students' ideas of pattern and ways of construing a collection. The discussion centers on the results of the toothpick-sampling experiment, which were presented on the board in the following frequency table:

# of yellow toothpicks	0	1	2	3	4	5
# of samples	0	0	1	1	5	3

Figure 5.3. Results of the toothpick-sampling experiment: selecting 10 samples of 5 toothpicks each from a large collection of red and yellow toothpicks.

Episode 2, Lesson 3:

Segment 2

16. I: Now, does this (Figure 5.3) look like there was any pattern?
17. Nicole: No.
18. I: No? We might have to reinterpret that word "pattern", but in some sort of general way. Nicole?
19. Nicole: What are you asking?
20. Peter: A pattern of, like, four and one going five times
21. I (to Nicole): A pattern in the samples (points to table in Figure 5.3)
22. Peter (continues): I mean, that's a pattern.
23. I: Uhh, right.
24. Nicole: I don't see any pattern there.

25. I: Ok, now patterns, uh the thing about patterns is they allow you to make predictions. That's why, that's what makes a pattern a pattern. You can say "oh, it's gonna repeat".
26. Nicole: Yeah, but nothing is repeating!
27. I: Suppose that we did this ten more times. Could you make any predictions about what might happen?
28. Peter: Yeah
29. Nicole: If you got the same numbers!
30. I: Uhh
31. Peter: Yeah, I could say you're gonna get maybe the majority of them are four to one, yellow toothpicks!
32. I: Alright, so, so it's not uhh, you're not gonna say that you're gonna get 1, 1, 5, 3, are you Peter?
33. Peter: No.
34. I: Ok, but you are saying something else, that's almost like that. You're saying that you're gonna get mostly fours and fives. All right? So this is called a statistical pattern. See, it's how things work out in the long run. That's what makes statistics different from most mathematics: it's not exact in that sense that "oh yeah, on the fourth sample, if we do this again, on the fourth sample here's what we get". We can't do that, we can't make that kind of prediction. But we can make a prediction about what's going to happen over the long run. We may be off, but at least we can say "yes, that suggests a pattern", a pattern that most samples will have more than three of yellow toothpicks. Ok, does that make more sense now, Cathy, I mean uhh Nicole?
35. Nicole: Yes.

The colored text highlights Nicole and Peter's utterances. The excerpt illustrates that Nicole, as before, still did not see a "pattern" in the data. Her conviction is steadfast, despite suggestions from the instructor to extend her idea of pattern (line 18) and suggestions from Peter on how to interpret a pattern in the data (lines 20 and 22). The interchange between the instructor and Nicole (lines 25-29) offers hints of her notion of pattern: it appears that Nicole had in mind that a pattern ought to be a sequence of elements/outcomes sufficiently definitive to allow one to make an almost certain prediction on its basis. When implored to anticipate what might happen if the sampling experiment were repeated (line 27), Nicole responded as though she believed that the second set of outcomes would have to be identical to the first ones in order to say that a pattern had emerged (lines 26 and 29). In other words, Nicole seemed to have understood the instructor, in lines 25 and 26, to be speaking of a *deterministic* pattern—that is, of repetition as *identicalness*.

The instructor's intended meaning was, however, quite different. In line 25 the instructor characterized a pattern as that which allows one to make predictions about what might be repeated. His elaboration in line 34, however, clarifies that he had in mind a repetition of outcomes in a statistical rather than a deterministic sense. That is, the instructor's was a notion of relatively *gross* patterns that emerge in sampling outcomes over the long run. This conception entails a sense of repetition tempered by a consideration of the expected variability among outcomes and is thus qualitatively different from a sense of repetition as identicalness. Rather, this statistical sense of repetition is more like an expectation that a certain class of outcomes might occur with similar frequency in the future under similar conditions (e.g., we expect that "*most* samples will have *more than* three yellow toothpicks").

It is worth noting that the instructor's elaboration (line 34) was highly consistent with Peter's notion of pattern in this segment (lines 20, 22, 28, and 31), and also with Cathy's sense of pattern elaborated in segment 1. Indeed, the instructor offered the characterization in line 34 as an explication of Peter's sense of pattern (lines 31-33). His aim was to publicly unpack Peter's line of reasoning and to share it with Nicole to the extent that she might begin to "see" his sense of pattern—one which entailed a *blurring* of the individual outcomes and a refocus instead on the relative frequency with which a *class* of outcomes occurred.

Nicole eventually succeeded in construing a collection of sampling outcomes in terms of a statistical pattern as characterized above.[7] I imagine that her engagement in the discussion highlighted in segment 2 of Episode 2 was instrumental in helping her do so. The question remains as to how Nicole was able to make a shift from her previously and steadfastly held sense of pattern to a statistical sense of pattern. I would speculate that she was able to assimilate Peter's sense of pattern and the instructor's detailed elaboration of it by accommodating the conceptual operations of blurring distinctions among individual outcomes and reconstituting the collection as classes of outcomes.

The way of operating on the data collections exemplified by Peter and Cathy in these discussion excerpts verges on a proportional line of reasoning. The discussion immediately following segment 2 of Episode 2 turned to examining relationships between the data collections and the underlying population proportion. The next episode is drawn from that discussion; it

[7] In the next section I present evidence of this in a subsequent discussion excerpt.

illustrates issues raised in students' investigation of the results of the coin-sampling experiment (see Figure 5.4). The excerpt, comprised of 3 contiguous segments, lasted approximately 3 minutes.

# of Pennies	**0**	**1**	**2**	**3**	**4**	**5**
# of samples	**0**	**0**	**2**	**3**	**7**	**3**

Figure 5.4. Results of the coin sampling experiment: selecting 15 samples of 5 coins each from a large collection of Pennies and Nickels.

Episode 3, Lesson 3:[8]

Segment 1

36. I: All right. So what does that suggest to you, the two of you—Sarah and uhh Nicole?
37. Nicole: Nicole!
38. I: you did it together?
39. Nicole (affirms): Hmm hmm
40. I: Ok. What does this suggest to you about what's in that bag?
41. Nicole: More Pennies than Nickels.
42. Sarah: More Pennies than Nickels.
43. I: And uhh, does this suggest anything to you—?

[…]

Segment 2

44. I: Now, what, what was that, what is your prediction about what's in the bag? Ok. Uhh, any, do you feel safe making any guess about the fraction of Pennies?

(2-second pause)

45. I: What percent of those coins might, if you had to make a guess, what percentage of coins might--?
46. Sarah: Like 4/5^{ths}
47. I: Like 4/5^{ths}. So, 4/5^{ths} would certainly fit this (points at data table, see Figure 5.4). Right?
48. Nicole (affirms): Hmm hmm

[8] The symbol "[…]" signifies text that has been omitted from the transcript. The central reason for text omission is to improve the readability of already very messy classroom discussions. Omitted text typically consists of a small number of utterances within a segment of a discussion excerpt. Occasionally, however, entire intervening discussions between analytically interesting segments are omitted because they are not of interest to the issue and analysis at hand. In any case, text omission occurs only if it does not compromise the nature of the data and analyses.

Segment 3
49. I: Would uhh, would 5/5^{ths} fit?
(3 second silence)
50. I: 5/5^{ths} of the coins being Pennies.
51. Nicole: No.
52. I: No, it wouldn't fit at all because we're in fact getting Nickels. Right?
53. Nicole: Yeah.
54. I: Uhh, would 2/5^{ths} fit this?
(3 second silence)
55. I: (inaudible) do you understand what I'm asking, Kit?
56. Kit: No.
57. I: Would the assumption that 2/5^{ths} of the Pennies in that bag, the coins in that bag are Pennies, would that uhh, they're saying that wouldn't fit this. In what way would that not fit this data? (points to data table in Figure 5.4)
58. Sarah: There's more than half (inaudible)
59. Cathy: Yeah.
60. I: Ok. Uhh, now here's a way to answer questions like that: we would—uhh, if in fact there are, 2/5^{ths} of the coins are Pennies, then results like that would be pretty unusual (points to table). We wouldn't expect that to happen very often when we took 15 samples. (4 second pause) Does that, does that make sense?

Nicole and Sarah's responses (lines 41 and 42) to the instructor's call for an inference to the population suggests their being oriented, as was Peter in an earlier episode, to making gross quantitative inferences. That is, at this point Nicole and Sarah were willing to claim only that the sampled population might contain more Pennies than Nickels. They did not provide justification for their claim. In the second segment we see that the move to making a specific quantitative inference required some prompting from the instructor. When asked to give a plausible specific percentage, Sarah inferred that the population might contain 4/5^{ths} Pennies. Here again Sarah did not elaborate her line of reasoning, but she might have based her estimate on the most frequent outcome in the data set: 7 of the 15 samples selected contained 4 Pennies (and 1 Nickel)—a result consistent with a population proportion of 4/5^{ths}.

It is productive to speculate as to why students seemed disinclined to make specific claims about the sampled population in this phase of the activity. On further reflection their disinclination seems quite sensible at that early stage of engagement: I suspect that students were then grappling with how to integrate statistical considerations—a sense of the variability and uncertainty among outcomes—with making a prediction in the way they might have been accustomed. The single-outcome-based inference style of reasoning elaborated in the first part of

the sampling activity was relatively unproblematic for students once it became accepted as normative. But when the activity entails making collection-based inferences, that line of reasoning can become problematic for several reasons. First, the "thing" upon which one is expected to base an inference is a very different kind of animal; it is not a single sampling outcome, but rather a collection of such outcomes. It conceivably entails learning how to construe a collection so as to infer a population proportion. A Second, and related, reason is that even if one learns how to structure such a collection, say, into classes having such and such outcomes (proportions), there remains the problem of deciding which class an inference should be based upon.

Thus, in my view, students' apparent disinclination to making quantitative inferences at this stage is consistent with their struggling with how to structure a collection of sampling outcomes so as to make an inference to the sampled population.

As an afterthought to this point, it is worth considering whether the development of collection-based inference might be facilitated by shifting one's perspective to reason in reverse. That is, in order to answer the question "what does this set of outcomes suggest about the sampled population?", it might be productive to first address the question "what population proportion might be reasonably consistent with the particular set of outcomes obtained?". Indeed, this was the instructor's rationale for conducting the questioning in the third segment of Episode 3 of Lesson 3. Sensing that students had little intuition about what sampling outcomes one might reasonably expect from a population having a given proportion, he set out to structure the discussion so as to occasion the development of some intuition. In lines 57-59 (third segment) it is Sarah who had a sense, apparently echoed by Cathy, that the sampling outcomes in Figure 5.4 are inconsistent with a population proportion of 2/5^{ths} Pennies because, as I interpret her utterance in line 58, more than half of the samples contained over 2/5^{ths} Pennies.[9]

The discussion in the third segment of Episode 3 of Lesson 3 concluded with the instructor again raising the issue of unusualness, but this time he referred to the unusualness of a collection of sampling outcomes relative to an assumption about the sampled population (line 60). He proposed this issue as an equivalent way to address the question raised in line 57: how to

[9] In my interpretation, the most coherent referent for Sarah's utterance ("There's more than half (inaudible)") was the collection of sample outcomes shown in Figure 5.4. It makes little sense to interpret this utterance as referring to the population proportion of 2/5^{ths}.

determine whether a particular collection of outcomes is consistent with a presumed population proportion? These questions moved the discussion toward statistical hypothesis testing. Indeed, the sequence of discussions described above as having unfolded from Activity 1 eventually led to other discussions that centered on testing a hypothesis. Those discussions are significantly different, in their focus and objects of discourse, from those in Activity 1 that I take them to constitute a second distinct activity: Activity 2 of the instructional sequence in Phase 1.

Discussion 3: Transitions to Activity 2

Before turning to Activity 2, I first describe some transitional discussions and issues raised within them that motivated its emergence. Immediately following the discussion in the third segment of Episode 3 of Lesson 3, the instructor had students repeat the second part of the sampling activity. After each group of students had selected another 10 samples of 5 objects from their respective populations and recorded their results, these results were presented on the blackboard in frequency tables that also displayed the group's first results. This section focuses on discussions centered on the outcomes of the candy-sampling experiments.

The instructor orchestrated these discussions in an effort to move students toward comparing the two collections of outcomes, raising these issues: their consistency or inconsistency, the plausible reasons for their differences, and the inferences one might make on the basis of each collection. The following illustrative excerpt, lasting approximately 2 minutes, is drawn from the beginning of these discussions.

# of red candies		**0**	**1**	**2**	**3**	**4**	**5**
# of samples	***(Result 1)***	**0**	**5**	**3**	**1**	**0**	**1**
	(Result 2)	**0**	**1**	**7**	**2**	**0**	**0**

Figure 5.5. Results of two iterations of the candy-sampling experiment: selecting 10 samples of 5 candies each from a large collection of red and white candies.

Episode 4, Lesson 3:

61. I: […] Let's talk about potential explanations for why the difference (motions with hand back and forth between two results in data table, see Figure 5.5 above). Now this would, ok, this would suggest that there are fewer reds than whites (points to Result 1 in table). Correct? Uhh, that's what we said. Would this suggest that there are fewer reds than white (points to Result 2 in table)?

62. Cathy: Hmm hmm (shakes head up and down as though affirming)
63. Kit: Hmm, yeah, sort of.
64. I: Let's see, this is, that's two, two reds (points to "7" in second row of data), correct?
65. Cathy: Hmm hmm (affirms)
66. I: All right. Does it suggest it as strongly as that (points to "3" in first row of data)?
67. (Kit & Cathy shakes heads from side to side, as though to mean "no")
68. Female student: More
69. I: Pardon me?
70. Female student: More so.
71. I: Uhh, more so?
72. Female student: Yeah
73. Cathy: No.
74. I: Ok, well here's the question: "how can we investigate whether or not this (points to Result 2 in table Figure 5.5) is inconsistent, uhh this in fact suggests that there are fewer reds than whites, or this (points to first row of data in table) suggests that there are fewer reds than whites?". That, that's the question—my question is about a question! How do we investigate that question? Ok? That's what, that's the issue that I'm raising: how do we investigate whether or not this suggests (points to last row of data in table) that there are fewer reds than whites?

The excerpt begins with the instructor first reiterating a previously accepted inference, drawn on the basis of Result 1, that the sampled population contains fewer red than white candies. He then asked students to reflect on whether they would conclude similarly on the basis of Result 2. Two students (line 62-63) had a sense that Result 2 also suggests fewer red than white candies in the population, but they did not justify their conclusions. Kit seemed unsure of this conclusion, appearing to waiver in her conviction. The uncertainty of students' responses incited the instructor to pose the question in line 66: his aim was to prompt students to reflect on, and articulate, their reasons for claiming that one or the other result suggests something about the sampled population.

The responses of Kit and Cathy in line 67 suggest that they viewed Result 1 as a stronger indicator that the population contained fewer red than white candies. Here again, however, they offered no justification for their conclusion. Though there is scant evidence of student's thinking in this excerpt, let me nevertheless elaborate a way of thinking about each sampling outcome that is consistent with Kit and Cathy's response: perhaps they were considering the number of samples in each collection that contained *at most* 2 red candies—the boundary value between a

majority and minority of red candies. In the first outcome, 5 of the 8 samples that contained at most 2 red candies contained only 1 red candy. In the second outcome, 8 of the samples also contained at most 2 red candies, but only 1 of those 8 contained 1 red candy. Thus, one might view Result 1 as more strongly suggesting fewer red candies in the population because it is more heavily weighted than Result 2 toward samples having only 1 red. This perspective might explain the responses of Kit and Cathy.

On the other hand, if one focuses on the number of samples containing *exactly* 2 red candies in each collection, then Result 2 is more heavily weighted toward such samples and it might be taken as stronger evidence that the population contains fewer red than white candies. This line of reasoning might have been at the root of the response of the unidentified female student in lines 68-72.

I reiterate my earlier caveat that there is little evidence of any such reasoning having occurred. However, my point in this analysis is not to make hard claims about the psychological reality of these conjectured ways of reasoning. Rather, my point is to raise the issue that there are diverse ways of structuring these collections of outcomes that might express themselves in students drawing different conclusions about the sampled population. Two additional and equally important points are: 1) at this stage of the experiment there was no consensus among students on how to structure these collections, and 2) there is a lot of evidence to suggest that students were generally not inclined to articulate their ways of reasoning. Concerning the second point, I attribute this to more than a mere insensitivity to the participatory and socio-mathematical norms (Cobb & Yackel, 1996; Cobb, 1998) that the instructor was trying to institute within this classroom. Rather, I take it also as an expression of students having been, as yet, insufficiently mindful of their own reasoning so as to operationalize it. Put slightly differently, students had not yet reflectively abstracted their ways of structuring these collections so as to be able to mentally "step back", as it were, to reflect on and describe these ways coherently (Glasersfeld, 1995).

In the last paragraph of Episode 4 of Lesson 3, the instructor raised what was to become a central issue in the next activity: how to investigate whether one or the other sampling result suggests that the sampled population contains fewer red than white candies. This question was followed by his briefly explaining the idea of sampling bias to the class and proposing it as a potential reason for differences in the sampling results. The instructor then divulged the sampled population proportion to the class: the bag contained 50% red candies and 50% white candies.

The following excerpt highlights developments that unfolded from this divulgence. The excerpt consists of 3 contiguous segments lasting a total of approximately 4.5 minutes.

Episode 5, Lesson 3:

Segment 1
[…]
75. I: Half and half. Ok? They're in fact half and half.
(4-second pause)
76. I: Now, so this result (points to Result 1 in Figure 5.5), with that knowledge, that they're half and half, is this surprising?
77. Cathy: Hmm hmm, yeah (shakes head up and down in acknowledgment)
78. (Kit, Sarah, & Peter, all shake heads up and down in acknowledgment)
[…]

As the first segment of Episode 5 of Lesson 3 indicates, a good number of students were surprised to learn that Result 1 was obtained from an evenly split population of candies. The source of their surprise was, presumably, their previous inference that the population contained fewer red than white candies (see Episode 4).

The instructor continued the discussion, in the second segment, by first asking for potential explanations of the surprising result other than sampling bias.

Episode 5, Lesson 3:

Segment 2
79. But, so let's come back here (points back to Result 1 in data table). Ok, that seems surprising now that we know that the split is half and half. (3 second pause) […] Let me ask you to speculate how this might have happened. Of course one is selection bias. There may have been something about the way that you selected the candies. Any thing else?
80. Peter: How the candies were placed in the bag. Like whether they were shook up enough, or something.
81. I: Ok. And that would introduce, so that would again introduce a selection bias. Right? So that that would be something about the process of the selection, not shaking them up enough. Ok. So, another example of a selection bias that might've happened?
(16-second silence)
82. I: What do <u>you</u> guys think? (motions to Group 1 members) Do you know, can you think of anything that you did that might've led to a bias toward picking more white?
83. Peter: (inaudible)
84. Cathy: Not really. We put them in there, we didn't separate 'em, and we shook it up every time.
85. I: You didn't separate them and you shook them thoroughly every time. So you can't think of anything.

86. Peter (under his breath): Whatever!
87. (Nicole chuckles at Peter's comment).
88. Cathy: No.
89. I: So it might just be luck of the draw.
90. Peter: Could be.

I note that students' explanations in the second segment of Episode 5 of Lesson 3 were restricted to sampling bias. The fact that no student ever raised the possibility that Result 1 was just "luck of the draw"—a perhaps rare, yet possible, outcome suggests that this was not a salient possibility for the students. That is, I take the second segment as an indication that at this point in their engagement, students had not yet developed a sense of distribution that entailed anticipating a range of possible outcomes and frequencies for a *collection* of sampling results. Upon reflection, this is not surprising, given the relative complexity of the object they now confronted—a collection of sample outcomes, rather than an individual sample outcome—together with the fact that instruction had only just begun to engage students in developing a sense of the possibilities for such outcomes.

The third segment of Episode 5 of Lesson 3 represents a turning point in the class toward a more systematic investigation of these possibilities.

Episode 5, Lesson 3:

Segment 3

91. I: All right. So the question is how could, uhh, how could we test whether or not something is the luck of the draw? (points to first row in data table)
92. Cathy: Test it a lot.
93. I: Sorry?
94. Cathy: Test it over and over again?
95. I: Test it over and over again, and see if, in fact, you get something like this (points to Result 1 in data table)—I mean, we wouldn't expect it a lot, but suppose that we got something like this three out of ten times. Would that then make this less surprising? (points to Result 1 in data table in Figure 5.5)

Cathy's suggestion to "test it over and over" (lines 92-95 of Segment 3) provided a direct segue into the second activity of Phase 1. Her idea prompted the instructor to suggest selecting multiple collections of 10 sample outcomes with the aim that students take the resulting collection of collections as a basis for determining whether outcomes like Result 1 (Figure 5.5)

are unusual, given that they were obtained from an evenly-split population.[10] This development ushered in the start of Activity 2.

Activity 2: Rationale and Description

The instructor's suggestion (line 97 of Segment 3) amounts to an operationalization of his interpretation of Cathy's idea, which he employed as a didactic strategy to move the discussion along an overarching instructional agenda. That agenda was to have students internalize and operationalize a method that might enable them to conceive *expectation* as a statistically quantifiable attribute. It is worth noting that though this was an a priori agenda, the instructor's decision to move the discussion in this particular direction at this juncture was unforeseen and momentary. The decision was based on his taking students' widespread surprise that Result 1 was obtained from an evenly split population as an opportune occasion to enact his agenda. Thus, the emergence of Activity 2 as an identifiably distinct entity from antecedent discussions was occasioned by complex interactions between a priori designed instruction, serendipitous classroom developments, and improvised adjustments to instruction.[11]

The underlying logic of the instructor's suggestion is as follows: if outcomes *like* Result 1 occur relatively infrequently under essentially identical sampling conditions, then this suggests that an outcome like Result 1 is statistically unusual under a given assumption about the sampled population proportion. In addition to articulating the logic of this method a priori, the instructor also reiterated it as the activity unfolded. The activity consisted of applying this method repeatedly: select 10 samples and obtain a collection of 10 outcomes; compare each collection of 10-sample outcomes with Result 1 to determine whether the two are similar or dissimilar; record the similarity decision. After a number of similarity decisions had accumulated, the class could

---

[10] I note, however, that the "it" that Cathy had in mind was never explicated. It remains questionable whether Cathy had a well- articulated image of what she meant to test repeatedly. Given her history of apparent engagement in the activity thus far, it is plausible that Cathy had, by this point, bought into the idea of repeating a random sampling process many times to collect many outcomes. However, it is unclear how Cathy was, at this juncture, structuring the selection process. Was she mindful that the unit of selection and aggregation was 10 samples, or did she have in mind selecting individual samples to aggregate a collection of 10 outcomes? In other words, it is unclear what experiment Cathy envisioned repeating. This contention is elaborated further in footnote 11.

[11] As mentioned in an earlier chapter, such interactions were enabled by the fact that the instructor and the instructional designer were one and the same person.

then draw a conclusion about Result 1's unusualness on the basis of the collection of similarity decisions.[12]

Instead of having students repeatedly draw actual samples from the population of red and white candies, at this point the instructor proposed using the sampling simulator *Prob Sim* (Konold & Miller, 1994) as an efficient way to simulate this repeated experiment. Figure 5.6 shows the program interface that was projected in class during the activity. The left-hand windows show the program parameter values set for simulating the experiment of randomly drawing 10 samples of 5 items each, without replacement, from a population consisting of 216 red and 216 white items.[13] The Data Record window (top right-hand side) lists the 10 sampling outcomes obtained in one iteration of the experiment.

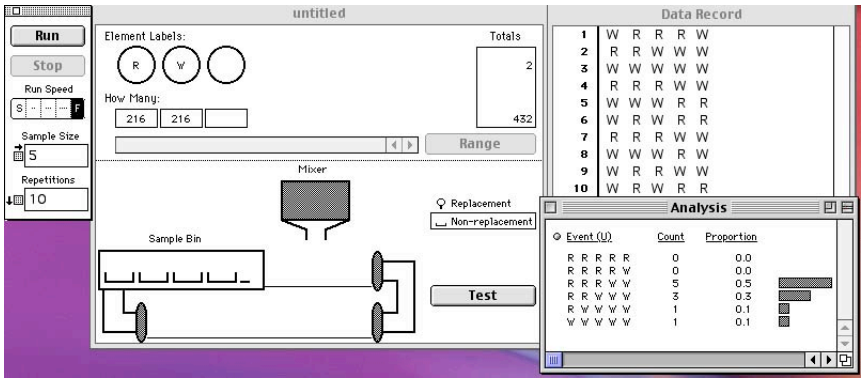


Figure 5.6. Prob Sim input (left) and output (right) windows corresponding to the candy-sampling experiment.

The Analysis window (bottom right-hand side) lists all possible unordered outcomes in the experiment's sample space (Event U column). Next to each outcome of the sample space is

[12] The doubt I expressed earlier (see footnote 8), concerning what process Cathy hand in mind, is rooted in her having been unclear on what to record and keep track of in this activity. After the instructor explicated the method to the class and assigned Cathy to be the official record keeper, she asked "what do you want me to keep track of, like, how many times we get five ones?". She thus seemed fixated on "sample containing 4/5^{ths} red candies" as the event of interest. This raises the possibility that she was oriented to a particular sample outcome rather than a collection of outcomes as a salient object of focus.

[13] This information, which was entered into the program by the instructor, was represented by the symbols "R" and "W", respectively, in the *Element Labels* field, and by 216 in each of the corresponding *How Many* slots. Prob Sim was chosen for several of its intended affordances: the program can function as a calculator in that it automates the selection of multiple samples and the presentation and analyses of a large number of sampling outcomes. The program's presentation format, as shown in the Analysis window, supports building imagery of distributions of sampling outcomes. Prob Sim uses the idea of a "Mixer" containing elements that can be labeled and that are randomly deposited into "Sample bins" when the program is run. This idea is intended as the basic metaphor for thinking about and modeling probabilistic situations in terms of a relationship between a population of items and randomly drawn samples of those items. This metaphor was first shared with students in a demonstration activity immediately preceding Activity 1.

shown the absolute number (Count column) of samples in the experiment having that outcome. The count of each outcome is expressed as a proportion of the total number of samples drawn in the experiment (Proportion column), and as a histogram bar's length (right-most column). With each iteration of the experiment, the information in these output windows is automatically updated to reflect the experiment's outcome.

Using Prob Sim, the instructor repeated the simulation of the sampling experiment five times. Each time, he time drew students' attention to the updated information displayed in the Analysis window. He encouraged individual students to share their interpretations of this information and to compare each outcome with Result 1, asking them "is this outcome like Result 1?". Each time, he took a class vote asking students to decide whether the outcome displayed in the Analysis window was similar ("Yes") or dissimilar ("No") to Result 1.[14] Cathy kept track of these decisions.

The outcome of each of the 5 iterations of the sampling experiment, as displayed in the Analysis window, occurred in the order shown in Figure 5.7:[15]

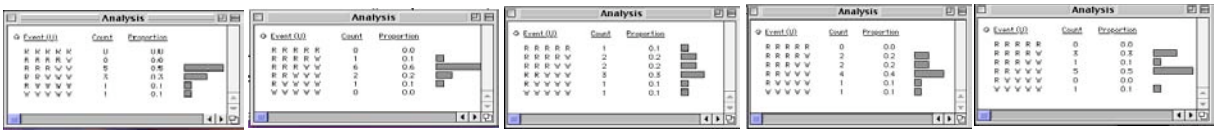


Figure 5.7. The sequence (from left to right) of approximate outcomes of the simulated candy-sampling experiments.

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

Figure 5.8. The reference result against which students compared outcomes of the simulated sampling experiment.

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

Figure 5.9. The class's similarity decisions made in response to the question: "Is this outcome like Result 1?".

[14] Result 1 remained displayed on the board as Activity 2 unfolded.

[15] Each result shown here is a very close approximation to the actual result. These approximations are used only for the sake of clarity. The actual results were only recorded in the classroom videotapes, screenshots of which are barely discernible.

Activity 2 discussion highlights

The discussions that unfolded from Activity 2 centered on deciding whether each simulated sampling outcome was similar or dissimilar to Result 1. Although individual students voiced their vote in these decisions, the group discussions did not expose each and every student's underlying rationale and reasoning at every turn. Instead, with each iteration of the simulated sampling experiment, different students were invited to share their ideas or some students volunteered their ideas. Thus, these discussions do not allow for a systematic inquiry into each and every student's thinking. However, the discussions do offer a cross section of ideas that emerged in the class. Despite these limitations, it has been possible to get a sense of certain individual students' development because of the prominence of their participation relative to other member of the class.

Perhaps the most important occurrence in these discussions was the emergence of a decision rule—a criterion for deciding whether a 10-sample outcome was similar or dissimilar to Result 1. In some sense, the criterion followed naturally from the last discussions in Activity 1 in which the instructor began moving students toward comparing collections of outcomes. However, its explicit introduction by the instructor within these discussions was inspired by his interpretation of a *Prob Sim* histogram as depicting a distribution of sampling outcomes that indicates where the collection's *weight* is concentrated. The criterion amounted to an informal characterization of Result 1 as "heavy toward white"—meaning that the collection was weighted more heavily toward samples containing more white than red candies.

More explicitly, the instructor characterized Result 1 as "heavy toward white" because a majority of that collection's samples contained a majority of white candies. This characterization of Result 1 can be understood if one considers its compliment. In Figure 5.8, Result 1 is expressed in terms of numbers of red candies: it shows that 8 of the 10 samples contained at most 2 red candies. But this is equivalent to saying that 8 of the 10 samples contained 3 or more white candies, and so a majority of the samples contained a majority of white candies. Thus, Result 1 was characterized as "heavy toward white".

Though the decision rule was mentioned only by the instructor, in the classroom interactions students seemed to readily accept the rule; they behaved as though they understood it and were able to apply it unproblematically. This is illustrated in discussion Episode 5 that follows. The episode is comprised of 3 segments, each highlighting a discussion that occurred around a

particular iteration of the simulated sampling experiment. These discussions occurred in the order presented here, but only the first two segments are contiguous. The episode took up part of the last 9 minutes of Lesson 3.

Episode 5, Lesson 3:

Segment 1: *first iteration*

305. I: […] well, let's do it again and, and I'll ask questions to make sure that we're all together (moves Data Record window so that all windows are fully visible as in Figure 5.6). All right, what does this window show, that I just moved?
306. Peter: The-uhh 10 samples.
307. I: Ok, the 10 samples that I just did! What does this window show us? (activates Analysis window, see Figure 5.10)

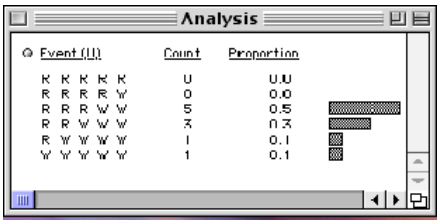


Figure 5.10. Outcome of the first iteration of the simulated candy-sampling experiment.

308. Peter: All together how many—
309. Nicole (utters something inaudible and laughs, presumably at Peter's aborted explanation)
310. I: Numbers of samples that have this many of the different colors (points to list of entries in Event column in Analysis window).
311. Peter: Yeah.
312. I: All right. This "3" tells us what? (points to the "3" in the Count column in the Analysis window)
313. Peter: There's—
314. I: Kit?
315. Peter (continues): three times two red came out.
316. I: Ok. Kit, what does "3" tell us? (points to "3" in window on screen)
317. Kit: Three times it came out with two reds and three white.
318. I: Two reds and three whites (points to corresponding row in Event column). Ok. All right, so, now is this one, is this outcome, set of 10 samples <u>similar</u> to this? (points to Result 1 displayed on board, see Figure 5.8)
319. Nicole: No.
320. I: Ok, why? Because the re--it seems kind of loaded to the reds rather than the whites. Correct? All right. So let's do it again.

Segment 1 suggests that some students had an unproblematic interpretation of the information shown in the window (lines 306, 308, 315, and 317). I note that Nicole was

definitive in her decision that the outcome was not similar to Result 1, and that her response came before the instructor's characterization of the outcome (line 320: "it seems kind of loaded to the reds rather than the white"). This suggests that Nicole was already able to compare collections of outcomes and to determine whether they were similar. In line 320 the instructor second-guessed Nicole's presumable justification and offered a characterization of the outcome that touched on the decision rule "loaded to the whites". His intention was to share with the class a presumable and useful way of characterizing each collection of outcomes. Though, there is no evidence that this characterization was indeed consonant with Nicole's perspective.

The second segment of Episode 5 of Lesson 3 highlights discussions occurring around the second iteration of the simulated sampling experiment.

Episode 5, Lesson 3:

Segment 2: second iteration

321. I (runs second iteration): Is that set of 10 samples (see Figure 5.11) like this one? (points to Result 1 displayed on board, as in Figure 5.8)

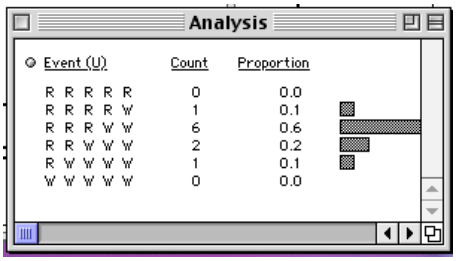


Figure 5.11. Outcome of the second iteration of the simulated candy-sampling experiment.

322. Cathy: No.
323. Peter: No.
324. I: Sarah?
325. Sarah: Uhh
326. Cathy: It's like the second one (referring to Result 2)
327. Sarah (to I): No.
328. I: Ok
329. Sarah: uhh well, yeah
330. I: You think it might be like this (referring to Result 1)
331. Cathy (?): Uhh it's not. It's—
332. Peter: No (shakes head from side to side)
333. I: Ok, this one's a little heavy toward the white (points to Result 1). Correct?
334. Sarah: (nods head in presumable agreement)
335. I (continues): This one is (points to outcome shown in Analysis window), is it-- if it's heavy, it's heavy toward what?
336. Cathy: Red
337. Peter: Red!

338. I: The red. Alright, so likeness will be when it's heavy to the white (points to Result 1)
339. Peter: Hmm hmm.
340. I: So, on this one then (points to histogram in Analysis window) now what would you say, Sarah? Ok this one's a little heavier toward the red. All right, so let's do it again. So we've done it twice and both times we said__?
341. Peter: No.
342. Female student: No.
343. I: Not similar.

Segment 2 shows that three students initially considered the outcome of the second experiment to be unlike Result 1. Cathy seemed to think the outcome was similar to Result 2, instead.[16] All but one student adhered to her/his initial assessment; Sarah was uncertain but did not share the reasons for her uncertainty. Sensing that Sarah was having difficulty deciding, the instructor invoked the "heaviness" metaphor as a way to help her think about and compare the two outcomes. He reiterated his characterization of Result 1 as "heavy toward the white" (line 333), with which Sarah seemed to agree. He then moved to engage students with this metaphor by asking them to characterize the current outcome in terms of its "heaviness" (line 335). Cathy's and Peter's responses (lines 336-337) were consistent with their initial assessments and suggest that, for them, thinking about the collection of outcomes in terms of its "heaviness" was unproblematic. The instructor then reminded students of the decision rule—two results are alike if they are both heavy toward white—and he re-emphasized that the current outcome is heavier toward the red. This sequence of interactions converged to the instructor pointing out, to Sarah in particular, why the decision should be "no" in this case.

Although Segment 2 contains little hard evidence of individual students' cognitions, it suggests that students' responses to the instructor's questions were, nevertheless, not inconsistent with their thinking of the collections in terms of the heaviness metaphor. Indeed, the central idea I draw from the first two segments in Episode 5 of Lesson 3 is that under heavy scaffolding by the instructor, students increasingly warmed up to the discourse of a collection's weight or "heaviness". Without any commitment, on my part, that students had common interpretations of

[16] Here again, there is no evidence of Cathy's underlying rationale for her claim. My only guess as to why she might have considered this outcome to be similar to Result 2 is that, as shown in the *Prob Sim* window, 7 of the 10 samples contained 3 or more red candies. The table displaying Result 2 (see Figure 5.5) showed that 8 of the 10 samples contained 3 or more white candies. Perhaps Cathy only eyeballed Result 2 and based her assertion on a mistaken interpretation of it.

the "heaviness" metaphor or that they shared ways of construing the collections, I take the group as a unit of analysis and speculate that this metaphor functioned much the same way that taken-as-shared (Cobb, Wood, & Yackel, 1992) meanings function; it helped generate a discursive space that enabled students' continued productive participation in the activity.[17]

A final observation I make about Segment 2, and indeed about all discussions in Activity 2, is that students were comparing the two collections across different representational formats: Result 1 was only ever presented in the tabular form shown in Figure 5.8, while the simulated sampling outcomes were presented as shown in the *Prob Sim* Analysis windows. This raises the possibility that however students were construing the collections, those ways were sufficiently robust to withstand these cross-representational comparisons.

The next and final segment of the discussion in Activity 2 illustrates Nicole's way of interpreting the information in the Analysis window. It suggests that her thinking developed, relative to an early part of the lesson, on how to construe collections of sampling outcomes.

Episode 5, Lesson 3:

Segment 3: fifth iteration

344. I: Let's do it again (runs fifth iteration, results immediately appear on screen as in Figure 5.12)

Figure 5.12. Outcome of the fifth iteration of the simulated candy-sampling experiment (left), and a close approximation of it (right).

345. I: All right. What does it show us? (points to Analysis window on screen) Nicole?
346. Nicole: Uhh "yes"!
347. I: Ok. Before you answer "yes", what is this showing us? (points to information displayed in Analysis window on screen)

[17] I should emphasize that the notion "taken-as-shared meaning" was developed as a participatory rather than a cognitive explanatory construct. Cobb et al. (1992) devised the construct to explain how communal mathematical discussions could proceed seemingly productively, or at least non-dysfunctionally, without committing themselves to asserting that participants actually share meanings and understandings of the objects of discourse.

348. (others laugh at Nicole's jumping the gun)
349. Nicole (to others): No, because--
350. Peter: Yeah
351. Nicole (continues): well there's like (1 second pause) 6 times you have, like, 3 whites.
352. I: Ok, that's what I wanted to know. All right, so is this a "yes, similar" or "no, not similar"?
353. Nicole: Yes
354. Peter: Yes
355. I: Ok. Everyone—if uhh, anyone uhh feel otherwise?
356. (no one responds)
357. I: Ok, so it's another "yes". What do we have so far (points to Cathy)
358. Cathy: Uhh three "yes"s and two "no"s.
359. I: Three "yes"s, two "no"s. So we have, so even though we thought this was really, this might be rare (points to Result 1 on board), so far we got three times out of five where we said "well yeah, what we got in those 10 samples is sort of like this"
360. Female student: Yeah.

The utterance highlighted in blue and red in Segment 3 above is Nicole's justification of her decision that the sampling outcome of the fifth iteration was similar to Result 1 (highlighted in grey). Nicole evidently interpreted the information in the Analysis window as being about a number of outcomes (white candies) in a number of times that a sample was selected. By way of a semantic analysis, I parse Nicole's utterance into two parts, each having a distinct referent: in saying "6 times you have, like, 3 whites", Nicole was mindful, on one level, of a number of *samples* containing, on another level, a number of sampled *items* of interest (white candies). Nicole thus appeared to have construed the collection of sampling outcomes in a way that entailed 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.

Figure 5.13 illustrates this analysis, highlighting the correspondence between Nicole's utterance and her presumed objects of discourse as they were displayed in the *Prob Sim* window in class:

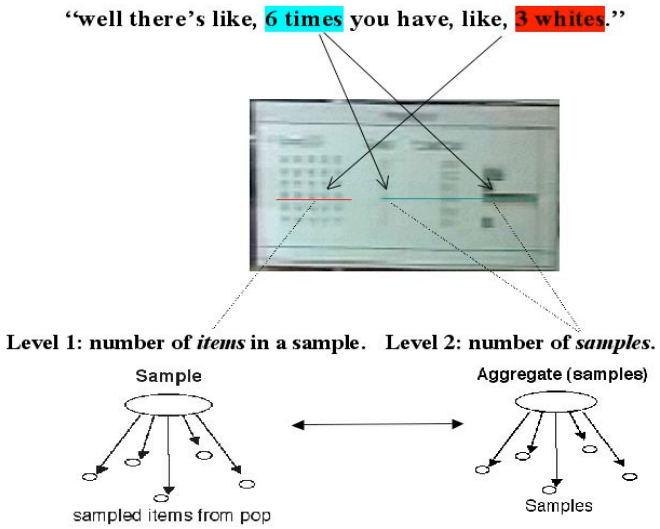


Figure 5.13. A semantic analysis of Nicole's utterance.

I should point out that Nicole's reasoning, from the perspective of this analysis, is consistent with the reasoning that the instructor had been promoting in the discussions. Though Nicole's utterance was not couched in the language of a distribution's weight or a collection's heaviness, it can be re-formulated in those terms and the two characterizations shown to be highly compatible:

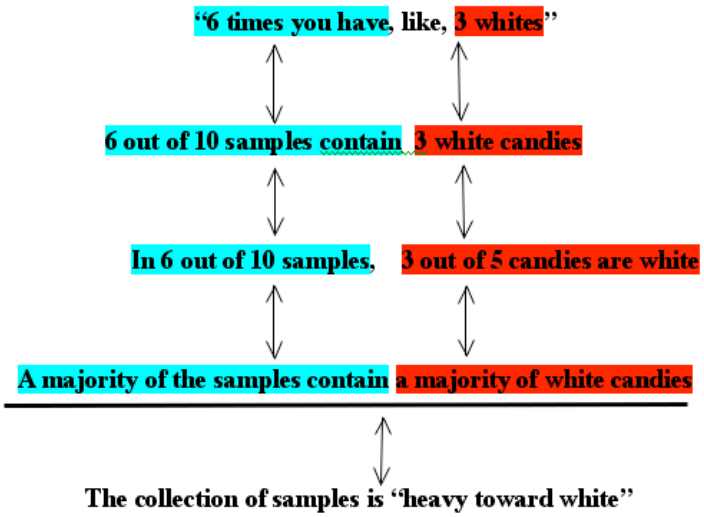


Figure 5.14. A sequence of re-formulations that suggests the compatibility between Nicole's and the instructor's characterizations. Each double-sided arrow links equivalent statements

In sum, I consider Nicole's way of construing the collections of sampling outcomes to be quite compatible with the line of reasoning that the instructor was trying to promote as normative in the classroom. I should also add that other students exhibited ways of reasoning about these collections that were consistent with Nicole's reasoning.[18] Though, no other student provided me with such a clear-cut illustration of having developed a way to construe a collection of sampling outcomes.

I conclude this commentary about the discussions of Activity 2 by reminding the reader of the agenda underlying the activity in the first place. On the one hand, there was the official agenda that was shared with students: the aim of the activity was to determine whether sampling outcomes *like* Result 1 are rare or statistically unusual. On the other hand, there was also an unofficial instructional agenda not explicitly shared with students: to use the activity of investigating a collection's possible unusualness as a context for orienting students toward thinking of collections as distributions. Once the activity was well under way and the discussions focused on how to interpret and construe the collections of sampling outcomes, the once pressing question that initially motivated the activity was overshadowed by the group's engagement in these discussions. Very shortly after the class had voted on the fifth and final collection of outcomes, the class dismissal bell sounded and students immediately began to disengage and depart. Unfortunately, the discussion did not continue beyond the instructor's summarizing comment (line 359 of segment 3), which was intended to suggest that outcomes like Result 1 might not be rare after all. Thus, it is uncertain whether this ultimate conclusion was salient for students.[19]

Activity 3: Rationale and Description

This final section of the chapter discusses a third activity that was designed as follow-up to the activities described so far. Activity 3 built directly on Activity 1 and was created to probe what students "took away" from their engagement in the classroom activities. From the perspective of this analysis, Activity 3 is distinct in that it occasions an unprecedented

[18] In the final section of this chapter I present and discuss evidence of this that emerged in the next (fourth) lesson.
[19] Activity 2 was revisited several lessons later, in a slightly different guise, and was followed through to its intended completion. Then, students agreed overwhelmingly, on the basis of a relatively high proportion of "yes" decisions, that outcomes like Result 1 were not at all rare, but were instead common.

opportunity to assess students' thinking, as it might have occurred outside of the classroom instructional interactions.

Activity 3 was assigned as a take-home activity near the end of Lesson 3. The instructor then conducted a whole-class discussion during the first 18 minutes of Lesson 4 focusing on students' ideas that emerged out of their engagement with the take-home activity.[20] In this section the descriptions of classroom discussions around Activity 3 are less extensive than those around prior activities. One of the aims in the previous sections was to delineate the tone of instructional interactions and engagements that characterize the teaching experiment in general. The central aim here is to present additional evidence of student thinking and engagement with respect to particular ideas and issues already highlighted in previous sections.

Figure 5.15 shows the Activity 3 guide from which students worked. As already mentioned, Activity 3 was designed as an extension of Activity 1. Question 1 amounts to a reformulation of Activity 1; part a) asks to make an inference to the population on the basis of an individual sampling outcome, and part b) asks to make an inference on the basis of the entire collection of outcomes. Rather than having students actually select samples, as in Activity 1, here they were presented with sampling outcomes (10 samples of 5 candies each) that the research team wanted them to understand as having been drawn from a large population of well-mixed red and white candies. The boxed statement in the activity guide expressed this intent.

[20] Couched in a more traditional discourse, I could simply say that "students were asked to complete a homework assignment and their responses were then discussed in class". However, I carefully choose a discourse that highlights my interest in student engagement and thinking. My occasional slippage into the traditional discourse should not create confusion about my interest.

Stat assignment 1

Answer these questions on a separate sheet of paper. Copy the question above your answer. Please write complete sentences and explain your thoughts fully.

A large jar is full of red and white jellybeans that are evenly mixed. Ten samples of 5 jellybeans each were selected at random from the jar, the samples had the following outcomes:

1	red	white	red	white	white
2	white	white	white	red	white
3	red	white	white	white	red
4	white	red	white	white	white
5	red	red	white	red	red
6	white	white	white	red	white
7	red	white	white	white	white
8	red	red	red	white	white
9	red	white	white	white	red
10	white	white	white	white	red

1. Examine how these samples are distributed with regard to the number of red jellybeans in them.

   a) Does any individual sample lead you to believe anything about what fraction of the jar's jellybeans are red? Please explain (i.e., if so, say what and why. Otherwise, say why not).

   b) Does the distribution of the ten samples lead you to believe anything about what fraction of the jar's jellybeans are red? Please explain your answer.

2. Tomorrow you will randomly draw ten samples from this same jar of jellybeans and record the color of each jellybean as you draw it. However, your samples will contain 8 jellybeans instead of 5. Make a list of 10 samples you can reasonably expect to draw. (Use "R" for "red" and "W" for "white".)

3. In making your list of 10 samples, did something occur to you that will be different about 8-bean samples than was the case with 5-bean samples? If so, describe it.

Figure 5.15. Written guide for Activity 3.

Question 2 of Activity 3 was intended to engage students in a sampling thought-experiment. The idea was to have students anticipate sampling outcomes that they might reasonably expect to draw from the same population of candies. Students' responses might then provide the research team with information about what ideas had been salient for them in the classroom sampling activities. In particular, the responses might shed light on the bases for students' predicted outcomes and on whether students were oriented to making connections between their anticipated outcomes and the population inference they made in Question 1.[21]

[21] The third question in the activity aimed to query students' intuitions and ideas that might relate to sampling variability. The question was not addressed in the class discussion in this phase of the experiment, nor will I address it here.

Activity 3 results and analyses

The next three subsections of the chapter each address one of the Activity 3 questions, analyzing students' responses in terms of what they suggest about their conceptions and elaborating this with evidence drawn from corresponding discussions in Lesson 4.

Question 1a

Table 5.1 shows students' responses for Question 1a: "Does any individual sample lead you to believe anything about what fraction of the jar's jellybeans are red?

Table 5.1. Students' written responses to Activity 3, Question 1a.

Student	**Response**
Nicole	No, because one individual sample does not prove a lot.
Sue	No, it doesn't because individual sample might be a rare sample; if so, a fraction of the red jellybeans will be different from the actual fraction.
Kit	Yes, if you just look at any one sample (example sample 4), it would lead you to believe there is about a 1:4 ratio (red:white).
Sarah	Yes, I would normally say the jellybeans are about 3/5 white. Most of the time there are either 3 or 4 white jellybeans out of 5.
Peter	No, not any sample leads you to believe anything about the fraction of red jellybeans. By just looking at one sample, can one be led to believe anything about the jellybeans? It could just be an unusual sample.
Lesley	Most samples, 8 out of 10, had fewer white than red jellybeans. Only 2 out of 10 times would you think there were more red than white jellybeans in the jar

Only Kit responded in the affirmative and employed the single-outcome inferential line of reasoning raised in the discussion surrounding the first part of Activity 1.

Three students—Sarah, Sue, and Peter—answered the question in the negative, implying that they thought individual outcomes don't provide enough information about the sampled population to make a trustworthy inference. Two of those students, Sue and Peter, appealed explicitly to the idea that an individual outcome could be unusual. Sue explicitly mentioned the possibility of making an erroneous or unreliable inference as a consequence.

Sarah's and Lesley's responses show that they were focused on the collection of outcomes as a whole; both referred to what happened in *most* of the samples as a basis for what might be true of the population. Sarah ventured to give a specific numerical estimate of the population proportion, thus making a full-fledged quantitative inference. Lesley fell just short of doing so, but her response suggests that she thought it obvious that the population contained a majority of

red candies. I might add that both Sarah's and Lesley's explanations are consistent with Nicole's way of construing a collection, as I proposed earlier (see Figure 5.13), and thus consistent with seeing the collection of sampling outcomes in terms of its "heaviness".

The difference in their degrees of elaboration notwithstanding, all but Kit's responses to Question 1a are consistent with their thinking that collections of sampling outcomes provide a useful, if not preferable, basis for making an inference to the underlying population.

Question 1b

Students' responses to Question 1b—"Does the distribution of the ten samples lead you to believe anything about what fraction of the jar's jelly beans are red?"—are displayed in Table 5.2. By virtue of their compatibility with Nicole's, these responses suggest that her structuring of the collection (see Figures 5.13 and 5.14) might not have been uncommon among students.

Table 5.2. Students' written responses to Activity 3, Question 1b.

Student	**Response**
Nicole	I'd say no more than 3/5 of the jellybeans are red because only one of the samples have more than that.
Sue	Yes it does. Count out the number of red jellybeans in each sample, and make a list which indicates how many samples in each 0 through 5 possible number of jellybeans. The list shows us which number will occur more likely. Then we can see a fraction.
Kit	It leads you to believe that the ratio would be less reds to more whites.
Sarah	Yes. Most of the time there are 2 or 1 red jellybeans out of five, meaning either 2/5 or 1/5 are red.
Peter	Yes it does. It leads me to believe that the majority of the bag is white and not red. In all samples except one a majority of white beans were taken.
Lesley	One might think that there are fewer reds than whites. Only 2 out of 10 (20%) of the sample have more red than white. Only 36% of the <u>total</u> samples

Lesley and Sue's responses, when considered in tandem with their discussion in the classroom, suggest interesting hypotheses about their perspectives. Consider the first sentence of Lesley's response: "One might think that there are fewer reds than whites. Only 2 out of 10 (20%) of the sample have more red than white". This is precisely what I claimed she was implying in Question 1a, but had fallen short of explicitly saying so. The next sentence of her response—"Only 36% of the <u>total</u> samples"—is, if taken at face value, perplexing because it seems to contradict her claim in the first sentence. However, a closer look at Lesley's copy of the Activity 3 guide revealed that she had enumerated the numbers of red and white candies drawn

in each sample and determined that there were a total of 18 red candies selected out of 50 candies (see Figure 5.16). That is, Lesley found that 36% of the total number of *candies* drawn—not samples—were red. She thus evidently mistakenly wrote "sample" instead of "candies" in her response.

						R	**W**
1	red	white	red	white	white	2	3
2	white	white	white	red	white	1	4
3	red	white	white	white	red	2	3
4	white	red	white	white	white		
5	red	red	white	red	red	1	4
6	white	white	white	red	white	*4	1
7	red	white	white	white	white		
8	red	red	red	white	white	1	4
9	red	white	white	white	red	1	4
10	white	white	white	white	red		
						*3	2
						2	3
						+1	+4
						18	32

Figure 5.16. A reconstruction of Lesley's work, written on her copy of the activity guide. The asterisks denote samples containing "more red than white".

Lesley's mistake, per se, is of little interest to me. What *is* interesting, however, is her idea to enumerate the number of candies: it suggests her having had an alternative way to structure the sampling outcomes as a *collection of candies*. Indeed, Lesley's strategy is further elaborated in the following brief discussion excerpt that ensued between her and the instructor (I).

Episode 1, Lesson 4:

1. I: Lesley? Did you summarize this? How, how was it that you looked at the whole set of all 10 samples?
2. Lesley: Do you want to me to read what I wrote down, or--?
3. I: No. I'm asking a different question
4. Lesley: Ok, well ask again 'cause I wasn't paying attention
5. Nicole chuckles
6. I: Ok. Uhh, what did you do to look at all 10 samples to uhh as one collection?

(3 second silence)

7. I: Did you just eyeball it? Or did you summarize it somehow?
8. Lesley: Oh, well I added up all the reds and all the whites,

9. I: Ok, so you count—
10. Lesley: and looked at the percentage
11. I (continues): so you looked at the numbers of each.
12. Lesley: Right.

Thus, in summing the total number of candies selected and looking at what percentage of them were red, Lesley evidently treated the collection of sampling outcomes *as though* it constituted one large sample. That is, Lesley seemed to collapse the collection of distinct samples into one large collection of candies. I presume that she took the 36% not as the *sole* basis for her assertion that the population contained fewer red than white candies, but rather as additional and supporting evidence for it.

Lesley's novel and apparently non-normative approach raises questions about what sense she had made of the classroom discussions leading up to this juncture. Those discussions had promoted partitioning a collection into classes of sampling outcomes as a basis for making an inference. As my analysis of Nicole's strategy asserts (Figure 5.13), this structuring entails coordinating two levels of imagery of the sampling process in a way that distinguishes sampled *items* from *samples* of items. Lesley's response to Question 1 clearly indicates that she was able to enact such a partitioning, but it also suggests that when left to her own devices she was inclined to take a perspective that seems to ultimately smudge over these distinctions. Accordingly, I pose two related questions: 1) what did Lesley think was the aim of selecting multiple samples, and 2) did Lesley view re-sampling predominantly as a way to amass one large sample?

These questions will, unfortunately, remain unanswered.[22] However, I reformulate question 2 as a conjectured schema that may have underlay Lesley's strategy: *re-sampling as a method for growing a sample*. Putting aside the issue of this conjecture's un-testability, such a schema entails a view of sampling that departs significantly from the multiplicative conception of sampling (MCS) already elaborated in an earlier chapter. Blurring over individual samples as distinct units of selection and quantification, and *dissolving* them into sampled items that amass into one large sample is a structuring that is almost antithetical to the imagery that supports conceiving a collection of sampling outcomes as a distribution. Figure 5.17 attempts to depict each of these conceptions pictorially.

---

[22] Lesley transferred out of the course two lessons later.

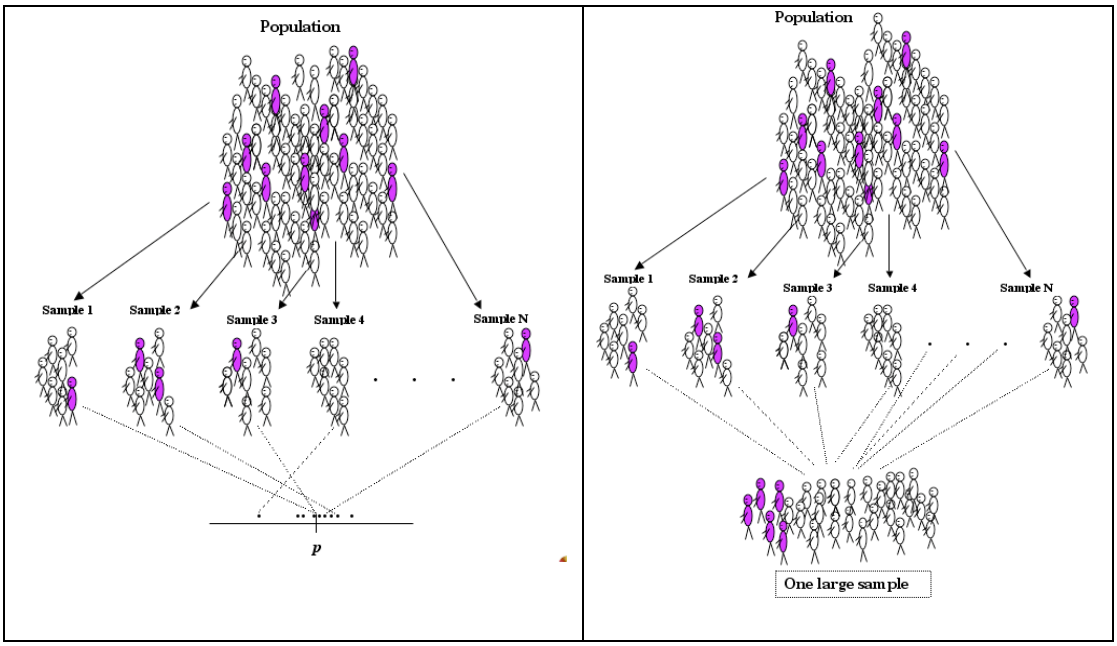


Figure 5.17. Multiplicative conception of sample (left) versus "grow a sample" schema (right).

The other student response to Question 1b that interests me is Sue's (Table 5.2). Note that Sue responded to the question affirmatively, but unlike the other students she did not conclude anything about the population. Instead, her response describes the general steps of the *process* by which one could come to make a conclusion about the population. Thus what appeared to be most salient for Sue was the method enacted in class, especially the tabular format used to represent sampling outcomes. A class discussion suggests that Sue had indeed assimilated the tabular format used in class as a way to impose a structure on the sampling outcomes. Though, interpreting that structure in terms of what it implies about the sampled population still entailed some effort for her. The discussion excerpt, lasting approximately 3.5 minutes, is comprised of 2 ordered but non-contiguous segments. It begins with the instructor asking Sue why she concurred with the conclusion of another student.

Episode 2, Lesson 4:

Segment 1

13. I: That, is that what you concluded, that the samples all—as a collection—suggest more white?
14. Sue: Yeah
15. I: Ok, and, and what led—if someone were to say "I don't see it", how would you explain to them how to look at it so that they would see that it suggests more white?
16. Sue: Oh, make a list.
17. I: Pardon me?

18. Sue: Make a list, like that (points to data table left on board from lesson 3)

The instructor constructed the following table on the board, which he filled in with sampling data read out by Sue.

# of whites	**0**	**1**	**2**	**3**	**4**	**5**
# of samples	**0**	**1**	**1**	**3**	**5**	**0**

Figure 5.18. Sampling data from Activity 3 organized in a frequency table.

Episode 2, Lesson 4:

Segment 2

[…]

19. I: All right, now so how do we, how should we look at this? Tell me how to look at this so that now I can see that it suggests we have more whites uhh than reds.

(5 second silence)

20. Sue: I uhh
21. I: Yeah, suppo—I, I'm pretending to be someone who says "I don't see why that suggests that there are more whites". So what would you say in response?
22. Sue: Uhh because there's more white in the, each sample.
23. I: There are more whites in each sample?
24. Sue: Yeah.

Though Sue appeared to have overestimated to all samples in the collection, her sense of the great frequency with which the outcome "more whites" occurred was her apparent warrant for concluding that the underlying population contained more white candies.

After Sue's explanation, a discussion quickly ensued in which other students engaged in explaining their reasoning with respect to the table in Figure 5.18. The following discussion excerpt, lasting approximately 2 minutes, is divided into two contiguous segments. The excerpt illustrates further how Nicole and Peter were structuring the collection of outcomes presented as in Figure 5.18.

Episode 3, Lesson 4:

Segment 1

25. I: Yeah, explain that to somebody who says "I don't see why that suggests that there are more whites".
26. Nicole: Ok, well, like (2 second pause), like half of the time—we took 10 samples—so half the time four out of the five were white. So that's, like,

pretty much—that's, like, a big number (bursts into laughter at her own response), isn't it? […] Like, I guess you could say that 80% of the time you have more (laughs at Peter's inaudible comments), three or more whites. Am I making sense to you?

[…]

27. I: So 80% of the time you had a majority of white.
28. Nicole: Right!
29. I: Ok. That, that, that's pretty compelling!

Nicole's structuring (segment 1) is consistent with that which she employed when interpreting the sampling outcomes as they appeared in the *Prob Sim* window (see Figure 5.13). Given that the data was presented in different formats each time, this suggests that Nicole's way of structuring the sampling outcomes was relatively stable and robust.

Episode 3, Lesson 4:

Segment 2

30. Lesley: Ok, but I don't understand how you can tell it just from that little thing, right there (points to data table in Figure 5.18). Ok, 'cause, like—
31. Peter (to Lesley): alright, see there are 10 samples
32. Lesley: Right
33. Peter (continues explaining to Lesley): and three—alright, from three, four and five (points at data table in Figure 5.18), if you have three, four, or five whites in—like if you have three whites in one sample that means you have a majority of whites, right?
34. Lesley: Right
35. Peter: Right? And three, four, and five (points to table in Figure 5.18 and motions with hand from left to right as though counting), so anything on the right side of the table (sweeps hand to right while pointing to table in Figure 5.18) means you would have a majority of whites. Like—
36. I: So, in these three samples (points to data value "3" in table in Figure 5.18)
37. Nicole (to Lesley): Like, eight of the samples
38. Peter (to Lesley): The numbers under three, four, and five.
39. David (to Lesley): Eight of the samples have more than three whites.
40. Lesley: Ok.

[…]

41. Peter: It's just eight outta ten, that's 80%.

Segment 2 illustrates that despite having participated in discussions over 3 lessons, many of which centered on interpreting such data tables, Lesley still found it problematic to interpret the table in a way that allowed her to make an inference to the sampled population (line 30). As indicated by her responses to Question 1 of Activity 3, Lesley was apparently able to structure the sampling outcomes coherently with Nicole's structuring. Furthermore, she did so when the

data was presented as a list of sampling outcomes as in Figure 5.13, rather than a numerical summary as in the tabular format. Given these considerations, I believe that Lesley's problem here was in decoding the particular tabular form (Figure 5.18). Thus, whereas Sue seemed to rely on this format as a way to impose structure on the collection, Lesley seemed to be hampered by this format.

Peter's explanation of how to consider the tabulated data (utterances highlighted in blue) strongly suggests his having structured the data in terms of a partition: the number or proportion of samples containing a majority of white candies. His imagery of this partition seemed to entail a vivid graphic component, as he noted, and motioned to in a sweeping gesture, that the entries in the right half of the table constituted 80% of the samples (see Figure 5.19). Thus, Peter's structuring seemed to entail coordinating the same two levels of imagery that I described in my earlier analysis of Nicole's reasoning (see Figure 5.13).

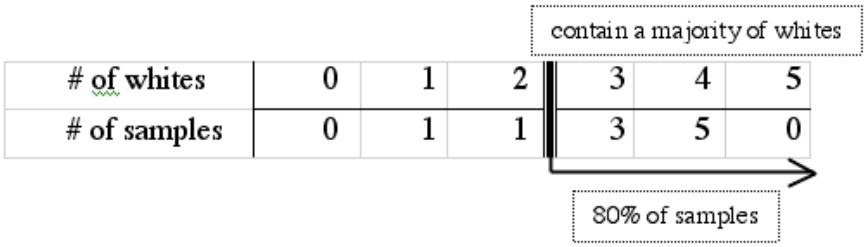


Figure 5.19. Peter's partitioning of the tabulated sampling outcomes for Activity 3.

Question 2

Table 5.3 displays students' responses to Question 2 of Activity 3. Recall that the task was to construct a list of 10 samples of 8 candies each that one could reasonably expect to draw from the same population of red and white candies. Moreover, the intention was that this list emerge out of engagement in a thought-experiment in which one imagines keeping track of individual outcomes as sampling items are selected.

Table 5.3. Students' written responses to Activity 3, Question 2.

Sample	Nicole *	Sue *	Kit *	Sarah *	Peter	Lesley	
						Red	White
1	RRRWWWWW	RRWWRWWW	RWWWWWRW	WWWWWRRR	WWWWWRRR	5	3
2	RRWWWWWW	WWWWRWRR	RWRWRWRW	WRWWRWWW	WWWWWWRR	5	3
3	RRRWWWWW	WRWWRRWR	WRRRRRRW	RWWWRWWR	WWWWRRRR	5	3
4	RRWWWWWW	RWWWWWWR	WWWWRWRW	RRRWWWWW	WWWWRRRR	2	6
5	RRRRRWWW	WWRRWRWR	RWWRWRWW	RWWWWWWW	WWWWWRRR	2	6
6	RRWWWWWW	WWWRWRWR	WWWRWWRR	WWRRWWWW	WWWRRRRR	2	6
7	RRWWWWWW	WRWWRWWR	RRWWWRWW	WWRWRWWR	WWRRRRRR	3	5
8	RRRRWWWW	RWWRRRRW	RWWWRRRW	RWWWWWWR	WWWWWWRR	5	3
9	RRRWWWWW	WRWWRWWW	WRWWWRWW	RWWRWRWW	WWWWRRRR	4	4
10	RRWWWWWW	RWRWWRWW	WRRWWRWW	WWWRRRRW	WWWWWRRR	4	4
						total R	total W
						37	43
						46%	54%

If we consider each list as a distribution of sampling outcomes, we can see that almost all are heavy toward white samples.[23] Peter's list is almost heavy toward whites; exactly half of his samples, instead of the majority of them, contain a majority of whites. Overall, students' anticipated outcomes were consistent with their inference, in Question 1b, that the sampled population contains more white than red candies. However, students did not justify or explain their choice in their responses.

The lists constructed by Nicole, Peter, and Lesley draw my attention because they are noticeably different from the others. Nicole's and Peter's sequencing of items in a sample appear to indicate a disregard for the order in which individual items might be drawn. One would presumably expect the random selection process to produce *some* mixing or alternation of

[23] The asterisk appearing next to a student's name in the table indicates that his/her list is heavy toward whites. It is easy to determine whether a list is heavy toward white by examining its frequency table (as in Figure 5.18).

individual outcomes. But this expectation is not at all reflected in neither Nicole's nor Peter's lists, which partition the red candies from the white ones.

Lesley's response is even more reductive than other students' in that it specifies only the numbers of red and white candies expected in each sample instead of the sequence of sampled items. Moreover, Lesley's tallying of the total number of red and white candies drawn in the samples is consistent with part of her response to the previous question; it suggests that she amalgamated samples of candies into one collection of candies, a majority of which was white. Again, I assert that this strategy is consistent with her having an underlying schema of re-sampling as an additive method for growing a large sample.

No additional information is available concerning either Nicole's or Lesley's thinking. However, a class discussion sheds some light on Peter's thinking and may provide a basis for speculating about another student's thinking. The discussion was prompted by the instructor's call for students to share the thoughts that occurred to them as they were engaged in the activity. The following discussion excerpt, spanning approximately 7 minutes, is comprised of 3 ordered but non-contiguous segments.

Episode 4, Lesson 4:

Segment 1

42. I: All right. Let's talk a little bit about the process that you went through, and anything that might have occurred to you as you were doing it. Did you do what it asked you to do?—write "W" or "R, and then "W" or "R", one at a--as if you were selecting these things one at a time? Did you do that?
43. Peter: I just did 'em all like—like, put all the "W"s and put all the "R"s together.
44. Nicole: Yeah, that's the way I did it
45. Peter (continues): I didn't really do, like one at a time, like picking them up. […]
46. Peter: I just put them all together, just like "W W W R R R …" (shows his work to I). I didn't go, like, "W" ,"R", "W", "R", "W", "W", "W", "R", type of thing.

Segment 1 shows that Peter did indeed disregard any plausible order in which items might have been selected. In addition, Peter's claim in lines 45-46 suggests that the item-by-item selection process was not a salient part of what he had imagined. Instead, he seems to have focused on the post hoc results of a sampling experiment.

Episode 4, Lesson 4:

Segment 2

47. I (to Kit): Now, as you were writing this one, where you had a whole bunch of "W"s (points to Kit's first sample). Do you remember what you were thinking as you wrote that "R"?
48. Kit: (Shakes head from side to side)
49. I: Like, maybe "gee, I'm getting too many Ws"?
50. Kit: No.
51. I: No? (2 second pause) Uhh, when you wrote (2 second pause), let's see (3 second pause), ok did you have one where you started out with a number of "R"s?

[…]

52. I (writes Kit's 7th sample on the board): All right, now was there anything that you were trying to do that was true of your samples as you made these?
53. Kit: No!
54. I: No?
55. Kit: I just randomly did it.

Segment 2 reveals that Kit also did not focus on an imagined item-by-item sampling process. In addition, she claimed that in generating her list she was not guided by any criterion for what might be true about the samples. Instead, she claimed that she simply generated the samples "randomly".

Episode 4, Lesson 4:

Segment 3

56. I: Ok. (walks over to right side of board and points to data table for Question 1b, see Figure 5.18). Now what does this, uhh what did we say that this suggests, about the jar?
57. Peter: A majority of them are white.
58. I: Majority of them are white. All right. So, uhh if we picked eight and, in fact, there are a majority of white in the jar
59. Peter: Like
60. I: then what would you expect over the long run?
61. Peter: what I was doing when I was making my samples was, I was—what I got from this that you gave me (picks up Activity 3 guide), there's a majority of white. So when I was doing mine, I was, I guess I was trying to make sure that there was, like, a majority of white.
62. I: Ok.
63. Lesley: Yeah.
64. I: All right. That's fair enough. Uhh, were you doing that, Kit?
65. Kit: No. I just wrote it down.

Segment 3 begins with the instructor following-up Kit's comment about having randomly generated the samples. He apparently sensed that Kit was using "random" in a non-stochastic sense, perhaps meaning outcomes that are not knowable in any way ahead of time, pre-determined, or guided by any deterministic considerations. His strategy was to draw students' attention to the issue of long-run expectation of sampling outcomes under an assumption about the sampled population. Peter immediately took to this issue, as though it suddenly jogged his memory that he had tried to make his list of sampling outcomes reflect what he had assumed about the population. Peter had apparently taken the list of samples given in the activity guide as evidence that the population contained more white candies, and he claimed to have used this as a criterion for generating his list of samples. However, as I already pointed out, his list falls a bit short of reflecting this assumption.[24] The segment also suggests that Lesley had constructed her list with the same rationale in mind as Peter. Kit, however, seemed adamant that no such rationale motivated the creation of her list. Yet, the distribution of her list is weighted more heavily toward white candies than either of these two students.[25]

Chapter Summary

It is useful at this point to take the class as a unit of analysis and to consider the instructional episodes of Activity 1 and Activity 2 in terms of broad phases that unfolded out of cycles of interaction between instruction and student engagement and thinking. These phases are delineated by key shifts in the group's focus of attention and objects of discourse. The first phase of engagement was marked by a focus on outcomes of individual samples and inferences to the sampled population based on them. A critical development occurred with students' acceptance—prompted by an orienting cue from the instructor—that the uncertainty of, and the

---

[24] It came out in a subsequent discussion that Peter had also interpreted the words "evenly mixed" in the boxed text in the activity guide (Figure 5.15) to mean that the population was evenly split between red and white candies. Thus, the tension he sensed between two conflicting implications about the population composition may have influenced the generation of his list of samples, 50% of which contain a majority of white candies. Perhaps Peter tried to create a list that could be consistent with both assertions about the population.

[25] Given the paucity of information concerning Kit's strategy for generating her list, it is difficult to surmise what might have been at play with her. However, a relatively innocuous conjecture is that Kit had indeed created her list to reflect her assumption that the population contained a majority of white candies, but that she interpreted the discussion in Episode 3, especially the instructor's questions, to have been about what she decided at the moment of writing each sampling outcome in her list. Thus, Kit, like Peter, may have approached the task as one of creating a post-hoc list of sampling outcomes.

variability among, sampling outcomes necessitated a consideration of multiple outcomes in order to obtain reliable information about the sampled population.

This development led to a second phase of engagement marked by re-sampling to accumulate collections of sampling outcomes and a focus on these collections. Here, instruction attempted to orient students toward thinking of these collections as distributions. Students moved toward structuring a collection of sampling outcomes in terms of the relative number of samples that contained a majority of one outcome (e.g., white candies). This structuring is consistent with that promoted in instruction which involved organizing sampling data in frequency tables. In the case of the candy-sampling experiments, this structuring appeared to form the basis of students' inferences to the sampled population. I hypothesize that this structuring entails coordinating two levels of imagery, each involving the quantification of a different attribute: one level considers individual sample compositions, while another level considers the relative weight of a collection's part satisfying some outcome criterion. These conceptual operations and their coordination can be taken to characterize aspects of imagining a collection of sampling outcomes as having a distributional structure.

The line of reasoning exemplified by this structuring of collections seemed to enable students' continued productive engagement in the instructional activities. I argue that it underlay their eventual ability to agree on a method of comparing entire collections of sampling outcomes and to use this method as the basis of a rule for deciding when two such collections are similar. These developments marked the emergence of a third phase of engagement that appeared to have been driven by interactions centering around two critical events: 1) the surprise that students experienced upon learning that Result 1 was actually obtained from an evenly-split population, and 2) instruction that capitalized on this surprise 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 the class borrowed the previously used strategy of selecting multiple samples and, using a sampling simulator, applied it to entire collections of samples. Instruction moved students toward structuring the simulated re-sampling results in collections of 10 and to compare the distribution of each 10-sample collection with the reference distribution given by Result 1. This move was facilitated by the sampling simulator's presentation format. The similarity decisions that emerged out of these comparisons were recorded and accumulated into a collection. The ultimate instructional aim was to have students view this collection as a

distribution of decisions and use it as a basis for deciding on the relative unusualness of the reference distribution (i.e., Result 1).

In sum, the class's focus and objects of discourse appeared to progress in complexity from individual sampling outcomes, to collections of sampling outcomes, and finally verged on a collection of similarity decisions arising out of comparing distributions. This hypothesized developmental trajectory is summarized in Figure 5.20.

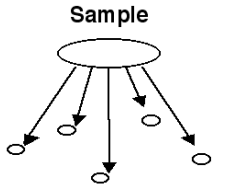
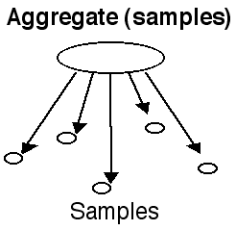
Phase	Activity	Central objects of discourse
1	Select items from population to aggregate a sample Estimate population proportion on basis of individual sampling outcome	Individual samples of 5 items 
2	Select multiple samples to aggregate a collection of samples Estimate population proportion on basis of distribution of collection	Collection of 10 samples 
3	Compare collections Use a decision rule to determine similarity of collections (Decide on relative unusualness of a collection of 10 sample outcomes)	Collections of 10 sample outcomes (Collection of similarity decisions)

Figure 5.20. A hypothetical trajectory of the class's development in Phase 1 of instruction.

CHAPTER VI

PHASE 2: MOVE TO CONCEPTUALIZE PROBABILISTIC SITUATIONS AND STATISTICAL UNUSUALNESS

This chapter describes parts of two activities that unfolded over Lessons 5 through 8 (see Figure 6.1). Activities 4 and 5 engaged students in designing sampling simulations as a means for investigating the unusualness of specified events. The instructional aim of this set-up was to provide a context that might support students' conceiving a given situation in terms of a probabilistic experiment, and conceiving an event's unusualness as a statistical quantity.

Phase 2:	**Move to conceptualize probabilistic situations & statistical unusualness**	
Lesson	**Activity (A)**	**Duration**
5 (08/20)	Preliminary discussions	16 m.
5	A4: Situation1—The Birthday Problem	12 m.
6 (08/23)	A4: Situation 1 revisited	14 m.
	A4: Situation 3—The Movie Theatre Problem	12 m.
7 (08/24)	A5: Problem 1—The Guts Scenario	40 m.
8 (08/26)	A5: Problem 3—The Gallup versus Harris Scenario	12 m.
Assessment session		

Figure 6.1. Chronological overview of activities in Phase 2.

The chapter begins by elaborating the rationale for Activities 4 and 5, drawing on conjectures about student engagement in activities of Phase 1 and their responses to an assessment item. The chapter then elaborates developments that emerged as students engaged in the simulation-design activities, highlighting classroom discussions around two tasks in Activity 4. The chapter concludes with analyses of student responses to a series of post-activity assessment questions.

Analyses foreground the nature of students' thinking with respect to two central issues:
1) conceiving an event's expectation, likelihood, or unusualness as a statistical quantity;
2) (re)conceiving a situation as an idealized probabilistic experiment.

Prelude to Phase 2

The emergence of the experiment's second phase was marked by a move to engage students in designing sampling simulations, using the *Prob Sim* program (Konold & Miller, 1994) as a template. I should point out, however, that designing simulations per se was not the overarching

instructional aim of the activities in this phase. Rather, design activities structured by the constraints of the software were intended as a context for helping students conceptualize a stochastic experiment. That is, the simulation design activities were intended as a context to occasion (re)conceiving certain situations in terms of a repeatable random sampling experiment —one entailing a sample, a population, and a random selection process—for which it makes sense to ask and answer the question "what fraction of the time might we expect such and such sampling outcome to occur in the long run?".

The research team's rationale for moving instruction in this direct at this juncture of the experiment was manifold. For one, the activities of Phase 1 made it unproblematic for students to think of a sample and a population because these components were transparent in the problem statements. Thus, students never really had to grapple with conceptualizing a scenario in terms of a population and a sample under possibly ambiguous conditions. The research team also felt that students' responses to the Activity 3 task, discussed in Chapter V, suggested that the activity of structuring collections of sampling outcomes might have overshadowed images of the repeated sampling process that produced those outcomes. A related issue was the fact that Activity 2 had not been taken to its intended conclusion, and this raised concern that students had not internalized re-sampling and the emergence of *long-run* data patterns as a viable statistical method for investigating hypotheses and as a basis for quantifying likelihood, expectation, and unusualness of outcomes. Students' responses to a short in-class assessment item, given at the end of Phase 1, provide evidence of the limits of their thinking with regard to some of these issues. Students were asked to interpret and respond to the following scenario:

> *Jamie surveyed 10 randomly selected groups of 5 students at BHS, asking them to rank all their favorite singing groups. The Backstreet Boys were ranked at least third in 6 of the 10 groups. In writing her report, she stated, "This suggests that the Backstreet Boys will be ranked at least third around sixth-tenths of the time." What does she mean by "six-tenths of the time?".*

By design, the scenario describes a sampling survey having a structure that could be seen as similar to the experiments in the previous activities: 10 samples of 5 items were selected at random and a common outcome was observed in a proportion of those 10 samples. The available student written responses are shown in Table 6.1:

Table 6.1. Students' written responses to the "Jamie scenario".

Student	Response
Nicole	Since she took 10 samples and 6 of the times the Backstreet Boys were ranked 3rd, then she concluded that 6 out of 10 times they would be ranked 3rd, or 6/10
Kit	6/10 as opposed to 10/10 = 100% 60% of the time
Sarah	~~Out of ten times a group is asked the will~~ ~~Ten groups are asked. Are~~ Six of those groups will rank the back street Boys as third or higher
Peter	6/10 of teenagers will list them in the top three
Lesley	60% 6 out of 10 students don't care for them
David	She means that six out of ten groups of students at BHS rank the BSB's 3rd best singing group, or better

Two types of responses strike me as especially suggestive of some students' thinking. Three students—Nicole, Sarah, and David—give the first type. Their responses to the question "What does she mean by six-tenths of the time" suggests that they were not interpreting the "six-tenths of the time" as a rate; they hadn't generalized this proportion to a percentage of any large number of times that a similar experiment might be repeated. Jamie's statement in the scenario was intended to highlight the subtle distinction between what she had *observed* and what she *inferred* would happen any large number of times she were to repeat the same experiment. Students' allusions to "6 out of 10 times" suggest that they were insensitive to this distinction: they seemed to focus on the observed outcome, or on the expected outcome were the experiment repeated exactly 10 times.

Attention to such seeming minutiae is, nonetheless, part of the subtle reasoning entailed in thinking of statistical inference as deeply tied to the idea of a sampling distribution. Developing sensitivity to such a detail presumably supports understanding that outcomes of random sampling experiments converge to stable values in the very long run, or equivalently, that distributions of sample statistics converge to stable theoretical distributions in the very long run. These are foundational ideas that undergird statistical inference.

The responses of Peter and Lesley typify another problem that echoes findings from a previous experiment (Saldanha & Thompson, 2002). Peter and Lesley evidently interpreted Jamie to be referring to a proportion of *people* in the sampled population rather than a proportion of *samples* selected. This suggests that Peter and Lesley had not interpreted the re-sampling experiment described in the scenario in terms of a coordinated two-tiered process: on one level 5 individual people are selected to form a sample, on a second level the first level process is

repeated to accumulate a collection of 10 samples. The outcome that Jamie observed—a proportion of samples in which Back Street Boys was ranked third or higher—quantifies an attribute of the collection of samples, at the second level of the process. But, because a ranking by an individual person need not be the same as a ranking by a group of persons, the proportion that Jamie observed does not simply translate into an inference about what proportion of people in the population ranked the Back Street Boys third or higher.[1] Thus, it would appear that Peter and Lesley confounded people with samples of people in their efforts to make sense of this scenario.[2]

Preamble to the Main Activities

Before describing the main activities of this phase of the experiment, I should mention that these activities were preceded by a whole-class discussion centering on how to use Prob Sim to simulate simple situations that involve a random selection process. These discussions were intended not only to help familiarize students with the program's interface and workings[3], but also to begin raising more substantive sampling issues such as the distinction between sampling with and without replacement. For instance, one situation was described as follows: "Select 10 pennies from a bag containing 100 pennies. Place them on a table, recording which side is up." It emerged in the discussion that, for the purpose of simulating this situation, it is unimportant to actually draw 10 pennies from a collection of 100 pennies. Instead, it suffices to select one penny with replacement from a collection of one penny or a collection of two items—a head and a tail, and to repeat this experiment 10 times and record the 10 outcomes. This, in turn, is like tossing a fair coin 10 times and recording the outcome of each toss.

---

[1] In a class discussion of this problem, students ended up constructing a counter-example that highlighted the distinction between a group's ranking and an individual person's ranking.

[2] This conflation between the quantification of two entities that need to be distinguished when imagining the re-sampling process emerged among other students in subsequent phases of the experiment. In later chapters I provide evidence that this proved to be a robust conceptual difficulty for students, despite concerted instructional attempts to help them disentangle these two levels.

[3] Unfortunately, computers were unavailable for students to use in the classroom before the 8th lesson. Thus, students did not have hands-on experience working with Prob Sim before this late date in the experiment. Instead, the instructor led class discussions around the Prob Sim interface that was projected on a screen in the classroom. He made changes to the program's parameter settings according to relevant ideas that emerged in these discussions and he engaged students in making sense of these settings and explicating their relationship to the situation being simulated.

A central issue that emerged in these preliminary discussions was that simulating a situation is not the same as replicating it. Whereas replication focuses on repeating the physical actions that are described in a situation, simulation requires looking beyond a situation's surface characteristics and distinguishing aspects of it that are important from those that are unimportant for designing a simulation of it. An overarching instructional aim of these preliminary discussions was to orient students to the idea that designing a simulation entails conceiving of the situation to be simulated in terms of a repeatable process that, when simulated, will produce results *as if* the described situation were actually enacted.

In Phase 2, students engaged in two clusters of simulation design activities that I will refer to as Activity 4 and Activity 5, respectively. Each activity was comprised of several situations described in non-statistical terms, and each situation raised an issue about an event of interest. The activities unfolded as whole-class discussions directed at having students think and describe how they would use Prob Sim to investigate and resolve these issues. After the situations in Activity 4 had been "unpacked" in such discussions, students were asked to consider the situations in Activity 5 on their own as a take-home assignment. The research team expected that the issues raised in the whole-class discussions centering on Activity 4 would be relevant for students when they attempted the take-home task.

Though the situations differed from one another, the structure of the tasks was common to all the situations in Activity 4 and Activity 5. Students were asked to:

1. Think through the assumptions for their investigation;
2. Plan the method of their investigation;
3. Implement the investigation (i.e., run the simulation of an experiment) and describe its result;
4. Draw a conclusion about the issue raised in the situation on the basis of the result of their investigation.

Because of this similarity between the two activities, this chapter does not give a full account of each activity and the discussions that emerged from it, as in the previous chapter. Instead, the chapter highlights selected issues that emerged within these discussions and provides illustrative evidence from engagement across the situations. Toward that end, the next section describes two situations from Activity 4 up front. The narrative in subsequent sections then refers back to these descriptions as needed.

Activity 4 Situations

The first situation that students considered in this activity was adapted from a problem often cited in the statistical reasoning research literature—the "birthday problem" (Figure 6.2). The birthday problem has been used as a context for estimating the probability of finding two people having the same birthday in a group of people.

Investigating "Unusualness"

Jill was in a class of 25 people. The teacher asked each student for his or her birthday to enter in the class calendar. The class was surprised when two people had the same birthday! Is this, in fact, an unusual event?

"Gut level" answer:

Assumptions for your investigation:

Method of investigation:

Result:

Conclusion:

Figure 6.2. Written guide for investigating the Situation 1 of Activity 4: the birthday problem.

As Konold noted (ibid., 2002), people are often surprised to learn that the probability of this event is relatively high. The task set for students in this study was to design an investigation to determine whether it is an unusual event to find two students in a class of 25 who share a birthday. Classroom discussions centering on this situation unfolded over the last 12 minutes of Lesson 5 and then again over the first 14 minutes of Lesson 6 (the next day).

The third and final situation that students considered (Figure 6.3), another oft-cited problem in the statistical reasoning literature, is similar to the first situation. The "movie theater problem" was adapted from Konold (2002). It asked students to design an investigation to determine whether it is unusual to see two or more acquaintances in a movie theater in a small town. Discussions centering on this situation occurred during the last 12 minutes of Lesson 6.

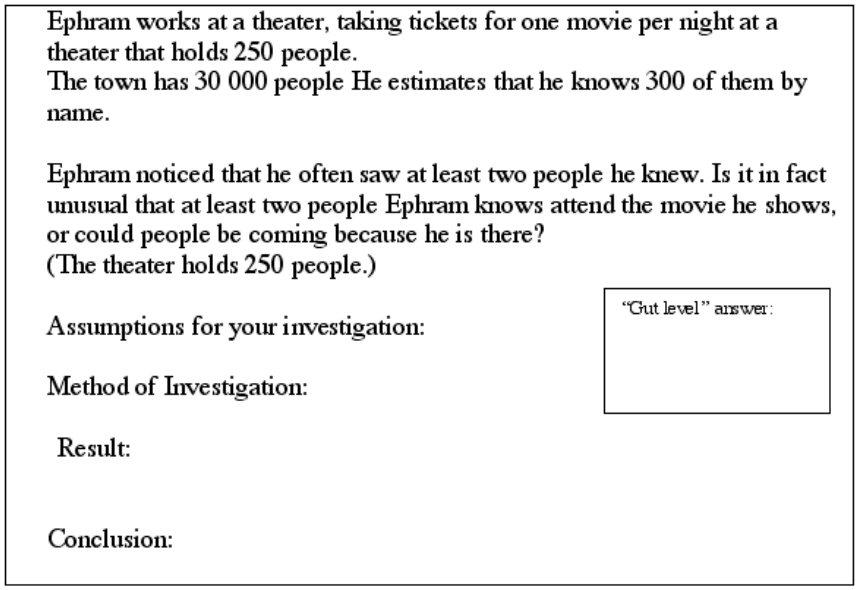


Figure 6.3. Written guide for investigating Situation 3 of Activity 4: the movie theater problem.

Conceiving expectation as a quantity

A significant difficulty that many students experienced in these tasks was to think about and describe expectation in quantitative terms. For instance, after spending about one minute reading the first situation (Figure 6.2), students had no questions about it. The instructor then began the discussion by having students interpret what it means to say that an event is unusual. With the instructor's help, students eventually offered interpretations that seemed to touch on ideas of surprise and expectation. However, students' descriptions generally suggest that their ideas were not quantitative and were focused on individual occurrences not embedded within a sequence of trials of a repeatable experiment. This is nicely illustrated in the following excerpt drawn from a 3-minute-long discussion during Lesson 5.

Episode 1, Lesson 5 (Situation 1):

1. I: […] So what does it mean for an even to be unusual?
2. Peter: You're not expecting it.
3. Kit: You don't expect it.
4. I: Yeah, it's unexpected, that in a large number of times that you do this (2 second pause) you, you, you expect to see it rarely. Ok? (2 second pause) All right, so in this particular case what does it mean that, to wonder if, what the event that's described is unusual?
5. Peter: You wouldn't expect 2 people in a cla—in a group of 25 people to have the same birthday.
6. I: Ok. Now, you're leaving out something. That's good as far as you went, but you're leaving something out.

[…]

7. I: Are we talking, when we talk about an event being unusual, are talking about just <u>one occurrence</u>?

(4-second silence)

8. I: Are we just talking about <u>that</u> class?
9. Luke: No. It's unusual--as many times as you did this test with the class, it would be unusual for 2 people to have the same birthday.
10. I: So you wouldn't expect it to happen very often.
11. Luke: Correct.
12. I: But the idea, the, the part that you, uhh both of you left out, uhh Peter and Nicole, was if you looked at a whole bunch of classes, of size 25, you wouldn't find very many. You see? You left out the part of looking at a whole bunch of classes. All right? And (2 second pause), and that's, that's, that's a key idea. The idea that what we're talking about is doing something a large number if times, looking at a class of 25 students. Ok?

As the excerpt indicates, ideas of (relative) frequency were largely absent from students' discourse. Students' descriptions suggest that they were interpreting the situation as it is described rather than re-conceiving and re-describing it in terms of a repeatable experiment that might support thinking about how to quantify expectation. It was the instructor who oriented the class to quantify expectation by explicitly pointing out what idea students were leaving out of their descriptions and by couching unusualness in terms of infrequency and re-describing the situation in terms of a repetitive process. Even Luke's attempt (line 9) to incorporate the idea of frequency into his way of thinking is tenuous, as suggested by the circularity of his description.

The instructor began the next lesson (6) by once again raising the idea of unusualness, soliciting students' thinking about it before discussing the other situations. The following brief excerpt is drawn from the 3 minute-long opening discussion of Lesson 6.

Episode 1, Lesson 6:

1. I: Uhh, in last class we looked at how, how, ways in which we could investigate whether or not something was unusual. All right? Now, what did it m—what did we mean by saying that something was unusual?
2. Luke: It's happens less than 50% of the time.
3. I: Or uhh, was that it? I mean, if it happened 49% of, like, is it unusual to get a tail if we toss a head?
4. Michelle: Unexpected
5. Luke: Yeah, it's unexpected.
6. I: All right, it's unexpected. And how, how would you quantify that?
7. Michelle: How would we what?
8. I: How would you quantify that, that it's unexpected?
9. Nicole: I mean

10. David: It doesn't usually happen.
11. I: That's not quantifying it. That's putting it—what, what does that mean –?
12. Luke: Rephrase it, David
13. I (continues): in terms of—see, quantifying unusualness saying, is saying "ok, if it happens a certain fraction of the time or less, then it's unusual". (2-second pause) Now, half the time or less, I don't think most people would call that unusual.

The excerpt illustrates, once again, how disinclined students were to thinking of expectation in quantitative terms. Indeed, students seemed stymied when pressed to explicate how they would quantify expectation (lines 6-11), as though they barely understood the question and could only appeal to an intuition or a feeling that an outcome was unexpected.

Luke's response (line 2 of Episode 1, Lesson 6) is interesting because it appears to center on the idea of relative frequency. However his focus was on a numerical value what might constitute a conventionally accepted cut-off for deciding that an event is unusual. This need not imply that Luke was mindful of expectation as a quantity, more generally.[4] Also of note, is that Luke's response is devoid of a description of the process by which one might arrive at a choice of 50% as a proportion below which we agree to call an occurrence "unusual".

In my view, Luke's was more a response to the question "below what cut-off proportion do we call an event 'unusual'?" than to "how can we think about an event so that we can determine how 'unusual' it might be". To clarify, I am not arguing that Luke misheard and incorrectly answered the instructor's question. Rather, I am arguing that Luke's response can be taken as indicative of his experiencing difficulty in conceptualizing the underlying situation in terms that support his conceiving expectation as a quantity.

Responses like David's "it doesn't usually happen" (line 10 of Episode 1, Lesson 6) amount to a non-quantitative rephrasing of "it's unexpected"; such responses were not uncommon in the earlier phases of the experiment. In a very real sense, the students and the instructor were speaking different languages: the former were living in the realm of feeling and intuition, while the latter was trying to map those intuitions onto an "objective" relative frequency perspective (Kahneman & Tversky, 1972).

---

[4] The distinction is akin to that drawn by Thompson (Thompson, 1994; Thompson & Saldanha, 2003) between a quantity and quantity's values. The use of numerical discourse need not imply being mindful of an underlying quantification process, which should entail seeing numerical values as expressions of a conceptualized object's attribute relative to some measuring unit.

Getting students to assume the latter perspective proved to be an instructional challenge, as their attempts to transition from the former to the latter seemed fraught with difficulties. The next episode is divided into 3 ordered segments, the last two of which are contiguous. As a whole, they illustrate both the instructor's and students' challenges. The episode, lasting approximately 4.5 minutes, is drawn from discussions that unfolded around the movie theater problem (Figure 6.3) during the later part of Lesson 6.

Episode 2, Lesson 6 (Situation 3):

Segment 1

400. I: Let's all make sure that we know what's going on. What is, what is it that's at issue?
401. (Peter and Lesley chatter inaudibly in the background)

(7-second silence)

402. I: Kit?
403. Kit: Whether or not it's unusual for him to see at least 2 people that he knows.
404. I: And what does it mean, what does "unusual" mean?
405. Kit: Not expected.
406. I: Ok. Go on and quantify that.

(5-second silence)

407. I: It means if he were, if he were to do this many many times he would expect some small fraction of the time for this to happen, to see—see, keep, you gotta, I want you to keep putting this idea of repeating an event over and over and over again. (3 second pause). Ok, it's not a matter of feeling like "gosh, I don't expect this to happen". That's not where likelihood is determined! Likelihood is determined in the actual repeating. Not your <u>feeling</u> about it, but rather the repetition of the event and the fraction of the time that something happens in those repetitions. Ok? So now, Kit, once more: what is, what does it mean to be unexpected?
408. Kit: uhh a small fraction of the time, when it's done several times.
409. I: Ok, and what is the <u>it</u> in this case?
410. Kit: Uhh, the people—uhh seeing more than 2 people or 2 people one time per night, that he knows.
411. I: Ok, so over many many nights, assuming he's there just once a night, over many many many nights, we have a small fraction of those nights where he sees 2 or more people that he knows. Now, do you see how that quantifies, brings, brings ideas of quantity into it? It's no longer just a feeling that he has about expectation. It's rather, we're talking about repeating something many many times and just looking at the fraction of the time that something happens! (3 second pause) Now you'll, it'll—it'll get so that this is second nature to you to start thinking this way. And you'll also wonder how you could've thought any other way. But it takes practice, and I, that's why I keep insisting that you bring this idea out in the open.

Segment 1 begins with Kit offering her interpretation of the issue raised in the situation. Kit recognized that the issue is whether the event in question is unusual, but her sense of unusualness seemed non-quantitative. When asked to quantify "unexpected" she offered no response, which suggests that quantifying expectation was still problematic for her. The instructor's response was to try and highlight the difference between a mere feeling that something is unexpected and thinking about expectation quantitatively, in terms of a process that can be repeated many times and by considering the fraction of the time that the event of interest is observed.

Kit's subsequent attempt to describe "unexpected" in quantitative terms (line 408) indicates that the idea of "a small fraction of the time" was salient for here. However, her description also suggests that she was unclear on the distinction between the repeated process and the event of interest (line 410), as though she thought that observing the event of interest repeatedly was the repeated process.[5] Thus, it would seems that, at that point, Kit was mindful of two images—repeating a process and observing the event of interest—but was not fitting them together to think of this scenario: *imagine having repeated the process of looking in the movie theater once per night, on many nights, and counting the fraction of those nights on which you saw at least two acquaintances.*

The segment concludes with the instructor describing this scenario and reiterating how this thinking quantifies expectation. Moreover, he tried to explicitly institute this way of thinking and describing as normative in the classroom and as requiring practice.

The second segment continues the same discussion, after a brief diversion, and illustrates another student's particularly robust difficulty:

Episode 2, Lesson 6 (Situation 3):

Segment 2
412. I: […] What does it mean, what is the event that we're talking about?
413. Nicole: Him seeing 2 people he knows at the movie.
[…]
414. I: What does it mean for that to be unusual?
415. Nicole: That, that's what they're trying to figure out, if it's unusual! I thought.

[5] This confusion is understandable if one expects the event of interest to be observed numerous times, since then there are two kinds of repetitive events to consider: 1) looking in the theater each night, and 2) observing the event of interest. Thinking of the event's expectation as a quantity entails distinguishing and coordinating these images. An inability to do this might be reflected in the way Kit expressed herself.

416. I: All right, but what does it mean? What are we trying, what is it that we're trying to figure out if it's unusual? What would that mean, that it in fact turns out to be unusual?
417. Nicole: That he actually doesn't see at least 2 people he knows every night.
(4-second silence)
418. I: Say that again?
419. Nicole: For to, if it was unusual for him to see 2 people he knows every night then it—wait (bursts into laughter at her own confusion)
420. (Others join in laughter and tension is broken)
421. Peter (to Nicole): You did great!
[…]
422. Peter: No, you're saying it, but you're not saying it all. You're saying, like, half of it. (4 second pause) Kind of—
423. Luke: Uhh (inaudible)
424. Nicole: I don't understand the question.

Nicole was clear on the issue raised in the situation. However, Nicole was confused by the instructor's meta-question "what does it mean to say that the event in question is unusual?". She responded to his call to operationalize unusualness as if she understood him to be asking what they were investigating in the activity (line 415 of Segment 2). In line 417, Nicole seemed to register a different question and attempted to incorporate a notion of frequency into her explanation; her response suggests that she was over-generalizing in thinking that the event is unusual if it does not occur *every* night. Nicole eventually abandoned her attempt and admitted not understanding the question after she and the other students became conscious of her confusion.

The final segment of the episode illustrates another student's struggle to forge ahead with quantifying unusualness.

Episode 2, Lesson 6 (Situation 3):

Segment 3
425. I: All right, go ahead, Luke.
426. Luke: In a collection of nights it would be unusual if a majority or uhh, it didn't, it would be unusual if most of the time he'd see, or 2 in (inaudible)
427. I: Ok, so you, there you're thinking suppose that we've got, we're looking back at the past month.
428. Luke: Right.
429. I (continues): We're looking back at 30 nights. And so it would be unusual, and so then you could sort of check them off: "saw at least two people", skip, skip, skip "saw at least two people" (gestures as though moving along days in a calendar and making a check mark), skip, skip. So out of those 30, it would

be unusual for him to see uhh two or more people if, if what was true about those 30 nights?

(3-second silence)

430. Luke: That he saw two or more people?
431. I: Yeah. No, what--?
432. (Kit, Lesley, I and others laugh at the apparent confusion)
433. Luke: I didn't follow your question.
434. I: All right. We've got those 30 nights. We go along and we check, we've got, like, a board that's numbered 1 through 30
435. Luke: a calendar!
436. I (continues): and we check those nights. Every night that he sees two or more people, he puts a check (motions with hand as though making a check mark).
437. Luke: All right.
438. I: All right. So what would it then mean about those 30 nights that it's unusual for him to see at least 2 people?
439. Luke: Most of the time he didn't see two or more people.
440. I: Yeah, or most of those nights aren't checked.

Luke started off (line 426 of Segment 3) by referring to a "collection of nights", which suggests his having had an image of collecting past data on which he would then base an assessment of unusualness. But Luke wasn't quite able to coherently describe what he would look for in those past nights; he seemed to confound "unusualness" with a tenuous description that appeared to be more consistent with a meaning for "usualness". The instructor offered support (lines 427-429) by describing a scenario intended to help Luke imagine what he would keep track of in those past nights. Luke's response, in line 430, however, suggests that he was indeed confounding unusualness with usualness. Eventually, in line 439, Luke was able to get things straight; his response suggesting that he understood "unusual" to mean that on most of the nights Ephram did not see two or more people he knew at the movie theater. We see that this response emerged in the context of some very heavy instructional scaffolding.

To summarize, Episodes 1 and 2 of Lesson 6 demonstrate how difficult it was for students to think of and describe expectation and related ideas of likelihood and unusualness in quantitative terms. These difficulties were not merely isolated or fleeting instances attributed to students' lack of linguistic resources for constructing precise descriptions. Rather, I assert that these difficulties were both pervasive and conceptually robust. Further, the interactions within which these difficulties emerged were not students' first encounter with these ideas; they had previously engaged in a protracted discussion-based activity (Activity 2) in which they helped invent and

implement a method for investigating whether an outcome is unusual. The current episodes suggest that students had not reflectively abstracted from that concrete experience the essence of their method to operationalize expectation as a quantity.

I should also point out that students' concerted efforts to think of and describe expectation as a quantity were not spontaneous. Rather, these efforts emerged out of interactions with an instructor who provided heavy prompting and scaffolding. The episodes considered here also exemplify the instructor's attempts, throughout the entire experiment, to have students describe likelihood, expectation, and unusualness, in terms intended to support thinking about them quantitatively—by appealing to ideas of repeating an experiment and calculating a relative frequency.

A common strategy employed by the instructor was to re-describe situations in terms of these ideas in an effort to provide students with an imagined process to which they might relate and imagine themselves. He also tackled the problem on a social plane by attempting to make it explicitly normative to always think about these ideas and employ this discourse in the classroom. This led to a normative style of classroom engagement and interaction that was sometimes uncomfortable for students. The discomfort seemed rooted in the fact that the interactions entailed having students become aware of the limits of their own thinking and understandings in a public forum. Another impediment to students' engagement in such discussions emerged later in the experiment; students seemed to grow tired of a style of engagement that forced them to describe these ideas carefully and precisely. They eventually began resisting participating in such interactions, as though they didn't appreciate their importance and thought they were just being held to a very pedantic standard of description. Yet at the same time, students would run into problems in situations in which it would have been powerful to reason quantitatively about these ideas.

In sum, this classroom community's move toward conceiving situations in ways that might support reasoning quantitatively about likelihood, expectation, and unusualness was an ongoing challenge both conceptually, for students, and instructionally, for the instructor. Results of an assessment item given to students at the end of Phase 2 indicate that about half of them were able to give a (quasi)-quantitative characterization of what it means to say that an event is unusual. Table 6.2 displays students' responses to this question:

"What does it mean, in statistics, that an event is "unusual"? (We know that *unlikely*, *unexpected*, and *rare* are synonyms of unusual, so mentioning them will not answer the question. Please explain the meaning, don't just give synonyms.)"

Table 6.2. Students' responses to an assessment item given at the end of Phase 2.

Student	**Response**[6]
Nicole	An unusual event would be one the most unlikely to occur. (I.E.- In 3-card poker it is unusual to get a 3 of a kind.) It's that something that occurs that wasn't predicted.
Sue	That mean there is a only few percentage of event occur during the longrun collection of samples.
Kit	When something happens a small % of the time.
Sarah	Of all the samples taken, or items tried, the usual occurance happens the least or close to the least
Peter	That an event is not likely to happen. If 1000 samples are taken and unusual event will happen about 5% or less of the time.
Tina	It means that., How unusual is it for this to occur? That it does not occur/show up as often
David	The event unusual means in statistics that there is a lower percentage ["chance" inserted here] that something unusual will occur. Like something unusual will probably only occur 10% of the time.
Luke	In statistics the word usually relates to the occurance that something occurs. When someone says that the results are unusual, then they mean that the results don't come out like this on a common occurance.

Conceptualizing scenarios as probabilistic situations

The activity of designing simulations with the structural constraints of Prob Sim in mind forced students to think of unusualness, expectation, and likelihood as quantities. Indeed, that entailment was part of the underlying instructional rationale for the activity in the first place. Conceiving of expectation as a quantity, however, was not the only problematic aspect of the activity for students. The discussions in Activity 4 brought to light other difficulties that students experienced in re-conceiving a described situation as a *probabilistic situation*—that is, as an idealized stochastic experiment.[7] These difficulties can be interpreted as problems of constructing a mathematical model. The tasks entailed construing and re-describing the given situations in terms of idealized assumptions, in terms of a population, a sample, and a random

[6] Tina, Kit, and Peter gave the clearest quantitative characterizations of *unusual*, making explicit reference to the idea of frequency. David's response is circular and therefore less clear, similarly with Luke's response. Sarah and Tina's reponses are perhaps quasi-quantitative because they only implicitly refer to the idea of frequency. Nicole's characterization makes no allusion to frequency.

[7] I do not mean to imply that these difficulties were separate and unrelated. On the contrary, I see them as arising in an interrelated way within the same activity of trying to construe and re-describe the task situations. However, I attend to these difficulties separately in my analyses for the sake of clarity.

sampling process. I emphasize again that the situations per se, as they were actually presented to students, did not describe such aspects and relations. Rather, students had to learn to construe situations in those terms, a process that entailed re-configuring and creatively interpreting information given in the situations. This turned out to be a significant challenge for most students; their progress was slow and tightly embedded within their interactions with the instructor in the classroom discussions.

The next discussion episode is drawn from classroom discussions centering on explicating the assumptions for simulating Situation 3 (Figure 6.3). The episode, lasting approximately 4.5 minutes, is comprised of 3 contiguous segments. The first segment illustrates David's sense of overwhelm by the possibilities for assumptions.

Episode 3, Lesson 6 (Situation 3):

Segment 1

428. David: I didn't get this question 'cause it, there are so many different things that could happen. Like, what if only half the town goes to see movies? Or uhh what if it's the same 2 people every night, that he sees? It says he knows 300, but couldn't, like, the same 2 people go see the movie every night?
429. Nicole: Yeah
430. I: Sure, that's right. So that's where you lay—
431. Peter (to David): Good thinking!
432. I (continues): you settle all of this in your assumptions. Like, one of the assumptions that you have to make in order to look at this in the abstract, without actually knowing him and the town, is that it's a random process by which the verand--theatre gets filled every week. (3 second pause) Now, it may not in fact be! But in, that's an assumption that you could make that will let you proceed.
433. David: Oh, ok.
434. Sarah: You also have to assume that he sees everyone that goes to the movie.
435. I: Very good! Because if he only sees a small fraction of the people going in, people could be there and he might not see them. (3 second pause) All right. So we're not saying he does, but we're saying in order to proceed we'll make this assumption. Ok, all right. Does that make sense, David?
436. David: Yeah.

Segment 1 suggests that David was disabled from deciding what to assume because he felt lost in a sea of possible choices. It seems that David may have thought of an assumption as a hard fact about the situation, rather than a working supposition—a reasonable hypothesis—upon which to proceed further. Thus, David's difficulty appeared to be in looking beyond the information given in the situation and reconfiguring the situation in terms of aspects that are not

explicitly given per se but which are nonetheless necessary to presume. The need to reason hypothetically about a situation as a starting point for designing an investigation of an issue, together with the absence of clear constraints on what could be hypothesized, made the tasks seem too open-ended and ambiguous to some students. The classroom discussions were intended to help students learn to deal with such ambiguity by providing them with opportunities to unpack their implicit assumptions (i.e., line 433) and create new assumptions.

The class's difficulties in deciding what to assume about the given situations were ongoing in these discussions. Decisions were rarely made in a clear-cut manner, instead they often emerged out of relatively arduous negotiations embedded within messy interactions. The next two segments of Episode 3 illustrate this. Segment 2 begins with David struggling to make sense of the underlined part of this statement given in the movie theater situation (Figure 6.3): "*Is it in fact unusual that at least two people Ephram knows attend the movie he shows, <u>or could people be coming because he is there</u>?*".

Episode 3, Lesson 6 (Situation 3):

Segment 2

482. David: Why did you throw in that last part that says "or could people be coming because he is there?". Why did you put that part? That was, that wigged me out (motions with hands above head), I didn't know what to do. It's, like, what is that?
483. I: Oh! Well—
484. David: It says (reads) "or could people be coming just because he is there?"
485. Nicole: Yeah! That's my point!
486. I: Or for some other reason or another.
487. David: Yeah. I was, like, what is <u>that</u>?
488. I: Well, if he always saw--
489. Peter: We have to assume that they're not?
490. I (continues): if he always saw 30 people that he knows—a tenth of the people that he knows in this town are there every night—then something's going on, right? (2 second pause). That, I mean (coughs)

[…]

491. David: Yeah, maybe he's sneaking them in for free
492. I: Perhaps. Something's going on (turns on laptop and window re-appears on screen). Would you expect him to see very many people that he knows? If he knows, if there are 30,000 people and it's a random draw to fill the theatre, would you expect him to see very many people that he knows?
493. Nicole: Well, how many movies are there a night?
494. David: He only knows, like, 1% of the town, so it's kind of weird that he'd see people, 2 people every single night.
495. Luke: Yeah.

496. I: Yeah, he knows 1% of the town.
497. I: Well, we're gonna simulate it, we don't know the answer to the question yet!
498. Kit (to David): I think there's one movie per night.
499. I (continues): It might be rare

The, evidently problematic, statement was intended to occasion reflection on the reasonableness of the randomness assumption in the case that the event in question turned out to be statistically unusual. So that, if the event turned out to be unusual, issues could then be raised about whether something other than a random attendance process was at play in the scenario. David and Nicole (lines 482-487) were not, however, interpreting things this way. All they seemed to see was an isolated question which made little sense to them and which they could not relate to the greater task. In retrospect, their difficulty is little surprising, as they were, in a sense, putting the cart before the horse. That is, because the class had not yet investigated whether the event in question might be unusual, these students could not interpret this statement as broaching a possibly relevant issue. The tension that these students experienced drove the instructor to start bringing issues of underlying assumptions out into the discussion.

In the ensuing interaction, David expressed his "gut level feeling" that the event in question—seeing two or more people at the movie theater each night—is inconsistent with the given assumption that Ephram knows only 1% of the 30,000 people in the town. Thus, David's intuition seemed to touch on the idea of drawing an unrepresentative sample from an underlying population. Moreover, David's intuition also suggests that he was mindful of the 300 acquaintances as a proportion of the entire population, but he did not seem to reason similarly about a sample to the think of the number of acquaintances as a proportion of 250 people selected. Had he done so and noticed that, say, two or three acquaintances out of 250 is close to 1%, he might have been less inclined to find such a result surprising.[8]

Nicole (line 493) also began entertaining possibilities for underlying assumptions: how many movies were they to assume were shown each night?—as though this had important implications

[8] David's intuition is inconsistent with what sampling theory would lead us to expect—that the distribution of sample proportions for samples drawn randomly from a population at least ten times as large will cluster tightly around the population proportion's value. Thus, samples of size n = 250 drawn from the population of 30,000 will tend to be representative of the sampled population. So it should not be unusual for Ephram to see two acquaintances in the theater, assuming random attendance. Moreover, the sample proportion 2/250 is very close to the sampled population proportion (i.e., we can expect this sample proportion to occur relatively frequently).

for how many acquaintances Ephram could reasonably expect to see at the movies each night. Thus, Nicole had not yet structured the situation in terms of an unambiguous sampling experiment in which the act of looking into a filled 250-person theater once per night is *as if* one were collecting a 250-person random sample from the town's population (assuming random attendance).

In the third segment of Episode 3 of Lesson 6 the instructor moved to engage students in choosing values for the Prob Sim parameters in order to model Situation 3. The discussion thus turned to making explicit connections between the situation and an idealized sampling experiment constrained by considerations relevant for using the software.

Episode 3, Lesson 6 (Situation 3):

Segment 3

504. I: Now, I'm going to, I'm gonna do this in a way that uhh the guy who wrote this program suggested (starts setting up Prob Sim to simulate Situation 3). Put a little tiny dot to represent a person in the town who he doesn't know, and a big dot to represent a person in the town that he does know. How many of these dots are there gonna be in this mixer? (points to small dot in left-most element label on screen, see Figure 6.3)
505. Luke: 30,000
506. I: No
507. (Others students chime in unanimously and simultaneously): "27,000".
508. I: Yeah, 27,000. No, twenty nine thousand seven hundred.
509. Peter: 29,700
510. (I enters this value into "how many" slot under first element label)
511. Peter (to others): Mathematical geniuses!
512. (Peter and Nicole laugh in background)
513. I: So how many people does—uhh that's because he knows 300 of those 30,000. Right?
514. Luke: Nods. He knows (inaudible) thousand.

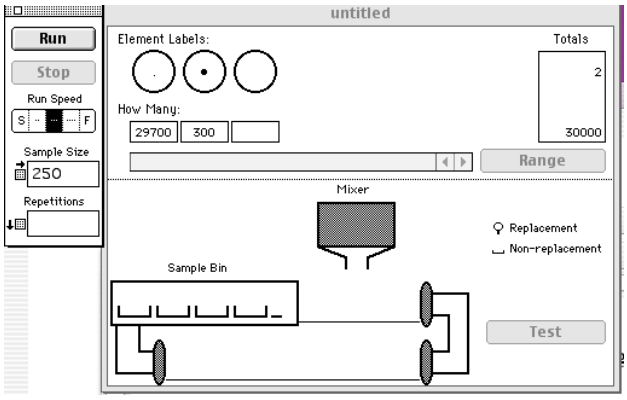


Figure 6.4. Prob Sim window set up for simulating Situation 3.

[…]
515. I: Ok, we're gonna take 250 of those people. Right? (sets this value as sample size in window on screen) Are we taking them with or without replacement?
516. Lesley: Without
517. Peter: With, with!
518. Luke: No, without replacement.
519. Lesley: With!
520. Nicole: No, you can't, it
521. Luke: You can't—
522. Nicole: If it's people it has to be—without
523. Peter: 'Cause they can come back the next night.
524. I: No, no, we're talking about one night.
525. Kit (to Peter): Yeah, but not on the same night.
526. Peter: Oh!
527. Luke: Repetitions is (inaudible).
528. I: One night. So, is it with replacement or without replacement? (has cursor pointing at "replacement" option in Prob Sim window on screen)
529. Nicole: With
530. Lesley: With
531. Luke: Without replacement.
532. Kit: Without
533. I: If a—can a person be in a theatre twice?
534. Kit: No.
535. Luke: No.
536. Lesley: No.
537. I: Ok, so it's without replacement
538. Peter: They snuck back in and watched it again.
539. Nicole: Wait
540. Kit (to Peter): not at the same time you can't
541. David (to I): You got it on without replacement
542. Nicole: it's one movie one night
543. I: Yeah.
544. Lesley: I don't understand.
545. I: Or, it's just one night. We don't know how many movies, but--
546. Nicole: Well then it's a difference! That's what I asked you
547. I: Ok, then let's say one movie one night. That's a good assumption. (2 second pause)

Segment 3 began with the instructor deciding to represent Ephram's acquaintances and non-acquaintances in the town with a small and a large dot, respectively. After entering these symbols into the program's element labels to represent the population items (see Figure 6.4), Luke proposed that the Mixer—the software's metaphor for a population—should contain 30,000 of the small dots. This suggests that in this context Luke had not conceived of the

population as comprised of two distinct classes: acquaintances and non-acquaintance. The instructor and other students immediately chimed in with a different answer (lines 506-509). After an estimation error was resolved, the instructor explained to Luke that the population is divided into 300 acquaintances out of 30,000 total people, to justify choosing 29,700 as the number of small dots in the mixer.

The discussion then turned to setting the sampling parameters. The instructor proposed that sample size should be 250. There were no objections to this and the issue then became whether to sample with or without replacement. Here there were different opinions, but little direct evidence of how students were thinking about the issue.[9] However, it seems that the source of the differences can be attributed to students having had different assumptions about how many movies are shown each night. Peter said enough to suggest some insights about his thinking. In line 523, Peter appeared to assume that people could return to the movie theater on different nights. This assumption suggests that Peter had not structured the situation in terms of what might occur on an individual night as a distinct unit. Rather, he seemed to be considering events that could occur across several nights. Peter's utterance in line 538 indicates that even when restricting himself to considering an individual night, he was thinking of contingencies that suggest he was unclear on *what,* in the situation, constituted a sample. "Could one sample be like one full audience, or should it be thought of as all audiences in one night?" These seemed to be the kinds of underlying questions that Peter was contemplating without resolution.

Peter's problem can be interpreted as a difficulty in making an idealized assumption about the situation. Reformulating a situation in terms of idealized assumptions is a hallmark of the mathematical modeling process. This entails choosing to ignore certain contingencies that, while perhaps realistically plausible, may make the situation too unwieldy to model. One has to consciously make discerning choices about how to "tame" and simplify a situation so as to make it amenable to model. Peter's comments presumably reflect an underlying difficulty in making

---

[9] I should mention that the distinction between sampling with and without replacement had been explicated in classroom discussions prior to this one, most notably in the preliminary discussions of this phase of the experiment. Those discussions established that sampling with replacement refers to replacing an individual sampled item before selecting another one from the population, and not the replacement of an entire sample of items. It was important to make this distinction because the simulations were of repeating a sampling experiment and this raised the possibility of confusing the unit of replacement. It was stressed that in a simulation of a repeated experiment one could think of the individual samples as always being replaced and that sampling with replacement does not refer to that, but rather to what we do with individual sampled items that aggregate to form a sample.

such judicious choices. Evidently, for Peter the constraints of Prob Sim were not enough to help him structure the situation as an unambiguous sampling experiment. The metaphor of population as Mixer may have been relatively transparent, but conceptualizing the situation in terms of a clear-cut sample, the selection of which he might imagine repeating, was problematic.

Although the best available evidence is of Peter's thinking, Segment 3 of Episode 3 illustrates that the group was generally indecisive about whether to sample with or without replacement. Those students who changed their minds (Lesley and Nicole), flitting from one sampling option to the other, were evidently unsettled on what to take as a sample in the situation; their assumptions were still formative and highly unstable. Eventually, the instructor and Nicole negotiated the assumption that a sample should be a single movie in a single night (lines 542-547)—a simplification consistent with sampling without replacement.

Episode 3 of Lesson 6 highlights the most problematic aspects of the activity with regard to Situation 3. Once the sampling parameters were settled, the activity unfolded in much the same way as Activity 2 (Chapter V) and without any apparent difficulties for students. Briefly, the instructor ran ten Prob Sim simulations of the experiment, one at a time. Each time the program displayed the results as in Figure 6.5, and students interpreted this as showing the seating of people in the movie theater, with the number of large dots representing the number of Ephram's acquaintances seen each night. Students voted "yes" or "no", after each display, as to whether Ephram saw two or more people. In seven of the ten iterations, the class voted "yes" and thus concluded that the event in question was not at all unusual.

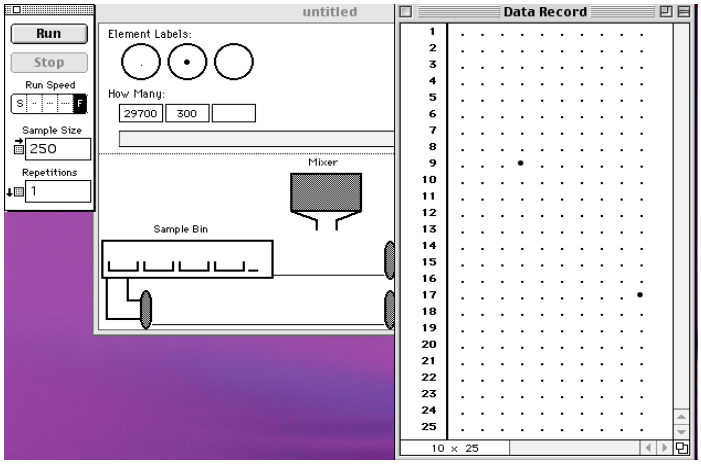


Figure 6.5. Prob Sim output displayed after each repetition of the simulated sampling experiment.

In sum, the discussion excerpts presented in this section suggest that construing the situations as statistical experiments can be a highlight non-trivial activity. Even under conditions of heavy instructional guidance, entailing a supportive environment in which it was normative to unpack ideas and the use of software constraints intended to structure the activity in productive ways, students experienced considerable difficulties. These difficulties extended to several aspects and various levels of detail of situations, but particularly to making discriminating decisions about underlying assumptions. The connections between a situation *as it is described* and what might be seen as its construal and re-description as a stochastic experiment was not transparent for students.

Activity 5 Situations

Though the discussions of Activity 4 clearly evidenced the difficulties that students experienced, the gravity of these difficulties did not become fully evident to the research team until the next activity. Activity 5 asked students to again design Prob Sim simulations of situations in order to investigate issues that needed resolution. However, students were now asked to do this independently as a take-home assignment, outside of tightly scaffolded instructional interactions. The research team assumed that students' engagement in the fairly protracted discussions of Activity 4 had enabled them to deal somewhat productively and independently with similar issues outside of an environment that provided direct instructional support. This assumption proved to be a *pre*sumption; it turned out that students were generally quite unable to deal with these supposedly similar tasks on their own. Beyond recognizing that they were being asked to answer similar questions, students appeared to draw few substantive connections between the *issues* raised in Activity 4 and how they might engage with the tasks in Activity 5. Paradoxically, compelling evidence of this is found in the almost total lack of student written data, since students were largely unable to even begin to attempt the tasks.[10] Figure 6.6 shows two of the Activity 5 tasks that students found to be insurmountable.

[10] Only 1 of 8 students was able to attempt Activity 5 along the lines of the requested task structure (as elaborated in the preliminary section of this chapter).

Investigations (again)

Use Prob Sim to investigate these situations. Use the blank screens as worksheets as you think through all the decisions and assumptions you need to make in order to have Prob Sim simulate the situation accurately.

Problem 1.

In all poker-like card games, the "most unusual" hand is the winner (i.e., the hand least likely to occur over the long run).

Horace, Hillary and friends were playing cards -- "guts". Guts is a game where people are dealt 3 cards each, and when the dealer says "1, 2, 3, drop." People thinking their hands aren't good enough to win drop. All the people thinking they have a good enough hand will keep their cards.

Horace and Hillary remembered that, in 5-card poker, a flush (all cards the same suit) beats three of a kind (three cards the same face value). But they didn't know if that should be the case in 3-card poker.

Use Prob Sim to resolve this question.

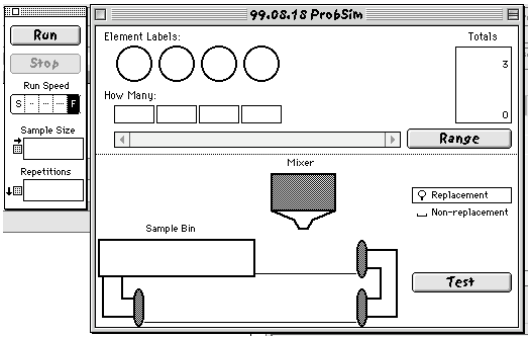


Figure 6.6. Two tasks from Activity 5.

Problem 3.

The Gallup company asked 21 adults, selected at random, whether they attended church or synagogue during the past week. Fifty percent said they had. The Harris company performed an identical survey. In their survey, only 33% responded "yes".

Use Prob Sim to investigate whether the two polls contradict one another.

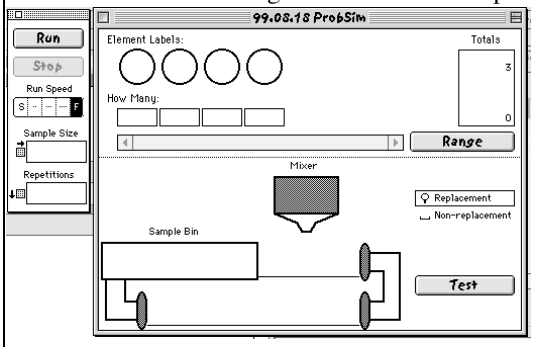


Figure 6.6.--continued.

In addition to the very real conceptual problems that students experienced with the reconstrual issues already discussed, my impression is that their difficulties were also exacerbated by a kind of "cultural tension" that they experienced.[11] This tension was between the norms of engagement in this class and the expectations that they had internalized from years of past experience in school mathematics. Students seemed to lack a "readiness" in their dispositions or attitudes to engage with a problem if it entailed effort that exceeded their realm of expectation for school mathematics (Schoenfeld, 1989). Thus, outside of relatively intensive classroom interactions intended to engage them substantively with a task, students easily gave up their attempts with similar tasks. Again, I am not attributing their difficulties solely to this, but merely surmising that it was a significant factor.[12]

Because of students' difficulties with these tasks, Activity 5 turned into a repeat of Activity 4. In Lessons 7 and 8, meanings for each problem of Activity 5 were negotiated within classroom discussions that entailed heavy prompting and instructional support of the kind exemplified in the discussion excerpts already considered here. These discussions indicate that students experienced essentially the same difficulties that I already described. Consequently, I do not elaborate those discussions further here.

Post-Activity Student Assessment

At the end of Phase 2, after the 8th lesson of the experiment, students' ideas related to the activities of this phase were assessed with an in-class written test.[13] The test questions were designed to query students' abilities to construe a complicated scenario (like those of Activities 4 and 5) as a probabilistic situation—that is, as entailing a population, a sample, and a random selection process that could be repeated as a means for investigating an issue of relevance raised in the scenario. In addition, the assessment asked students to explicate the logic of the design of

---

[11] This impression emerged out of my interaction with students during Lesson 7, when their work in Activity 5 was supposed to have been discussed in class.

[12] Evidence for this claim is drawn from students' comments about how different this class was from any other mathematics class they had ever taken. Moreover, in response to students' extreme difficulties with this particular activity, the research team explicitly laid out the norms and expectations for out-of-class assignments. Students were expected to engage in class to the extent that discussions and issues were meaningful for them when they attempted similar tasks on their own. Thus, they were expected to be proactive participants and to consider homework in terms of issues raised in the class discussions. Toward that end, students were encouraged to ask questions relating to their own or others' understandings of ideas under discussion and to seek clarification about take-home activities *before* taking them home.

[13] There were 8 students enrolled in the course during that time, all of whom took this test.

the investigation they would conduct using Prob Sim. The scenario and sequence of assessment questions are shown in Figures 6.7 and 6.8.

I. Here is a scenario: A consumer agency purchased 20 pairs of RoadRunner in-line skates from sports retailers chosen at random across the U.S. It found that 10% of the pairs (i.e., two) had defective wheels. However, the skate manufacturer claims that its tests show that no more than 2% of all RoadRunner pairs it distributes have defective wheels.
a) Was there a sample chosen in this scenario? If so, identify it. A sample of 20 *pairs* of skates was chosen. Thus, a pair of skates constitutes a sampled item. This has implications for what one takes as the sampled population in part b)
b) If a sample was chosen, what was the population from which it was drawn? In order to be consistent with a), the sampled population would have to consist of all pairs of RoadRunner skates distributed to retailers in the U.S.
c) In this scenario: 1) "10%" refers to 10% of what?, 2) "2%" refers to 2% of what? 10% of the 20 pairs (i.e., 2) that comprise the consumer agency's sample 2% of all pairs of skates distributed to retailers in the U.S. That is, 2% of the sampled population.
d) What is at issue between the consumer agency and the manufacturer? Ways to construe what is at issue: Does the difference between the outcome of the sample and the manufacturer's claim leads us to question whether something is amiss? Are the observed sampling outcome and the manufacturer's claim discordant? Is it in fact unusual to obtain a sample outcome of 10% defective items from a population that is presumed to contain only 2% defective items? If so, then this is reason to suspect that either the manufacturer's claim is wrong or that the sample is unusual.

Figure 6.7. Part I of post-activity assessment.

e) Here is a question. *Don't answer it!*

How unusual would a result like the consumer agency's be if, as the skate manufacturer claims, only 2% of all RoadRunner pairs have defective wheels?

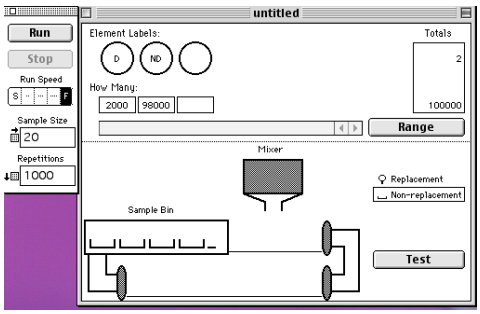
Describe, in principle, how you would use Prob Sim to investigate the question. (You don't need to provide details of setting up Prob Sim. You'll provide that information in the next question. Instead, focus on the overall logic of your strategy.

To investigate the question is to investigate what fraction of the time one can expect to draw a sample containing 10% defective items from a population containing 2% defective items. To do so, one could simulate the following experiment and examine the result:

1) Assume a large population containing 2% defective items and 98% non-defective ones; 2) select a sample of 20 items from this population w/o replacement and record the number/percent of defective items in the sample.

Repeat this experiment a large number of times, recording the number of times that a 10% defect rate is observed. Calculate the relative frequency, f_o, of this outcome and, on its basis, decide whether results like this are rare or not so rare [use convention: rare $\Leftrightarrow f_o \leq 5\%$]

f) Fill in the blank Prob Sim window below as you actually would to investigate this question.



"D" stands for defective items, of which there are 2000 (2%) in a population of 100,000 items

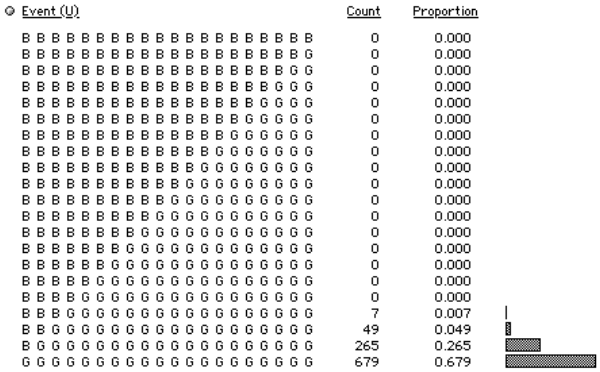
"ND" stands for non-defective items, of which there are 100,000 – 2000 = 98,000 in the population.

Samples are of size 20 and sampled items (i.e., pairs of skates) are selected without replacement.

The simulated experiment is repeated 1000 times.

Figure 6.7--continued.

II. The screen below shows results of having analyzed data generated by Prob Sim to investigate the in-line skates issue. The data was generated under the assumption that the manufacturer's claim is accurate. In this screen, "B" stands for "bad wheels".



Event (U)	Count	Proportion
B B B B B B B B B B B B B B B B B B B B	0	0.000
B B B B B B B B B B B B B B B B B B B G	0	0.000
B B B B B B B B B B B B B B B B B B G G	0	0.000
B B B B B B B B B B B B B B B B B G G G	0	0.000
B B B B B B B B B B B B B B B B G G G G	0	0.000
B B B B B B B B B B B B B B B G G G G G	0	0.000
B B B B B B B B B B B B B B G G G G G G	0	0.000
B B B B B B B B B B B B B G G G G G G G	0	0.000
B B B B B B B B B B B B G G G G G G G G	0	0.000
B B B B B B B B B B B G G G G G G G G G	0	0.000
B B B B B B B B B B G G G G G G G G G G	0	0.000
B B B B B B B B B G G G G G G G G G G G	0	0.000
B B B B B B B B G G G G G G G G G G G G	0	0.000
B B B B B B B G G G G G G G G G G G G G	0	0.000
B B B B B B G G G G G G G G G G G G G G	0	0.000
B B B B B G G G G G G G G G G G G G G G	0	0.000
B B B B G G G G G G G G G G G G G G G G	0	0.000
B B B G G G G G G G G G G G G G G G G G	0	0.000
B B G G G G G G G G G G G G G G G G G G	7	0.007
B G G G G G G G G G G G G G G G G G G G	49	0.049
G G G G G G G G G G G G G G G G G G G G	265	0.265
	679	0.679

a) How many samples were drawn?

A total of 1000 samples (of 20 pairs each) were drawn (as indicated by the sum of counts for each (unordered) outcome in the sample space).

b) How large was each sample?

Sample size is 20 items (as indicated by the number of elements in a row list, which denotes a sampling outcome).

c) Interpret the line that is second from the bottom.

The line shows that the sampling outcome "1 defective item in a sample of 20 items" or "1 pair with bad wheels in a sample of 20 pairs of skates" occurred 265 out of 1000 times, or 26.5% of the time. This proportion is also expressed as a histogram bar's length.

d) Based on these results, answer this question:

How unusual would a result like the consumer agency's be if, as the skate manufacturer claims, only 2% of all RoadRunner pairs have defective wheels?

The third row from the bottom shows that the outcome "*2 pairs with bad wheels in a sample of 20 pairs of skates*" — that is, a sample of which 10% of items are defective — occurred 49 times out of 1000, or 4.9% of the time. Thus, it would appear to be a statistically rare event to obtain 2 defective pairs in a sample of 20 pairs from a population of which presumably only 2% of pairs are defective [by the 5% convention explicated in question *Ie*)]

Figure 6.8. Part II of post-activity assessment.

The blue text below each question is offered as a normative or "model" response that guided my evaluation of the clarity and coherence of students' thinking on the question.[14]

I analyzed students' responses to each question in terms of whether they expressed a central idea, interpretation, or understanding that is *consistent* or *inconsistent* with the normative response. Responses whose status I found questionable were coded as *ambiguous.* Table 6.3 displays the frequency counts for student responses in each category.

Table 6.3. The frequency distribution of student responses compared to the normative responses. For instance, in the first row "7" denotes that 7 out of 8 students' responses to Question *Ia* were consistent with the normative response.

Question	*Consistent*	*Inconsistent*	*Ambiguous*
Ia	7	0	1
Ib	2	5	1
Ic_1	7	0	1
Ic_2	6	2	0
Id	4	1	3
Ie	2	2	4
If	1	1	6
IIa	5	3	0
IIb	8	0	0
IIc	6	1	1
IId	1	4	3

Students' responses to items *Ib*, *Ie*, *If*, *IIa*, and *IId* are notable for their relatively widespread inconsistency or possible inconsistency with the normative response. It is worthwhile to consider some of these responses more closely.

[14] Though these model responses were crafted to reflect interpretations and understandings that the research team deemed as coherent and desirable instructional endpoints, I emphasize that they are not taken as rigid criteria for judging the "correctness" of students' responses. My concern is not with the responses "correctness", but rather with what responses might suggest about the clarity and coherence with which students interpreted a question and its crucial issues. These model responses guide my evaluation of students' interpretations by providing a rough benchmark for assessing, say, what aspects of the situation were problematic for students and what ideas and issues were salient to them.

I coded 5 students' responses to Question *Ib* as inconsistent because they all identified a population that was not consistent with their identification of the sample in Question *Ia*. Here are one student's responses that typifies this problem:

Question *Ia*: "the sample was the 20 pairs of skates"

Question *Ib*: the population was "random sports retailer across the U.S."

I take these two responses to be mutually inconsistent because the population is not identified as a collection of the same kind of items that make up the sample. Instead, these responses may suggest an inclination to think of the *container* of sampled items—that is, all sports retailers in the U.S.—as the population. Thus, this inconsistency seems to suggest an underlying difficulty in conceptualizing a population as a collection of items of the same kind as those that comprise a sample.

Question *If* asked students to specify the Prob Sim parameter values they would select in order to investigate the question raised in the scenario (see Figure 6.7). Difficulties associated with conceptualizing a sampled population are suggested also in students' responses to this question. I coded 6 responses to this question as ambiguous because those students chose population proportions that are inconsistent with the normative response and, in some cases, with their own responses to Questions Ic_2 and *Ie*. Here are Peter's responses that typify this problem:

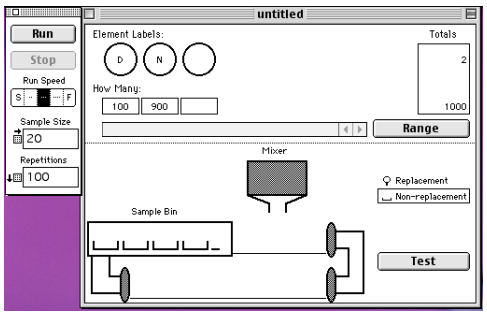
Question *If*:	**Question *Ie*:**	**Question Ic_2:**
	You need to take a big number of samples to see if in fact that getting 10% defective wheels is unusual. That most of the time if samples are taken the percentage of defective wheels will be around 2%	2% out of all the company produces

Figure 6.9. Peter's responses to three related assessment questions.

Peter's choice of the Prob Sim parameters suggest that he thought the population contained 10% defective skates, rather than 2% as he indicated in his response to Question Ic_2. Other students chose numbers of element labels, in Question *If*, that suggest they thought the population was evenly split or they chose the number of defective items to be 10% of the number of non-defective items. These inconsistencies might also reflect students' underlying difficulties

in understanding how population and sample are distinguished in the Prob Sim interface, as though they were uncertain how to interpret the program's element labels or the metaphor of a "mixer" as a population.[15]

In Question *1e*, only 2 students gave reasonably clear explications of the logic of their investigation. Here is the clearest of those:

> *"Set up a large population of the skates, with 10% of the population's skates defective. Choose random samples of 20 pairs of skates, multiple times. Find at what the percentage of the skates chosen* were defective and see if getting 2% of them defective is usual or unusual. Then you will know whether or not the 10% was more accurate, or the 2%.*
>
> **in the samples"*

I note that despite having a clear sense of the strategy for the investigation and its connection to the issue at hand, this student interchanged the population and sample proportions. This suggests a difficulty in placing and coordinating all of the information given in the scenario.

Some of the ambiguous and inconsistent responses to Question *1e* differ markedly from the response above; they typically lack the richness of detail and coherence, and they suggest inability to operationalize the logic of the investigation. Here are two such examples:

> *"You would apply all of your information to ProbSim and just repeat the test many times. This way you can see how unusual 10% actually is. Or if, in fact, it is unusual at all."*

> *"I would make a circle for defective wheels and o.k. wheels to see of the probability of the consumer agency getting a 10% defective rate if only 2% of all Roadrunners wheels are defective."*

Regarding Question *IIa*, each of the three inconsistent responses suggest difficulties in interpreting the information displayed in the Prob Sim data analysis window (see Figure 6.8). The common response was "21 samples drawn", which is, in fact, the number of unordered outcomes listed in the sample space. Thus, these students evidently interpreted a *class* of sampling outcomes as a particular sample. This suggests that they did not consider the displayed list of sample space outcomes in coordination with the information contained in the "Count" column, which lists the corresponding number of samples obtained in each class. Thus, these

---

[15] This is plausible, given that students had little hands-on experience with Prob Sim before the 8th lesson of the experiment.

students seemed to have interpreted the information in the display in isolation from the other parts, rather than as an integrated whole.

The inconsistent responses to Question *IId* also suggest difficulties in interpreting the display information as a distribution of sampling outcomes; those students all judged the consumer agency's 10% result as "not unusual", citing evidence from the display as a warrant. Those students evidently did not see the histogram in the display as indicating otherwise.

All of the ambiguous responses to Question *IId* judged the consumer agency's result to be unusual. Two of those did not include a justification for the claim, and a third one—Luke's response—entailed an interesting argument:

> *"The results of the consumer agency would be unusual because they should have generally received the same results as the manufacturer. The results may have varied slightly due to the small sample size taken."*

Luke's argument does not appear to be based on an interpretation of the display information as a distribution of outcomes. Yet, it certainly suggests his having a strong sense that the consumer agency's sample is expected to reflect the manufacturer's claim, with some allowance for slight deviation attributed to the small sample size.

The results of the assessment seem to indicate that students did interpret parts of the scenario in ways that are consistent with the normative interpretation. But they also suggest that many students did not integrate these parts to construe the scenario as a coherent and holistic probabilistic situation. This is reminiscent of diSessa's (1988) metaphor of "knowledge in pieces". Moreover, it seems consistent with Schwartz et al.'s (1998) research-driven hypothesis that children do not possess an abstract schema they can use to understand all statistical situations. Instead, their intuitive statistical understanding is comprised of a collection of overlapping, and even incongruent, schemas that are differentially evoked depending on the particular problem context. Similarly, many of our students seemed not to have constructed a stable web of interconnected and internally consistent meanings that might be expressed in their construing scenarios as probabilistic situations.

Chapter Summary

The second phase of instruction moved to engage students in designing sampling simulations as a means for investigating the statistical expectation of an event. These design activities were intended as a context for helping students conceptualize scenarios as probabilistic situations—that is, as stochastic experiments entailing a population, a sample, and a repeatable random selection process.

Analyses of the classroom discussions centering on a first design activity suggest that students experienced significant and robust difficulties in conceptualizing expectation as a statistical quantity. Students could not easily operationalize expectation in terms of relatively frequency—the fraction of the time that an event of interest might occur. Moreover, these discussions suggest that it was highly problematic for students to re-construe a scenario in terms of simplified assumptions that might facilitate their designing a simulation of it. These difficulties were significant enough to disable students from engaging in a similar design activity independently, outside of instructional interactions.

Analyses of students' written responses to post-activity assessment questions suggest that they were able to interpret parts of a given scenario in ways that are consistent with a normative interpretation targeted in instruction. However, many students did not appear to easily coordinate and integrate those parts into a coherent and holistic view of the scenario as a probabilistic situation.

CHAPTER VII

PHASE 3: MOVE TO CONCEPTUALIZE VARIABILITY AND DISTRIBUTION

This chapter describes a sequence of two multi-part activities that unfolded over Lessons 9 through 13 (see Figure 7.1). Broadly speaking, the activities addressed ideas of sampling variability, accuracy, and distribution. Specifically, Activity 6 aimed to engage students in exploring relationships between sample size and sampling accuracy. Activity 7 aimed to support students' developing a conventional interpretation of frequency histograms.

Phase 3:	**Move to conceptualize variability and distribution**	
Lesson	**Activity (A)**	**Duration**
9 (08/30)	A6: Favorite Musicians sampling scenario, Part 1	25 m.
10 (08/31)	A6: Favorite Musicians sampling scenario, Part 2	20 m.
11 (09/01)	A6: Favorite Musicians sampling scenario, Part 3 Transitions to A7	25 m. 20 m.
12 (09/02)	A7: Part 1—Constructing histograms A7: Part 2—Interpreting histograms	24 m. 12 m.
13 (09/03)	A7: Part 2—Interpreting histograms	12 m.

Figure 7.1. Chronological overview of instructional activities of Phase 3.

The chapter begins by elaborating the rationale for the design and implementation of Activity 6, drawing on a key insight into students' thinking that emerged in Lesson 8. The chapter then characterizes discussions that unfolded around Activities 6 and 7, typically following their temporal order and analyzing students' thinking that emerged within them.

Analyses elaborate several issues: students' interpretations of Activity 6; students' ideas related to sampling accuracy; students' problematic interpretations of histograms in relation to their difficulties in conceiving a sampling distribution.

Prelude to Phase 3

The emergence of the experiment's third phase was motivated by a development that strongly suggested that students' conceptions of variability were not rooted in an image of distribution. This development occurred within the class discussion of Problem 3 in Activity 5 (see Figure 7.2), in the 8th lesson.

Investigating unusualness (again)

Problem 3.

The Gallup company asked 21 adults, selected at random, whether they attended church or synagogue during the past week. Fifty percent said they had. The Harris company performed an identical survey. In their survey, only 33% responded "yes".

Use ProbSim to investigate whether the two polls contradict one another.

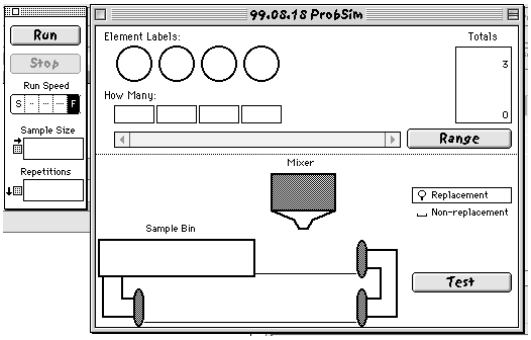


Figure 7.2. The "Gallup versus Harris" scenario.

Problem 3 is a scenario designed to engage students in investigating the unusualness of an event and thus conceptualize a probabilistic situation. Specifically, the "Gallup vs Harris" scenario described the results of two identical random surveys, each conducted independently by the Gallup and the Harris polling companies, respectively. The two identical surveys obtained different results with the question "did you attend church or synagogue during the past week?". In the Harris survey, 50% responded "yes", while the Gallup survey obtained a 33% "yes" response rate. The question that students were asked to investigate was whether the two polls were contradictory.

From the researcher team's perspective, the underlying statistical issue in this question was whether the two polling outcomes are sufficiently different to think that their difference is unusual. Equivalently, the issue is whether the variability among such sampling outcomes is broad enough to account for two such seemingly divergent outcomes. Thus, the *intended* issue in the task was framed with an underlying sense of variability in mind that is rooted in ideas of distribution. For students, however, this was unanimously and unequivocally a *non*-issue. This is illustrated in the following set of ordered excerpts drawn from a classroom discussion occurring halfway into Lesson 8 and lasting approximately 11 minutes.

Episode 1, Lesson 8 (Problem 3):

Segment 1

1. I: […] what is the issue that's being raised in that situation?
2. Peter: Whether you go to church or not? […] Whether you go to church or not? Or what uhh, what, whether the two polls contradict each other, or something? (2 second pause) I don't know.

(2 second silence)

3. I: Well, any body else?

(3 second silence)

4. I: Luke?

(2 second silence)

5. Luke: I don't have no idea.
6. I: You don't—no idea about what? About what I'm asking or about what they're describing?
7. Luke: About what you're asking.
8. I: Ok, I'm asking what is the fundamental sit—what's the situation that we're being asked to investigate?
9. Luke: Uhh, it's between the Gallup company and the Harris company, and they're survey's didn't add up.

(3 second silence)

10. I: Uhh, didn't agree?
11. Luke: Uhh (reads situation)
12. I: Is that what you mean by "add up"?
13. Luke: Uhh (reads), nnnno.

Segment 2

14. Tina: The Harris company that performed the identical survey, does that mean that they chose 21 adults too?
15. I: Hmm hmm (affirms)
16. Tina: Ok.
17. Nicole: Did they choose the same adults?
18. I: Huh huh (negative), they couldn't have (inaudible)
19. Tina: No, just random
20. Lance: Selected at random
21. Tina: Yeah.
22. I: Yeah, they're selected at random.
23. Tina: So
24. Luke: so is that not the reason why they didn't, why the results did not come out as the same?
25. Nicole: Yeah, right.
26. Tina: Yeah,
27. Peter: it's a different kind of
28. Tina: cause it's a different group of people
29. Nicole: different (inaudible)
30. Luke: 'Cause, it's gotta be, if you ask the same group of people you get the same results every time

[…]
31. Tina: So that's why! 'Cause--.
32. I: Ok. But could they--
33. Tina (continues): everybody's different
34. I: Everybody's different, that's right. But those results are very different, aren't they?
35. Tina: Right
36. Nicole: Yeah, but they were asking different people!

[…]

Segment 3

37. I: Ok, so uhh (2 second pause) it sounds like you're saying "ok, they asked different people, so what? They got different results". Is that sort of what you're trying to say?
38. Nicole: Like--
39. Peter: Pretty much
40. Nicole (continues): What exactly are you trying to—ask? Obviously they-- I mean, like what do you mean "do they contradict one another?" ?
41. Tina: Meaning the difference in numb—percentages; one was – (turns to Nicole and explains, pointing to her paper) there's such a big difference between them
42. Nicole (speaking to Tina): I know but they went and asked different people (inaudible).

As is clearly evidenced in these excerpts, students did not interpret the task question from the perspective intended by the research team. In fact, students did not understand what was at issue in the scenario and were thus disabled from engaging with the task in a meaningful way. Instead, their thinking was along these lines: "ok, Gallup and Harris asked different people, because the polls were random, and they got different results. So what? That's what you get when you poll different people!"

This development is rather compelling evidence that students' images of sampling did not entail a sense of variability that extended to ideas of distribution of sample statistics. Instead, their sense of variability seemed restricted to "differences between outcomes". Figure 7.3 depicts students' sense of variability in comparison to that entailed in a multiplicative conception of sampling (MCS).

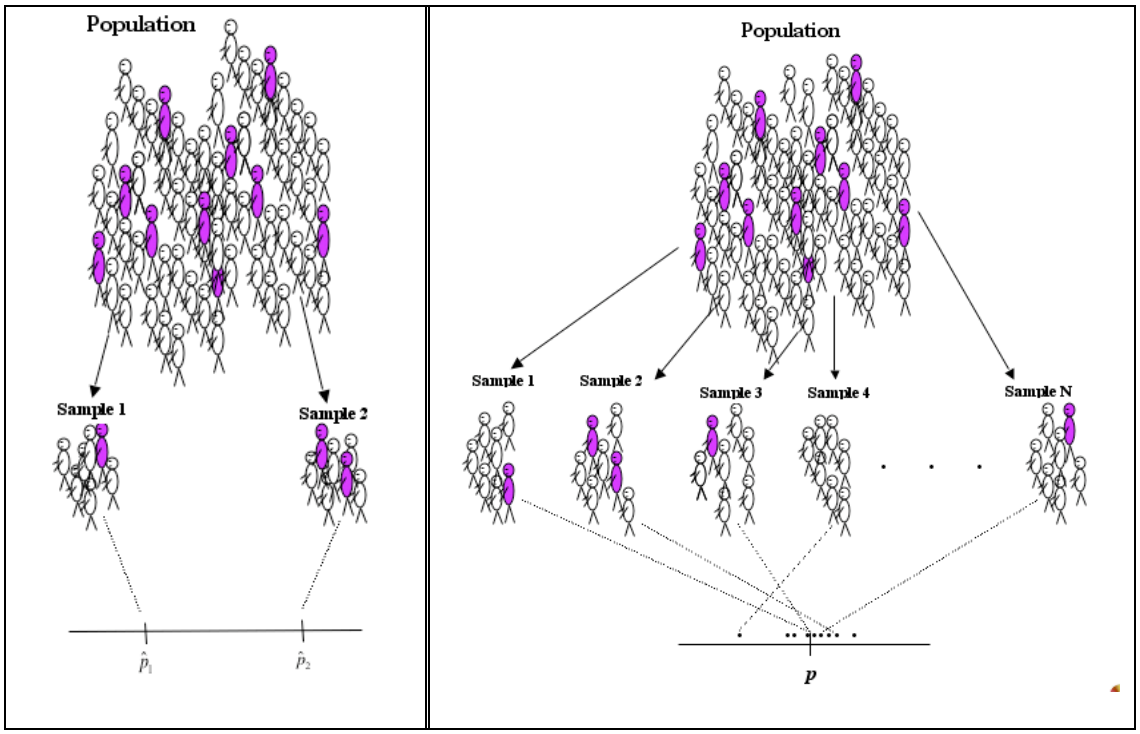


Figure 7.3. Students' image of variability (left panel) in contrast to a MCS (right panel).

From discussions in previous activities, students evidently understood that if a sampling process is repeated then individual sample proportions are expected to vary. But the discussion above suggests that they did so only to the extent that if more samples were drawn and their proportions computed, those values would differ from the ones for the samples already drawn. Thus, students' conceptions seemed not to entail an image of these values' *accumulation* and an anticipation of a range of outcomes that induces a bound on their possible differences.

The research team's realization of the extent of students' conceptions of variability constituted the point of departure for a new phase of instruction. A sequence of discussion-based sampling activities was designed with the aim of moving students toward developing a sense of a sample proportion's dispersion around the sampled population proportion's value.[1] The basic idea of these activities was to give students the experience of exploring how values of a sample proportion, generated by computer simulations, compare to the sampled population parameter. By systematically investigating these values' deviations from the population parameter, the

[1] The entire sequence of activities was not designed a priori. Rather, the first two parts of Activity 6 were designed as attempts to address students' limited ideas about variability, as characterized above. The other parts of activities in the sequence were, similarly, the research team's efforts to address issues that arose within classroom discussions. As such, the activities as a whole constitute an *emergent* sequence.

research team reasoned, students might develop a sense of a range of differences and possible patterns of dispersion that occur within this range.

The research team intended that students would reflectively abstract from these experiences an image of sampling variability that is embedded within ideas of, at least, a *proto-distribution*—that is, an orientation to thinking of a collection of sample statistics' as dispersed and therefore constituting a *cluster* of values within an interval. Moreover, this orientation entails a sense of the cluster's denseness within an interval, but it need not be an operational sense that enables a quantification of the denseness.

Activity 6: Favorite Musicians Scenario

The "Favorite Musicians" scenario entailed a simulated population consisting of 41,588 teenagers' favorite pop music performer, selected from a common list of 10 performers, and simulated samples of various sizes drawn from this population. Figure 7.4 shows the written guide used during discussions centering on the activity.

The setting

A group of 41,588 high school students were asked to select their favorite from among 10 performers. The performers were presented in a list. Students could make just one choice by checking a box. Here are the performers.

- ☐ Aerosmith
- ☐ Backstreet Boys
- ☐ Brittney Spears
- ☐ Dave Matthews
- ☐ Korn
- ☐ Limp Bizkit
- ☐ Madonna
- ☐ Mariah Carey
- ☐ nsync
- ☐ Stevie Wonder

The table below shows students' choices. It shows that 419 students, or a little more than 1% of the students, selected Aerosmith; about 10.5% of the students selected Backstreet Boys. Seventy-four (74) students left their forms blank.

41588 total cases of which 74 are missing

Total Cases	41514
Number of Categories	10

Group	Count	%
Aerosmith	419	1.009
Backstreet Boys	4379	10.548
Brittney Spears	26	0.063
Dave Matthews	23300	56.126
Korn	2231	5.374
Limp Bizkit	4169	10.042
Madonna	1020	2.457
Mariah Carey	779	1.876
nsync	4833	11.642
Stevie Wonder	358	0.862

We will accept these opinions as fact. CONSIDER THESE RESPONSES AS A POPULATION. DO NOT CONSIDER THEM AS A SAMPLE.

In this activity we will randomly select samples from this population and compare the samples with the population to see how representative the samples are. We will do this with samples of size 40, 200, 400, and 1000. For example, we will randomly select 40 questionnaires from the 41,588 questionnaires and then examine the percent of the sample choosing Aerosmith, the percent of the sample choosing Backstreet Boys, and so on. We will repeat this process many times for samples of size 40. Then we will repeat the entire process taking samples of size 200.

Remember, the purpose of this is to see how accurately samples of various sizes reflect the population's composition.

It might be helpful to organize the comparison in a table (like the one attached).

Sample size: 103	**Actual**	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
Aerosmith	1.01%															
Backstreet Boys	10.55%															
Brittney Spears	0.06%															
Dave Matthews	56.13%															
Korn	5.37%															
Limp Bizkit	10.04%															
Madonna	2.46%															
Mariah Carey	1.88%															
nsync	11.64%															
Stevie Wonder	0.86%															

Figure 7.4. The written guide for the "Favorite Musicians" sampling activity.

Students were asked to consider these 41,588 teenagers' selections, obtained from a simulated survey, as a population of selections rather than as a sample.[2] The composition of this population was described both in terms of the absolute counts and the percentages of the 41,588 teenagers who selected each of the 10 performers. For instance, the table in the center of Figure 7.4 shows that 1020 teenagers (2.457% of the 41,588 surveyed) chose Madonna as their favorite pop musician. The scenario, as intended by the research team, entailed investigating the non-trivial issue of sampling accuracy: students were asked to investigate how accurately samples of various sizes reflect the population's composition. Moreover, students were asked to consider as an issue the competing tension between sampling accuracy and sample size, and to make some judgments about optimal sample size and reliability that would balance this tension.

I should stress that sampling accuracy was not formally mentioned or defined by the instructor in the classroom a priori. Rather the aim was that the inquiry-based discussions around the activity would support the emergence of students' ideas about sampling accuracy and that these ideas might then serve as a basis for moving toward some consensual meaning of it. I should also reiterate that this scenario was designed as a context for helping students enrich their, apparently limited, ideas of variability by engaging them in exploring the behavior of sample percents relative to the sampled population percent. This activity thus differed from the sampling activity of Phase 1, where the underlying population parameter value was unknown to students and its estimation was the name of the game.

Activity 6, Part 1

Activity 6 unfolded in a sequence of three distinct parts over Lessons 9 through 11, each part itself a series of phases determined by the foci of discussions. Part 1 lasted approximately 25 minutes. It began, in a first phase, as a whole-class discussion of the Favorite Musicians sampling scenario in which the instructor oriented students' attention to the sampled population of teenagers' favorite musicians and to what samples drawn from it might look like.[3]

In the second phase, the discussion moved on from one student's (Luke) suggestion that a sample size of 10,000 would adequately mirror the population. This suggestion was in response

[2] The simulated population was generated in a LISP environment using a random number generator programmed to reflect the preferences of a small informal sample of real teenagers. This data file was then imported into the *Data Desk* (Data Description, 1999) statistical analysis program, from which the drawing of multiple samples of various sizes was easily done. The frequency table shown in the center of Figure 7.4 was produced in Data Desk.

[3] Students had the written activity guide in hand when the discussion began.

to a question posed by the instructor: “how big a sample should we take in order to be somewhat assured that its composition will roughly mirror that of the population?”. Following Luke’s suggestion, the instructor then used the Data Desk program to simulate drawing five samples of size 10,000 from the population, each time showing a sample’s composition as in Figure 7.5 and asking students to make informal “eye-ball” comparisons between the sample percentages and the population percentages (Figure 7.4).[4]

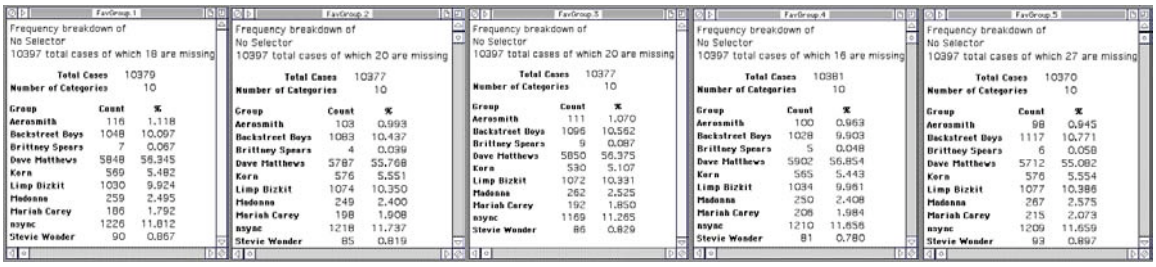


Figure 7.5. A sequence of Data Desk windows showing the compositions of samples of size 10,000 like those selected from the population of teenagers’ favorite musicians.

While making their comparisons, students noticed that in most samples most of the sample percentages were close to the population percentages. Though, students had no formal criterion for measuring the proximity. This observation prompted Luke to suggest that they consider samples of size 100 to see how their compositions compared with the population.

The instructor followed Luke’s suggestion and drew two samples of size 100 from the population. Students again made eye-ball comparisons with the population percentages. Upon examining the percentages of the first sample, students immediately noted how much farther they were from the population percentages than those in the samples of size 10,000. Students also noticed that there was variation among a sample’s percentages in how much they differed from the population percentages: percentages for some performers differed from the corresponding population percentages by more than others. In their comparisons, students seemed to use absolute deviation as an informal metric, they spoke of sample percentages as being more or less “off” the population percentages.

In the discussion, the instructor proposed that students consider the percentage difference relative to the population percentage—that is, percentage error—as a measure of deviation

[4] These figures were shown on a projection screen in class. Only the instructor had access to a computer and only he manipulated Data Desk to produces the simulated samples. The discussion focused on the tables showing the sample percentages and frequency breakdowns.

between a sample percentage and corresponding population percentage. Using this metric, the instructor roughly estimated the range of deviations among sample percentages relative to the population percentages to be between 5 and 10 percentage points. On the basis of having eyeballed the differences in absolute deviations[5] between the samples of size 10,000 and 100, Luke made a conjecture: *"the larger your sample size, then the more accurate you're gonna be to get a true reading of the population"*. Thus, in addition to having an astute intuition about the relationship between sample size and representativeness, Luke seemed oriented to thinking that the overarching goal of the activity was to get samples whose compositions were as close as possible to that of the population. The way in which Luke framed this issue suggests that his sense of "accuracy" was wrapped up in ideas of prediction and proximity.[6]

The discussion then moved to a more systematic examination of sample percentages; this phase lasted approximately 5 minutes. Individual students were each assigned to track one performer and to record their sample percentages in a table (see Figure 7.4, bottom panel) as the instructor drew each of 15 simulated samples of size 103 from the population of favorite performers. Once this data had been collected, the instructor oriented students' attention to the range of differences between sample percents and corresponding population percentages. Some students followed by volunteering the maximum differences they observed between their performer's percentages in the samples and the population.

The discussion concluded with the instructor attempting to informally orient students toward a statistical interpretation of accuracy: accuracy is not how "far off" any *one* sample percentage is from the population percentage. Rather, accuracy is how close to the population percentage we expect sample percentages to be in the *long run*, on the basis of where they tend to place themselves along a range of possible values relative to the population percentage. I reiterate that this was deliberately offered only as an informal, and not an operational, characterization of accuracy. It was intended to orient students to consider accuracy from a different perspective in the next part of the activity.

---

[5] The instructor's use of percentage difference was not appropriated by students.

[6] Taken at face value, the way in which Luke expressed his conjecture suggests that he expected the composition of *all* samples of one size to be more accurate (i.e., closer to population composition) than that of *all* samples of smaller sizes. Accordingly, Luke's intuition seems like a naïve or deterministic version of the law of large numbers; it did not entail a sense that the variability among sampling outcomes leads one to expect *some large proportion*, rather than *all*, of sample percents to be within some distance of the population percent.

Activity 6, Part 2

As an extension to the last discussion in Part 1 of Activity 6, students were given a homework activity that asked them to consider the sample percentages in 14 samples of size 831,[7] record them in a table (see Figure 7.4), and examine the range of their deviations from the population percentages to see how they were distributed. The instructor plotted points on a number line on the blackboard and suggested this as a way of showing how sample percentages are distributed relative to the population percentages. Students were given this homework activity near the end of Lesson 9, with an understanding that there would be a whole-class discussion in the next lesson about their work and ideas.

Part 2 discussion highlights

The class discussions around Part 2 of Activity 6 unfolded over approximately 20 minutes during the early part of Lesson 10. Highlights of these discussions are elaborated in this section

The discussion began with the instructor orienting students to the greater issue at hand, of which Part 2 was but one of several activities entailed in its investigation: "how small can samples be and their percentages still be relatively close, over the long run, to the population percentages?" The ensuing discussions centered on two students' strategies for determining how well samples of size 831 reflect the sampled population of students' favorite musicians.

Peter's idea

Peter proposed the following strategy:

"Maybe we can like uhh, like take averages of these 12 samples and put them next to the sample that came out, the man—(inaudible), and see how close they are. Maybe if they're not close enough, maybe we should take more samples."

When other students asked for elaboration, it emerged in the discussion that Peter had in mind applying this strategy to each of the distinct population/sampling categories:

"And the average of Aerosmith, I got .98 […] and the regular [population] percentage was 1.009."

[7] Students received the sampling data in the same form as that used in class—pictures of Data Desk windows like those shown in Figure 7.5.

In other words, Peter 's idea was to compute the average of a number of the 15 sample percentages for, say, Aerosmith, and to compare that average to the population percentage for Aerosmith. Similarly for the other sample categories. Moreover, as suggested by the last sentence of his explanation in the first quote, Peter thought that the average of the sample percents could be made closer to the population percent by taking more samples and, presumably, incorporating their appropriate percents into the computed average. Indeed, it seems plausible that Peter had interpreted the task this way: the purpose of considering multiple samples is to enable closer and closer approximations to the population percent. It appeared that for Peter the interesting relation was between sampling accuracy and numbers of samples, where "sampling accuracy" meant something akin to "proximity of a sample percent to the population percent".

Peter's strategy of taking the average of a collection of sample percents is a sensible thing to do for someone trying to assimilate multiple samples into this "proximity-based" conception of accuracy, since the average collapses many percents into a single numerical index. This would also explain why Peter considered only 12 of the 14 sample percents available to him; he obtained a "good enough" approximation using 12 of them.[8]

Another student, David, responded to Peter's idea in a brief but revealing interchange with the instructor:

Episode 1, Lesson 10:

1. David: So you could do that for all of them and then compare the sample sizes to see if the lower one's accurate to the higher one and see which—
2. I: Now if you did that for uhh, well if you did that for samples of size 40, you know average, average the percents, what would you expect in the long run?
3. David: For the averages to be off
4. I: You'd expect the average of the, of all the samples that you take to be off. Ok, uhh if you, for larger samples would you expect the average to be off less?
5. David: Yeah.

This discussion excerpt suggests that, in contrast to Peter, the salient relation for David was between accuracy and sample size. David saw that Peter's strategy could be applied to samples of different sizes and that by comparing the proximity of each to the population percent, one could determine which sample size yields higher accuracy. Though David's still seemed like a

[8] This difference represents a percentage error of only 3%.

proximity-based conception of accuracy, it entailed a consideration of different sample sizes and an expectation that greater sample size produces more accurate percentages, on average.

Nicole's idea

After discussion of Peter's and David's ideas, the conversation turned to Nicole's method for considering the accuracy of sample percents. Nicole articulated her idea in stages in the course of a 5-minute-long discussion with the instructor, a substantive excerpt of which is shown below.

Episode 2, Lesson 10:

101. I: Anybody have another way of looking at this? What we're trying to do is get a handle on the, the long-term accuracy that we get by taking samples of a particular size. […] Ok, this is Peter's criterion for determining accuracy […] Any other suggestions?
102. Nicole: Why can't you just take a lot of samples and compare those results to each other?
103. I: And, and how w--, ok what--, how would we compare the results?

[…]

104. Nicole: Just see if you have a pattern in the percents
105. I (says Nicole's suggestion out loud as he writes it on board):

"***2) Nicole's suggestion: Take a lot of samples and see if you have a pattern in the percents***"

106. I: […] So what, in here, do you think might take might take more explanation?
107. Nicole: To see if you have a pattern.
108. I: Yeah. What do you look for to see if you have a pattern? What do you have in mind? […] What would <u>you</u> look for?
109. Nicole: To see if the numbers look about that same, or if they're like really different or how big—I don't know what you're asking me! (chuckles nervously)
110. I: I'm just saying I don't know, I don't know what to do. I, I can get lots of samples
111. Nicole: Ok. Ok, take a lot of samples and look at the (hesitates and laughs) sample—what is it, percents? Is that, I mean is that what we're calling it?
112. I: Yeah. Sample, alright, sample percents. Alright, now that, that's better
113. Nicole: right
114. I (continues): now we know what to look at in those samples
115. Nicole: And then compare the percents of each sample you take to see if they are approximately close.

[…]

116. I: (Paraphrases Nicole as he writes it down on board) and 'see if they are close together'. Is that what you were saying?

"***2) Nicole's suggestion: Take a lot of samples and see if you have a pattern in the percents, see if they are close together***"

117. Nicole: Sure

This discussion excerpt shows that, in contrast to Peter's idea of taking the average of sample percents and comparing it to the corresponding population percentage, Nicole thought to compare sample percents to each other to get a sense for their proximity. Nicole thus had a sense that the dispersion or clustering of a collection of sample percents was indicative of the sample accuracy. Though, her sense seemed informal, pre-quantitative, and still only proto-distributional, for it apparently did not entail an operational image of proximity or accuracy. That is, Nicole did not specify any method or criterion for determining the closeness of the collection of sample percents. For instance, she did not suggest parsing the collection into proportions of it that lie within various percentage ranges of the population percent or within sub-ranges of the entire collection's range.

Despite the limitations of Nicole's idea of sampling accuracy, her image of sample percents' closeness was vivid enough to make a useful criterion for determining the accuracy of different sample sizes on the basis of rough visual comparisons of their clusterings. Indeed, the instructor promoted Nicole's image, which clearly resonated with several students who equated her sense of closeness with sample representativeness. This is illustrated in the following excerpt drawn from a discussion lasting approximately 3 minutes.

Episode 3, Lesson 10:

71. I: So, you could compare samples of different sizes by looking at how close together these samples are — say, of size 40 — looking at how close together the sample percents are from the samples of size 800. And whichever samples, whichever groups of samples seem to be closer together, what would you conc—what, what would that, what would you call those samples that seem to be closer together?
72. Luke: So they'd be more representative of the population.
73. I: Yeah. They tend to be more representative of the population because the percents would vary less (moves arm in sweeping motion) from the actual population percent.

[…]

74. I: […] these suggestions allows us to make comparisons among, say 40 samples of size 15 and 40 samples of size 800, and 40 samples of size 2000. Alright? Now, s-- let me suggest something here 'see if they are close together' (reading Nicole's description on the board), you could do that by eyeballing the numbers. […] You could do that visually, by plotting little dots (motions with hand as though plotting dots with pencil) because then you can get a visual sense of the pattern. […] Say if you had these two patterns (draws figure 7.6 on the board) […] So you have the same range; you have one down here at .52 and one up here at .60, just like up here (points to first clustering).

But which would you say, which pattern suggests that the samples, the sampling method produces more accurate results?

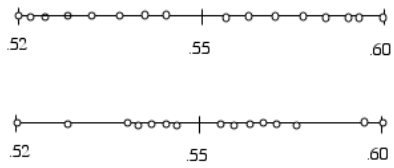


Figure 7.6. The instructor's sketch of two clusterings of sample percents around the sampled population percent.

75. Kit: The last one.
76. I: Why Kit?
77. Kit (softly): Because they're closer together, More of them—
78. I: Because the pattern is — using Nicole's idea — the pattern is that, that they tend, the tendency is — even though you might get some far out — the tendency is to be closer in there. Ok, so the tendency is that they're less spread out (spreads hands apart). Ok, so that 's (points to Nicole), so this idea of a pattern is a good one. That's a good suggestion.

In this excerpt the instructor's characterization of Nicole's idea in terms of a collection of sample percents' "spread out-ness" was intended as an imagistic basis for students to think about accuracy. The next part of the Favorite Musicians activity aimed to extend the discussions in Part 2 to have students draw on this idea in a more systematic investigation of the relationship between sample size and accuracy.

Activity 6, Part 3

Part 3 of Activity 6 was assigned for homework near the end of Lesson 10, with the understanding that students' work and ideas would constitute the topic of discussion in Lesson 11 (the next day). The written guide for Part 3 of Activity 6 is shown in Figure 7.7; it summarizes some, and alludes to other, big ideas that emerged in the discussions of Part 2. The goal of this structure was to make those potentially useful ideas readily available to students, but without explicitly asking them to connect the ideas to the assignment task.

Variation amongst Samples: Homework

We started this part of the course asking this question:

How large does a randomly chosen sample have to be so that we feel assured it is a fair representation of the population?

We looked at samples from a population of students who chose their favorite singers from a list, and then we made an "eyeball comparison" of the percents from the samples with the percents from the population.

JV came up with this observation:

> ***When we compare a population percent (such as the percent of the population who are Dave Matthews fans) with the equivalent sample percent (such as the percent of a sample drawn from that population who are Dave Matthews fans), larger samples tend to be more accurate than smaller samples.***

We gave a very special meaning to the phrase "tend to be more accurate." We said it means that *the percents calculated from larger samples tend to vary less, over the long run, from the actual population percent than do the percents calculated from smaller samples.*

In class today, we devised ways to compare "variation over the long run" for samples of various sizes.

Your assignment is to write a short essay that responds to this question:

> ***We are going to take a sample from a population having approximately 40,000 individuals in it. How small can the sample be so that, over the long run, samples of this size accurately reflect the population's composition?***

Your response is important, but the major part of this assignment is that *you justify your response*. Your justification must be based on the data presented in the attached sheets.

Also, you should make clear that there are two competing motives. One motive is that we want the sample to be very small. This will make it easier to actually collect it. The other is that we want the sample size to be such that there is relatively little variation, amongst samples of this size, from the actual population percents.

Figure 7.7. The written guide for Part 3 of Activity 6.

Along with the activity guide, students also received simulated sampling data for samples drawn from the population of 41,588 teenagers' favorite musicians. The data consisted of the percentages of a sample that selected each music performer as their favorite. These sample percentages were already organized in tables like those that students had previously filled in with sampling data (see Figure 7.8). Students received 12 such tables ordered in increasing sample size, each displaying the data for 15 samples of a common size.[9]

[9] The 12 different samples sizes were n = 10, 41, 103, 205, 414, 1038, 1659, 2492, 4155, 6230, 9129, 12,451.

Sample size = 414	Actual	S1	S2	S3	S4	S5	S6	S7	S8
Aerosmith	1.01%	0.97%	1.45%	0.48%	1.93%	0.72%	0.00%	1.21%	0.00%
BackstreetBoys	10.55%	9.90%	9.18%	10.39%	9.66%	10.14%	8.94%	9.42%	10.14%
BrittneySpears	0.06%	0.00%	0.00%	0.24%	0.00%	0.00%	0.00%	0.00%	0.00%
Dave Matthews	56.13%	53.14%	57.49%	52.90%	52.66%	56.04%	59.18%	59.42%	55.31%
Korn	5.37%	5.31%	6.04%	5.56%	6.04%	4.59%	6.76%	3.62%	4.35%
Limp Bizkit	10.04%	8.94%	9.42%	9.90%	11.11%	13.53%	12.56%	11.35%	11.59%
Madonna	2.46%	2.66%	2.42%	3.14%	2.66%	2.66%	1.93%	2.42%	3.62%
MariahCarey	1.88%	1.93%	1.69%	2.66%	2.66%	1.21%	1.69%	1.21%	0.48%
nsync	11.64%	16.43%	11.84%	13.53%	12.56%	10.87%	8.21%	10.14%	13.04%
StevieWonder	0.86%	0.72%	0.72%	0.97%	0.97%	0.24%	0.48%	0.97%	1.69%

Sample size = 414	Actual	S9	S10	S11	S12	S13	S14	S15
Aerosmith	1.01%	1.45%	0.00%	0.48%	0.72%	0.72%	0.72%	1.21%
BackstreetBoys	10.55%	9.66%	11.59%	10.39%	9.90%	8.70%	11.35%	11.35%
BrittneySpears	0.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Dave Matthews	56.13%	56.76%	57.97%	58.45%	58.94%	58.21%	57.25%	54.35%
Korn	5.37%	3.38%	5.56%	4.83%	7.73%	5.07%	5.07%	7.00%
Limp Bizkit	10.04%	11.84%	8.45%	10.63%	8.70%	10.63%	7.73%	11.35%
Madonna	2.46%	2.17%	2.90%	2.66%	2.66%	1.69%	2.90%	1.69%
MariahCarey	1.88%	1.21%	2.90%	0.97%	1.45%	1.93%	2.42%	2.42%
nsync	11.64%	13.29%	9.90%	10.87%	9.42%	12.80%	12.32%	9.66%
StevieWonder	0.86%	0.48%	0.72%	0.97%	0.72%	0.48%	0.48%	0.97%

Figure 7.8. One of 12 data tables showing the sample percentages for 15 simulated samples.

Students received this activity near the end of Lesson 10, whereupon the instructor led a brief class discussion intended to clarify how to interpret these data tables and the goal of the activity. The instructor stressed that the goal was to analyze the data from samples of different sizes with an eye toward answering the question posed in the boxed text of the activity guide: “How small can a sample drawn from a population of roughly 40,000 people be so that, over the long run, samples of this size accurately reflect the population’s composition?”. Moreover, the instructor stressed that the aim of the activity was not to answer the question definitively, but rather to give a persuasive analysis that might form the basis of a judgment about how small a sample to choose and still be confident that samples of that size tend to be representative.

Part 3 discussion highlights

Part 3 of Activity 6 and the issue with which it aimed to engage students is admittedly non-trivial. Indeed, the classroom discussions that emerged around this activity in Lesson 11 readily reveal that it was problematic for many students simply to construe what was at issue. These discussions unfolded in phases over a time period of approximately 25 minutes, each phase typically marked by the instructor’s renewed efforts to help students clarify issues that emerged

in the preceding segment of the discussion.[10] This section characterizes students' conceptions as they emerged within these interactions. Analyses foreground two central issues: 1) students' problematic interpretations of the activity, and 2) students' implied sense of accuracy.

The class discussion around Part 3 began with the instructor asking how students had understood the question posed and the issue and logic of the activity. The following brief excerpt from the opening discussion provides insight into some students' ideas.

Episode 1, Lesson 11:

1. I: […] So let's make sure that uhh we're all on the same page on the question. How did you understand the question? Nicole […] tell me about how you understood the question before you say your essay.

[…]

2. Sarah: Ok. Well, you wanted us to, ok you gave us uhh a big, all the sheets, the sheets, (refers to sampling data in handout) and on the left side of all the boxes (refers to tables of data) was the original experiment (refers to actual population percentages), which had forty thousand people […] and we had all the samples with different sized, sizes, and you want us to compare them and uhh, see how, see how accurate we could get for the lowest number, like the highest accuracy with the lowest number.
3. I: Ok, now why is that, why is that an issue?
4. Sarah: Because, we want to make, [what] we want is to find out how, uhh, small samples we can get and still be accurate.
5. I: Ok. Now, so—
6. Tina: You don't want to hassle with a lot of samples.
7. I: Go ahead, Tina?
8. Tina: You don't wanna, like, hassle with a lot of samples?

This discussion excerpt illustrates two students' different interpretations of the issue.

Sarah's idea

Sarah's interpretation was in line with that intended in instruction: she had a sense that the issue at hand was to balance the competing tension between sample size and sampling accuracy. Sarah's sense of accuracy, however, is unclear in this excerpt. In a later part of the discussion Sarah read her written response to the activity task aloud, almost verbatim. Her explanation

[10] It is useful to think of these discussions as being driven by cycles of discursive interactions between students and the instructor (Davis & Simmt, 2003; Bowers & Nickerson, 2001), where the former presented their ideas and asked questions and the latter, in turn, responded with elaborations and clarifications intended to align students' understandings along those of the instructional agenda.

offers some clarification of both her interpretation of the activity and her conception of accuracy. Her utterance is shown below, the text in square brackets was present only in her written response:

> *"Though it may be more accurate to use a lot—a very large sample, the smaller sized sample would be much easier to obtain and analyze. Through ob—through observing the provided samples I've become aware of this process.*
> *I've compared the results of different sized samples and their accuracy to the original experiment of the population. I observed how much of a difference*[*] *there was between samples and the* [then] *original experiment percentages.*
> *I concluded that*—[a sample size of 1,038 was the smallest size with the best accuracy].
>
> [**In the percentages*]*"*

In the first sentence of the second paragraph of this explanation, I interpret Sarah to mean that she compared sampling results (percentages) for samples of different sizes with the population percentages, not with each other. Further, I interpret the second sentence of the second paragraph as her explication of "accuracy", which she apparently meant as *difference* between sample percentages and population percentages.

According to the last sentence of the explanation, Sarah concluded that a sample size of 1,038 was the smallest that gave the "best accuracy", presumably meaning the smallest difference. Moreover, Sarah reportedly arrived at this conclusion by comparing the sample percents in the given samples of each size with the population percents, with an eye toward balancing the smallest differences between them with small sample size. Presumably, Sarah did not collapse a collection of sample percents into an average, as I interpret her to have meant that she compared individual sample percents with the corresponding population percent.[11] In my interpretation of Sarah's explanation, there is little evidence that she was oriented to thinking about the collection of these differences or their distributional character.

Evidence from other students' written responses to this task shows that Sarah's strategy, and thus her sense of accuracy, is representative of two other students'. One student referred to

[11] It is not clear, nor did it arise in any discussions, how Sarah handled cases where these differences for a collection of percents for samples of one size are not *all* smaller than those for a collection of another size. In other words, it is not apparent how Sarah dealt with the aggregate nature of the random data. She might have considered the proportion of each collection of 15 differences that were within some cut-off value, in order to make *overall* comparisons. This is, however, pure speculation on my part.

Peter's strategy of comparing the average of a collection of sample percents to the population percent. It is plausible that Peter employed this strategy himself, though his written response is inconclusive.[12]

Tina's idea

Lines 6 and 8 of Episode 1 of Lesson 11 indicate that Tina thought the issue was to keep the number of samples, rather than sample size, small. Tina's concern thus seemed consistent with Peter's thinking, in Part 2 of the activity, that the salient relation was between accuracy and numbers of samples. Unfortunately, no more information is available in the discussions to elaborate Tina's thinking. However, her evident focus on, and concern with, number of samples as a salient issue raises questions about what sense she had made of the activity of simulating drawing many samples from a population. Tina may have thought that the simulation was of conducting a real survey, rather than a context for studying the behavior of sample percents. If so, this might partially explain her concern. Another possibility is that Tina had confounded sample size and the number of samples.

As the discussion unfolded, other students asked questions that strongly suggest they were not distinguishing the aims of these two different activities: simulation of conducting a real survey versus simulation of repeated sampling from a known population as a context for exploring sampling distributions.[13] For instance, Luke asked:

> *"If you were gonna do a survey or something, why would you wanna come sort of close to the right answer and not just go ahead and do forty thousand surveys and know the exact answer?"*

And Nicole asked:

> *"[…] if you've already taken the forty thousand why are you gonna do it again?"*

These questions emerged after students had already participated in re-sampling activities and discussions, during Lessons 9 and 10, in which the goal was to study the behavior (e.g., variability) among sample percents for samples drawn from a certain population. Yet these

[12] The majority of students did not complete this written analysis for Lesson 11, as requested. Some students eventually completed it for the next lesson.

[13] The research team detected a similar confusion among many students who participated in a similar activity in the first teaching experiment (Saldanha &Thompson, 2002). Analyses of students' difficulties in that regard, however, have not yet been published.

students evidently did not have a clear sense of that goal. Their comments seem consistent with attempts to assimilate the activity of re-sampling into a "polling scheme"—that is a non-probabilistic conception of sampling, where *one* subset of the population is selected so as to accurately determine the desired population parameter. It appears that students experienced unresolved tensions between these two ideas—they were unable to make them "fit". This left them with an incoherent interpretation of the activity, which in turn disabled them from productively engaging with the task.[14]

Impelled by these questions from students, the instructor steered the discussion toward clarifying the activity's aim. He began by lecturing about the overarching logic of the activity, elaborating the distinction between re-sampling from a population having a known parameter and conducting a real survey of a population having an unknown parameter value. He stressed the big idea that the goal of the former activity was to study how collections of sample percentages for samples of various sizes aggregate around the population percentage. Moreover, he stressed that we study such patterns of dispersion with an eye toward inferring relationships between them and sample size that might transcend the underlying population. This last point was only intended to orient students to the overarching logic of the activity, as students were not expected to have made such a generalization at that stage of engagement.[15]

Students responded to the instructor's lecture with comments that suggest they continued to struggle with how to interpret the activity. For instance, Peter wondered whether the goal of the activity was to find the relationship between optimal accuracy and sample size relative to the population size—as in "what *proportion* of the population will yield the most accurate sampling results?". David was oriented to choosing the largest samples because they yield the most accurate results, thereby ignoring the call to balance accuracy and minimal sample size. Some students moved toward elaborating their emerging sense that the issue was the trade-off between sample size and accuracy: *"you're trying to look for the smallest amount of people in your sample that you can do so that it'll be accurate"*.

These comments, in turn, impelled the instructor to cast the issue in terms of a *point of diminishing returns* metaphor; if cost was no object, we would simply select the largest possible

[14] Tina, Luke, and Nicole were among the majority of students who did not complete the requested essay describing the analysis that underlay their choice of an "optimal" sample size.
[15] This last point foreshadowed activities in the next phase of the teaching experiment that aimed to support students' making this very generalization.

samples, which would ensure maximum accuracy. However, cost *is* a consideration and we must therefore balance the need for maximum sampling accuracy with smallest sample size, and thus lowest possible cost of conducting a survey. There comes a point beyond which the increased cost for a larger sample is not worth the accompanying small gain in accuracy.

Students' reactions to this metaphor at the time of the discussion revealed that, much to the surprise of the research team, the trade off between cost and accuracy was not a real issue for them. Numerous students were very surprised to learn of the considerable costs entailed in conducting a representative survey and so had little reason to consider cost as a constraint. Consequently, these students experienced great difficulty making sense of the activity; they were essentially disabled from engaging in it because they could not construct a clear and authentic goal.[16]

In addition to these aforementioned issues, these discussions also indicate that most students understood that Part 3 of Activity 6 activity entailed comparing the sampling data with the values in the "actual" column of the data table (see Figure 7.8). However, it emerged near the end of these discussions that many students had problematic interpretations of the values in the "actual" column; some thought the value in each row of the column was the average of all 15 sample percentages for that category. Other students referred to these values as the "actual sample" or the "original" percentages, thus suggesting that they had not conceptualized the 41,588 sample percents as a population from which the other samples were drawn. Still others did not make the connection between drawing samples from the 41,588 sample percents and the accuracy of sample percents drawn from "a population having approximately 40,000 individuals in it", as mentioned in the activity's written guide (see Figure 7.7). These students evidently thought the 40,000 individuals was a large sample selected from the 41, 588 individuals and for which they had no sample percentages.

Overall, Activity 6, particularly Part 3, was extremely problematic for most students. Even after discussions in which the instructor had tried to convince students of the reality of the tension between sampling accuracy and sample size or cost, the issue of balancing the tension between these constraints, as it was framed in the activity, seemed to entail too many parts and

[16] The non-reality of this issue for students was unanticipated by the research team. Indeed, the design of Activity 6 presumed this to already be a real issue and built on this presumption. In retrospect, Activity 6 would need to be re-designed to entail a significant component aimed at helping students conceive this as a real issue *before* engaging them with statistical ideas. Such a re-design might also entail having students understand the significance of statistical reasoning in helping to resolve a real issue.

relations for most students to manage and make sense of as a coherent whole. This is evidenced not only in the classroom discussions, but also in students' responses to the following assessment question administered two lessons later: "What did the instructor mean when he said 'We reach a point of diminishing returns' when we try to increase sampling accuracy by making samples larger and larger?". Table 7.1 displays students' written responses to the question:

Table 7.1. Students' responses to an in-class written assessment item administered in Lesson 13.

Student	**Response**
1. Nicole	This means that you're really not getting more accuracy for the added size. The accuracy does not improve enough to increase the size. — It's not worth it.
2. Sue	It doesn't show a lot of difference between the each sample size.
3. Kit	The accuracy does not increase as much as the # of items in the sample. Say sample size 10 — 20% Size 100 — 90% ← smallest higest % Size 200 — 92% not enough of % increase to matter
4. Sarah	When the sizes get larger and larger, there will be fewer results that could be not accurate and therefore the accuracy and the percentages get closer to the correct percentages as if the survey was taken from a whole population.
5. Peter	You put in too much without getting enough in return. You return isn't worth the cost of the extra accuracy.
6. Chelsea	He meant that getting those extra samples may not be worth the cost, etc. Your getting very little for the time, energy, and cost it takes to get it.
7. David	It makes the sampling process harder and harder if you get the sizes larger and larger. If you payed a company to sample something you would want to get the smallest sample size that would be the most accurate so you wouldn't have to pay them as much.
8. Luke	When taking samples the size of the sample needs to be the smallest possible. When you take a very large sample you are almost surveying the whole population, which is not what you are trying to do.

On the face of it, these responses suggest that the *point of diminishing returns* metaphor resonated with students, for it was evidently still salient to them two days after having been mentioned in only one discussion.[17] However, careful scrutiny of these responses indicates that relatively few students were able to coherently articulate the issue as one of balancing the

[17] This assessment item was part of one of several unannounced in-class quizzes. These quizzes were designed to provide the research team with information about what sense students were making of ideas raised in classroom discussions.

considerations of cost and accuracy—where "accuracy" goes beyond the proximity-based meaning of "not far off", instead having a proto-distribution-based meaning of "statistics from repeated samples are clustered densely close to the population parameter". Peter's and David's responses explicitly mention both considerations, Kit did so only implicitly. Other students—Sarah, Nicole, Sue, and Luke—made reference only to accuracy in relation to sample size. Chelsea seemed to be concerned with the cost of selecting *more* samples rather than larger samples, as though she believed that gains in accuracy are related to the number of samples selected—an idea that resonated with Peter's thinking in Part 2 of the activity.

Due to the difficulties that students experienced in making sense of Part 3 of Activity 6, contrary to the research team's intentions, the activity proved un-useful in helping students develop the nascent ideas of accuracy and variability that emerged in discussions of Part 2 into operationalized notions.

Transitions to Activity 7

The move to engage students in the next activity of the sequence emerged out of the discussions characterized above. Discussions leading into Activity 7 progressed during the final 20 minutes of Lesson 11; they addressed accuracy and variability of a collection of data values and the graphical depiction of such collections.[18] This section highlights developments that emerged within those discussions.

In the final phase of the classroom discussion in Part 3 of Activity 6, instruction broached the idea of sampling accuracy with regard to an entire collection of sample percents and it moved toward representing distributions with stacked dot plots. The following discussion excerpt began with my (L) pointing out to students that the activity entailed comparing not just one but a collection of 15 sample percents with the actual corresponding population percent. I raised this point because I sensed that students were fixated on comparing individual sample percents with the population percent. My concern was that their focus of attention was not on a collection of sample percents and that they were thus not oriented to thinking of accuracy in terms of how the collection as a whole was dispersed around the population percentage. The classroom teacher (T) then interjected to elaborate my point and the discussion unfolded between him and Tina. The excerpt lasted approximately 2 minutes.

[18] Discussions in Activity 7 itself occurred during the early part of Lesson 12 (see Figure 7.1).

Episode 2, Lesson 11:

9. L: Don't forget, we select 15 samples of each size and there are many samples of the same size to compare
10. Nicole: right
11. T: So in other words, don't just look at 'actual' and S1 and make your decision based on that. You have to look at all of the S1s through S, let's look at one specific one, uhh sample size of 10, we'll go right back to the beginning and we'll look at Backstreet Boys, for example. When we did our actual sample, about ten and a half percent, 10.5% of the people said they preferred Backstreet Boys as a group.

[…]

12. T: Now, when we took, instead of taking forty thousand we just took ten. One time we did it 30% said they liked the Backstreet Boys, the next time nobody! As a matter of fact the next couple of times, nobody. Then we got a 10%, a 30%, a 20%, a 40, a 0, 20, 20, 30, 10 and 10.
13. Tina: So that's not very accurate.
14. T: And why?
15. Tina: Because—there's nothing close—[…] to the actual
16. T: Sure there is. There's a couple of tens
17. Tina: Well there's, but overall there's not

[…]

18. T: Look at the actual numbers. How close to the number we wanted, are they?
19. Tina: Not very close
20. T: Not close at all. They range anywhere from 40, I think in one case, down to zero. Now fine, if we take an average of them we may get 10,
21. Tina: That's not what (inaudible)
22. T: but they're, you know it's 10 plus or minus 20 or 30. So they're scattered all over the place [meanwhile the instructor begins sketching a dot plot of sample percentages on the blackboard]
23. T (continues): Uhh, the same with, well Dave Matthews. Anywhere from, it should be around 55 or 56 and we're seeing anywhere from, well
24. Tina: It's skipping from 30 to 70
25. T: big variation again

What began to emerge in this discussion was Tina's sense of accuracy of a collection of sample percents. At the start of the interaction, Tina seemed to lack the language to express this sense with precision (lines 13-15). As the teacher confronted Tina with counterexamples to her claim that none of the sample percents were close to the population percentage, Tina clarified that she was thinking about the *overall* accuracy of the sample percents. It was as though Tina had used the term "none" to mean relatively few when compared to the total number of sample percents. This was not an operational notion of accuracy of a collection, but one that would

nevertheless seem to provide a basis for talking about proportions of the total number of sample percents that might lie within vicinities of the population percentage.

This move toward speaking about a collection, or proportions of it, propelled the discussion in that direction. Indeed, the instructor, who had been listening intently to this interchange between Tina and the teacher, used the discussion as an opportunity to introduce the use of graphs as a way of depicting how collections of samples percents are distributed. On the blackboard, he constructed two stacked dot plots of the sample percents for Backstreet Boys, for 15 samples of size 10 and 414, respectively. Figures 7.9 and 7.10 show pictures of the actual dot plots.[19]

Figure 7.9. The instructor's dot plot of 15 sample percents for Backstreet Boys (for n = 10).

Figure 7.10. The instructor's dot plot of 15 sample percents for Backstreet Boys (for n = 414), contained within the boxed region of the graph, was superimposed on the dot plot of sample percents for n =10.

The instructor used these two stacked dot plots to help students compare how percents of samples of two sizes were dispersed. The following short excerpt highlights the culmination of this phase of the, approximately 4-minute-long, discussion.

Episode 3, Lesson 11:

26. I: Now. Look at how, which of those two sample sizes induced more variability in the samples?
27. Tina: The first one (points to graph on board)
28. I: Which one had more variability in the percents?
29. Tina: The first one
30. I: The first one. See, it's visually evident, they were scattered all over that. So, with size of, with a sample size of 414 you have a lot better accuracy. And what does that mean, to get better accuracy?

[19] Students played a peripheral role in the construction of these dots plots; they called out the values and the instructor plotted them in the appropriate place on the number line. Peter suggested that points be stacked vertically.

31. Nicole: They're closer together
32. I: Yeah
33. Nicole: not as much range
34. I: Less variability, uhh smaller range, ok? So the idea is that there's, they're not so spread out. And 414 is a heck of a lot better than forty—uhh looking at 414 questionnaires is better than looking at 41,000 questionnaires.

The excerpt suggests that Tina and Nicole had a sense of variability and accuracy that seemed to be based on the range of sample percents' values and their concentration or "closeness together" within a range.

The discussion around the use of stacked dot plots provided a segue into the next phase of the transition to Activity 7, which lasted approximately 9 minutes. In this phase the instructor employed a brief dynamic computer animation to show students how histograms are used to represent and organize a collection of values of a random variable. The animation was part of the *ActivStats* program (Velleman, 1999)[20]; it showed a histogram emerging—its bars "growing" as random data values dropped from a slot above the graph and were organized into numerical intervals along a horizontal axis. The animation entailed a concurrent real-time narration of what was happening, explaining what the histogram's horizontal and vertical axes represented (Figure 7.11).

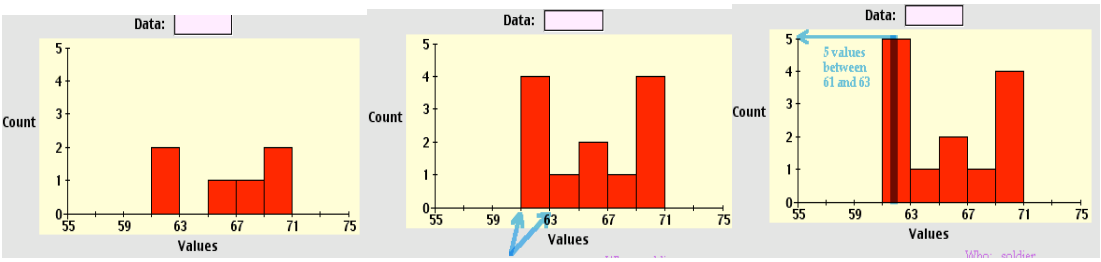


Figure 7.11. A sub-sequence of the displays from the ActivStats histogram animation.

After viewing this animation and a brief classroom discussion about the scale of the histogram axes and what these axes represented, the connection between histograms and the previously used stacked dot plots was made by Sarah. She saw a histogram as a dot plot of which

[20] *ActivStats* is a cross-platform interactive statistics course structured as a sequence of lessons in a virtual textbook. Each of the participants in this experiment received a personal copy of ActivStats to use on their home PC. ActivStats was used occasionally as a resource in this experiment, and was increasingly used in subsequent experiments involving the same student participants.

intervals between stacks had been covered with boxes: *"it's pretty much the same, if you put, like, boxes over those things"*.

The instructor moved to build on Sarah's image, explaining how a histogram's bars increase by one unit of area each time a data value is assigned to the interval that determines that bar. Thus, students "saw" how a histogram can be conceived as a natural extension of a stacked dot plot.[21] This idea provided a final segue into the next activity in the sequence.

Activity 7: Making Sense of Histograms

Activity 7, Part 1: Constructing a histogram

The next activity in the sequence was designed as a follow-up to Part 3 of Activity 6.[22] Part 1 of Activity 7 was assigned for homework, near the end of Lesson 11. Students' work was discussed in Lesson 12 (the next day).

The aim of this activity was to introduce students to histograms as a conventional way to depict distributions and the variability among sample percentages. In an effort to have students relate this inscription to the ideas of variability raised in Activity 6, the activity was set in the context of the "Favorite Musicians" scenario. It entailed constructing two histograms, by hand, of 200 sample percentages for samples of size 700 and 2000, respectively. Each was the percentage of people in a sample that selected "Dave Matthews" as their favorite music group. Students were supplied with this sampling data in a list (Figure 7.13), in addition to the written activity guide shown in Figure 7.12. The activity guide included a description of the sequence of menu selections that students needed to use in order to run the ActivStats computer animation showing how to construct and interpret histograms.[23]

---

[21] I do not mean to imply that students' exposure to the computer animation and the related discussion was sufficient for them to have internalized this imagery. Perhaps a better way to put it is to say that students were privy to a set of interactions that ostensibly primed them to develop the same imagery as Sarah.

[22] Activity 7 had actually been prepared before this lesson (11). The instructor steered the final phase of the classroom discussion in Part 3 of Activity 6 so as to segue naturally into Activity 7.

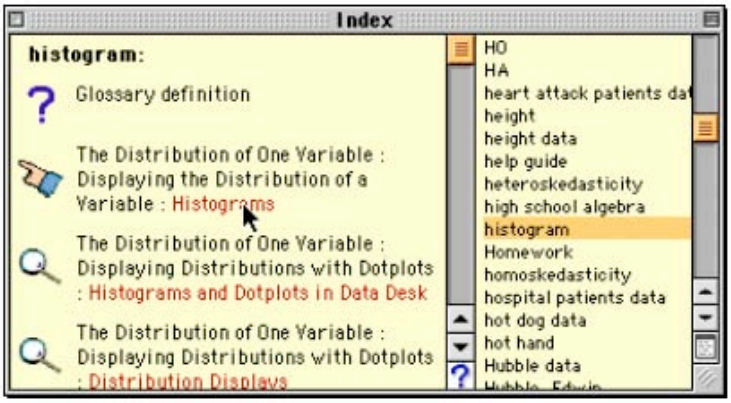
[23] This is the same animation shown in class of how histograms emerge from the aggregation and organization of randomly generated data values.

Using Histograms to Depict a Distribution

This assignment continues the "Favorite Musician" setting. It has a dual purpose – to get a sense for the variability among samples of various sizes and to introduce a conventional way of depicting that variability.

Your assignment is to make two histograms, by hand – one for each of the two sets of data. The first data set is of samples of size 700 taken from the Favorite Musicians data base; the second is of samples of size 2000 taken from the Favorite Musicians data base. The actual values are the percents of a sample that checked "Dave Matthews".

You can run the animation we viewed in class by going to your ActivStats CD, selecting Index from the Tools menu. Scroll down, click on Histogram, then click on the word "Histograms" (as shown below).



Then click on the Histogram icon (shown below).

Know how to make and interpret a histogram.
Histograms display distributions with bars that count

One of the first things we want to know about a quantitative
the range of common values.

Figure 7.12. The written guide for Part 1 of Activity 7.

200 Samples of size 700; Percent of each sample selecting "Dave Matthews"

55.14	58.29	55.86	55.57	56.71	58.14	56.29	55	54	55.57
53.86	57	58.29	55.29	55.14	56.43	53.71	53.14	59.29	59.29
50.14	56.86	52.14	54.43	52.71	53.86	57.71	56.57	57	56.29
52	57.57	55.43	55.71	53.14	53.29	54.71	58.86	58.86	58.57
55.14	54	57.14	53.43	60.86	59.71	57.86	54.57	53.57	54.29
58.57	56.57	57	54	57.71	56.71	52.57	58.43	57.29	55.29
56.71	55	55	55.57	56.43	53.14	56.71	59.86	54	57.43
56.71	55.71	55.29	55.14	54.71	54.57	55	55.29	59.86	55.57
53.86	56.57	56.71	52.71	56.86	52.29	54.29	60.29	52.71	54.14
54.57	55.71	57.71	58.43	56.14	56	56.43	57.29	55.71	54
52.86	55.86	58.43	56.86	56.29	59.14	54.43	56.14	58.29	53.71
57.14	52.29	59.29	57.29	56.14	54.71	54.86	54.29	55.57	56.29
56.86	53.86	57.71	54.43	55.29	55.43	56.43	55.86	55.14	53.29
56.57	57.14	55.29	56.57	54.57	55.43	53.57	55.57	53.71	55.29
53.43	54.57	58.43	54.57	52.57	54.86	54.71	56.57	56.14	53.57
57	56.57	54.57	56	52.29	56.86	55.14	58.71	58.71	57.71
53.43	53.86	57	53.29	54.86	55	58.71	52.71	54.14	54.71
60	57.57	56.86	55	55	54.14	57	55.14	53.71	57.14
57	58.71	58.29	56.71	52	57.71	58.57	58.86	57.14	56
53.43	55.14	58.29	58.14	58.71	58.29	50.71	57.14	59.43	54.57

Figure 7.13. One of the two simulated sampling data sets presented to students.

Constructing histograms of 200 sample percentages "by hand" is a tedious affair. However, the research team felt that it would benefit students to have the concrete experience—this one time—of organizing moderately large data sets. Indeed, it was intended to force them to grapple with issues entailed in representing a data set with a histogram (e.g., constructing classes of data values, deciding what interval size to choose, choosing a scale for axes, clarifying what each axis represents, constructing a descriptive title). This activity was also, in part, intended to prepare students for engagement in future activities in which histograms would be used extensively.

The research team had envisioned that the classroom discussion around this activity would unfold in two phases. Issues surrounding a histogram's construction and what it represents would be discussed first. The instructional agenda was to then move the discussion toward comparing histograms for the two sample sizes, with the aim of engaging students in thinking about what the histograms suggest about the variability and accuracy of samples of each size.

When Part 1 of Activity 7 was assigned, near the end of Lesson 11, the instructor led a brief discussion aimed at having students anticipate some of the issues they would consider when constructing the histogram. These issues included what range to select for the horizontal axis, what interval size and how many intervals might be needed. In addition, there was some discussion about where to place particular data values and the convention of placing boundary values in the higher interval was established.

Activity 7, Part 1 discussion highlights

It turned out that the classroom discussion in Lesson 12 never moved into the second intended phase because significant issues having to do with interpreting histograms emerged in the first part of the lesson (lasting 24 minutes) and the remainder of the lesson (lasting 12 minutes) was devoted to their discussion and resolution. This section focuses on the first part of Lesson 12, highlighting aspects of students' histograms and ideas that emerged within class discussion of them.[24]

Students' work in Part 1 of Activity 7 suggests that two related aspects were salient for them: 1) the process of organizing the data values one at a time into intervals, and 2) an image of corresponding histogram bars growing by small rectangular "slices". Indeed, almost all students'

[24] In contrast to the narrative strategy I employed in previous parts, I will not characterize the unfolding of these discussions here. Suffice it to say that the discussions were structured around showcasing three students' histograms; Luke, Nicole, and Tina each sketched one histogram on the blackboard and explained what it depicted.

histograms were evidently constructed in this "building-up" fashion and it came out in the classroom discussion that several students thought this was a necessary feature of a "correct" histogram. Figures 7.14 and 7.15 show two students' histograms which typify this universally shared feature.

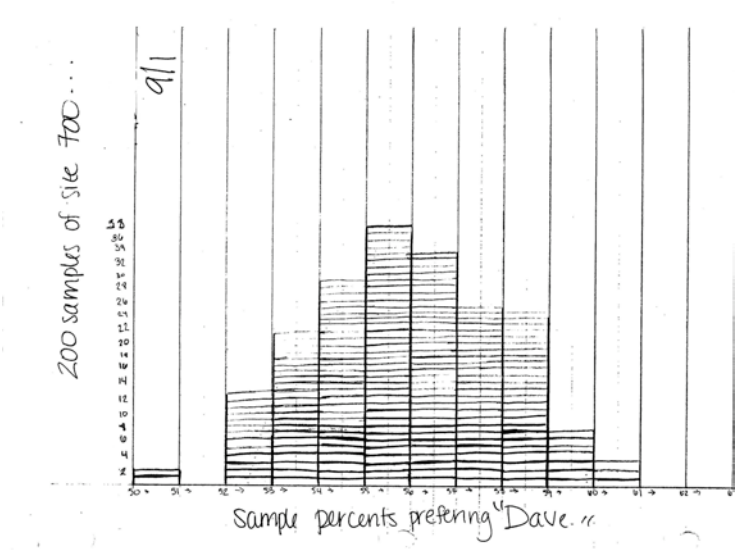


Figure 7.14. One of Nicole's histograms, a prominent feature of which are vestiges of its construction.

That these aspects were salient for students presumably reflects their having been impressed by the ActivStats histogram animation and by the part of the transitional discussion in which Sarah and the instructor related the construction of histograms and stacked dot plots.

Another prominent feature of several students' histograms was either a lack of or problematic labeling of the axes. The classroom discussions suggest that this was not merely an oversight on the part of students; in some cases this seemed to be a reflection of their level of engagement with the task. Some students appeared not to have attended to the meaning of the values in the data tables (see Figure 7.13), thus raising the possibility that they had proceeded in a somewhat mechanical manner when constructing their histograms. In other cases, students' labeling of axes suggest their difficulties in making coherent sense of the underlying re-sampling process. For instance, in Tina's histogram a slice of a bar represented a sample percent. However, she labeled the vertical axis of a histogram as "the number of votes" (see Figure 7.15). I take this inconsistency as evidence that Tina tended to confound a *number of samples* with a *number of people* in a sample who selected Dave Matthews as their favorite music group.

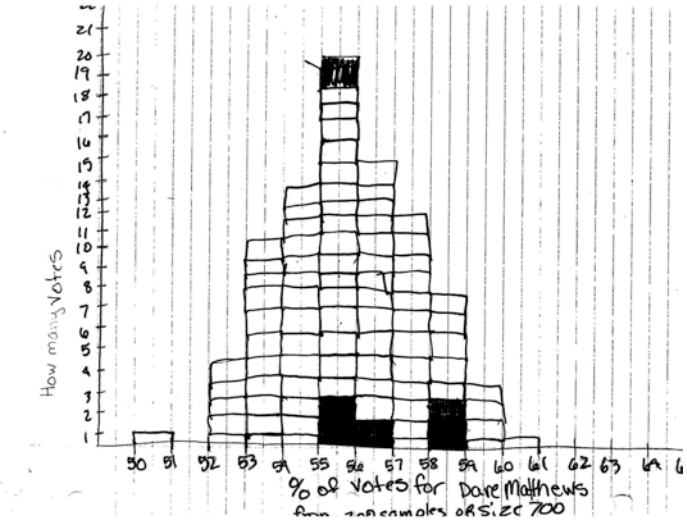


Figure 7.15. Tina's histogram for samples of size 700.[25]

A similar confounding is illustrated in the following brief discussion excerpt focusing on the histogram that Luke had sketched on the blackboard (Figure 7.16). The discussion was about creating descriptive labels for histogram axes.

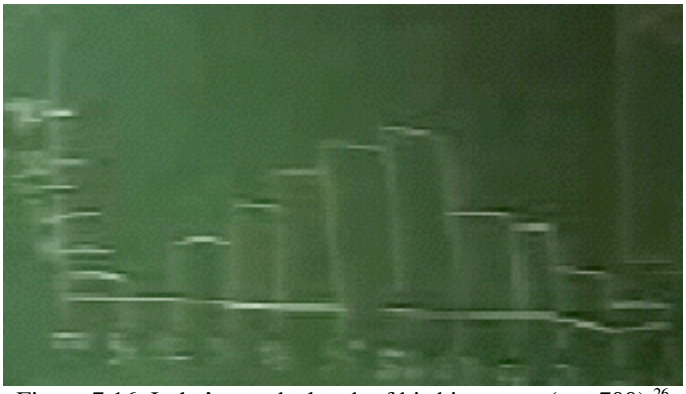


Figure 7.16. Luke's rough sketch of his histogram (n = 700).[26]

Episode 1, Lesson 12:

1. I: How would we label this axis (runs hand along unlabeled horizontal axis of Luke's graph—see Figure 7.16). Luke? How would we label the horizontal axis? What do these numbers stand for? That's the way to label it, to say what these numbers stand for.
2. Luke: It's the intervals, but

[25] The bottom line of Tina's label for the horizontal axis reads "from 200 samples of size 700".

[26] Unlike his paper copy, Luke sketched his histogram on the backboard without constructing each bar by accumulating unit slices. Instead, he seemed to make a rough sketch by eyeballing its overall shape in the paper copy. This suggests that Luke had interiorized the process of constructing the histogram.

3. I: No, this number right here stands for (points to 52 on horizontal axis), any number on here stands for what?
4. Luke: Oh, it's 52, uhh out of a sample of 200, it's each shown
5. David: (inaudible)
6. Luke: Right, out of the sample—what?
7. Tina: Of any interval
8. David: a particular sample
9. I: Ok, so any one number stands for, in general, stands for?
10. Luke: The number of people in that sample
11. I: Not number of people
12. Luke: The percent of people in a sample, that preferred Dave Matthews.

The excerpt illustrates Luke's struggle to make sense of the numerical values on the horizontal axis; his first inclination (line 4) was to think of "52" as 52 people out of a sample of size 200. However, 200 is, in fact, the number of samples selected and not the sample size. Thus, Luke tended to confound sample size and number of samples selected. Constrained by his interaction with the instructor, Luke seemed to move toward some clarification (lines 9-12), by first thinking of "52" as an absolute number of people and finally as a percent of people in a sample. However, it remains unclear whether Luke resolved his ambiguity between sample size and number of samples selected.

Another common feature among most students' histograms was an impoverished title, or no title at all. A typical histogram title, when provided, was a relatively dull label (e.g., "200 samples of size 700") that did not adequately capture the richness of the underlying scenario or clearly describe what the graph depicted.[27] The classroom discussions in this and the activity that followed it show that constructing a descriptive title for the histograms was highly problematic for students. Their attempts, replete with false starts and hesitations, suggest not only that they were uncertain about what a histogram described but also that they were not oriented to seeing a histogram as embodying an underlying re-sampling process. The next discussion excerpt illustrates this point; it foreshadows the difficulties that many students would experience in teasing apart the different quantities and levels of the underlying re-sampling scenario. The excerpt, lasting approximately 2.5 minutes, centers on Nicole's attempt to explain what her

[27] It is a challenge to construct a clear and coherent description of what such a histogram shows. A reasonably coherent and sufficiently descriptive title might be: *The frequency distribution of 200 sample percents, each of which is the percent of people who selected "Dave Matthews" as their favorite musician in a random sample of 700 people*.

histogram depicted.[28] Nicole had sketched her histogram of 200 sample percents, for samples of size 2000, on the blackboard. The instructor then asked her to explain to the class what her histogram showed.

Episode 2, Lesson 12:

1. Nicole: […] well it's kind of self-explanatory after that, and then numbers between 53 and 54 six times (points to bar above this interval). 54 and 55, twenty five times (points to bar above that interval). The greater numbers (chuckles), between 55 and 56 'cause I got that seventy three times (points to tallest bar). 56 and 57, fifty two times, 57 and 58 thirty five times. 58, 59, nine times, and 59, 60, one time.
2. I: what are those numbers?
3. Nicole: What are those numbers?
4. I: Yeah.
5. Nicole: That's the percent that chose Dave Mathews.
6. I: The percent of what?
7. Nicole: Percent of (pauses)—2000! (2 second pause) I think. They polled, ok they checked, they asked 2000 people who, like which was their favorite band and this percent (points to first interval on graph), uhh wait, ok (pauses) between, ok
8. (Peter chuckles at Nicole's difficulty)
9. Nicole (laughs): no, it is 25
10. Peter (facetiously to Nicole): Think!!
11. Nicole (to I): 25 times they got between 54 and 55%. Is that what you're trying to say?
12. I: Of what?
13. Nicole: Of what what?
14. (Peter and Nicole laugh)
15. I: 56% is 56% of something. So what is the something that that 56% is of?
16. Nicole: Out of this total, 2000! (points to histogram bar above the interval [54,55])
17. I: It's the same 2000?
18. Nicole: Out of how many people they—
19. I: So there was just <u>one</u> group of 2000?
20. Nicole: No. They did 200—wait, no there's 200 samples of 2000.
21. I: All right. Ok, so one number, fifty six point, say 56.3% is 56.3% of what?
22. Nicole: Of the 2000—no! Wait, I don't know!
23. I: Of one of those 200 groups.
24. Nicole: Right. So those should add up to 200 (points to histogram bars)
25. I: Yeah. It's of <u>some</u> group of 2000.
26. Nicole: Ok.

[28] The histogram that Nicole's sketched on the board differed from her paper copy in that it contained no vertical axis. Instead, Nicole inscribed a value in the upper part of each bar to denote the number of samples it represented. A picture of Nicole's classroom sketch is not included here because the image was barely discernible in the video camera.

27. I: Not the group of 2000, 'cause there's not one group of 2000.
28. Nicole: No.
29. I: You see the difference?
30. Nicole: Yes.

At the start of the excerpt (line 1) Nicole's description was exclusively in terms of numerical values. This numerical description suggests her having been oriented toward a calculational rather than a quantitative interpretation of the graph (Thompson et al., 1994; Cobb, 1998)—she did not appear inclined to unpack the numbers and speak of the histogram in terms of values of underlying quantities. The instructor evidently picked up on this and asked Nicole to describe what the numbers represented (lines 2-4). Nicole's response in lines 5-7 indicates that she was aware that the numbers represented percents of people who had selected Dave Matthews as their favorite band. But when pressed further, Nicole began to exhibit doubts about precisely what each percent was a percent of—as though she was unsure of the underlying quantity. Nicole had an inkling that each number was a percent of 2000 people, but she had to re-imagine the sampling scenario in order to convince herself (line 7).

In line 11, Nicole reverted back to a numerical mode of description that leaves it unclear whether she understood the "25 times" as 25 samples of 2000 people each. The instructor pressed Nicole further for clarification about the meaning of 56%, suspecting that she might have thought there was only one sample of 2000 people. What ensued from that point on was a gradual unraveling of the scenario for Nicole; at first she seemed certain that there was not one, but 200 samples of 2000 people selected. However, she continued to speak of "the 2000", as though there was only one sample, or as if she were only able to focus on one sample at a time. Nicole then seemed to confound the 200 and 2000, as if she was losing track of which is a number of samples and which is a number of people in a sample. Nicole eventually lost her sense of grounding and became unsure of what the 56% is a percent of (line 22).

The excerpt illustrates a kind of instability that many students experienced in their thinking when trying to interpret the histogram in terms of the re-sampling scenario. When asked questions that required them to shift their focus from one item or quantity (e.g., number of samples) to another (e.g., number of people in a sample), students would easily loose sight of which was which and would confound items. They seemed to lack a well-coordinated set of

images of each part of the scenario that could withstand shifts in their focus of attention between the parts.

Activity 7, Part 2: Interpreting a histogram

The research team had anticipated that students might find it problematic to interpret histograms of sampling distributions. The instructor thus came to the lesson prepared with an activity designed to engage students in interpreting histograms. The classroom discussions around this activity constituted Part 2 of Activity 7 (see Figure 7.17).

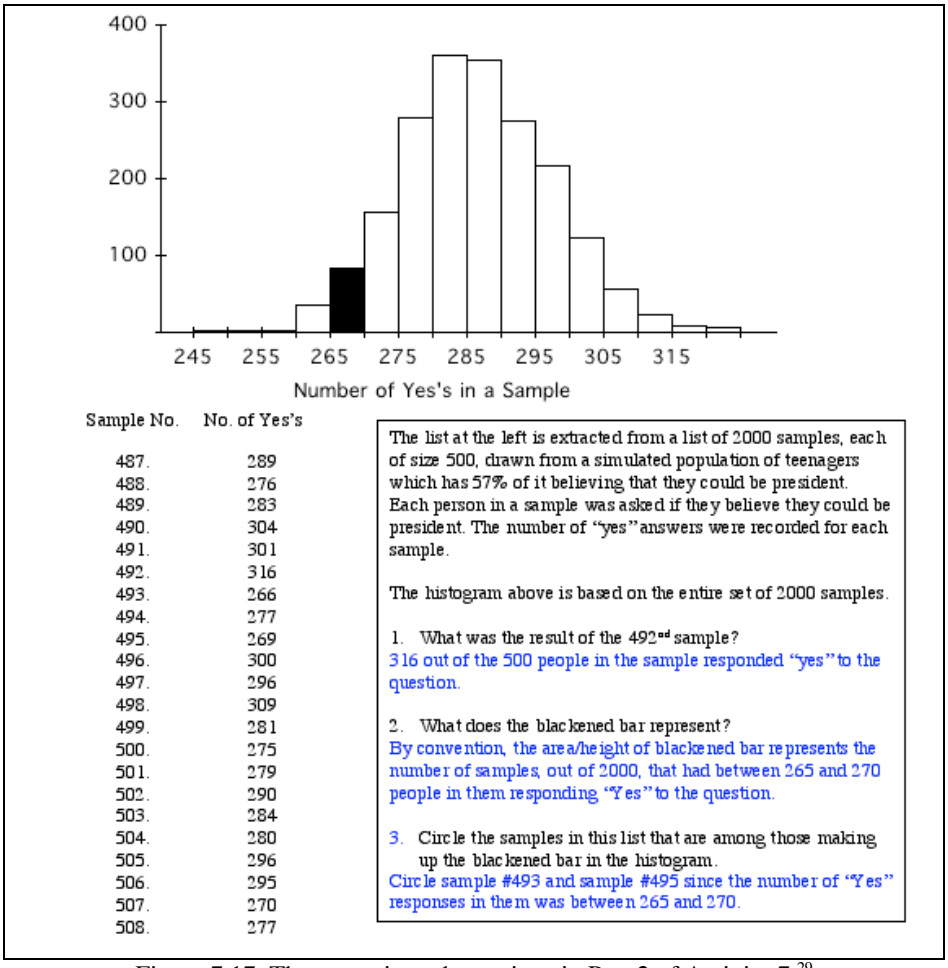


Figure 7.17. The scenario and questions in Part 2 of Activity 7.[29]

The activity described a repeated sampling scenario and it presented simulated sampling data and a corresponding histogram about which students were asked questions. The sampling data

[29] The blue text is not part of what students received. Instead, these are "model" responses expected from someone having an unproblematic understanding of the histogram and its relation to the described scenario.

was presented in a two-column list; values in the left-hand column were ordered indices of each of a subset of 2000 simulated samples drawn. Corresponding values in the right-hand column were of the outcome of each sample (i.e., the number of "Yes" responses in that sample of 500 people asked the same question). The histogram showed the frequency distribution of the 2000 simulated sampling outcomes (frequencies).[30] The histogram and its vertical axis were deliberately untitled. Only the horizontal axis was labeled. The histogram bar representing the number of samples in which between 300 and 305 people responded "Yes" was highlighted in dark ink.

The three questions in the activity guide were intended to help students build their understanding of the histogram by having them work to interpret features of the graph in relation to the meaning of the data values in the list. Another task, not shown in the activity guide, was to create a descriptive title for the histogram.[31]

Activity 7, Part 2 discussion highlights

The classroom discussions around Part 2 of Activity 7 unfolded over segments of two consecutive lessons. The class spent approximately the last 12 minutes of Lesson 12 discussing and answering the activity questions. Students then completed the questions, in writing, for homework. During Lesson 13 (the next day) the activity was re-visited in a discussion lasting approximately 12 minutes. This section highlights parts of those discussions and students' conceptions that are evidenced within them.

The instructor steered the opening discussion in Lesson 12 toward having students understand the scenario at hand. Those students who participated in this discussion were able to clearly distinguish what the values "500" and "2000" represented. No evidence emerged in the opening discussion to suggest that students had problems understanding the scenario. The discussion then turned to the questions of how to name the histogram and what the blackened bar represents. What unfolded were protracted and messy discussions that contain even more

[30] A similar configuration would be used in later activities of the instructional sequence. Thus, engagement in Part 2 of Activity 7 was also intended to prepare students for those activities.

[31] I suggest the following title as a reasonably clear and coherent description of what the histogram depicts: *The frequency distribution of 2000 values of a sample statistic. The statistic is the number of people in a sample of 500, randomly selected from a common population, that answered "Yes" to the question "Do you believe you could be President?".*

pronounced evidence that students were struggling with issues similar to those that emerged in Part 1 of Activity 7.

In order to make a compelling case for the robustness and persistence of students' difficulties, I include a rather lengthy sequence of representative discussion excerpts spanning the two consecutive lessons. Each excerpt is followed by summary analyses. The first two excerpts focus on the task of naming the histogram. The second two excerpts focus on that of interpreting the meaning of the highlighted histogram bar. Taken as a whole, these representative episodes provide telling evidence of students' difficulties.

The first discussion excerpt lasted approximately 2 minutes and follows below.

Episode 3, Lesson 12: Naming the histogram:

Segment 1: Peter's attempt

522. I: […] how might we label this histogram?" To give it a name, to say what it's showing.
523. Peter: Uhh, (hesitates) the sample size
524. (Nicole laughs at Peter's apparent sluggishness)
525. Peter (continues): if they are, 500 samples
526. David: No, that's wrong
527. Peter: 2000 samples of size 500 of the percentage, no not percentage. 2000 samples of size 500 of people that said that they wanted to be President. That they thought they could be President. Something like that.

Segment 2: David's attempt

535. I: All right, David you want to add to that?
536. David (looking at his handout sheet): Histogram. This is a histogram of uhh 2000 samples, sizes of 500 teenagers that think they can be President. Or that uhh believe they could be President.
537. Peter: That's what I just said.
538. I: Ah, ok. So is this histogram about teenagers?
539. David: Yes
540. I: Is that what is in the histogram? Teenagers?
541. Nicole: Yeah
542. Tina: Yes
543. Female student: percentage
544. I: I don't think so
545. Tina: No?
546. Nicole: That's what it said!
547. David: It's a population of teenagers
548. Tina: It's yes and no
549. I: It's samples of teenagers.
550. Nicole: Oh.
551. I: What's one unit in here? Teenagers?

552. David: Did I not say teenagers?
553. Tina: 500. No, 500 teenagers
554. David: I said samples of teenagers, didn't I?
555. I: Yes. One unit in here is 500 teenagers.
556. David: It's what, that's what I said.

In the first segment, Peter begins sluggishly, as though he isn't sure how to fit the information together; he knows there is a sample size involved, but he seems to have interpreted the "500" as the number of samples (lines 523 and 525). Peter eventually begins to piece things together coherently (line 527): there are 2000 samples, each of size 500. Peter quickly corrects himself, realizing that the statistic is not a percentage. His second sentence (line 527) suggests he understood that people answered "yes" to a question, and that there were 2000 samples of 500 people asked. However, his parsing makes it ambiguous as to who answered "yes" to the question; was it just people or *some* number of people in each of the 2000 samples of 500 people? In other words, it is questionable whether Peter conceived the scenario as having a structure like this: *there are 2000 values, and each one represents the number of people who answered "Yes" in a distinct group of 500 people asked the question*. It is equally questionable whether Peter conceived the histogram as embodying this structure.

In the second segment David was invited to contribute his ideas; he ended up repeating Peter's description almost verbatim. The ambiguity in these students' descriptions incited the instructor to ask for clarification: "is the histogram about teenagers? … Is that what is in the histogram?" (lines 538 and 540). Several students' automatic "yes" responses are telling; they were thinking of the histogram as showing teenagers, rather than groups of 500 teenagers. Moreover, students seemed oriented to consider this distinction *only* under prompting by the instructor (line 544). Nicole's utterance in line 546—"that's what it [scenario description] said!"—suggests she was confounding what constitutes a unit in the histogram with what constitutes a unit in the sampled population. In other words, she was not distinguishing between "people" as unit items in the population and "group of 500 people" as unit items in the histogram. Nicole seemed surprised when the instructor explicitly made this distinction (line 550). The remainder of the segment (lines 552-557) shows David flitting back and forth between thinking that he mentioned people and groups of people in his description, thus suggesting an instability in the way he conceived the sampling scenario.

Here is a brief excerpt from the Lesson 13 discussion, in which the same question was revisited.

Episode 1, Lesson 13: Naming the histogram

65. I: What did you name the histogram?
66. Chelsea (reads her description): A histogram showing the number of teenagers who answered "yes" to the question "do you believe you can be President?" of the 2000 samples of size 500.
67. I: Say that again.
68. Chelsea (repeats her previous statement verbatim):
69. I: Could you repeat what she said, David?
70. David: Oh, No I can't do that (inaudible)
71. Peter: Do you want me to repeat what she said?
72. I: Yeah, I'm uhh, that's one way to uhh, that we're uhh to ensure that we're communicating is if someone can repeat what the other person says.
73. Peter: Uhh, she said a histogram uhh of teenagers responding "yes" to the question "do you think you can be President?" of (hesitates) 2000 samples of size 500.
74. I (to Chelsea): Is that what you said?
75. Chelsea: Basically […] yeah
76. I (to Chelsea): Ok, equivalent. All right. It's a histogram of teenagers—that, is that correct?
77. Chelsea: Yeah
78. Nicole: Hmm hmm
79. Tina: That's what I put down.
80. I: So one unit in that histogram is a teenager.
81. Tina: No, 500 teenagers.
82. I: All right. So it's not a histogram of teenagers. […]

The excerpt illustrates Chelsea's ambiguous parsing of the description of the histogram; she evidently thought of the histogram as showing numbers of teens that answered "Yes" to the question. But it is unclear how the number of samples and the sample size fit into this scheme for her. Peter followed suit with an almost verbatim restatement of Chelsea's description, and Chelsea confirmed that it was essentially the same as her description. Chelsea, Nicole, and Tina all agreed that the graph is a "histogram of teenagers" (lines 77-79). It wasn't until the instructor asked whether a teenager constitutes one unit in the histogram that Tina began to make a distinction between teenagers and groups of 500 teenagers. This interchange illustrates that students were prone to smudging over the distinction between people and groups of people when thinking about the histogram and the underlying sampling scenario.

The next two excerpts illustrate the recurrence (in the first excerpt) and, for some students, the resolution (in the second excerpt) of these issues within discussions of Question 2: "What does the blackened bar represent?". The first excerpt lasted approximately 1 minute and comprised the tail end of the Lesson 12 discussion.

Episode 4, Lesson 12: Interpreting the highlighted histogram bar:

Segment 1

623. I: Ok. What does the blackened bar represent?
[…]
624. Tina: What does the blackened bar represent?
625. I: Yeah.
626. Tina: That uhh (interrupted)
627. Kit: "yes"s
628. Tina: Hmm?
629. Kit: It's the number of "yes"s between 265 and 700, uhh 270
630. I: Go ahead Kit, say that again.
631. Kit: The number of "ye—yes" responses to the question between 265 and 270
632. I: Ok, not quite. It's not the number of "yes"s between 265,
633. Tina: percent
634. I (continues) it's a number of samples. Remember, each thing is, one unit is a sample, not a "yes", it's a sample
635. Luke (simultaneously with I): [inaudible] the number that responded

Segment 2

636. David: Number of samples between 265 and 75
637. I: No, it's not the number of samples between 265 and 270
638. David (continues): that said "yes"
639. I (continues): it's the number of samples that had between 265 and 270 students in it saying yes.
640. Luke: Out of 500
641. I: out of 500
642. David: number of samples that had
643. Luke: All right, I follow you now. I, I'm with you (inaudible)

The first segment illustrates that Kit took the highlighted bar to represent a number of "yes" responses to the question, rather than a number of samples. This was not a mere slip of the tongue, as her re-iteration of this description in line 631 suggests. Kit's tagging of the statement "between 265 and 270" to her claim is difficult to make sense of; she might have understood "between 265 and 270" to be the range of the number of "yes" responses in a sample, in which case hers would appear to be a one-dimensional interpretation of the histogram. That is, Kit's

interpretation would differ from a conventional interpretation in that it did not distinguish the histogram bar (or its height) from its defining interval on the horizontal axis. It's as though Kit conflated the two entities. Put in quantitative terms, Kit's interpretation did not entail the coordination of two quantities—a number of "yes" responses in a sample, or range thereof, and a number of samples containing that number of "yes" responses. Now, the latter quantity may well have been non-existent or non-salient for Kit, in which case there would be nothing to coordinate. This would be consistent with Kit's evidently salient image of numbers of "yes"s in *one* sample. Another possibility is that Kit took the label of the horizontal axis as a description of the histogram, so that each bar was seen as a number of "yes"s in a sample.

The instructor responded to Kit by alerting her to the convention that a number of samples, rather than a "Yes" response, constitutes one unit in the histogram. In the second segment we see David struggling, as he had in other instances, to fit these ideas together coherently (lines 636 and 638). His attempted description, if taken literally, would suggest that he thought the highlighted bar shows a number of samples in which every person responded "Yes" to the question. The instructor immediately countered with his description (lines 639), trying to clarify that the range "265 to 270" denotes numbers of students who responded "Yes" in the number of samples represented by the highlighted bar. Luke's reaction (line 643) suggests that this was a clarification for him.

The next, and final, excerpt is from discussions of the same question that occurred again in Lesson 13.[32] The excerpt lasted approximately 2 minutes.

Episode 2, Lesson 13:

Segment 1

83. I: Ok. Uhh "what does the blackened bar represent?" Nicole, uhh no, I'm sorry. Kit?
84. Kit: The number of samples that have between 265 and 270 that said "yes"
85. I: Very good. So those, are there any others that between 265 and 270? I mean, are there any other samples that had between 265 and 270 other than those that make up the blackened bar?
86. Kit: No (softly)
87. I: Ok, so that's all of them. Ok, good.

[32] In the intervening time between Lessons 12 and 13, students produced written responses to the activity question as part of a homework assignment. Half (4) of these descriptions were clearly consistent with the interpretation promoted by the instructor in the class discussions. One student did not respond to that question only, suggesting that it was problematic for him. The other three descriptions suggest a fixation on thinking that a histogram bar represents a number of people or number of responses.

Segment 2

88. Nicole: So is it the number of samples or the number of "yes"s?
89. I: Oh, good question. Is it the number of samples or the number of "yes"s?
90. Resounding collective response: number of "yes"s in each sample
91. Kit: It's the number of samples
92. Female student: that had that many
93. Luke: that responded "yes"
94. David: It's the number of samples that had that many "yes"s.
95. I: That's right David. Very good! It's the number of samples that had that many "yes"s—between 265 and 270 "yes"s.
96. Nicole: Oh (adds this to her written response)

Segment 3

97. Tina: Can I read mine and you tell me if I worded it right, please?
98. I: Go ahead.
99. Tina: Ok, "Out of the 2000 samples drawn from each sample, between 265 and 270 students answered 'yes' to the question".
100. I: In ea—ok you, what you just said claimed that in every sample, between 265 and 270 students answered "yes". Is that true?
101. Tina: Probably
102. I: In all, in every one of those 2000 samples, between 265 and 270 said "yes"?
103. Tina: Yeah
104. I: No sample had 300 saying "yes"?
105. Tina: Ok, I, yeah I get it now (sighs in apparent frustration)
106. I: Ok. So it's not in every sample. It's just, those are the samples in which that happened. (3 second pause)

Segment 4

107. I: Sarah, what did you write for that one?
108. Sarah: What number?
109. I: Uhh, two.
110. Sarah: That there are just under 100 samples that had between 265 to 270 that said "yes".
111. I: in each sample?
112. Sarah (nods in affirmation)

In the first segment we see a significant change, from the lesson before[33], in Kit's interpretation of the histogram's highlighted bar. Here (line 84), Kit gave a coherent and succinct description of what the bar represents, with an apparent understanding that it stands for *all* samples drawn that had between 265 and 270 people in them responding "yes" to the question.

[33] See Segment 1 of Episode 4, Lesson 12.

In the second segment we see Nicole continuing to experience uncertainty (line 88) about how to interpret the histogram bar, despite having participated in numerous sustained classroom discussions in which the conventional interpretation had repeatedly been made public. Line 90 of the discussion indicates that a number of students evidently continued to think that the histogram bar represents a number of "yes" responses in each sample. Kit was unwavering in her new conviction (line 91). David (line 94) now appeared to have developed the same coherent interpretation as Kit—a stark difference from the significant difficulties he had previously experienced.[34]

In the third segment Tina's description suggests that she thought each and every one of the 2000 samples drawn contained between 265 and 270 students in them who responded "Yes" to the question. Tina's subsequent exchanges with the instructed show that she evidently continued to have a very problematic interpretation of the sampling scenario. Moreover, Tina's opening question (line 97) raises my suspicion that she had construed the task as one of "wording it right"—that is, describing things "correctly"—rather than one of making sense of the scenario.

The fourth segment illustrates Sarah's descriptions (line 110), one of the few unproblematic ones; it is coherent, succinct, and entails a rough estimate of how many samples the bar represents.

To summarize, representative highlights from the discussions around Part 2 of Activity 7 indicate that many students were strongly oriented to thinking of the histogram as showing something about people rather then samples of people. Their interactions with the instructor show that, when pushed, they could distinguish between the two entities. However, students' attempts to tease apart images of people and samples of people were fraught with a persistent slipperiness that suggests they did not know why or when they should think of one rather than the other.

A plausible hypothesis is that students were generally insensitive to the *significance* of the distinction between thinking of people and thinking of samples of people when interpreting the histogram and the accompanying sampling scenario. It is as though students thought that "2000 samples of 500 people is still just people" (see Figure 7.18).[35] Moreover, students' orientation to

[34] This is also starkly different from David's written description, which clearly suggests he was oriented to interpreting the histogram bar in terms of people rather than samples.

[35] This hypothesized conception is remarkably reminiscent of a line of reasoning that emerged in an earlier phase of this experiment. Recall Lesley's approach, in Activity 3 of Phase 1, of apparently *dissolving* a collection of distinct

thinking of individuals in the population—a kind of default schema to which they seemed to easily revert—is consistent with their having assimilated the described resampling scenario and the histogram into an image of sampling as drawing individual items from a population.

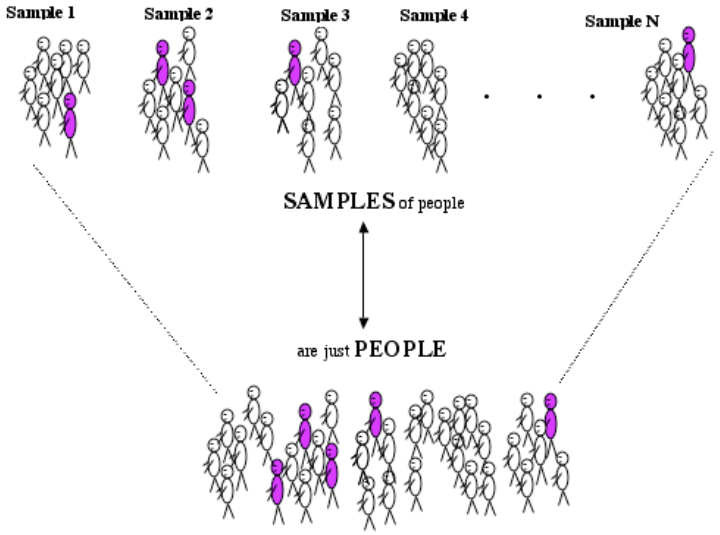


Figure 7.18. Two different structurings of a collection of people, the significance of which I hypothesize was not salient to students.

If this hypothesis is viable, it would suggest that students had not conceptualized the sampling scenario described in Part 2 of Activity 7 as having a hierarchical structure that emerges from a two-level scheme of conceptual operations centering around the images of repeatedly sampling from a large population, recording a statistic's value, and aggregating a collection of the statistic's values (Saldanha & Thompson, 2002):

Level 1 image: Randomly select items from the population to accumulate a *sample* of items of a given size. Record the value of a sample statistic of interest.

Level 2 image: Repeat Level 1 process a large number of times to accumulate a collection of values of a statistic.

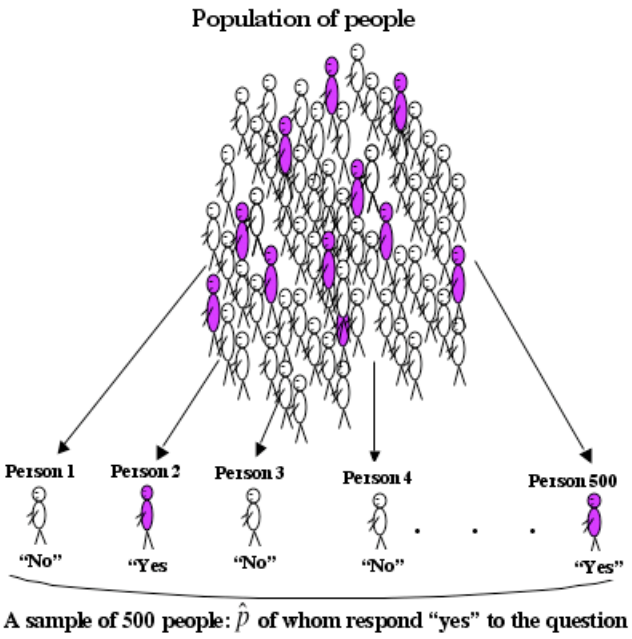
In this structuring, there is a clear distinction between individual items in the population and samples of those items as repeatable units of selection. Moreover, the focus of one's attention is on samples and the quantification of an attribute of them, whose values then aggregate as the

---

samples into one large sample (see Figure 7.18). In doing so, Lesley appeared to smudge over the distinction between a collection of items from the population and their organization into a collection of distinct and equal-sized samples. An intriguing question is whether a similar underlying image was at play in these students' conceptions as well.

process is repeated. Figure 7.19 depicts these levels pictorially, in terms of the particular context of the sampling scenario described in the activity.

Level 1 image:



Level 2 image:

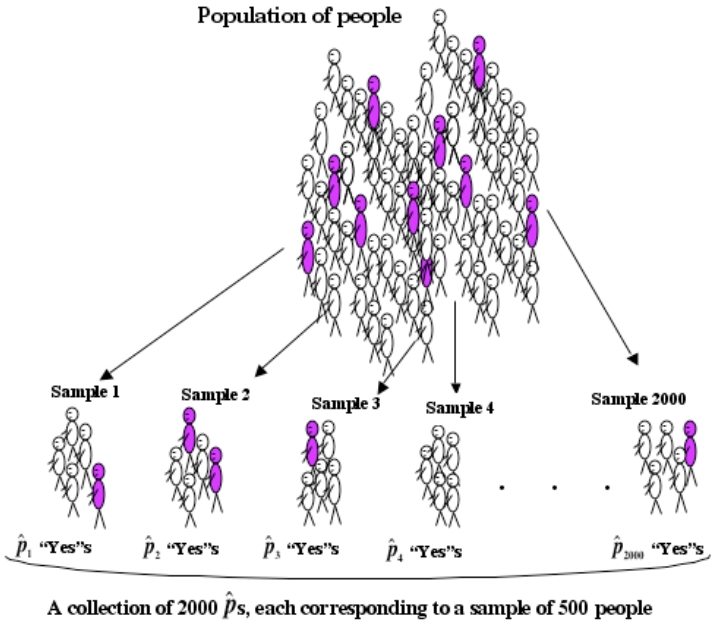


Figure 7.19. Two levels of imagery entailed in conceiving the sampling scenario as a collection of sample statistics.

What students struggled with in the activity was reifying a *sample* of people as a repeatable object of selection. They appeared to experience difficulty moving beyond thinking of selecting

individual population items to selecting entire groups of them. In their attempts to negotiate that move, they tended to slip into a "pre-Level 1" image, the focus of which was individual items that comprised the population.

Chapter Summary

The instructional activities employed in Phase 3 of the teaching experiment emerged out of the research team's efforts to help students extend their sense of sampling variability to ideas of distribution.

Activity 6 unfolded in a sequence of 3 interrelated parts that engaged students in investigating collections of sample percents, with the aim that they broach ideas of long-run sampling accuracy and representativness. The context of Activity 6 framed the issue of the investigations as one of balancing competing tensions between two constraints: minimal sample size (cost) and maximal sampling accuracy.

Discussions across all parts of Activity 6 suggest that images of proximity or "degree of closeness" were salient in students' ideas of sampling accuracy and representativeness. Two prominent ideas emerged out of students' engagement in Part 2 of Activity 6, each suggesting a different orientation to thinking about relationships between proximity and sampling accuracy:

Peter's idea to collapse a collection of sample percents into their average suggests an attempt to deal with aggregation and variability among its constituents by reduction to a single numerical index. By comparing the proximity of this average to the underlying population percent, Peter was able to determine how well samples of that size reflect the sampled population. This approach resonated with at least one other student, who suggested applying it to collections of samples of various sizes as a way of investigating whether samples of one size are more accurate than those of another size.

In contrast, Nicole's idea of selecting many samples of a given size and looking for patterns in the proximity among their percents suggests a focus on the aggregate as an object of attention and scrutiny. This idea was promoted in instruction as an imagistic basis for thinking about sampling accuracy in terms of how a collection of sample percents' values are clustered or "packed" within a range of possible values. This proto-distributional idea resonated with several students; they appeared to identify sample representativeness and variability with the "closeness together" of a collection of samples percents.

Discussions around Part 3 of Activity 6 revealed that most students experienced tremendous difficulty making sense of the activity. Numerous students did not understand the overarching issue of the activity as the research team had intended because they could not conceive cost of sampling as a real constraint. Moreover, there are suggestions that a number of students interpreted the activity by assimilating it into a non-stochastic view of sampling, in which the goal is to select a single sample of appropriate size that will yield an "accurate" result—that is, a value of the statistic that is as close as possible to that of the true underlying population parameter.

Recasting the issue in Part 3 of Activity 6 in terms of a point of diminishing returns metaphor appeared to make it more accessible for students. Nevertheless, students' written work suggests that the issue continued to be somewhat intractable for them, seeming to entail too many pieces to manage and fit together into a coherent whole.

Due to students' difficulties in interpreting Part 3 of Activity 6, the activity appears not to have been useful for helping extend their emergent ideas of accuracy into a statistical sense of accuracy.[36]

The instructional interactions leading into Activity 7 oriented students to the use of stacked dot plots and frequency histograms as ways to organize and depict collections of randomly generated sampling data.

Activity 7 introduced students to the frequency histogram, with the intention that they interpret it as a conventional way to depict distributions and variability. Discussions around Part 1 of Activity 7 indicate that a salient image for students was that of the process of constructing a histogram from the set of numerical data values provided. Moreover, students were not oriented to "step back" from this process to interpret the histogram as embodying an underlying re-sampling process and an emergent sampling distribution.

Indeed, discussions around both parts of Activity 7 provide striking evidence that students experienced robust and persistent difficulties in interpreting the histograms as showing

[36] In a statistical sense of accuracy the focus is not on the proximity of *particular* sample percents to the population percent. Rather, the focus is on the proportion of sample percents, for samples of a given size, that are expected to lie within various ranges of the sampled population percent *over the long run*—as such sample percents accumulate under repeated sampling. This distinction has to do with the difference between quantifying an attribute of an *individual* sample, in the former case, and quantifying an attribute of a *collection* of samples, in the latter case. Moreover, a statistical conception of accuracy is deeply tied to ideas of distribution. In the next chapter I elaborate this connection and hypothesize what essential elements are entailed in such a conception

something about samples. Students tended to confound people with samples of people and easily slipped into thinking that the histograms showed the latter rather then the former.[37]

A concluding hypothesis is that students did not conceptualize the sampling scenarios associated with the histograms as entailing a repeated two-level sampling process that gives rise to a collection of sample statistics, so that one sample is seen to correspond with one unit in a histogram bar.

As a final remark, students' widespread difficulties interpreting histograms in Activity 7 should caution us strongly against over-attributing their earlier images of a cluster of sample percents' "closeness together" to a sense of distribution. Instead, these difficulties give reason to suspect that those earlier images were not nearly elaborate enough to support seeing histograms as also depicting the density of sample statistics' dispersions.

[37] Given the evident difficulties observed among these students in interpreting the frequency histograms in Activity 7—even those they had seemingly constructed unproblematically themselves—it is natural to reflect on what prior experiences might have textured their difficulties. We know from their work in Algebra II, that students had already encountered frequency histograms prior to their engagement with Activity 7. Given the Algebra II curriculum, however, I presume that such histograms were not of distributions of sample statistics, but instead showed how elements within individual samples are distributed. Additionally, a cursory inspection of current popular news publications, which students presumably frequently encountered outside of school contexts, indicates that histograms presented in such venues typically display the distribution of elements within an individual sample (such as one drawn for a poll). Such prior experiences would seem to jibe with the difficulties documented here: it would be natural for students to assimilate the frequency histograms of Activity 7 into these presumed prior experiences with histograms, which typically describe non-stochastic aspects of sampling. Indeed, I assert that students' difficulties are not attributable simply to problems in understanding frequency histograms *per se*, but rather to conceiving frequency histograms as depictions of distributions of sample statistics.

CHAPTER VIII

PHASE 4: MOVE TO QUANTIFY VARIABILITY AND EXTEND DISTRIBUTION

This chapter describes a final sequence of 3 activities that unfolded over Lessons 13 through 18 (see Figure 8.1). Broadly speaking, these activities extended the explorations of Phase 3 by aiming to move students toward developing a sense of distribution that is intimately bound up with quantifying the variability among sample percents. Specifically, Activities 8 and 9 engaged students in systematically exploring relationships between sampling variability, population parameter, and sample size. Activity 10 sought to have students develop a statistical interpretation of the notion of margin of error.

Phase 4:	**Move to quantify variability and extend distribution**	
Lesson	**Activity (A)**	**Duration**
13 (09/03)	Preliminary discussion to A8: quantifying variability	12 m.
14 (09/07)	A8: Investigating effect of p on sampling variability—Discussion 1	25 m.
15 (09/09)	A8 revisited—Discussion 2 A9: Investigating effect of n on sampling variability—Discussion 1	25 m. 25 m.
16 (09/10)	A9 revisited—Discussion 2	50 m.
17 (09/13)	A10: Margin of error—Discussion 1	16 m.
18 (09/14)	A10 revisited—Discussion 2	22 m.

Figure 8.1. Chronological overview of instructional activities of Phase 4.

The chapter begins by elaborating a conception of distribution targeted in instruction, highlighting how it extends ideas of distribution that emerged in Phase 3 and providing a rationale for Activities 8 and 9. The chapter then details Activities 8 through 10, characterizing discussions that unfolded in their temporal order, and highlighting students' thinking that emerged within them. The chapter concludes with analyses of students' written work on post-instruction assessment tasks.

Prelude to Phase 4

The first two activities detailed in this chapter were designed for use in a previous teaching experiment (Saldanha & Thompson, 2002). They were re-employed in Phase 4 of the current experiment as an extension to the activities of Phase 3. Recall that in Phase 3, emergent ideas of

a collection's variability, accuracy, or dispersion were largely informal and grounded in perceptual features discerned from dot plots and cluster diagrams of sample percents. The research team reasoned that these images might provide an imagistic underpinning for a more elaborate sense of distribution—an *operational conception of distribution.*

By "operational conception of distribution" I mean a structuring of a collection of randomly generated data values that emerges from partitioning the collection's range into classes, each of which is determined by the proportion of the values that are contained within some sub-range of it or within some vicinity of the sampled population parameter. Another way to think of such a structuring is as the quantification of the variability among the collection's values. By determining what fraction of all values are contained within various sub-ranges of the whole, or within intervals around the parameter, one is essentially giving a measure of the collection's dispersion relative to a reference range, or value, respectively. Thus, distribution and quantified variability can be seen as conceptual isomorphs (Thompson & Saldanha, 2003), in that each one induces a sense of the other; having one in mind essentially leads to having the other in mind as well.

Hereafter, I shall refer to this "operational conception of distribution" simply as *distribution.* The activities of this phase moved to support students' developing this sense of distribution by engaging them in quantifying the variability among sample percents. The context of these activities framed the task as one of investigating the effects of two factors on the long-run behavior of sample percents: Activity 9 investigated the effects of the sampled population percent, and Activity 10 investigated the effects of sample size.

A Preliminary Discussion

Activity 8 was preceded by a preliminary discussion—lasting approximately 12 minutes during the later part of Lesson 13—in which the instructor introduced the idea of quantifying the variability among a collection of sample percents. The aim of this discussion was to orient students to the possibility of moving beyond perceptually-based judgments of a collection's variability and toward measurement-based judgments.

The preliminary discussion focused on two dot plots (Figure 8.2), each showing the dispersion of a collection of sample percents. The instructor had spontaneously sketched these on

the blackboard and deliberately constructed each to have different numbers of elements. This was intended to problematize comparing their variability on the basis of visual judgments.

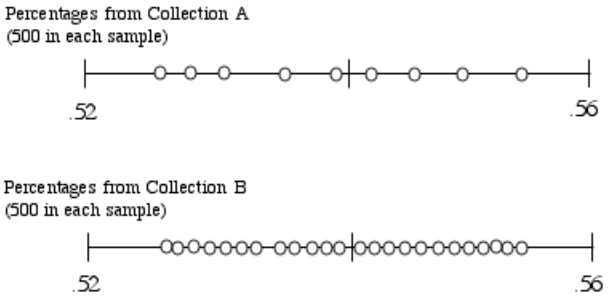


Figure 8.2. The distribution of each of two collections of sample percents.

The instructor then seeded the discussion by asking the class to judge, on the basis of a rough visual estimate, which collection is more variable. Several key ideas emerged among students in the course of the ensuing discussion. They are summarized below:

- The two collections have different numbers of sample percents
- The two collections have the same range, so if we go strictly by range they appear to be equally variable
- Collection A *looks* more spread out. If we go by how far apart the individual sample percents are, then Collection A is more variable
- Collection B contains more sample percents, and more of them are farther away from the center. So if we go by how many percents are far away from the center, Collection B is more variable
- Nicole observed that each collection is "evenly dispersed throughout the range", leading her to believe that the two collections were equally variable

The instructor took Nicole's observation (in the last bullet) as occasion to introduce a conventional way of quantifying variability, that is to look at what fraction of an entire collection of sample percents are contained within various ranges. He drew a common interval around each distribution's center (Figure 8.3) and asked students to determine what fraction of each collection's percents were contained within that interval.

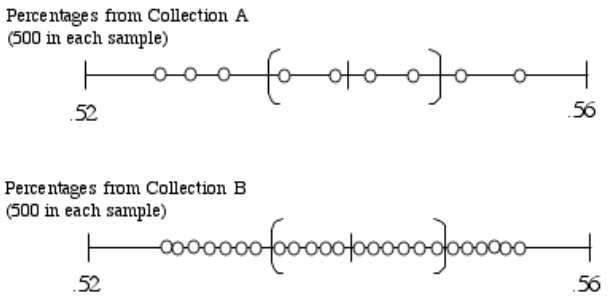


Figure 8.3. The distribution of two collections of sample percents, each having some fraction of it contained within a common interval around the center.

On the basis of a quick calculation, the class determined that in both collections a little less than half of the percents were contained within the common interval. The instructor then stressed that by this conventional measure, the two collections were about equally variable.

The discussion concluded with the instructor highlighting that the proportions students had just calculated and compared were measures of the variability of each collection. While students did not object to the idea that a collection's density within regions around a location provided a measure of its variability, there was also no direct evidence that they generally appropriated this idea as a way to measure variability. The concluding segment of the discussion does, however, suggest that two particular students were broaching this idea. For instance, Peter was mindful of the attribute being measured as akin to the tightness of a collection's clustering relative to its midpoint. In addition, both Peter and Nicole believed that the tighter the clustering then the smaller its measure would be, thereby expressing an intuitive sense that variability and its measure are inversely related.

Activity 8: Investigating Effects of the Population Parameter on Sampling Variability

Activity 8 was the research team's attempt to engage students in systematically quantifying the dispersion of sample percents relative to the population percent. It comprised the table shown in the left panel of Figure 8.4 and four displays like the one shown in the right panel. The histograms and data lists corresponded to collections of samples drawn from four populations (having, respectively, 57%, 60%, 65%, and 32% "yes" on an issue, e.g. "Do you believe you can be President?") from which 2000, 2000, 2000, and 3000 samples were drawn, respectively. All samples contained 500 individuals. The histograms and lists showed how "percent of a sample who responded *yes*" were distributed.

The attached pages show the distributions of samples taken from various simulated populations. Use the attached sheets to complete this table.

Percent of Yes in Population	Number of Samples Drawn	% of Sample Percents within 1 Percentage Point of Population %	% of Sample Percents within 2 Percentage Points of Population %	% of Sample Percents within 3 Percentage Points of Population %	% of Sample Percents within 4 Percentage Points of Population %

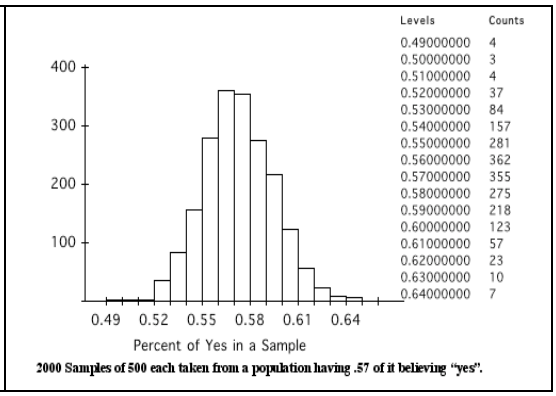


Figure 8.4. One of four distributions of sample percents (right panel), drawn from a simulated population, for which students investigated and quantified the variability (one row of table in left panel).

Students were to fill in the rows of the table shown in the left panel of Figure 8.4 with information drawn from the histograms and data lists like that in the right panel.

Each row of the table in the left panel corresponded to one of the populations. For each row students were to specify the population percentage and number of samples selected. They were to then fill in a row's remaining 4 cells with the percent of sample percents, for samples drawn from that population, which were within one, two, three, and four percentages points of the population percentage.

The aim of this set-up was to structure students' investigation of the sampling data so as to orient them to patterns in the data's dispersion around the population parameter. These patterns might then facilitate quantifying the density of the dispersion in the conventional way proposed in the preliminary discussion—that is, by determining what fraction of sample percent's values are contained within small intervals around the population percent's value. Moreover, by examining these patterns of dispersion across various populations, students might discern a relationship between sampling variability and the population parameter. Indeed, an instructional endpoint of the task was to have students realize that for a given sample size, the sampling variability—that is, the patterns of dispersion that emerge—across the four populations is relatively constant, thus suggesting a general relationship: for samples of a given size, sampling variability is relatively unaffected by underlying population percentage.[1] Evidence for this generalization can be seen in Figure 8.5, which shows the completed table.

[1] This last point was not formally asked of students in the written activity guide. Rather, it was a culminating part of the activity that the instructor raised during classroom discussions. Also, while this generalization is not technically true, the research team took it as a suitable line of reasoning for pedagogical purposes. In fact, the dispersion will

Percent of "Yes" in population	Number of samples drawn	% of sample percents within 1 percentage point of population %	% of sample percents within 2 percentage point of population %	% of sample percents within 3 percentage point of population %	% of sample percents within 4 percentage point of population %
57%	2000	36%	64%	82%	93%
60%	2000	36%	65%	85%	93%
65%	2000	37%	65%	85%	95%
32%	3000	37%	66%	85%	94%

Figure 8.5. Percent of sample percents contained within 1 through 4 percentage points of the sampled population percentages. Note that patterns of dispersion hardly vary across changes in population percentages.

It goes almost without saying that Activity 8 is not uncomplicated. Understanding this scenario coherently requires fitting together a lot of information, clearly distinguishing and relating a host of objects and quantities. Among the most fundamental of these for any one population are: a population of items, samples of items drawn from that populations, a population percentage, ranges of sample percentages, groups of samples, numbers of samples comprising a group, numbers of items comprising a sample. Moreover, students were asked to look for patterns across scenarios to discern stable and varying relationships within them. On the basis of its experience with Activity 8 in a prior teaching experiment (Saldanha & Thompson, 2002), the research team anticipated that students in the present experiment might experience difficulties in composing these ideas.

Activity 8 was assigned for homework immediately after the preliminary discussion, at the very end of Lesson 13. Due to lack of class time, students were not substantively briefed on the activity itself.[2] Consequently, the research team anticipated that students would not engage substantively with it as homework. The team thus planned to continue the activity during the next lesson.[3] In Lesson 14, students engaged in the activity and a discussion (Discussion 1) unfolded over approximately 25 minutes during the later part of the period. Once again, lack of time cut the activity short. In Lesson 15 (the next day), the activity was revisited and taken to its intended conclusion, whereupon a second discussion (Discussion 2) unfolded over approximately 25 minutes during the early part of the lesson. The next two subsections highlight

---

actually be less for samples drawn from populations having very small or very large percentages—a point that is elaborated further in footnote 7 of this chapter.

[2] The instructor only had time to briefly point out, while distributing the Activity 8 handout to students, that in completing the table with information from the four histograms and data lists they were to look at the percent of sample percents within various percentage intervals around the population percent.

[3] Additionally, a school holiday introduced a four-day lapse between Lessons 13 and 14.

aspects of Discussions 1 and 2, highlighting student engagement and thinking that emerged within them.

Activity 8, Discussion 1 (Lesson 14)

Lesson	Activity (A)	Duration
13 (09/03)	Preliminary discussion to A8: quantifying variability	12 m.
14 (09/07)	**A8: Investigating effect of *p* on sampling variability—Discussion 1**	**25 m.**
15 (09/09)	A8 revisited—Discussion 2	25 m.

Figure 8.6 Chronological overview of discussions surrounding Discussion 1 of Activity 8.[4]

The opening discussion of Activity 8, in Lesson 14, revealed that students had attempted the activity on their own. Students had attempted to calculate the percentages of sample percents specified in the table headings, but they had experienced difficulties with three issues: 1) reading and interpreting the data table appearing next to each histogram, 2) understanding what they were calculating, and 3) understanding the purpose of the activity.

With regard to the first two issues, students knew they needed to add values appearing in the "Counts" column, and they understood that these values corresponded to entries in the "Levels" column. They also seemed to understand that they needed to consider those values in the "Levels" column that deviated from the population percentage by 1, 2, 3, and 4 points. However, students appeared to have attended largely to the surface characteristics of the numerical information in the data list and proceeded somewhat mechanically in their calculations. They added together "Counts" values in a way that suggests they were not interpreting each value in the "Levels" column as the included left endpoint of an *interval* denoting a range of sample percents. For instance, by convention, the pair of table entries "0.56000000, 362" denotes that 362 of the 2000 samples had a sample percentage, $\hat{p}$, somewhere in the interval [0.56000, 0.57000), so that $0.56000 \leq \hat{p} < 0.57000$. However, when calculating the percent of sample percents within 1 percentage point of the population percent, many students added together 362, 355, and 275—one too many values (see Figure 8.7, below). They calculated other counts similarly.

[4] Figures like this one are employed throughout this chapter in an effort to help the reader keep track of the chronological order in which discussions unfolded.

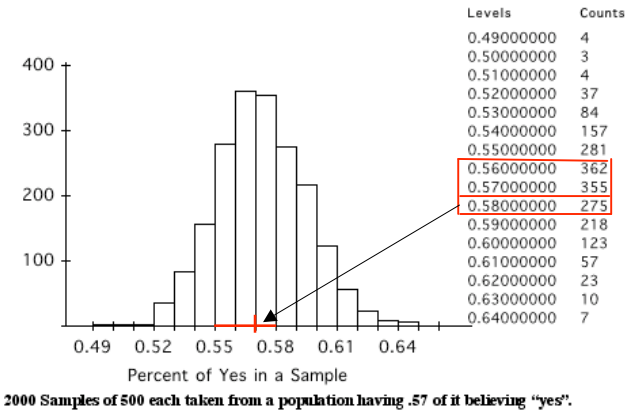


Figure 8.7. The histogram and data table of sample percents for samples drawn from a population having a parameter value of 0.57. Values encircled in red are those that students used to calculate the percent of sample percents contained within 1 percentage point of the population parameter. Corresponding intervals are highlighted in red on the histogram's horizontal axis.

Sensing students' difficulties, the instructor clarified the convention used in the data tables. He led a whole-class discussion aimed at having students coordinate their interpretations of various items shown in the activity sheets. He began the discussion by having students parse the statement in the table headings "% of sample percents within 1 percentage point of population %", pointing out how many times the word "percent" appears in it and explaining what each "percent" was a percent of in the described sampling scenario.

The instructor then discussed how to calculate the table entries of the first row, corresponding to the population having 57% of it believing "Yes". He sketched a diagram like that shown in Figure 8.8 to aid students in coordinating the meaning of the various values read from the data list with that of the required calculations.

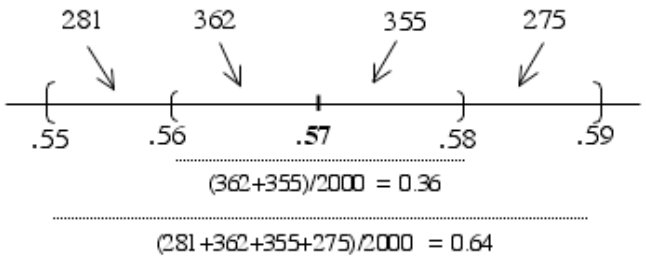


Figure 8.8. A diagram used by the instructor to help students keep track of and coordinate the various quantities and calculations involved in Activity 8. At this stage, the diagram shows the information entailed in calculating the percent of sample percents contained within 1 and 2 percentage points of the population percent (0.57).

The diagram was built up a bit at a time as the discussion progressed, pieces being added as students explained their meaning. After each cell of the table was filled with the calculated percentage, the instructor asked students to explain the value's meaning. Students typically struggled to coordinate all the pieces and express this meaning coherently without reading the table heading. The following excerpt illustrates this in the case of Peter, who seemed to experience the least difficulties. In the excerpt, the instructor asks Peter to explain the value of 0.64 that Peter had just calculated.

Episode 1, Lesson 14:

1. I: And what did we, what have we just found here (points to 0.64)? What is this?
2. Peter: It's the percent of uhh percent of percents (3 second pause)—all these percents everywhere (chuckles)—uhh within two percentage points of the population's percent.
3. I: Can you say that whole thing again?
4. Peter: All right. It's the percent of sample percentages, uhh two percentage points of the population's percent.
5. I: Ok.
6. Peter: Yeah.
7. I: Ok, that's good. Now what's the background to that statement? We have all these samples, right?
8. Peter: Yeah.
9. I (continues): drawing from a population having 57% of it saying "yes"
10. Peter: Yeah
11. I: What size samples are we drawing?
12. Peter: Uhh, 500
13. I: and how many samples did we draw?
14. Peter: 2000

The discussion proceeded along these lines until the table's first row was completed (see Figure 8.5), and the diagram in Figure 8.8 was extended to show all pertinent information and calculations.

Concerning the third issue, although students were apparently engaged in the activity the discussions contain evidence that the broader goal of it escaped them. For instance, when asked what they thought the activity was about, several students echoed Sarah's comments: *"I don't get the point of it",* and *"the only thing I don't understand is are we just doing this so we'll be able to do it or are we gonna use it or something?"*. The instructor responded by explaining that the activity's aim was to investigate the variability of the sample percents, and that they would discuss this variability once students had compiled the necessary information in the table.

In the remainder of the lesson students worked in pairs, filling in the remaining rows of the table (Figure 8.5) with the appropriate values. The instructor presumed that by that point students were capable of calculating and, perhaps, interpreting these values. The intention was to have students regroup for a whole–class discussion of the results and variability, pointing out the broader goal of the activity. However, due to lack of class time the discussion was postponed until Lesson 15.

Activity 8, Discussion 2: Part 1 (Lesson 15)

Lesson	Activity (A)	Duration
14 (09/07)	A8: Investigating effect of p on sampling variability—Discussion 1	25 m.
15 (09/09)	**A8 revisited—Discussion 2: Part 1**	**4 m.**
15 (09/09)	A8 revisited—Discussion 2: Part 2	21 m.
15 (09/09)	A9: Investigating effect of n on sampling variability—Discussion 1	25 m.

Figure 8.9. Chronological overview of discussions surrounding Part 1 of Discussion 2, Activity 8.

The second discussion around Activity 8 occurred during Lesson 15. The discussion unfolded in two parts. The first part, lasting approximately 4 minutes and occurring near the beginning of the lesson, revolved around having students clarify their sense of variability.[5] This part began with the instructor asking what students thought the activity was getting at. Several students knew that "variability" was the issue, but they offered no further elaboration. Students were then asked to unpack "variability"—to describe precisely *what* was varying and *how* it was varying. In the ensuing interchange, students agreed that sample percents were varying. When

[5] This part was immediately preceded by the lesson's opening discussion, lasting roughly 2 minutes, in which the instructor oriented students' attention to the activity table that they were supposed to have filled out for homework. One student was evidently unclear on what he was supposed to have done, and another student had not completed the table in its entirety.

prompted for elaboration, two main ideas then emerged: 1) Nicole thought the variability was in the way the sample percents "move away" from the population percent; 2) Chelsea thought the sample percents "vary with each other", in the sense that the sample percent calculated for each sample varies from sample to sample.

The instructor followed Chelsea's idea to conclude the first part of the discussion by highlighting a subtle point: once a particular sample has been drawn, its calculated sample percent is fixed and unvarying. The variability occurs in the sample-to-sample differences, the respective percents of which are also different and thus *vary* amongst each other. This elaboration was intended to have students distinguish between non-statistical and statistical variability, situating the latter within the idea of repeated random selection.

Activity 8, Discussion 2: Part 2 (Lesson 15)

Lesson	Activity (A)	Duration
14 (09/07)	A8: Investigating effect of *p* on sampling variability—Discussion 1	25 m.
15 (09/09)	A8 revisited—Discussion 2: Part 1	4 m.
15 (09/09)	**A8 revisited—Discussion 2: Part 2 (in 3 phases)**	**21 m.**
15 (09/09)	A9: Investigating effect of *n* on sampling variability—Discussion 1	25 m.

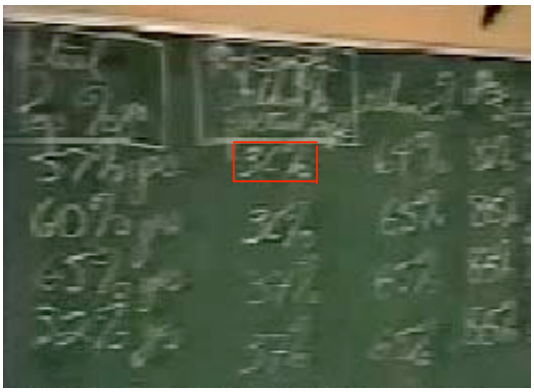
Figure 8.10. Chronological overview of discussions surrounding Part 2 of Discussion 2, Activity 8.

The second part of Discussion 2, lasting approximately 21 minutes during the middle of Lesson 15, revolved around interpreting the results of the table entries that students had calculated during the previous lesson (Figure 8.11). This part unfolded in 3 phases.

Part 2: Phase 1 (Lesson 15)

The second part of Discussion 2 began with the instructor soliciting students' interpretations of the "36%" in the first row of the table (Figure 8.11).[6]

[6] This completed table had been sketched on a side black board, almost in its entirety, by the classroom teacher in the early part of the period.



Actual Pop. %s	% of samples within 1% of percent pop.	Within 2%	Within 3%
57% yes	36%	64%	82%
60% yes	36%	65%	85%
65% yes	37%	65%	85%
32% yes	37%	66%	85%

Figure 8.11. A sketch (left) of the table of calculated percentages, and its reproduction (right).

It emerged within this first phase of the second part of Discussion 2 that though students had calculated the correct values, many remained uncertain as to how to interpret these values, despite having participated in the instructional discussions of Lesson 14. Of particular significance is that in their attempts to describe a value's meaning, students seemed to struggle fitting together all of the pieces of the sampling scenario and easily lost sight of the value's meaning in the process. Their struggles suggest that explaining a value's meaning entailed coordinating too many parts for them to manage comfortably. The following discussion excerpt, lasting approximately 2 minutes, illustrates this in the case of Sarah, one of the more engaged students. In the excerpt, Sarah is attempting to explain the meaning of "36%" highlighted in red in Figure 8.11.

Episode 1, Lesson 15:

1. I: […] One more, let's get one more person to try. Sarah? Can you try to say what this stands for?
2. Sarah (clears throat): Uhh, well, uhh first is it, is each thing one sample or is it just a whole lot of samples?
3. I: Go ahead, say it, uhh, ask your question again.
4. Sarah: Is it one sample--?
5. I: What is "it"?
6. Sarah: Ok, you have like the population percent is uhh going up, you know you have those going up and going down one (gesticulates with hands up and down, pointing at table on blackboard), is that one sample or that instead of--
7. (Peter continues explaining activity to David in background throughout Sarah's subsequent description)
8. I: (points to Sarah) I'm sorry, you're doing lot's of this and (moves a hand up and down, mimicking Sarah's gesticulations), and--
9. Sarah (chuckles): Ok, I'll just try. Uhh, we have a population percent, and (clears throat) you have a number of samples, and uhh you wanna

know—and they're all in order (motions with hand as though running down a vertical list), and population percent's in the middle. And you wanna know uhh what one population perc—wait, what one percentage (laughs)
10. (Nicole and Sarah both chuckle)
11. Sarah (continues): Ok, you have a percent of samples and you wanna know what uhh one step above is and one step below is
12. (Nicole nods)
13. Sarah: Well no, wait—just one step above one (inaudible), and you wanna know what percentage of those altogether is part of the whole (inaudible)
14. (Nicole and Sarah turn to each other and smile)
15. I: All right. Ok. And so what does this 36% end up showing about all of that?
16. Sarah: That 36% of uhh the samples were in 1 percentage point of the whole, all the samples (inaudible)
17. I: Very good! Now let me point something out to you: what you just did is you said "ok, this 36% is 36% of the samples that are within 1 percentage point of the actual", so that's what you ended up saying (chuckles)
18. Sarah: Oh
19. I: You ended up saying that <u>this</u> (points at "36%" in table) is <u>that</u> (points at the column heading "Percent of samples within one percentage point of the population percent")
20. Nicole (chuckles)
21. Sarah: Is that what it is?
22. I: That's what it is! So, you, when you were, you were describing how to, how to answer this question ["What does 36% stand for?"]. In your answer of what this stands for, you told, you, your answer described how to get this number [points to "36%"]. What you hadn't realized yet, it sounds like, was that this number answers that question. This number <u>is</u> the percent of the samples that lie within 1 percentage point of the population percent.

The instructor then moved the discussion toward clarifying details of the underlying sampling scenario, asking the class how many populations were involved, how many groups of samples there were, and whether the groups were taken from the same population. The ensuing discussion indicated that many students were unclear as to whether all groups of samples were taken from the same population. In particular, they seemed prone to confounding samples and groups of samples, as evidenced by a unanimous "yes" response to the question: "were two <u>groups</u> of samples taken from the same population?". The instructor appeared to have resolved this confusion by explicitly pointing out to students that each of the table's rows, and thus

collection of samples, corresponds to distinct populations. Nevertheless, the confusion suggests that distinguishing the various objects in the sampling scenario—sampled items, sample of items, and group of samples—was precarious for students.

Part 2: Phase 2 (Lesson 15)

The second phase of part 2 of Discussion 2 was marked by the instructor's attempts to have students interpret the pattern of results displayed in the table (Figure 8.11). Once assured that students had a sufficiently unproblematic understanding of what the table values represented, the instructor asked "what about the variability of these samples, from these different populations?". In response, several students characterized the behavior of the percents for samples within a group. For instance, Peter gave this description: *"most of the samples clustered closer to the population and spread out as you get out further from the population percent"*. Sarah made allusions to what seemed like an imagined bell curve or a histogram showing part of a Normal distribution. She traced out part of its shape in the air with her hand as she explained that the percents seemed to follow a pattern: *"they got bigger at the top"*, she said, presumably referring to how numbers of sample percents increase with increasing distance from the population percentage.

Following these comments, the instructor moved to shift students' focus from particular groups of sample percents and orient them to a next level—the behavior of the 4 collections of sample percents as a whole, *across the different populations*. He turned students' attention to the percent of sample percents within 1 percentage point of each population percent, shown in the table's second column. Students noticed that this value hovered around 36%, across all of the different populations and that it thus seemed to be independent of underlying population percentage. The instructor concluded the discussion by highlighting that this observation was the basis of a generalization: *the variability among samples of a given size is largely independent of the sampled population parameter*.[7]

[7] The generalization "largely unaffected" holds for population parameter values, p, in the approximate range [0.30. 0.70]. It relies on the fact that the underlying population variance is maximal when p is 0.5 and decreases most rapidly near the tails, thus allowing us to vary p within this range without significant effect on the distribution of sample percents. Students' investigations were restricted to populations having parameter values within this range.

Part 2: Phase 3 (Lesson 15)

This concluding generalization provided a segue into the final phase of the discussion, which culminated with one student implicitly extending the generalization a step further. The instructor began this phase of the discussion by describing a hypothetical sampling scenario in which thousands of samples of size 500 were selected from a population whose percent was unknown. He then solicited students' ideas about what fraction of those samples they expected would be within 1 percentage point of the unknown population percent. Peter and Luke immediately anticipated that the proportion would hover around 36%, on the basis of the pattern deduced from the table. Moreover, and significantly, Luke extended his anticipation to reason that *"so we could almost tell what the population percent was?"*. Thus, Luke appeared to be developing a sense that one can reason in reverse from knowledge of how sample percents generally cluster about any population percent to infer an estimate of an unknown parameter.[8]

The implicit reverse reasoning expressed by Luke's idea was the highpoint of the discussions around Activity 8. Luke's idea was not, however, developed further in Discussion 2. Instead, it foreshadowed an intended instructional endpoint of the next activity.

Activity 9: Investigating Effects of Sample Size on Variability

The structure of Activity 9 was identical to that of Activity 8: students were to complete a table with percentages of sample percents contained within 1 through 4 percentage points of a sampled population percent. The calculation of these percents entailed interpreting information from 6 pairs of histograms and data lists, each showing the distribution of sample percents for 2500 samples drawn from a common simulated population. Figure 8.12 displays a representative part of the activity guide that students received.

[8] In this way, Luke was, in effect, informally broaching the logic of confidence intervals.

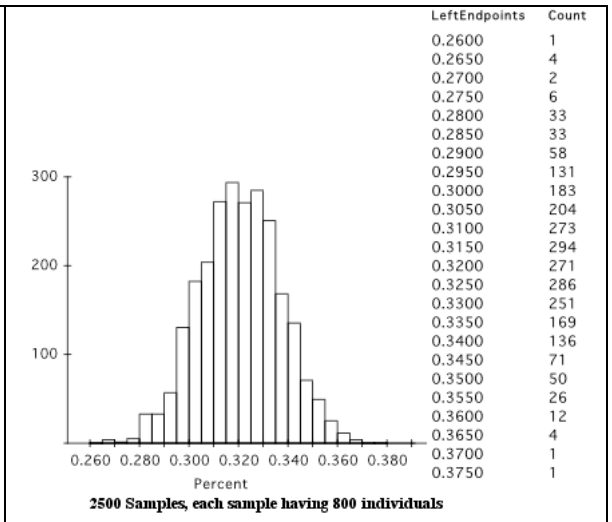
The attached pages show collections of 2500 random samples. One collection contains samples of size 100. The other collections contain samples of sizes 200 400, 800, 1600, and 3200. Every sample was taken from a large population having .32 of it believing "yes" on a subject. (Each sample was replaced before selecting another). Use the attached sheets to complete this table.

Number of samples drawn: 2500
Population Percent: 0.32 (32%)

Sample Size	Percent of Sample Percents within 1 percentage point of Population Percent	Percent of Sample Percents within 2 percentage points of Population Percent	Percent of Sample Percents within 3 percentage points of Population Percent	Percent of Sample Percents within 4 percentage points of Population Percent
100				
200				
400				
800				
1600				
3200				

After completing the table:

What can you say about variability amongst samples of a given size in relation to sample size per se?



LeftEndpoints	Count
0.2600	1
0.2650	4
0.2700	2
0.2750	6
0.2800	33
0.2850	33
0.2900	58
0.2950	131
0.3000	183
0.3050	204
0.3100	273
0.3150	294
0.3200	271
0.3250	286
0.3300	251
0.3350	169
0.3400	136
0.3450	71
0.3500	50
0.3550	26
0.3600	12
0.3650	4
0.3700	1
0.3750	1

Figure 8.12. One of six distributions of sample percents (right panel), each drawn from a common simulated population, for which students investigated and quantified the variability (left panel).

In contrast to Activity 8, in this activity each of the six groups of sample percents was computed for samples of a different size (100, 200, 400, 800, 1600, and 3200), and samples across the six groups were all drawn from the same population (p =.32*)*. Having established, in Activity 8, that sampling variability is largely unaffected by the underlying population percentage, there was no need in this activity to consider different populations. The aim now was to investigate how changes in sample size played out in this variability. Thus the written question posed at the bottom of the table in Figure 8.12.

Activity 9 was assigned for homework at the very end of Lesson 14. There were two discussions around the activity, preceding much along the same lines as those of Activity 8 in their tone and nature. A first discussion (*Discussion 1)* unfolded during the last 25 minutes of Lesson 15 (the next day), immediately after the conclusion of Activity 8. A second discussion (*Discussion 2*) occurred for a period of 20 minutes during Lesson 16. The next two subsections elaborate these Activity 9 discussions and central issues that arose within them.

Activity 9, Discussion 1: Part 1 (Lesson 15)

Lesson	Activity (A)	Duration
15 (09/09)	A8: Investigating effect of *p* on variability—Discussion 2: Part 2 (in 3 phases)	21 m.
15 (09/09)	**A9: Investigating effect of *n* on variability—Discussion 1: Part 1**	**5 m.**
15 (09/09)	A9—Discussion 1: Part 2	20 m.

Figure 8.13. Chronological overview of discussions surrounding Part 1 of Discussion 1, Activity 9.

The first discussion around Activity 9 unfolded in two parts. It began with the instructor prompting students for their interpretation of information shown in the data lists corresponding to each histogram. For instance, with regard to the distribution of sample percents for samples of size 800, the instructor tried to ensure that students knew how to read the data list, asking them where they would look to find how many samples had a percent between 0.29 and 0.30 (see right panel of Figure 8.12). Conversely, students were asked to explain what a given value in the "Counts" column represented. These first discussions arose out of one's student's need to resolve her confusion about the partitioning of the histograms' horizontal axes. As can be seen in the right-hand panel of Figure 8.12, each distinct interval along the horizontal axis represents not one, but one half of a percentage point. This feature required clarification.

Activity 9, Discussion 1: Part 2 (Lesson 15)

Lesson	Activity (A)	Duration
15 (09/09)	A9: Investigating effect of *n* on variability—Discussion 1: Part 1	5 m.
15 (09/09)	**A9—Discussion 1: Part 2 (in 3 phases)**	**20 m.**
16 (09/10)	A9 revisited—Discussion 2: Parts 1-4	50 m.

Figure 8.14. Chronological overview of discussions surrounding Part 2 of Discussion 1, Activity 9.

The second part of the discussion revolved around a slide presentation that the instructor had prepared. The presentation unfolded in three phases. The first phase culminated with the table displayed in Figure 8.15, showing the sample counts from the data lists that accompanied the histograms.

Six Groups of 2500 Samples Each, All Taken from a Population of which 0.32 believe "Yes"

	±4%							
		±3%						
			±2%					
				±1%				
# in Samp	.28-.29	.29-.30	.30-.31	.31-.32	.32-.33	.33-.34	.34-.35	.35-.36
100	142	184	202	214	232	202	176	181
200	142	223	283	267	306	311	230	190
400	139	204	346	381	444	361	229	164
800	66	189	387	567	557	420	207	76
1600	10	79	395	761	774	372	95	13
3200	0	17	253	999	961	248	22	0

Figure 8.15. Culminating slide of the first phase of the presentation in Part 2 of Discussion 1, Activity 9.

Figure 8.16 shows the correspondence between the information contained in this table and the data lists. The data for each sample size are collapsed into distinct rows of the table.

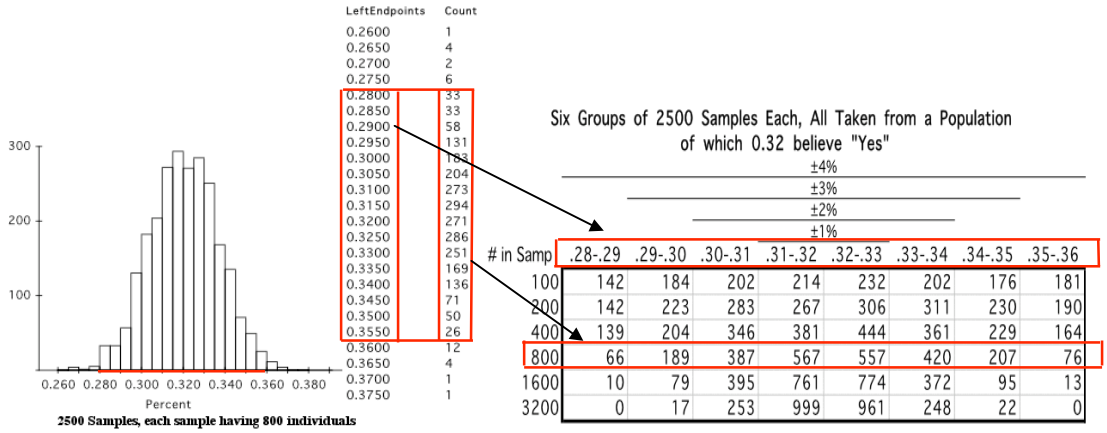


Figure 8.16. Intended connections between the data lists and the table in the culminating slide.

Each column of the table corresponds to samples whose percent of "yes"s are in the interval specified above it. For instance, the value "387" in row 3 of column 3 denotes that 387 of the 2500 samples selected, of size 800, had a percent of "yes" in them somewhere in the interval [0.30, 0.31).[9] Each of the four stacked horizontal lines above the table highlights a percentage range around the population percent—1%, 2%, 3%, and 4%, respectively.

[9] The table incorporates features of the instructional diagram used in Lesson 14 (see Figure 8.8), the table that students filled in (Figure 8.12, left panel), and the data lists accompanying the histograms (Figure 8.12, right panel). The aim of presenting all of this information in one inscription was to facilitate students' comparison and composition of the information drawn from these various inscriptions.

Part 2, Phase 1 (Lesson 15)

The discussion in the first phase, lasting approximately 4 minutes, focused on ensuring that students understood how to read this table and what the values in it denote.[10] This is no minor point, since the table contains a lot of information and presumably entails some effort to decode and understand. The following brief discussion excerpt illustrates potential difficulties in interpreting the table entries, particularly how easy it can be to loose one's way when trying to coordinate the various pieces of information to make sense of an entry. The excerpt is of Peter trying to describe what the value "381" represents.

Episode 2, Lesson 15:

430. I: […] So, tell me what is this (aims pointer at "381" in table), what is this number right there?
431. Peter: 381?
432. I: Yeah, that 381. What does that number stand for?
433. Peter: That means out of uhh (3 second pause) the, the sample group, where you're taking sample size of 400, uhh between 31% and 32%, 381 samples said "yes".

(4 second silence)

434. I: Can you say that again Peter?
435. David: 381 samples?
436. I: Start with the 381
437. Peter: 381—
438. David (chimes in as though to finish Peter's sentence): people in the sample
439. I: Ok, 381. And what does, that 381 is 381 what?
440. Peter: 381 "yes"s (2 second pause), between 31
441. Nicole: No
442. David: No
443. Peter: All right, uhh, blahh (ridicules himself for getting tongue tied)
444. (Nicole laughs)
445. Peter (continues): in those 381 samples, 31% to 32% of the people in the samples said "yes".
446. I: Very good!
447. Peter: that's what I'm trying to say

As indicated in lines 445-447 of the excerpt, Peter eventually produced a very coherent and articulate interpretation of "381". But notice that this did not come easily and entailed some support from the instructor. Of particular significance is that, as Peter began his attempt, he

[10] This table was actually the last in a sequence of versions of it. The difference between subsequent versions in the sequence was the addition of another horizontal line above the table, denoting the width of a larger percentage interval around the population percent. This pattern culminated in the stacked lines shown in Figure 8.15.

seemed inclined to think of "381" as a number of people in a sample that answered "Yes" (lines 433 and 440), rather than a number of samples. There is some suggestion, in lines 435 and 438, that this distinction was precarious for David as well.

It seems plausible, from Peter's utterance in line 433, that his confusion was rooted in his image of samples and his image of people having "spilled" into each other as he tried to coordinate 2 quantities that did, in fact, refer to numbers of people (i.e., "400" and "0.31-0.32"). In other words, Peter's first attempt at explanation is consistent with his having easily lost sight of the "381" as a number of samples because he attempted to simultaneously interpret other quantities associated with "381" that did refer to people.[11]

The instructor followed Episode 2, above, by suggesting that students use a linear "building-up" strategy to interpret the table's entries: first say what the "381" is a number of, then move to incorporate the other associated quantities into the interpretation by saying what is true about those 381 samples. He concluded this first phase of the presentation by highlighting that 4 percentage points within the population percent is the same as the range 28% to 36%.

Part 2, Phase 2 (Lesson 15)

In the second phase of the presentation, the instructor turned students' attention toward elaborating patterns in the data, asking them what those patterns imply about variability. Luke pointed out that variability among samples of a given size could be discerned by tracking the data horizontally, across columns within a row. The instructor highlighted Luke's observation for the class and then turned students' attention to the next slide, displaying only the data for samples of size 100 and 200 (Figure 8.17).

[11] The table in Figure 8.15 is a far more complicated inscription than I have let on in this analysis. A powerful interpretation of the table entails viewing any of its individual entries as a multi-dimensional quantity formed by the composition of 3 other associated quantities that are also described in the table. For instance, one should interpret the "381" as a number of samples (percents) satisfying several constraints that are also represented in the table: each of these 381 samples has size 400; the fraction of people who responded "yes" in each sample is in the interval [0.31, 0.32); this interval is within 1 point of the sampled population proportion. This multi-dimensionality mirrors that of sampling distributions and the compositions entailed in a coherent conceptualization of them. Students' observed difficulties in interpreting the table's entry can be taken as a reflection of their difficulties in conceiving sampling distributions.

Six Groups of 2500 Samples Each, All Taken from a Population of which 0.32 believe "Yes"

±4%
±3%
±2%
±1%

# in Samp	.28-.29	.29-.30	.30-.31	.31-.32	.32-.33	.33-.34	.34-.35	.35-.36
100	142	184	202	214	232	202	176	181
200	142	223	283	267	306	311	230	190

Figure 8.17. Sampling data for samples of size 100 (first row) and size 200 (second row).

Students were asked to describe what the two rows of values displayed in Figure 8.17 indicate about the relative variability of samples of those sizes. At this juncture Chelsea appeared uncertain that she "saw" variability in this data. This prompted the instructor to take an abrupt and momentary diversion from the slide presentation. The diversion, lasting approximately 2 minutes, was to show a computerized animation of how the histograms were made, using the Data Desk program (Data Description, 1999). The animation showed a histogram's bar heights growing as values got highlighted in a separate data file that was hot-linked to the graph (see Figure 8.18).[12]

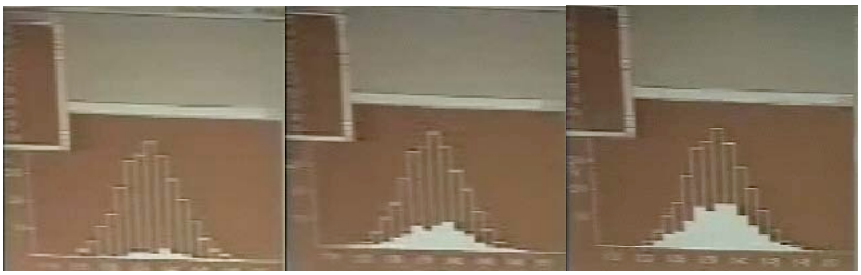


Figure 8.18. Three states in a histogram's emergence as it appeared on a projection screen.

As the animation unfolded, the instructor described what was happening, as above. He stressed that students should think of the histogram not as existing all at once, but rather as being built up as samples are selected and their percents get dispersed along the emerging graph's horizontal axis.

At the animation's conclusion, after the histogram had emerged in its entirety, the instructor immediately went back to the slide in Figure 8.17 and resumed the discussion of variability that

[12] This animation was very similar to the one that students viewed in Lesson 11.

he had seeded just prior to this diversion. In the ensuing discussion, it is implicitly evident that the instructor's diversion was out of concern that students did not have a strong image of how patterns in the data they were examining might have emerged. He sensed that students had been looking at these collections of data purely as after-the-fact entities and that they lacked a way of describing them that might facilitate comparing their variability. The diversion, then, was a deliberate effort to help students infuse their thinking with images of sample percents' emergent dispersions. Significantly, the subsequent discussion suggests that these images served to occasion a discourse about variability that had not occurred prior to that. The following excerpt is from that discussion

Episode 3, Lesson 15:

23. I: All right. Well, the reason that I'm (projects slide shown in Figure 8.17), the reason that I did that was, where is the variability that this shows? (points to table appearing on screen)

(3 second pause)

24. I: Do you—did these 142 samples, did they, did they come before those 223 samples? (points to these values in table's second row, see Figure 8.17)

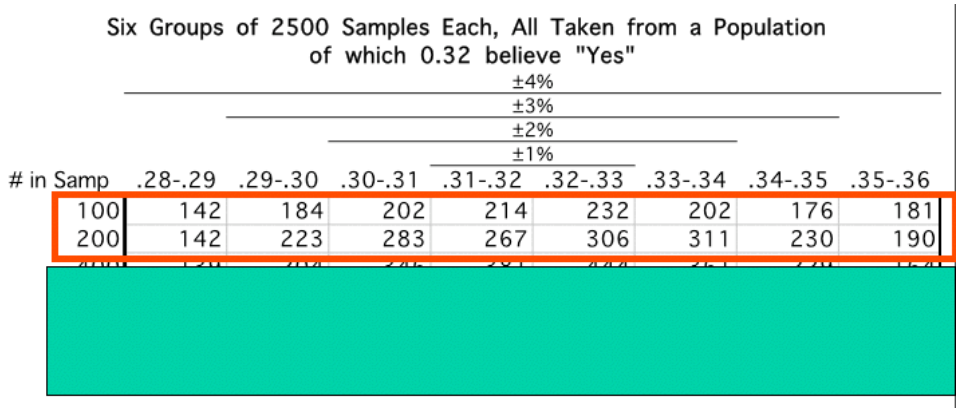


Figure 8.17. Sampling data for samples of size 100 (first row) and size 200 (second row).

25. Peter: No
26. I (continues): Were they selected before those 223?
27. Peter: No
28. Chelsea: No (inaudible)
29. David: It was at random
30. Chelsea (continues): randomly selected
31. I: It was at random. So, is there—none of these were, you, you can't, you just can't tell when those, you know, when the samples came, when they were selected. But as they were coming, where did the samples start to build up more? Can you tell that by looking at the rows? (points to table shown in Figure 8.17)
32. Peter: Towards the middle

33. David: Yeah, towards—
34. I: Toward the middle?
35. David: Towards 30 and 33
36. I: Ok, did these--
37. Sarah: Toward (inaudible)
38. I (continues): did these (points to 200 sample size row) start building up toward the middle at a higher rate than those (points to 100 sample size row) started building up toward the middle?
39. David: Yes.
40. Nicole: Yes.
41. I: Yeah. So you can see that, even though they were both building up and it was kind of scattered, what about the second row?

(4 second silence)

42. Peter and Kit: It was building up faster
43. I: Pardon me?
44. Peter: It was building up faster.
45. I: It was building up faster in a particular place!
46. Peter: Yeah.
47. I: Not the whole table!
48. Peter: But that, that row right there (points to table)
49. I: The ro-- (points to table and runs forefinger along highlighted rows)
50. Peter: The whatever, the (motions with hand up and down) column, vertical
51. I: Yeah, within, within this range of the uhh hit (motions with hands as though denoting a small interval), of the uhh population percent, it was building up faster, samples of size 200 were building up faster toward the middle. Ok, so, would you call that more variable or less variable, than the samples of size 100?
52. Kit: Less.
53. I: Yeah, less. They're less variable. They tend not to depart so far from the population (motions repeatedly with hands away from a middle point as though denoting departure of sample percents from pop. parameter). They still depart! But more of them bunch up closer! (motions with hands to denote "crowdedness" of sample percents in a small interval).

The excerpt is particularly interesting in that it illustrate how the instructor used a discourse about rate to occasion the emergence of ideas about how the data were aggregated around the population percentage. Here, he evidently was not referring to a time rate, but rather to rate in the sense of "relative accumulation per interval around the population percentage"—that is, the percent of sample percents contained within each of the increasingly wide intervals around the parameter. This seemed to provide a way for students to begin thinking about the rows of data values in terms of patterns of their accumulation. This enabled comparisons of these rows of data, across different sample sizes, in terms of the relative degree of their accumulations. For

instance, Peter's claim (lines 42-46) that the sample percents for sample size 200 were building up "faster" than those for sample size 100 was presumably not a reference to accumulation with respect to time, but rather with respect to spatial intervals around the population percentage.[13]

The discussion continued along these lines. As the instructor showed subsequent slides—each displaying the table in Figure 8.17, but with data for the next larger sample size uncovered—he pointed out that the relative clustering patterns students had described between samples of size 100 and 200 continued with subsequent increases in sample size. The discussion culminated with his summarizing this idea for the class: as sample size increases, a greater portion of sample percents tend to cluster closer around the population percent and are thus less spread out. In other words, the larger the sample size, the less is the variability among sample percents.

This second phase of the presentation culminated with the slide shown in Figure 8.19, in which the lower table expresses the sampling data as percentages of the sample percents contained within 1 through 4 percentage points of the underlying population percent. The instructor stressed that this table captures the pattern of variability that he had just summarized for the class. He also showed the histograms for different sample sizes, pointing out how these patterns are visually evident in them—distributions for bigger sample sizes have more sample percents bunched up closer to the population percent, as indicated by the smaller ranges and higher histogram bars.

[13] This presumption is based on Peter's implied agreement with the instructor's comments in lines 24-31 of Episode 3, Lesson 15.

Six Groups of 2500 Samples Each, All Taken from a Population of which 0.32 believe "Yes"

# in Samp	.28-.29	.29-.30	.30-.31	.31-.32	.32-.33	.33-.34	.34-.35	.35-.36
				±1%				
			±2%					
		±3%						
	±4%							
100	142	184	202	214	232	202	176	181
200	142	223	283	267	306	311	230	190
400	139	204	346	381	444	361	229	164
800	66	189	387	567	557	420	207	76
1600	10	79	395	761	774	372	95	13
3200	0	17	253	999	961	248	22	0

	±1%	±2%	±3%	±4%
100	17.8%	34.0%	48.4%	61.3%
200	22.9%	46.7%	64.8%	78.1%
400	33.0%	61.3%	78.6%	90.7%
800	45.0%	77.2%	93.1%	98.8%
1600	61.4%	92.1%	99.0%	100.0%
3200	78.4%	98.4%	100.0%	100.0%

Figure 8.19. The sampling data expressed as percentages of the sample percents contained within 1 through 4 percentage points of the underlying population percent (lower table).

Part 2, Phase 3 (Lesson 15)

The conclusion that sampling variability decreases with increasing sample size set the stage for the third and final phase of the slide presentation. The discussion in this phase centered on the slides shown in Figure 8.20.

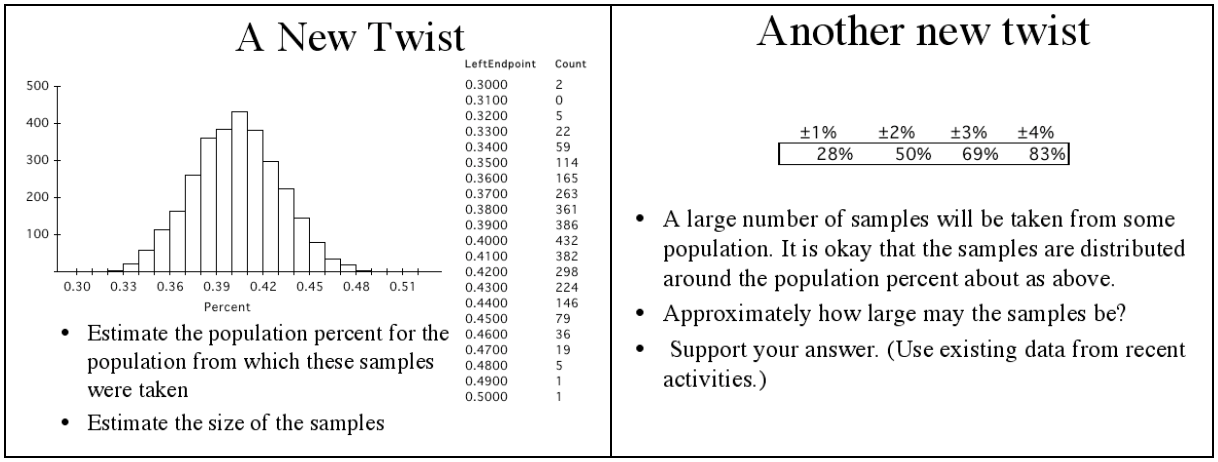


Figure 8.20. The distribution of sample percents for samples in which neither size nor sampled population percent are specified.

Each panel in Figure 8.20 shows information about the distribution of sample percents for a collection of samples for which neither size nor underlying population percent are specified.[14]

[14] This population, like all those considered in previous activities, was assumed to be dichotomous (e.g., consisting of "Yes" and "No" responses to a question posed).

The aim of the discussion around these slides was to push students to extend the conclusions they had drawn about variability, on the basis of their investigations in Activities 8 and 9, to reason in reverse and estimate the unknown population percent and sample size.

The logic of this reverse reasoning is as follows: past sampling results would indicate that sample percents tend to cluster symmetrically around the sampled population percentage. Hence, a reasonable estimate of the population percent, for the distribution in the left panel, might be that it lies in the interval [0.40, 0.41). Moreover, by quantifying sampling variability, we determined variability to be inversely related to sample size. It is thus possible to estimate the sample size in this case by comparing the variability of this distribution with that of the distributions for different known sample sizes (see lower table in Figure 8.19).[15]

The discussion around the slides in Figure 8.20 was rather brief, coming at the very end of the lesson. Nevertheless, it suggests that several students were broaching this reverse line of reasoning. For instance, with regard to the distribution shown in the left panel, David immediately responded to the first question and estimated the population percentage to be "about forty percent". Chelsea suggested that this estimate was justified on the basis of the past patterns, but she did not elaborate. In response to the second question, Peter proposed checking the variability of the distribution—comparing it to results shown in the table in Figure 8.19.

Similarly, with regard to the next slide (right panel of Figure 8.20), David responded without hesitation to the first question and estimated that *"It's a small sample size … because there's only 28% within 1%"*. The instructor concluded the presentation by unpacking the logic of David's idea for the class, stressing that David was saying that the distribution is highly variable, which in turn suggests a relatively small sample size.

While these ideas were still very informal and not well elaborated, their emergence at this juncture would seem to suggest that Activity 9 was useful in helping students move toward reasoning in this way. In particular, the disposition to make an inference about an unknown

[15] The variability of this distribution can be estimated with a quick calculation that entails an estimate of the population percent. Say we use $p = 0.40$. Then the sample percents are distributed around this estimate in the proportions shown below. Comparing this distribution to those in the table appearing in the bottom of Figure 8.17 we estimate the sample size to be between 200 and 400. A reasonable estimate is 300, given the magnitude of the percentages below relative to those for samples of size 200 and 400 shown in Figure 8.17.

±1%	±2%	±3%	±4%
27%	52%	71%	84%

The same strategy leads to a similar inference for the distribution shown in the right panel of Figure 8.20.

sample size on the basis of having quantified patterns of dispersion in past sampling data was a new development in the teaching experiment.

At the end of the final discussion students were asked to respond in writing to the questions in the slides of Figure 8.20, as a homework assignment. Their ideas would constitute the basis of a second discussion around Activity 9, during Lesson 16 (the next day). Highlights from that discussion are elaborated in the next section.

Activity 9, Discussion 2: Part 1 (Lesson 16)

Lesson	Activity (A)	Duration
15 (09/09)	A9: Investigating effect of *n* on variability—Discussion 1: Part 2 (in 3 phases)	25 m.
16 (09/10)	**A9—Discussion 2: Part 1**	**7 m.**
16 (09/09)	A9—Discussion 2: Part 2	8 m.
16 (09/10)	A9—Discussion 2: Part 3	25 m
16 (09/10)	A9—Discussion 2: Part 4	9 m.

Figure 8.21. Chronological overview of discussions surrounding Part 1 of Discussion 2, Activity 9.

The second discussion around Activity 9 unfolded in four parts, taking up the entire duration of Lesson 16 (approximately 50 minutes). The discussion's first part began with the instructor soliciting students' ideas about how they answered the questions in the slide "A new twist" (see Figure 8.20). With regard to the first question—"Estimate the population percent for the population from which these samples were taken"—students agreed unanimously on an estimate of 40%. Furthermore, students gave essentially the same justification, that the histogram shows most sample percents to be clustered around this value.

With regard to the second question—"Estimate the size of the samples"—things turned out quite differently. The discussion around this question contains much evidence that the reasoning employed by Peter and David near the end of the last lesson was not shared by many other students. Indeed, the discussion reveals that this reverse line of reasoning was highly problematic for many students, and seemingly rooted in their significant difficulties tying together ideas and information from Activity 8 and 9. In particular, and somewhat surprisingly, the distinction between samples and people emerged once again as precarious for some students. The following series of discussion excerpts illustrates a variety of these problems. The first, and very brief, excerpt is of Kit responding to the instructor's call for an estimate of the sample size:

Episode 1, Lesson 16:

53. Kit: […] obviously if 40% is the population percent it would be uhh pretty high number of samples, like probably 2000 or something. 'Cause they're mostly around that.
54. I: Oh, you're estimating the number of samples that were taken?
55. Kit: No, the sample size.
56. I: So you think that size was 2000
57. Kit: Size 2000
58. I: Ok, say it a little louder.
59. Kit: Sample size 2000, if the 40% was the population percent.

Notice how, at the start of the excerpt, Kit appeared to unwittingly slip into thinking about the number of samples rather than the number of people in samples (line 53). After clarifying that she was referring to sample size (line 55), Kit again confirmed her sample size estimate of 2000 (lines 57 and 59).

Now, Kit's estimate could not have been based on any past data from Activities 8 and 9. A coherent interpretation of the distributions shown in the tables in Figures 8.5 and 8.19 does not suggest that a sample size estimate of 2000 is at all warranted for a population proportion of 0.40. In fact, the discussions around Activity 8 had established only that sampling variability was independent of underlying population, while those around Activity 9 had established that variability and sample size were inversely related. Here, Kit was apparently relating population and sample size, but it is not clear whether she related them through variability. Unfortunately, the discussion abruptly moved to another students' idea, so Kit's reasoning remains a mystery. Nevertheless, I would speculate that Kit was actually thinking of the number of samples and that she confounded it with the label "sample size". The number of samples in this particular collection was 3000, while that in each collection shown in the tables of Activity 8 and 9 was 2000, 2500, or 3000. The consistency among these values' magnitude might have been the rationale for Kit's estimate of "sample size". In sum, Kit's (mis)estimate might reflect her difficulty in placing and coordinating the sampling information from the Activities 8 and 9.

Another student, unclear as to how to proceed from her experience in Activities 8 and 9, guessed that the sample size should be close to that used in previous activities. This is illustrated in the following excerpt.

Episode 2, Lesson 16:

60. I: Ok, Nicole?
61. Nicole: Uhh (3 second pause) I just guessed 500 ‘cause that’s the normal sample size that we used.
62. I (chuckles at Nicole’s response): So you, but you didn’t get any clue from the data that was presented?
63. Nicole: I don’t, no I don’t know how you’d figure that out.

The next discussion excerpt is of Sarah justifying her estimate of the sample size. The excerpt, lasting approximately 2.5 minutes, is parsed into three contiguous segments for ease of analysis. In her dialogue with the instructor, Sarah was referring to the two slides in Figures 8.19 and 8.20, which she had in front of her. These slides are re-presented in Figure 8.22 for the reader’s convenience.

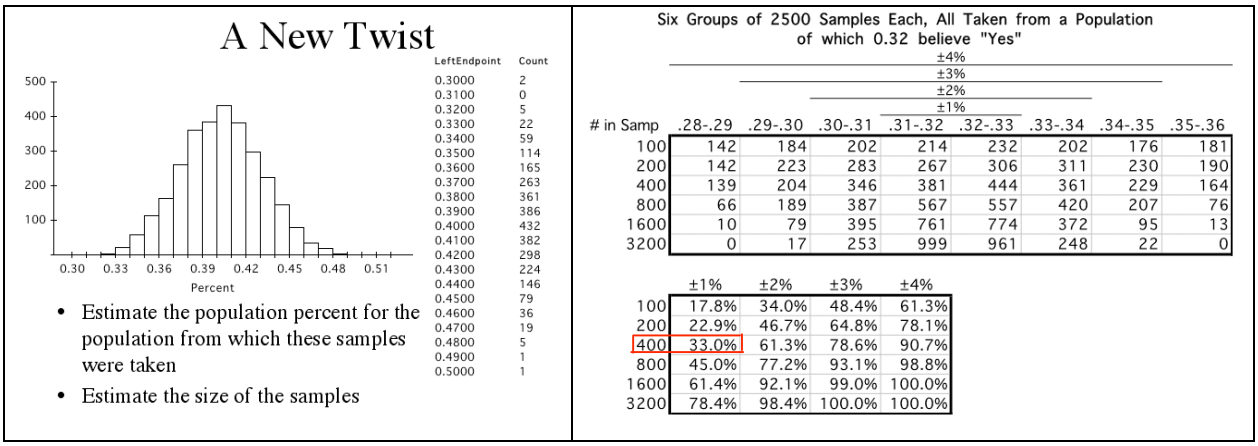


Figure 8.22. Two slides to which Sarah referred.

Episode 3, Segment 1, Lesson 16:

66. Sarah: I said 400, because uhh when we looked at uhh these ones we did the other day (raises left slide in her hand, Figure 8.22)
67. I: Huh huh
68. Sarah: That was about, it was about the same. No, wait (stairs at left slide, Figure 8.22)
69. I: What, what was about the same?
70. Sarah: Hold on, let me think (continues to stare at left slide in Figure 8.22 for 3 seconds). Yeah, because uhh the sample, the percent, was about the same as the one we got right here (points at right slide, Figure 8.22). Ok, look. We have this (holds up left slide in left hand, Figure 8.22) and this is 400 and this number is 33 (points to highlighted values in lower table in right slide, Figure 8.22)

71. Pat: And that number is what?
72. Sarah: And this is 40 (holds up right slide in right hand, Figure 8.22), and they're kinda close, so—

As indicated in Segment 1, Sarah's approach was to appeal to past sampling results. Her sample size estimate of 400 was apparently based on having compared the population estimate of 40%, from the distribution in the left slide, with the "33%" highlighted in the right slide. In her thinking, these two values—40% and 33%—were sufficiently close to warrant inferring that the samples drawn from the 40% population were also of size 400.

The instructor immediately asked Sarah for clarification of the meaning of these two values. The ensuing interchange follows in the next segment of the excerpt.

Episode 3, Segment 2, Lesson 16:

73. I: […] but what number are you pointing to? You said "this is 33". What is "this"?
74. Sarah: These are the samples within 1 percentage point of the population percent
75. I: All right. So—
76. Sarah: I don't [know] if that has anything to do with—
77. I (continues): percent of samples within 1 percent of the population percent
78. Sarah: Yeah
79. I: And that was about 33%, you said?
80. Sarah (affirms): Hmm hmm
81. I: And what did you find in this collection of samples? (refers to left slide, Figure 8.22)
82. Sarah: 40% was uhh population percent. I thought it would have something to do with it, so I put 400.

The second segment supports my analysis of Sarah's strategy. Moreover, and significantly, lines 74 and 78 suggest that Sarah understood the "33%" to stand for the percent of sample percents that were within one percentage point of the population percent. Line 82 suggests that Sarah understood the "40%" to be the estimated population percent (for population in left slide, Figure 8.22). It would thus appear that Sarah had compared two incomparable amounts, and that this had not dissuaded her from making her inference about the sample size. The instructor then continued the discussion, seeking further clarification from Sarah. The ensuing discussion is shown in the excerpt's third segment.

Episode 3, Segment 3, Lesson 16:

83. I: Ok, 40% was the population percent.
84. Sarah: Yeah.
85. I (continues): and so you used that to compare to the 33% that was (points to left slide in Sarah's hand, Figure 8.22)—All right, is 40%, is it, uhh, that … what does that 40% stand for?
86. Sarah: Population percent (inaudible)
87. I: So that's the percent of the population who believed "yes".
88. Sarah: Yeah.
89. I (continues): All right. What does the 33% in that table stand for? (points to the right slide that sits on Sarah's desk, Figure 8.22)
90. Sarah: The percent of the population--this, that was 1 percentage point uhh away from the percent of the population that said "yes".
91. I: The percent of the population that was within 1 percentage point.
92. Sarah (affirms): Hmm hmm, so it's not the same?
93. I: No, I don't think tha—that is—I don't think that's what it stood for. Nicole?
94. Nicole (looking at sheet in hand): It's the sample percents within 1 percentage point
95. Luke: Got it.
96. I (to Sarah): Yeah. It's the percent of sample percents that are within 1 percentage. See you were comparing a percent of a number of people to a percent of a number of samples.

The utterance in line 90 of the third segment is key evidence of where Sarah's thinking was problematic. Whereas previously she appeared to have interpreted the "33%" as a percent of sample percents, she now seemed to have slipped into interpreting it as the percent of the *population*. This shiftiness in her object of focus, from thinking that a value stands for a number of samples to thinking that it stands for a number of sampled items (people), is of the kind that many students experienced when trying to interpret histograms, in Phase 3.[16]

It is difficult to say, from these excerpts, whether one or the other interpretation was more stable for Sarah. However, the apparent instability of her interpretation certainly raises the possibility that she might have been thinking of the "33%" as a population percent, even when she called it a "percent of sample percents". This would explain her having thought it unproblematic to compare "33%" with "40%". Indeed, as Sarah explained shortly after Segment 3, *"… I didn't realize this [40%] was from people and this [33%] was from samples"*.

[16] See the analyses elaborated in Part 2 of Activity 7, in Chapter VII.

Activity 9, Discussion 2: Part 2 (Lesson 16)

Lesson	Activity (A)	Duration
16 (09/10)	A9: Investigating effect of *n* on variability—Discussion 2: Part 1	7 m.
16 (09/10)	**A9—Discussion 2: Part 2**	**8 m.**
16 (09/10)	A9—Discussion 2: Part 3	25 m
16 (09/10)	A9—Discussion 2: Part 4	9 m.

Figure 8.23. Chronological overview of discussions surrounding Part 2 of Discussion 2, Activity 9.

Prompted by students' difficulties in the first part of the discussion, the instructor made an impromptu move to help students clarify the distinction between quantities that refer to samples and those that refer to people. This marked a second part of the discussion. On the blackboard, the instructor made a very rough sketch of part of a hypothetical table of values (see Figure 8.24), mimicking the table that students had completed in Activity 8 (Figure 8.5). He then proceeded to engage students in interpreting values shown in the sketch, having them say very explicitly what the values represented.

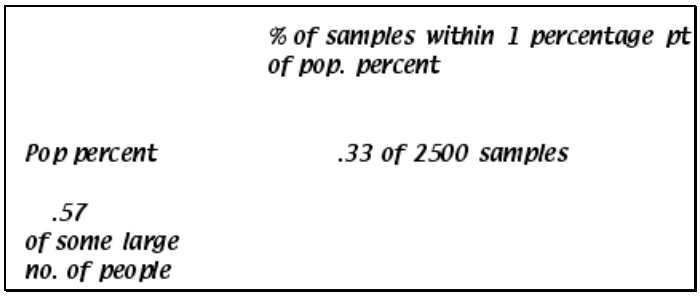


Figure 8.24. A rough sketch of part of a hypothetical data table mimicking the format used in the table of Activity 8.

The discussion culminated around the question of whether values in the sketch's "Pop percent" column can be compared with those in the "Percent of sample percents" column. The following excerpt, lasting approximately 1.5 minutes, is drawn from that culminating discussion. It illustrates the continued fragility of one student's thinking on this question. In the excerpt the instructor is leading students in explicating values shown in Figure 8.24.

Episode 4, Lesson 16:

134. I: If it said "57%", it's 57% of what?
135. Peter: People.
136. I: People! Uhh, do we know how many people?
137. Nicole and Peter: No.
138. I: Ok, so some large number of people (writes "of some large number of people" under ".57" on the board). All right. So, this 33% (points to .33 on the board), can we compare that to 33%, if 33% appeared over here (points to "Pop percent" on the board) are they comparable?
139. Nicole: No.
140. I: Chelsea?
141. Chelsea: I'm— (shakes her head from side to side)
142. I (to Chelsea): If we have, ok, now suppose this was .33 (erases .57 and writes .33 in its place). Population percent is .33. Over here in this column we get .33 (points to .33 that is % of sample %s …). Do these stand for the same things?
143. (Nicole silently shakes her head as though answering "no" to the question)
(7 second silence)
144. Chelsea (very hesitantly): Yeeeah, I think so because they're both a uhh a percent of a large number of people.
145. I: They're both a percent of a large number of people?
146. Chelsea: Uhh
147. I: Ok, this says 33% of what? (points to 33% of sample percents on the board)
148. Chelsea: Of the samples within—
149. I: Of 2500 samples.
150. Chelsea: Oh, no they're not. Because—
151. I: Ok. All right, so yeah, they're not. They're not percentages of the same thing.
152. (3 second silence)
153. Chelsea: 'Cause, that's 33% of the samples and that one's 33% of the population.

The excerpt began with the instructor clarifying that the ".57" represents a percent of some large number of people (lines 134-138). He then asked whether, if that value were changed to ".33", it would be comparable to the ".33" appearing in the "percent of sample percents" column of the table (lines 138 and 142). This line of questioning was intended to test whether students were oriented to comparing the table values on the basis of their magnitudes instead of their meanings.

When asked whether the two ".33" values now stood for the same thing, evidence of Chelsea's thinking emerged (line 144). After taking some time to ponder the question, Chelsea seemed inclined to think that the values do represent the same thing. Her underlying rationale is

interesting: she was not simply comparing the values' magnitudes, instead she evidently was thinking that both values were a percent of "a large number of people" (line 144).

After some prompting from the instructor, Chelsea reversed her decision and appeared to realize that the two values were, indeed, percents of different things (lines 147-153). Thus, Chelsea's confusion appears to have been momentary and therefore arguably inconsequential. I would, however, argue differently. My point is not that Chelsea had an *enduring* image of the 33% as a percentage of people. Rather, my point is that Chelsea's thinking, with regard to the meaning of these values, seemed labile and sensitive to perturbations. This suggests that there was an instability to her thinking.

It is useful to speculate about the potential sources of this instability in Chelsea's thinking. One obviously plausible problem is that interpreting the tables in Activities 8 and 9 entails the discrimination and coordination of many values. Understandably, it is evidently easy to loose sight of the meaning of one value when trying to coordinate that with the meaning of another one. But the unwitting conflation of samples with people was documented in Chapter VII as pervasive and robust in a context different from Activities 8 and 9. This impels me to entertain the possibility that perhaps a more foundational problem was at play—one that goes beyond the particular representational formats used in these activities, but that might nonetheless be exacerbated by them and express itself in Chelsea's particular difficulty. The hypothesis I proposed in Chapter VII seems plausible here: many students' image of a collection of samples of people tended to easily dissolve into an image of just people because they were not oriented to the significance of the distinction between these two structurings (see Figure 8.25).

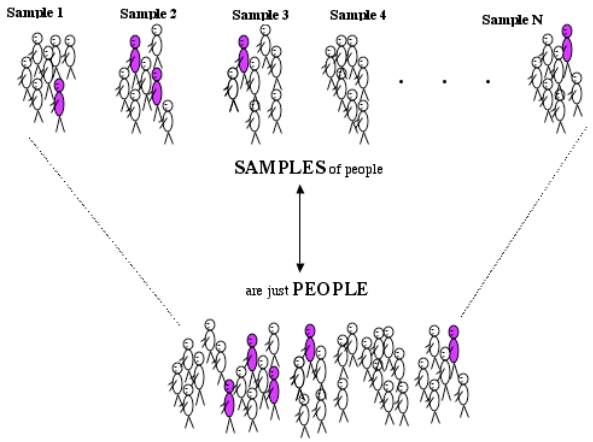


Figure 8.25. Two different ways to structure a collection of people, the significance of which is hypothesized not to have been salient to students.

This hypothesis would explain the slippage in Chelsea's thinking, which seems to indicate her having conflated a large population of people with a large number of samples of people selected from that population. This conflation would be understandable in the context of the task, were the significance of the distinction between these two structurings not salient in one's mind.

Activity 9, Discussion 2: Part 3 (Lesson 16)

Lesson	Activity (A)	Duration
16 (09/10)	A9: Investigating effect of *n* on variability—Discussion 2: Part 1	7 m.
16 (09/10)	A9—Discussion 2: Part 2	8 m.
16 (09/10)	**A9—Discussion 2: Part 3**	**25 m**
16 (09/10)	A9—Discussion 2: Part 4	9 m.

Figure 8.26. Chronological overview of discussions surrounding Part 3 of Discussion 2, Activity 9.

The discussion excerpts above, drawn from the first two parts of Discussion 2 in Activity 9, unfolded over a period of approximately 15 minutes. During this time the classroom teacher sat quietly at the periphery of the room listening to students' comments. He recognized in these discussions significant and widespread confusion among students about the ideas under discussion. This prompted him to interject the observation that students were exhibiting many disconnections in their understanding of the ideas in Activities 8 and 9 that required clarification. He then asked the instructor for permission to make an impromptu effort at tying ideas together

for students. He proceeded with a 22-minute-long lecture tying together ideas discussed over the last 4 lessons.[17] This lecture marked the start of a third part of the lesson.

Briefly, the classroom teacher's lecture highlighted these ideas:

- We repeated drawing a sample of a given size from a population whose parameter value we knew (i.e., percent of population who believed "Yes" on some issue)
- We did this activity for samples of different sizes and with different populations
- We computed the percent of people in each sample who believed "Yes" on the same issue
- We noted these sample percents and began to anticipate their values after having selected large numbers of samples
- We represented those sample percents in histograms and used the histograms to gives us a sense of patterns that could emerge in collections of sample percents
- We noted that the patterns of variability among sample percents for samples of a given size were independent of the sampled population
- We noted that the variability among sample percents tended to be greater for smaller samples
- The bottom table shown in Figure 8.19 establishes how far sample percents are expected to deviate from the underlying population percentage, in general, for various sample sizes. This table can therefore be used to predict sample size given information about the variability of a collection of sample percents.

The instructor followed up the teacher's lecture by summarizing some of these points. He, like the classroom teacher, also realized that students were very unclear about how to proceed with the questions posed in the slides "A new twist" and "Another new twist" (Figure 8.20). He recognized that many were evidently not catching on to the approach of reasoning in reverse to infer a sample size from past information and conclusions. He concluded this part of the discussion by framing the task as being akin to playing a game of detective; the goal was to determine a missing bit of information—population percent and sample size—on the basis of a

[17] This development was neither planned nor part of the research team's instructional agenda. Nevertheless, it is worth summarizing what the teacher said here because of its potential influence on students' subsequent thinking, and because it influenced how the instructor wrapped up the lesson.

number of available clues—namely, information provided about a particular distribution of sample percents and its variability—and by relating these clues with what is already known.

Activity 9, Discussion 2: Part 4 (Lesson 16)

Lesson	Activity (A)	Duration
16 (09/10)	A9: Investigating effect of *n* on sampling variability—Discussion 2: Part 1	7 m.
16 (09/10)	A9—Discussion 2: Part 2	8 m.
16 (09/10)	A9—Discussion 2: Part 3	25 m
16 (09/10)	**A9—Discussion 2: Part 4**	**9 m.**

Figure 8.27. Chronological overview of discussions surrounding Part 4 of Discussion 2, Activity 9.

The instructor concluded the discussion in Lesson 16 with a final part aimed at having students tie together relationships among 3 key ideas addressed in Activities 8 and 9: sampling variability, sample size, and population parameter. This part served as a capstone of sorts—a review of these key ideas and a drawing together of the conclusions that emerged out of Activities 8 and 9.[18] It culminated in the establishment of the network of dependence relations depicted in Figure 8.28. In fact, this very diagram emerged out of the discussion as a schematic representation of the most important relationships between the ideas.

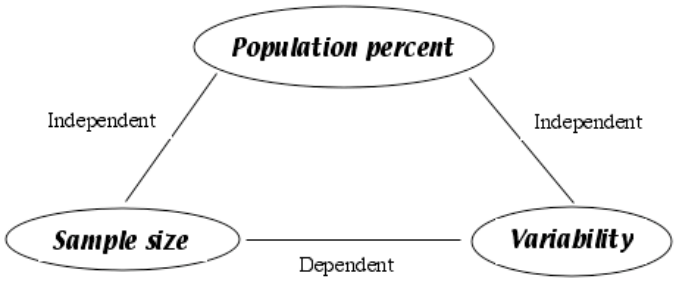


Figure 8.28. A network of dependence relations among population percent, sample size, and variability.

As each dependence relation shown in the diagram was elaborated, the instructor summarized that part of the discussion by explicitly characterizing that dependence. For instance, sample size and population percent are independent in the sense that either could have any value

[18] I should stress that this was not a lecture. Rather, it was more like a "consensus-building" discussion that involved students as active participants and co-developers of these relations in interaction with the instructor, who enacted his agenda. I do not provide details of this discussion here because they offer nothing more of interest to the analyses already elaborated.

and knowing one tells us nothing about the other. Similarly, variability and population percent are independent in the sense illustrated in Activity 8. That is, for any given sample size, the percent of sample percents contained within a given number of percentage points from the population percent is about equal for different population percents, so that a change in population percent has no significant effect on how variable are sample percents for samples of the same size.[19] Finally, sample size and variability are dependent in the sense that as sample size increases, the variability—as characterized above—decreases.

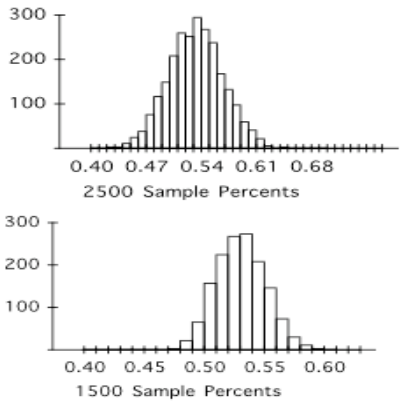
The instructor concluded this phase of the discussion by highlighting that the only factor affecting the variability of a collection of sample percents is the size of the samples.

At the end of Lesson 16 students were given a take-home activity entailing a series of questions around a sampling scenario.[20] The activity's culminating question, shown in Figure 8.29, was intended to query their dispositions to reason in reverse from distributions to estimate an unspecified sample size, and to make the connection between sample size and variability explored in instruction. Students' responses to this question are displayed in Table 8.1 (shown below Figure 8.29). They provide additional evidence of students' reasoning, above and beyond the evidence that emerged in classroom discussions.

[19] Again, I acknowledge the mathematical inaccuracy of this statement and I refer the reader to footnotes 1 and 7, in this chapter, for a qualification of it.

[20] This was intended as an in-class quiz designed to assess whether students were making connections among various ideas entailed in the reversible line of reasoning. However, the unexpected events that emerged during the lesson resulted in a lack of class time and the item was assigned as a homework activity.

Here are histograms showing sample percents that were calculated from two collections of random samples. All samples within a collection are the same size; samples in different collections are of different sizes. All samples were selected by computer from one (large) population of which 53% prefer spiked hair to crew cuts.



Which of these collections contains larger samples? Explain how you know this.

In previous activities, we established that sample size is inversely related to sampling variability, and that variability is indicated by the degree of a distribution's "spread out-ness" (i.e., density of its dispersion in an interval). Comparing the two histograms, it appears that the collection of 1500 sample percents is the less variable one — since 100% of its sample percents are clustered within a 13-percentage point range, whereas the collection of 2500 sample percents is clustered within a 23-percentage point range. This suggests that the collection of 1500 sample percents corresponds to a larger sample size.

Figure 8.29. A post-Activity 9 assessment question.[21]

Table 8.1. Students' written responses to the post-Activity 9 assessment question.

Student	Response
Nicole	I'd say that the collection with 2500 sample percents had a larger sample size. This is because the range in which the sample percents cover is larger/more spread-out.
Sue	The collection of 1500 samples contains larger samples because it has a small range of percents than collection of 2500 samples. As variability decrease, sample size increase.
Kit	The 2nd one Because there is less variability between samples. [Note: above top histogram Kit wrote "100% within 11% pts", and above the bottom histogram she wrote "100% within 7% pts"]
Sarah	The second one because there is less variability among the samples. (In other words, there are more samples whose percents of the people that said spiked hair that are closer to the population's percent that said spiked hair)
Peter	The one with 1500 samples percents because the sample percents are less variable "Less variable" = "the samples are more closely grouped toward the population percent"[22]
Chelsea	The one with 1500 sample percents. We know this because the distribution is less spread out around the graph. That means there is less variability. Less variability generally goes along with a larger sample size as we've discovered by looking at past experiments.
David	They are both the same sample size, the histograms are different
Luke	The first collection contains larger samples because the sample percents vary the most.

[21] The blue text did not appear in the activity questionnaire that students received. It expresses a line of reasoning consistent with that targeted in instruction, indicating the kinds of connections the research team intended would emerge from engagement with the activities. This "normative" explanation serves as a rough benchmark against which to consider students' responses.

[22] The red text was not part of Peter's written explanation. It is a quote from his oral explanation given during a brief class discussion in Lesson 17.

A central observation I make about these responses is that 6 of the 8 students recognized that the histogram showing 1500 samples percents is less variable than the one showing 2500 samples percents. On this basis, those students all concluded that the smaller collection (i.e., the one containing 1500 sample percents) contains larger samples. It would thus appear that most students had, at least implicitly, made the connection between sample size and variability that emerged in Activity 9: sample size is inversely related to variability. Two students (Chelsea and Sue) made explicit reference to this result.

The other two students reached a different conclusion: David concluded that there was no difference in sample size, despite having apparently noticed that the histograms differ. Nicole noticed that the percents in the top histogram had a wider range or were more spread out, from which she erroneously concluded that samples in that collection must be larger. It may be that Nicole relied on memory, which failed her and led her to a false conclusion.

In light of difficulties that students experienced earlier in Activity 9—in making the intended connection between variability and sample size and reasoning in reverse around this connection—it seems significant that at this juncture of the experiment three fourths of the students evidently employed a coherent line of reasoning that was highly compatible with that targeted in instruction. No less significant is the fact that they did so independently and outside of instructional interactions. This development is consistent with something having "clicked" or suddenly fallen into place for students. It may well be that Parts 2 and 3 of Discussion 2 helped solidify connections that perhaps were still only fragmentary for students prior to then. At the very least, this development suggests that the central ideas promoted in instruction around Activity 9 had some staying power for students. In the final section of this chapter, analyses of students' responses to post-experiment assessment questions will provide further insight into the extent of that staying power.

Activity 10: Exploring Margin of Error

The final instructional activity of the teaching experiment explored the idea of margin of error. Activity 10 differed from Activities 8 and 9 in that is was structured as a set of discussions directed at having students grapple with this idea in the context of interpreting a real polling scenario. The polling scenario and associated discussion questions are displayed in Figure 8.30.

Bradley In Dead Heat With Gore In NH, Poll Shows

Updated 11:03 AM ET September 5, 1999

BOSTON (Reuters) - Bill Bradley is in a statistical dead heat with Vice President Al Gore in the race for the Democratic presidential nomination in the key primary state of New Hampshire, a poll released Sunday by the Boston Globe and WBZ-TV found.

"The Democratic primary race really appears to be a race that could go either way," Gerry Chervinsky, president of KRC Communications, which conducted the survey of 800 likely voters, told the Globe.

Among those surveyed between Aug. 27 and Aug. 31, Gore led Bradley, the retired New Jersey senator, by 40 percent to 36 percent -- a statistical tie because the poll's margin of error was plus or minus 5 percentage points, the Globe said.

Other recent surveys showed Gore with a larger, albeit shrinking, lead over his only Democratic challenger for the party's presidential nomination.

Discussion Questions

1. How many samples were taken?
2. Was the sample a random sample? That is, is it reasonable to expect that it is somewhat representative of Democratic voters in New Hampshire?
3. In the phrase " Gore led Bradley ... by 40 percent to 36 percent", to what do 40% and 36% refer?
4. What do they mean by "the poll's margin of error was plus or minus 5 percentage points"?

Figure 8.30. Report of a public opinion poll and related questions that formed the basis of Activity 10 discussions.

Before describing this part of instruction, it will be useful to first elaborate the relatively sophisticated idea of margin of error, as the research team intended it be understood.

Margin of error: A conceptual analysis[23]

"Margin of error" is a technical term meant to convey a quantitative sense of how variable are sample statistics calculated from randomly drawn samples of a particular size. The "error" in margin of error conventionally refers to the expected deviation between a sample's statistic, or the *inferred* population parameter, and the *actual* underlying population parameter.

The conventional definition of margin of error is based on the idea of confidence interval. A "95% confidence interval" for a statistic is a range of values that is calculated from an estimate of the statistic's standard deviation (calculated from the sample's standard deviation) and centered at the statistic's value as calculated from that sample. "95%" means that we expect 95%

[23] I am indebted to Patrick Thompson for many of the central ideas elaborated in this analysis.

of the confidence intervals calculated from independent samples of the same size will contain the population parameter.[24,25]

The research team created an (almost) equivalent definition of margin of error in order to make the idea accessible to students without having to enter into the technicalities of sampling error, which itself depends on a sophisticated understanding of statistical estimation and sampling distributions. Instead of focusing on the sample statistic and calculating a confidence interval, we focused on the population parameter and the distribution of sample statistics in relation to it. The idea of margin of error used in this study comes from asking the question "In what range of the actual population percent will we find at least *x*% of the sample statistics when we randomly draw samples of a common size from the population?" The reader may recognize this as essentially the same question explored in Activities 8 and 9. Indeed, the procedures used in those activities to answer this question are precisely those used to compute a margin of error. Thus, margin of error is not about how accurate is any one sample (statistic), as in "by how much does this particular sample percent deviate from the true population percentage?". Rather, margin of error is about how accurate samples of a common size tend to be over the long run under repeated random selection, as in "what fraction of sample percents are expected to aggregated within certain percentage ranges of the sampled population percentage?".

In sum, margin of error refers to a particular sampling result only insofar as it is a statistical measure of the confidence or trustworthiness that the *sampling process* that produced that result will produce similar results in the future. Inferential statistics is fundamentally concerned with what happens over the long run, were an essentially identical selection process repeated a large number of times. Judgments of individual results' trustworthiness are based on how *collections* of such results behave. Moreover, the technical meaning of margin of error characterized here entails the quantification of two distinct attributes: on the one hand there is the *deviation* between the particular polling result and the intended population parameter's true value—this is what the "$\pm$5%" refers to in the Gore-Bradley poll (Figure 8.30). On the other hand there is the *likelihood*

[24] The interval length calculated from sample to sample will change according to each sample's standard deviation. The statement is that *this method* will produce confidence intervals that contain the population parameter 95% of the time.

[25] There is abundant confusion in both the lay and technical literature about margin of error. Some say that 95% of the calculated intervals will contain the statistic calculated from the original sample (American Statistical Association, 1998) while others say that 95% of the time the entire population is surveyed, the population parameter will be within the confidence interval calculated from the original sample (Public Agenda, 2003).

or *confidence* level of that deviation. This is a statistical sense of likelihood that is rooted in an operational conception of distribution; it is obtained by quantifying the dispersion of a collection of a sample statistic's values relative to a population percent's value, for samples of the same size as that used in the particular poll.

The difference between thinking of margin of error as referring to a *particular* sampling result (i.e., a particular value of a statistic) and thinking of it as referring to a *collection* of results that emerges in the long run, for samples of a common size, is hugely significant. It is, in fact, directly analogous to the important distinction drawn in the research literature between outcome and distribution-based perspectives on likelihood and inference (Kahneman & Tversky, 1982, Konold, 1989).[26]

It should be clear from this analysis, that the conception of margin of error targeted for instruction in Activity 10 is intimately tied to the notions of variability, accuracy, and distribution explored and developed in previous activities, Activities 8 and 9 in particular. In a sense, those activities were designed to lead up to this meaning of margin of error, helping to put into place conditions and conceptions that might support its emergence in this final activity of the teaching experiment. Indeed, having students engage with the idea of margin of error as the final activity of the instructional sequence was, obviously, not by happenstance.

To close this conceptual analysis, I offer a caveat and argue that the manner in which margin of error is typically presented in non-technical publications (e.g., newspapers and magazines) virtually disables anyone who does not already have the technical meaning in mind from developing that meaning. First, notice in the Gore-Bradley poll—and scores of others like it—the use of the definite article, underlined in the phrase "the poll's margin of error …". This can easily lead a reader to think that the margin of error is somehow saying something about the *particular* polling result obtained, thereby disorienting her or him from the intended issue. Second, reports of margin of error, at least in American media publications, typically mention only a measure of the error but not a measure of its confidence level as well. Consequently, readers who do not already have the technical meaning in mind or the knowledge or experiences to help them construct that meaning are unlikely to develop the intended interpretation of margin of error just from having read a statement like the one presented in the Gore-Bradley poll.

[26] In Chapter VI, and elsewhere (Saldanha & Thompson, 2002), I argued that the former is a relatively disabling conception. The analysis presented here implicitly extends those arguments to assert the importance of the distributional perspective for a coherent interpretation of margin of error.

Activity 10 Discussion Highlights

There were two separate but related discussions around the Gore-Bradley poll shown in Figure 8.30. A first discussion (*Discussion 1*) unfolded in two parts over approximately 16 minutes during the later part of Lesson 17. The activity was then revisited in the early part of Lesson 18, during which a second discussion (*Discussion 2*) unfolded, also in two parts, over approximately 22 minutes. The subsections that follow characterize those discussions and highlight students' ideas that emerged within them. The questions for Activity 10 are repeated here for the reader's convenience:

1. How many samples were taken?
2. Was the sample a random sample? That is, is it reasonable to expect that it is somewhat representative of Democratic voters in New Hampshire?
3. In the phrase, "Gore led Bradley … by 40 percent to 36 percent," to what do 40% and 36% refer?
4. What do they mean by "the poll's margin of error was plus or minus 5 percentage points"?

Activity 10, Discussion 1: Part 1 (Lesson 17)

Lesson	Activity (A)	Duration
16 (09/10)	A9: Investigating effect of *n* on sampling variability—Discussion 2: Parts 1-4	50 m.
17 (09/13)	**A10: Exploring margin of error—Discussion 1: Part 1**	**7 m.**
17 (09/13)	A10—Discussion 1: Part 2	9 m.
18 (09/14)	A10 revisited—Discussion 2	22 m.

Figure 8.31. Chronological overview of discussions surrounding Part 1 of Discussion 1, Activity 10.

The first discussion around margin of error introduced students to the Gore-Bradley poll scenario. The discussion unfolded in two broad parts. The first part, lasting approximately 7 minutes, centered on making sense of what the poll was about and answering Questions 1 through 3.

Students agreed that the poll was, as one student framed it, *"a pre-voting type of sample to see what people thought"*. They recognized the report as a public opinion poll that attempted to anticipate the outcome of the race for the Democratic party leadership between Al Gore and Bill

Bradley in the New Hampshire primary. Furthermore, there was a consensus in class that the poll consisted of a single sample. This point was also strongly stressed by the instructor.

When the conversation turned to Question 2, several students agreed that the poll was presumably representative. This agreement seemed to be based on an implicit assumption that such opinion polls typically use random sampling. The issue of the poll's presumed representativeness raised questions in the minds of some students about what constitutes an unbiased poll. This, in turn, led to the questions of *who* was sampled and *what* was the intended population. The instructor drew students' attention to the intended population mentioned in the report —"the survey of 800 likely voters", and asserted that not just anyone was polled. Students agreed with this assertion and began to consider who might have been polled. Eventually, after some possibilities had been proposed which suggested that "likely voter" could entail a combination of characteristics, the class began to realize that identifying the sample and the population was no straightforward matter. This issue was never fully settled in the discussion, though the class seemed content to consider Democratic voters in New Hampshire as the intended population.

In discussing Question 3, a few students offered their interpretations of the phrase "Gore led Bradley … by 40% to 36%", suggesting that they understood the "40%" and "36%" to be referring to percentages of people polled who said they would vote for Gore and Bradley, respectively. No one proposed a different interpretation.

Finally, the discussion turned to Question 4, what the research team considered to be the heart of the activity. Students were asked to share their interpretation of the statement "the poll's margin of error was plus or minus 5 percentage points". The question sparked a small controversy among a number of students who apparently interpreted this statement as contestable. Interestingly, it turned out to be difficult for the instructor to pin students down to articulate their interpretation of this margin of error. In other words, a number of students apparently had some meaning in mind, as evidenced by their having interpreted the reported margin of error as somehow problematic. However, it was difficult for students to articulate that meaning.

The most coherent articulation of the statement's intended meaning of margin of error was offered by Sarah, who read her written response: *"the poll could've been, could've actually ranged from 5 percent points below to 5 percents above the actual percentage that resulted from*

the survey". When queried by the instructor, Sarah confirmed that she meant that the actual population percent could be 5 percentage points higher or lower than the result obtained in the poll.

Luke reacted to Sarah's interpretation by claiming it implied that the poll *"doesn't mean anything"*. He went on to explain: *"when it's five percentage points and there's only four percentage points separating the two, then that could completely change the outcome of people who use this poll"*. Luke apparently meant that a 5 percent margin of error for polling results that differed by only 4 percentage points renders the poll useless as a reliable predictor of the election results. It appears that Luke thus interpreted the margin of error as an indicator of that *particular* polling result's representativeness.

Nicole, in turn, responded to Luke in the following dialogue:

481. Nicole: Like, this poll's just taken to see if like the candidate has a chance
482. Luke: How close the race will be
483. Nicole: Yeah. So the margin of error, I mean, it means it's, this isn't gonna tell you exactly who's gonna win
484. I: That's right, it isn't.
485. Nicole: So, the margin of error really doesn't make a difference. It's just gonna show you whether or not they're gonna be close or not.
486. Peter: Like everyone--
487. Nicole: I think!
488. Peter (continues): everyone that said they were voting for Gore could all change their minds. They could all vote for Bradley

It is difficult to read from these comments precisely what Nicole or Peter understood margin of error to mean. However, the comments suggest that these students thought it had something to do with predicting how the actual vote *might* turn out. Their idea seemed to entail a strong sense of the uncertainty of the outcome of the actual vote, but it also seemed to lack a sense of how the particular poll results are related to the reported margin of error. If Peter and Nicole were mindful of a connection between the two, they were not articulating that connection in a transparent way.

The instructor reacted to these comments by asking whether the reported margin of error was a statement about the particular poll that was conducted: *"does it say anything about the particular poll's accuracy?"*. Two students answered "*no"* to this question. The instructor then went on to explain that the accuracy of a particular sampling outcome (percent) cannot be known

without also knowing the sampled population percent—the very thing the poll was designed to estimate. These comments incited several students to ask how a margin of error is actually determined, thus suggesting that they had, understandably, an opaque sense of the idea's intended meaning. These questions set the stage for a second part of the discussion.

Activity 10, Discussion 1: Part 2 (Lesson 17)

Lesson	Activity (A)	Duration
16 (09/10)	A9: Investigating effect of *n* on sampling variability—Discussion 2: Parts 1-4	50 m.
17 (09/13)	A10: Exploring margin of error—Discussion 1: Part 1	7 m.
17 (09/13)	**A10—Discussion 1: Part 2**	**9 m.**
18 (09/14)	A10 revisited—Discussion 2	22 m.

Figure 8.32. Chronological overview of discussions surrounding Part 2 of Discussion 1, Activity 10.

Prompted by these last questions, the instructor made a first effort to orient students to the intended meaning of margin of error by embedding the notion within the ideas of Activity 9. He began by following-up Sarah's idea, in Part 1 of the discussion, by saying that the accuracy of a particular polling result (percentage) can never be assessed with certainty, due to sampling variability and lack of information about the underlying population percent. Instead, he said, the most that can be said is what we expect will occur over the long run in many samples *like* this particular poll—that is, of the same size and selected at random from a common population.

The instructor then turned students' attention to the table, created in Activity 9 (Figure 8.33), showing the distribution of sample percents for various sample sizes.

	±1%	±2%	±3%	±4%
100	17.8%	34.0%	48.4%	61.3%
200	22.9%	46.7%	64.8%	78.1%
400	33.0%	61.3%	78.6%	90.7%
800	45.0%	77.2%	93.1%	98.8%
1600	61.4%	92.1%	99.0%	100.0%
3200	78.4%	98.4%	100.0%	100.0%

Figure 8.33. The distribution of sample percents around the population percent, for n = 800, encased in red.

He pointed out that 800 people were surveyed in the news poll and asked students to consider the distribution of sample percents for the collection of samples of size 800 in the table. Here, he focused students' attention on the pattern in the percents of that row of the table, pointing out

how it indicates that 99% of the sample percents were contained within 4 percentage points of the population percent. He then summarized the intended implication of this observation for the class: this tells us we can expect that almost every time—99% of the time—we choose a sample of this size, its percent will be within 4 percentage points of the underlying population percent. He then asked, "What would you expect to be the case for 5 percentage points within the population percent?". Peter immediately responded that practically 100% of sample percents would be contained within this interval.

Out of concerned that few students were on the same wavelength as Peter, the instructor moved to reinforce the results shown in the highlighted row of the table (Figure 8.33) with the aid of a visual representation. He projected a histogram of the distribution of 2500 sample percents, for samples of size 800, on the white board. He then proceeded to quickly talk students through the process of compiling the results shown in the highlighted row of the table, marking the histogram as shown in Figure 8.34 and pointing out to students how it appeared that practically *all* of the sample percents were contained within a 5-percentage-point range around the population percent.

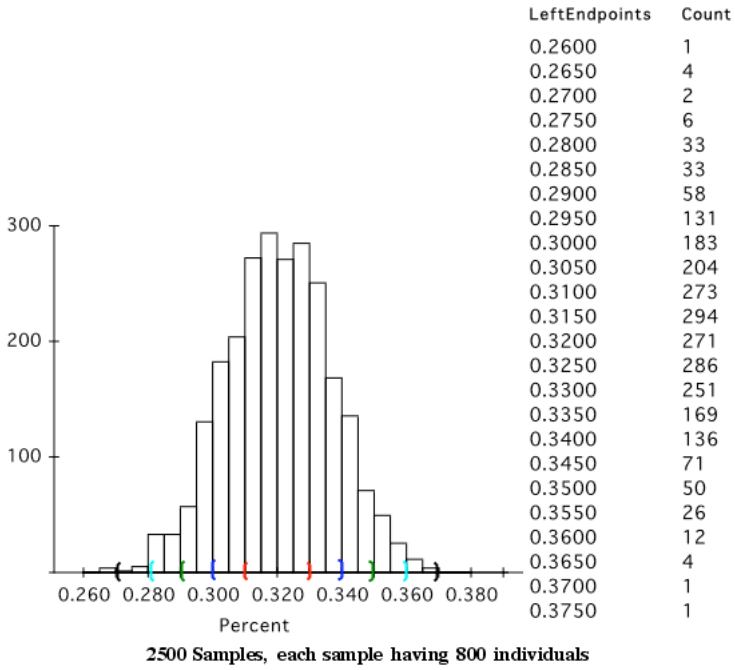


Figure 8.34. The various percentage intervals around the population percent highlighted directly on the histogram's horizontal axis.

Luke reacted immediately to this demonstration, seeming to make an important connection between margin of error and variability:

515. I: […] Do you see what I'm getting at?
516. Luke: So margin of error is the—almost has, almost has something to do with variability.
517. I: It sure does! Yeah, it addresses how variable we can expect our results to be, if we were to repeat them over and over.
518. Luke: So margin of error is the same thing as the, same thing as what you just did right there. You just found the margin of error. It's what we've been doing in class.
519. I: Uhh, basically. Yeah. But the important thing to note is that margin of error doesn't say anything about the one sample you took. It says "If I were to repeat this process of taking samples this way, how confident could we be that our results will be within 5 % of the population percent?" Well we saw, both from the table and this example, that 99% of the time you can expect your sample's percent to be within five percentage points of the population percent.
520. CT: As long as you took 800 people.
521. I: As long as you keep the sample size the same—yeah, you'd have to simulate this situation drawing the same size (3 seconds pass while he looks at another sheet). Do you have a sense of what this means? It's important that it's not about that one sample you took. It's about if you were to repeat this many many times, what percent of the time would your samples be within a range of 5% of the population percent.

The lesson culminated with Luke's realization of an intimate connection between margin of error and the activities that the class had been engaged in during the past four lessons. This development occurred near the very end of the lesson and no time remained for the instructor to involve other students. The idea of margin of error was revisited in Lesson 18, the final instructional session of the experiment. The next section highlights aspects of the discussion that took place then.

Activity 10, Discussion 2: Part 1 (Lesson 18)

Lesson	Activity (A)	Duration
17 (09/13)	A10: Exploring margin of error—Discussion 1: Part 1	7 m.
17 (09/13)	A10—Discussion 1: Part 2	9 m.
18 (09/14)	**A10 revisited—Discussion 2: Part 1**	**8 m.**
18 (09/14)	A10 revisited—Discussion 2: Part 2	14 m.

Figure 8.35. Chronological overview of discussions surrounding Part 1 of Discussion 2, Activity 10.

The discussion in this final lesson unfolded in two parts and was structured much like Discussion 1. The first part, lasting approximately 8 minutes, focused on soliciting students' interpretations of margin of error. A cross section of students' ideas is nicely illustrated in the following excerpt from this part of the discussion. In the excerpt, lasting approximately 3 minutes, students are responding to the instructor's invitation to share what they remembered about margin of error from Discussion 1. The excerpt is divided into 5 consecutive segments for ease of identification, each revealing information about a particular student's interpretation.

Activity 10, Discussion 2, Episode 1, Lesson 18:

Segment 1: Luke

120. I: right. Here's a test for, to, to see how much sense you made of the conversation yesterday, All right? Luke, what would that, what does that [margin of error] mean to you?
121. Luke: Well yesterday we had like a discussion, we were talking the Gore and Bradley, how it said that there's a margin of error of plus or minus 5 percent, and it showed that Gore uhh led by 40 percent to Bradley's 36 percent
122. I (affirms): Huh huh
123. Luke: So I was arguing that the poll didn't really mean anything because if it, if it varied uhh 5 percentage points then they could just be the exact same.
124. I: Or, it could be
125. Luke: vice versa, yeah
126. I: vice versa. All right, but what does that plus or minus 5 percent mean? What is, it's 5 percent of <u>what</u>?
127. Luke: Uhh, the people they surveyed uhh it could change. Like
128. I: Ok
129. Luke: like Gore's 40% could be 45% or 35%.

Segment 2: Kit

130. I: Ok. Kit, what did you understand it to mean? Based on the conver—on the discussion last, yesterday
131. Kit: How variable they were (waves hand in a horizontal motion).
132. I: How variable what is? The sample they took?
133. Kit: The samples (makes an encompassing motion with hands)--uhh, how they can (motions with hands on horizontal plane as though denoting a "crossing over" or "overlapping"), uhh the different percentages of the samples.
134. I: Of what samples?
135. Kit: Well, it, we, the Gore and Bradley, 5 percent
136. I: Ok, uhh so, so that,
137. Kit: (inaudible) variable
138. I (continues): that particular sample.
139. Kit: Yes

Segment 3: Sarah

140. I: Ok, Sarah.
141. Sarah: Hmm?
142. I: What's your interpretation of the plus or minus 5 percent from your disc—uhh the discussion yesterday?
143. Sarah: Uhh, oh the uhh the percent that said "Gore" or "Bradley", whatever, could've been uhh—I don't know why, I don't understand why— but it could've been 5 percent more of the people or 5 percent more

Segment 4: Chelsea

144. I: Ok. Chelsea, what was your understanding of the—?
145. Chelsea: Uhh, I wasn't here yesterday but I put down that it was like anywhere between 5 percent, so meaning that the 40 percent could've been anywhere between 35 percent and 45 percent.
146. I: What do you mean "it could've been"? Like, I mean is that like saying "I have 3 quarters in my pocket (jiggles change in pocket), but those 3 quarters could've been 5 quarters"? Or "those 3 quarters could've been 2 quarters"? What is the it that could have been?
147. Chelsea: The uhh percent, 40 percent
148. I: Of what?
149. Chelsea: Of the population
150. I: Of the population, you're saying?
151. Chelsea: Right.

Segment 5: Sarah and David

152. (Sarah raises hand)
153. I: Ok. Sarah?
154. Sarah: Are they saying that they're uhh sample isn't an exactly accurate uhh representation of the population?
155. I: Uhh (ponders a response)
156. David: Could it be up to 5 percent off? Like it could be 2 percent off or 1 percent off, or (inaudible) 5 percent?
157. I: All right. So those are all—I mean, that could be what it means. The question is, is that uhh the person who wrote that had in mind what it meant, and so in a sense what we're trying to do is guess what meaning they had in mind, or what meaning is conventionally given to that. All right, now here's--, ok here's how it's typically intended. That plus or minus … all right let me give you an analogy, ok? An analogy and a real brief activity.

These interpretations of margin of error seem highly compatible with those that students expressed in Discussion 1. With the exception of Kit's segment, a commonality among these interpretations is the idea that margin of error says something about the uncertainty of the particular polling results obtained, or the uncertainty of the inferred population percentages. Students seemed to focus on the idea that a poll result of 40% having a margin of error of 5

percentage points means that the inferred population percent could actually be anywhere from 35% to 45%. Thus, what was salient for students was the numerical bound on the possible deviation between the sample percent and the true population percent. Significantly, their sense of this bound was apparently not accompanied by a sense of the statistical confidence or likelihood of it. Indeed, there is little evidence in these discussion segments to suggest that students were thinking about this bound relative to anything other than the particular sample drawn.

Activity 10, Discussion 2: Part 2 (Lesson 18)

Lesson	Activity (A)	Duration
17 (09/13)	A10: Exploring margin of error—Discussion 1: Part 1	7 m.
17 (09/13)	A10—Discussion 1: Part 2	9 m.
18 (09/14)	A10 revisited—Discussion 2: Part 1	8 m.
18 (09/14)	**A10 revisited—Discussion 2: Part 2**	**14 m.**

Figure 8.36. Chronological overview of discussions surrounding Part 2 of Discussion 2, Activity 10.

The instructor sensed that students were missing this crucial part in their interpretation of margin of error. He thus steered the discussion into a second part, lasting approximately 14 minutes, in which he engaged students with an analogy designed to help them understand the intended meaning of margin of error. The analogy was presented in the form of a story that students might easily relate to. It centered on elaborating two different perspectives on accuracy and error in a linear measurement. Consider a building contractor who has a crew of carpenters working under his charge. Now, suppose the contractor is asked how accurate is a specific measurement made by one of his crew. There are two perspectives from which to consider this question.

1. The carpenter's perspective considers a *specific* item and is concerned that a *particular* measurement of the item is within a specified tolerance of its actual measurement.
2. The contractor's perspective considers *all measurements* taken by that carpenter and is concerned with *what percent* of those measurements are within a particular range of the items' actual measures. That is, the contractor knows about this carpenter's general performance but knows nothing about that particular measurement.

Thus, a particular carpenter might be able to answer how accurate is one of his measurements by estimating how far off the measurement is from the item's true measure. The contractor, on the other hand, has no information about particular measurements made by particular carpenters. He or she does not know how accurate specific measurements are. The most the contractor can say is something like this: "When we've studied this issue in the past, 95% of this carpenter's measurements were within plus or minus 1 millimeter of the items' actual measures, as determined by a much more accurate measuring instrument. So while I cannot say how accurate this particular measurement is, I can say that because 95% of the time measurements made by this carpenter were within $\pm$1 millimeter, I have great confidence that this measurement is very accurate".

After sharing this story with students, the instructor moved to have them relate it to the Gore-Bradley poll (Figure 8.30). He began by asking students whether anyone could know how accurate the particular poll's result was. Sarah contended that the only way to determine such accuracy would be to compare the result with that obtained from sampling the entire population. Students were then asked whether the question "how accurate is this sample?" was like a question for the contractor or for a carpenter. It was in response to this question that students seemed to begin making sense of the intended analogy and broaching some important connections. This is illustrated in the following discussion excerpt, lasing approximately 3 minutes.

Activity 10, Discussion 2, Episode 2, Lesson 18:

Segment 1

179. I: Now, asking "how accurate is this sample?" is like what, in regard to the story I just gave you? (2 second pause) Is this a question to the contractor or is this a question to the carpenter? (2 second pause) "How accurate this sample is?"
180. Kit: Carpenter
181. Peter: Carpenter
182. Nicole: Carp—
183. I: It's a question to the carpenter, right? But who are we like?

(3 second silence)

184. Nicole: Contractor
185. Sarah: Contractor
186. I: We're like the contractor! (2 second pause) We can't say how accurate that measurement is. But what can we say?

(5 second silence)

Segment 2

187. Peter: About this? (points to the Gore/Bradley poll on his desk)
188. David: About, most of the time he
189. Nicole: approximately
190. I: What can we say about samples that have 800 people in them?
191. Sarah: Most of the time they're between, they have a percent, well, most of the—(inaudible)
192. (Nicole chuckles at Sarah's abandon)
193. I: well, if we're talk--
194. Peter: Well all we can say is what we have information on right here (points to Gore/Bradley activity sheet on his desk)
195. I: That's all we can say about that particular sample.
196. Peter: Yeah.
197. I: It's the information that we have right there (points to Peter's sheet).
198. Nicole: 99%
199. Peter: Yeah.
200. I: Ok?
201. Sarah: Most of the time they're uhh at least five--close to 5%, close uhh away from the actual population percent?

In segment 2, we see evidence of the precariousness entailed in moving beyond thinking of the particular sample drawn in the poll to thinking of that sample as one of a class of samples of 800 people. In addition, the allusions made to "about", "most of the time" and "approximately" (lines 188, 189, 191, 201) indicate that many students sensed that a sampling outcome within 5 percentage points of the population percent would occur frequently. But these allusions also indicate that students were satisfied to express their sense of expectation non-quantitatively, thus perhaps suggesting that they were not yet oriented to quantifying this expectation. This raises questions about whether students made a connection between their experiences in Activities 8 and 9, interpreting the tables showing distributions of sample percents around the sampled population percents, and the quantification of expectation and likelihood developed in the earlier phases of instruction. The third segment of this episode follows.

Activity 10, Discussion 2, Episode 2, Lesson 18:

Segment 3

202. I: Yeah. Act … if you went to those tables that we had, then you can say, "well you know what? 99% of the time when you take an 800-person sample, it's within plus or minus 5%" (3 second pause). That's what that plus or minus 5% means.
203. Sarah: So then they're just saying, they're not just saying in this sample. They're saying from, like other type of sample they've taken.

204. I: That's right! From other things that we know! Things that we know about taking samples of size 800. But they don't know anything about this particular sample. (3 second pause) See, so it's kind of weird because they're saying there's a margin of error, plus or minus 5 percent. But they're not saying anything … they're not saying that this sample certainly is within plus or minus 5 percent, of the actual population percent. They're not saying that!
205. Peter: So they didn't, they didn't get that information anywhere from taking this survey.
206. I: That's right! That's good, Peter! They did not get that information from taking this survey! This plus or minus 5%.
207. Luke: So they're using this from past reference.
208. I: They're using it from the kinds of things that we've been doing with these uhh simulations. Looking at how samples, you know, how samples group themselves around the population parameters.
209. Peter: So
210. David: So they just figured out how accurate it could be.
211. I: Well, they figured out how, you know, what percent of the time you get samples within plus or minus 5%
212. Peter: So if they would've had a bigger sample size then that would've been different.
213. I: If—that's right! If they had had a larger sample size, then what might the margin of error have been?
214. David: Like 7 or 8, it went up, sample size would go up?
215. I: No.
216. Nicole: It would be like 2 or 3
217. Chelsea: No, it would go down
218. I: it might be two—so,
219. David (to Nicole): That's what I meant
220. I (continues): So, and, and how, and in what way would the margin of error go down?

(3 second silence)

221. I: Ok, Sue?
222. Sue: Uhh, as the sample size increase the margin of error decrease
223. I: That's right. As the sample increases, the margin of error will decrease—meaning, that you'll get a larger percentage of the samples closer to the population percent. Ok. All right, so that's, that's why this idea of margin of error is a really complicated idea. Because, it's not like the Carpenter's margin of error. It's like the Contractor's margin of error. He doesn't know how accurate any particular measurement is. He just knows that most of the time, most of the measurements are pretty accurate. And he can get bounds on how accurate, you know, on how close they will be over the long run. But he can't give any information about any one particular measurement.

Segment 3 illustrates some students making a breakthrough (lines 203, 205, 207) in realizing that the reported margin of error is not determined from the particular sample drawn. It appears that this was an important realization for them, and that they were on their way to making the intended crucial dissociation between the particular polling results and the reported margin of error. The instructor capitalized on these realizations to stress this as an overarching idea to the entire class, forcing the connection between this realization and the table of results that emerged from their work in Activity 9. This, in turn, seemed to impel Peter to broach the further important realization that margin of error is related to, and thus effected, by sample size. A brief dispute about the direction of this relationship then culminated with Sue's formalizing the correct relationship (line 213-222).

Episode 2 of Lesson 18, above, evidenced the most substantive connections that students made during class discussions in Activity 10. The instructor's public summarization of their connections concluded the instructional discussions around the topic of margin of error. The remainder of Lesson 18 was devoted to a review of ideas discussed throughout the entire experiment.[27] This review was in preparation for a written assessment that students took during the next class.

Post-Instruction Assessment: Additional Information About Students' Thinking

The assessment questions were designed to provide the research team primarily with information about students' understandings of ideas explored in Phase 4 of instruction. In this final section of the chapter, I consider students' responses to a subset of these questions in some detail. To the extent possible, analyses of students' responses draw on the individual post-interviews that provided students with opportunities to elaborate on their written responses.[28]

The assessment questions are displayed in Figures 8.37 through 8.40. The blue text appearing below each question expresses a "model" response to the question. Model responses attempt to capture well-developed and powerful understandings, and coherent expressions of them, consistent with intended instructional endpoints.

---

[27] Appendix A contains a summary of the ideas reviewed. A written version of this summary was also given to students.

[28] The post-interviews were conducted within 7 to 10 days of Lesson 18. Interviews were semi-structured, however, and not all students had occasion to respond to the same questions.

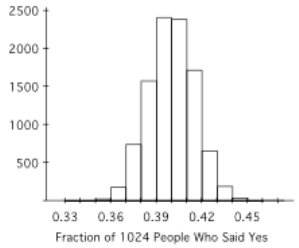
Students' written responses to the questions were coded according to their degree of consistency with the model responses. Degree of consistency was assessed qualitatively, according to a four-code scheme that emerged from inspection of the written responses:

- "Highly Consistent" response: entails *all* of the central ideas elaborated in the corresponding model response.
- "Somewhat Consistent" response: entails only *some* of the central ideas elaborated in the model response.
- "Inconsistent" response: departs significantly from the central idea(s) elaborated in the model response and/or entails a line of reasoning the leads to an erroneous conclusion.
- "Ambiguous" response: difficult to assess as any of the above without additional information.

Below each of Figures 8.37 through 8.40, a table displays the distribution of students' written responses, determined by applying this coding scheme. Each such summary is followed by more detailed analyses of students' responses to a subset of the questions. These analyses provide a cross section of students' ideas and they elaborate important distinctions between students' thinking and the conceptions targeted in instruction.

Assessment Question *2*

2. **The histogram and table at the right show the simulated results of selecting 10,000 random samples of 1024 people, each sample taken from a large population of which .40 believe "yes" on some subject.**



Interval Start	Counts
0.33000000	1
0.34000000	3
0.35000000	32
0.36000000	180
0.37000000	749
0.38000000	1585
0.39000000	2421
0.40000000	2393
0.41000000	1725
0.42000000	662
0.43000000	197
0.44000000	50
0.45000000	2

a. What does "0.38" stand for in the column under "Interval Start"?

It stands for a range, [0.38, 0.39), of proportions of people in a sample that believe "yes" on that subject. That is, it denotes proportions, $\hat{p}$, of people in a sample, satisfying the constraint $0.38 \le \hat{p} < 0.39$. For instance, the "1585" appearing in the corresponding row of the "Counts" column denotes that 1585 samples each had a proportion (of people who believe "yes") somewhere in the range [0.38, 0.39).

b. What does "50" stand for under in the column under "Counts"?

It stands for 50 samples, or more precisely, 50 sample *proportions*.

That this "50" appears in the same row as the value "0.44", in the "Interval Start" column, indicates that each of those 50 sample proportions were somewhere in the interval [0.44, 0.45).

c. We are going to simulate collecting 3000 random samples of size 1024 people, but from a population of which .60 believes "yes". Approximately what fraction of those samples will have between .59 and .61 of the people in them believing "yes"?

From our previous sampling activities, we established that for samples of the same size the variability among sample percents (proportions) is independent of underlying population percent (proportion). In particular, we noticed that for samples of the same size the percent of sample percents contained within 1 percentage point of the population percent is about the same across various populations. We can therefore expect these 3000 samples of size 1024 to exhibit patterns of dispersion very similar to those shown in the histogram above (since sample size is common). Hence, we estimate that about (2421+ 2393)/10,000 of these 3000 samples will have proportions in the interval [0.59, 0.61).

Figure 8.37. Question *2c* queried students' dispositions to employ connections developed in Activity 9.

Table 8.2. The distribution of students' written responses to Question *2*.

Question	*Highly consistent*	*Somewhat consistent*	*Inconsistent*	*Ambiguous*
2 a	Chelsea (1)	Nicole, Sue, Kit, Sarah, Peter, David (6)	(0)	Luke (1)
2 b	Nicole, Chelsea (2)	Sue, Kit, Sarah, Peter, Luke, David (6)	(0)	(0)
2 c	Peter, Nicole, Chelsea, Sue (4)	Sarah (1)	Kit, David, Luke (3)	(0)

This distribution of responses to Question *2* (Table 8.2) indicates that relatively few students gave responses that are highly consistent with the model responses.[29] It is useful to consider those points of inconsistency, as they suggest ways in which students' ideas might depart significantly from the intended instructional endpoints.

Questions *2a* and *2b*

Questions *2a* and *2b* aimed to assess how students interpreted entries in the data lists used in the activities. Question *2a* queried their interpretation of entries in a list's left-hand column, each of which conventionally denotes a half-open interval representing a *range* of sample proportions.[30] The 6 responses to Question *2a* that were coded as "somewhat consistent" (see Table 8.2) share a common feature: they all glossed over this subtle aspect and interpreted the "0.38" as a particular sample percent rather than as the included lower endpoint of a range of sample percents. Here is Peter's response, which is representative of that group*:*

"1585 samples out of 10,000 samples had 38% of each of the samples believing yes"

Compare this with Chelsea's response, the only "highly consistent" one:

"It stands for a marking point where between 38% and 39% said yes in 1 sample would be marked. In this case 1538 samples had between 38% and 39% saying yes"[31]

Both of these responses suggest that the students were able to read the data list with relatively little difficulty. But Chelsea's response suggests an attendance to detail that the others do not.

The distributions of students' responses to Question *2a* and *2b* are almost identical (see Table 8.2). Indeed, the distinction between the 6 responses coded as "somewhat consistent" and the 2 coded as "highly consistent" turns on precisely the same issue I elaborated for Question *2a,* above. Here is David's response, a representative from the "somewhat consistent" group:

"It stands for 50 samples of 1024 people, of those people 44% of the people in each separate sample believed yes on a subject"

Compare this with Nicole's, "highly consistent", response:

"That is the number of times (samples) that between 44% and 45% of the people believed 'yes'"

---

[29] The numeral appearing in parenthesis in each cell denotes the number of responses so coded.

[30] See the blue text below Question *2a*.

[31] Chelsea's response contains what appears to be a copying error ("1538" instead of "1585"). This error was considered to be relatively insubstantial for coding purposes.

The two responses thus differ in their interpretation of the "44%"—the constraint that the 50 samples satisfy. In the former, the 50 samples are seen as all having the same proportion of 0.44, whereas the latter entails an understanding that each of the samples has a proportion only *somewhere* in the interval [0.44, 0.45).

Recall that an interpretation similar to the former one emerged in the early part of Phase 4, during the first discussion around Activity 8. There, we saw how such interpretation was implicated in students' incorrect quantification of sampling variability, thereby disabling them from moving toward thinking of distribution. In these final assessment tasks, the distinction between these two interpretations might not express itself as important. However, it might play out in important ways more broadly and at a later time.[32]

The above distinction aside, a notable feature of the responses to Questions *2a* and *2b* is their reference to information in the data list's left column. This suggests that students were interpreting entries not as isolated values, but were instead oriented to making sense of them by coordinating with their interpretation of information contained in other parts of the data list. For instance, students evidently understood the "50" to stand for more than just "50 samples or percents"; they took it to stand for "50 samples that satisfied a particular constraint specified in the corresponding row of the left column". This coordination suggests that students were oriented toward a cohesive view of the data lists.

I might add that the lack of such coordinations seemed to be a crucial missing element in many students' thinking in Phase 3, when they experienced severe difficulties interpreting frequency histograms. That students were, by the end of Phase 4, evidently making such coordinations suggests that the structure of sampling distributions might, in some respects, have become less problematic for them to keep in mind. It is plausible that the instructional strategies and interactions in Activities 8 and 9 directed at having students carefully disentangle the meaning of entries in these data lists were useful in that regard.

[32] For instance, were the latter interpretation an expression of an encompassing notion of distribution as the enumeration of the actual values in a collection, rather than as a relatively course partitioning of a collection into classes of values, this could erect serious epistemological obstacles (Sierpinska, 1994) to understanding big ideas like probability density functions and cumulative density functions later on.

Question 2*c*

Question *2c* asked about a future collection of 3000 samples and represents the conceptual crux of this set of questions; its purpose was to query students' dispositions to use key connections, developed in instruction, between sample size, variability, and population as a basis for making an inference.[33] Specifically, the question queried whether students saw common sample size as the crucial link between the two sampling scenarios described in Question *2*, and how they used this realization to estimate something about the distribution of sample proportions in the second scenario.

As indicated in Table 8.2, half of the students gave a response that is "highly consistent" with the intended instructional endpoint. Here is Chelsea's response to Question *2c*, a representative from that group:

> *"Since this has the same number of individuals in a sample it can be compared to the information in the histogram above. Since about 1/2 or 48% were within 1% of .40 it is reasonable to believe that approximately 1/2 will have between .59 and .61 believing yes"*

A key feature of this representative response is the implied realization that the crucial link between the two sampling scenarios is a common sample size. Chelsea was evidently able to navigate through the two scenarios, unperturbed by the fact that population percents and numbers of samples selected were different in both. This suggests that the relationship between sample size and variability was both salient and enduring for students who responded similarly.

Contrast this response with one of the "inconsistent" responses, elaborated by Luke in the following excerpt from his post-interview. In the excerpt, Luke is explaining what connections he saw between the two sampling scenarios:

1. I: But you know what the question is asking, you understand what the question is asking?
2. Luke: Right. It's what fraction of the samples will-- will have bet-- between 59 and 60% or 59 and 61% of the people in them believing yes.
3. I: Right.
4. Luke: So you basically come back up to the histogram and you find--
5. I: (points to histogram in Question *2*) This histogram?
6. Luke: Uh, no. This histogram will not work cause this is for people where 40%-- this is a pop … this is taken from a population where 40% uh believe yes on a certain subject.

[33] See the blue text (Figure 8.37) below Question *2c*.

7. I: So you couldn't-- because it's 60% here (points at Question *2c*) we couldn't use this histogram, is that what you're saying?
8. Luke: That's correct.
9. I: Ok. What about the fact that it's uhh 3,000 random samples here (points to Question *2c*) and 10,000 over here, is that...?
10. Luke: Uh...yeah, that definitely was a big difference.
11. I: So if this was, if this was 60% here, not 40% (points to Question 2 opening scenario) and all other things being the same, could we then use this histogram?
12. Luke: Uh, no you couldn't because of the samples aren't the same, cause you got--
13. I: The samples?
14. Luke: Or the- the-- the number of samples, I'm sorry. Cause you got 10,000 random samples here and 3,000 here and there's no way that you can compare 10,000 samples and 3,000 samples.
15. I: Ok, ok. Even though the sample size is the same?
16. Luke: Well, sample size has nothing to do, it's-- it's the number of samples that are taken.

As the above excerpt clearly indicates, what mattered to Luke was that the number of samples and the population percents were different in both scenarios. To him, the common sample size was inconsequential for answering Question *2c*. Luke went on to explain that he expected the histograms for the two scenarios to look different because of the difference in these parameters. It would appear that after instruction, the salient and enduring relationship for Luke was between the size of a collection of samples and the shape of its corresponding histogram.

Not all students were as resolute as Luke in their conviction. Here is Sarah's, "somewhat consistent", response to Question *2c*; it combines ideas from both ends of the spectrum exemplified by the two previous responses:

> *"About 48%. This is just like the simulation above, except with a smaller number of samples. The population size is irrelevant in comparison to the above simulation. The population percents are irrelevant in this matter as well. The only factor here is the # of samples taken, at since they are both large numbers, they will have the same results. In the above simulation, 48% lied within 1% point of the pop. percent. This will be the same for this simulation"*

Sarah employed the intended line of reasoning, clearly noting the irrelevance of the different population percents for the variability of the two distributions. However, she also seemed to mistake comparably large numbers of samples as the crucial link between the two scenarios. The post-interview revealed that this was not simply a slip of the pen. Rather, Sarah had somehow

internalized the ideas that number of samples—as long as that number was at least 2000—*and* sample size were both determinants of a distribution's variability.[34]

[34] The basis of Sarah's first idea appeared to be some confusion in her recollection of how the two tables in Activities 8 and 9 (Figures 8.5 and 8.12) were distinguished. She seemed to have blended the numbers of samples—2000 and 3000—in Figure 8.5 with the relationship between sample size and variability, discerned from Figure 8.12, as two crucial factors. This blending is consistent with someone trying to "place" a value whose meaning is unclear to them among a web of relationships.

Assessment Questions *3a-3d*

3. **Here is a table similar to one we filled out in an in-class activity. Use it in answering parts *a* through *e*.**

Percent of Yes in Population	Number of People in a Sample	Number of Samples Drawn	% of Sample Percents within 1 Percentage Point of Population %	% of Sample Percents within 2 Percentage Points of Population %	% of Sample Percents within 3 Percentage Points of Population %	% of Sample Percents within 4 Percentage Points of Population %
65%	500	2500	36.7%	64.5%	84.8%	91.5%
32%	500	2000	37.1%	65.8%	83.9%	91.1%
57%	500	6800	36.2%	64.9%	84.2%	91.3
60%	500	5500	36.1%	65.2%	84.3%	91.4%

a. To how many populations does this table refer?

The table refers to four populations. Each row of the table tells the population percent—65%, 32%, 57%, and 60%—of each population and gives information about samples drawn from that population.

b. The entry in column 5, row 3 is 64.9%. That refers to 64.9% of what (be specific)?

"64.9%" refers to 64.9% of 6800 sample percents, each computed for a sample selected from the population of which 57% of it is "Yes"s.

The value indicates that 64.9% of these 6800 sample percents were within 2 percentage points of the population percent—that is, in the interval [0.55, 0.59).

c. The entry in column 1, row 4 is 60%. That refers to 60% of what (be specific)?

"60%" refers to 60% of people in a population. The value indicates that 60% of people in that population believe "Yes" on some issue.

d. All the percents in each of columns 4 through 7 are approximately the same. What can we conclude from that?

This pattern indicates that no matter what the underlying population percent, the fraction of sample percents contained within 1 through 4 percentage points of that percent — that is, the variability among sample percents — is about the same. This suggests the following generalization: sampling variability is independent of underlying population percent (for a given sample size).

Figure 8.38. Questions *3a-3d* queried students' understanding of ideas developed in Activity 8.

Table 8.3. The distribution of students' written responses to Questions *3a-3d*.

Question	*Highly consistent*	*Somewhat consistent*	*Inconsistent*	*Ambiguous*
3 a	Chelsea, Nicole, Luke, David, Kit, Sarah, Peter, (7)	(0)	Sue (1)	(0)
3 b	(0)	Sue, Kit, Nicole, Peter, Luke, David, Chelsea (6)	Sarah (1)	(0)
3 c	Peter, Nicole, Sarah, Chelsea, Sue, Kit (6)	Chelsea (1)	David (1)	(0)
3 d	(0)	Nicole, Sarah, Peter, Chelsea (4)	Sue, Kit, David, Luke (4)	(0)

Questions *3a* through *3d* queried students' understandings of the information displayed in the tables of Activities 8, and of the patterns in the distribution of sample percents.

Questions *3a*, *3b*, and *3c*

Students' responses to Questions *3a* through *3c* were largely unproblematic (see Table 8.3). The coding of 7 responses to Question *3b* as "somewhat consistent" turned on an arguably contentious issue: the 7 responses did not explicitly refer to details beyond those already mentioned in the description given in the column 5 table heading (e.g., see corresponding model response). This raises the possibility that students regurgitated that description—indeed, it appears that some did—which in turn makes it difficult to assess how well they understood what they wrote. Here is Nicole's response, representative of that group:

"64.9% of sample percents within 2 percentage points of the population percent"

Fortunately, there was occasion to pursue this issue with five students during the post-interviews. When pushed for elaboration on Question 3b, these students all gave expanded descriptions that were highly consistent with the model response. Moreover, under prompting from the interviewer, they often gave descriptions that entailed vivid imagery of the re-sampling process. Below is an illustrative example. It begins with Chelsea giving her description (line 1) and then unpacking it (lines 6-20) under prompting from the interviewer (L):

1. Chelsea: […] the 500 were drawn from, yeah, a big population of 57% saying "yes" and uhh 64.9% were within 2 percentage points of 57%.
2. L: Now, can you tell me a little about that, can you explain that statement to me, what does that mean, to say that "64.9% are within 2 percentage points of the population"?

[…]

3. L: To calculate 64.9%, uhh do you have to calculate anything else before you do that?
4. Chelsea: Uhh (2 second pause)
5. L: What do you do with each sample? You draw, you pick one sample of 500 people
6. Chelsea: Ok, say you have one sample of 500 people
7. L: Yeah
8. Chelsea: and you find out of that sample what percentage they have in "yeses"
9. L: Ok
10. Chelsea: and then they go in a certain, like, category
11. L: Ok
12. Chelsea: Like, between 56 and 57 percent saying "yes" or something

13. L: for instance?
14. Chelsea: yeah
15. L: Ok. So you do that for each of these 6800 samples?
16. Chelsea: Hmm hmm, right.
17. L: And that's, so again just clarify for me
18. Chelsea: and so the 64.9% are all the samples that fell within, between 55% and 59%
19. L: 55 and 59?
20. Chelsea: Hmm hmm, or, yeah because they were within 2 percentage points

Here is Sarah's response to Question *3b,* the only one coded as "inconsistent":

"2 percentage points away from the population percent is 64.9% of the people in each sample that said yes"

As the underlined portion of Sarah's response illustrates, the distinction between number of samples and number of people in a sample that emerged as precarious among many students in Phase 3, evidently remained precarious for her. There is little evidence in the post-interviews for the continued prevalence of such precariousness.[35]

Question *3d*

Before examining students' responses to Question *3d* more closely, it will be useful to elaborate this question's aim. Question *3d* was designed to query whether the independence relation between sampling variability and population percent (for a given sample size) developed in Activity 8 was salient to students. The table appearing in Question *3d* is similar to the one filled out in Activity 8, but the number of samples in each collection was now also made to vary greatly across the different populations (see 3rd third column). This particular adjustment was intended to elicit information about what students thought was important or unimportant, in relation to the invariance in the distribution of sample percents shown in the table and addressed in the question.

Students' responses to Question *3d* were equally divided between "inconsistent" and "somewhat consistent" (see Table 8.3). All of the "inconsistent" responses share an interesting common feature: they do not mention population percent as unimportant for the distribution of

[35] Although it is tempting to take this as compelling evidence that this problem was finally resolved, it is plausible that tasks different from the assessment questions could elicit such problems. It is thus difficult to conclude, one way or another, as to the status of this particular difficulty by the end of instruction.

sample percents. Significantly, three of those responses explicitly mention *number of samples* as unimportant for the distribution of sample percents. This suggests that the independence between variability and population percent might not have been as salient a relation to these students, despite it having been centrally developed and highlighted in instruction. Here is one of those responses, given by Nicole:

> *"You can conclude that the number of samples drawn will not affect the variability as long as sample size remains the same"*

Here is another of those responses, given by Chelsea:

> *"We can conclude that the size of the sample has more to do with variability then the number of samples drawn. Meaning getting variability depends more on the size of the sample rather then the number of samples drawn"*

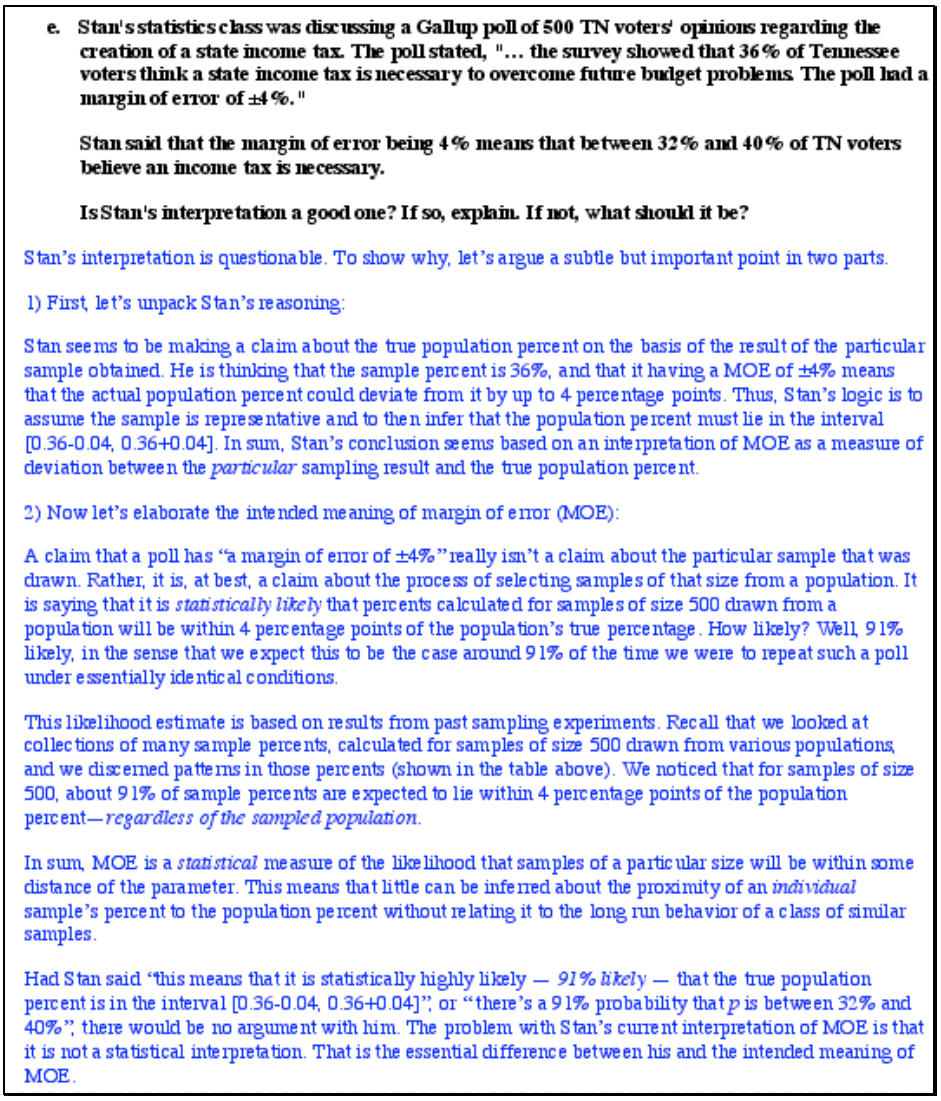
For Chelsea, the dependence of variability on sample size was evidently quite salient, for she invoked this relation in the face of information (in the table) that did not support making this conclusion.

All of the "somewhat consistent" responses to Question *3d* alluded to population percent as unimportant for the distribution of sample percents. Three of those responses did so explicitly, while two did so implicitly by describing how the distribution of sample percents was relatively invariant across the different populations. One response mentioned both number of samples and population percent as unimportant factors.[36]

In sum, responses to Question *3d* suggest that about half of the students were oriented toward population percentage as the salient unimportant factor for the distribution of sample percents, while the other half were oriented toward number of samples as the salient unimportant factor.

[36] "Somewhat consistent" responses differed from the model response largely in their degree of elaboration rather than in substance.

Assessment Question *3e*



e. Stan's statistics class was discussing a Gallup poll of 500 TN voters' opinions regarding the creation of a state income tax. The poll stated, "… the survey showed that 36% of Tennessee voters think a state income tax is necessary to overcome future budget problems. The poll had a margin of error of ±4%."

Stan said that the margin of error being 4% means that between 32% and 40% of TN voters believe an income tax is necessary.

Is Stan's interpretation a good one? If so, explain. If not, what should it be?

Stan's interpretation is questionable. To show why, let's argue a subtle but important point in two parts.

1) First, let's unpack Stan's reasoning:

Stan seems to be making a claim about the true population percent on the basis of the result of the particular sample obtained. He is thinking that the sample percent is 36%, and that it having a MOE of ±4% means that the actual population percent could deviate from it by up to 4 percentage points. Thus, Stan's logic is to assume the sample is representative and to then infer that the population percent must lie in the interval [0.36-0.04, 0.36+0.04]. In sum, Stan's conclusion seems based on an interpretation of MOE as a measure of deviation between the *particular* sampling result and the true population percent.

2) Now let's elaborate the intended meaning of margin of error (MOE):

A claim that a poll has "a margin of error of ±4%" really isn't a claim about the particular sample that was drawn. Rather, it is, at best, a claim about the process of selecting samples of that size from a population. It is saying that it is *statistically likely* that percents calculated for samples of size 500 drawn from a population will be within 4 percentage points of the population's true percentage. How likely? Well, 91% likely, in the sense that we expect this to be the case around 91% of the time we were to repeat such a poll under essentially identical conditions.

This likelihood estimate is based on results from past sampling experiments. Recall that we looked at collections of many sample percents, calculated for samples of size 500 drawn from various populations, and we discerned patterns in those percents (shown in the table above). We noticed that for samples of size 500, about 91% of sample percents are expected to lie within 4 percentage points of the population percent—*regardless of the sampled population*.

In sum, MOE is a *statistical* measure of the likelihood that samples of a particular size will be within some distance of the parameter. This means that little can be inferred about the proximity of an *individual* sample's percent to the population percent without relating it to the long run behavior of a class of similar samples.

Had Stan said "this means that it is statistically highly likely — *91% likely* — that the true population percent is in the interval [0.36-0.04, 0.36+0.04]", or "there's a 91% probability that p is between 32% and 40%", there would be no argument with him. The problem with Stan's current interpretation of MOE is that it is not a statistical interpretation. That is the essential difference between his and the intended meaning of MOE.

Figure 8.39. Question *3e* queried students' understanding of margin of error.

Table 8.4. The distribution of students' written responses to Question *3e*.

Question	*Highly consistent*	*Somewhat consistent*	*Inconsistent*	*Ambiguous*
3 e	(0)	Sarah, Peter, David (3)	Nicole, Kit, Chelsea, Luke (4)	Sue (1)

Question *3e* queried students' understanding of the idea of margin of error by having them comment on a particular interpretation of the reported margin of error for a public opinion poll of 500 people. It was of particular interest to the research team whether students would make any

connections between this scenario and the distribution of sample percents for samples of the same size shown in the table of Question *3*. This information can be used to determine the reported margin of error's confidence level, were one oriented to look for it.[37]

As indicated in Table 8.4, no student expressed an understanding of margin of error that I would construe as highly consistent with that targeted in instruction. Specifically, no student's interpretation of margin of error entailed both of these crucial elements: 1) a sense of the possible error or deviation between the particular sample percent and the actual population percent, and 2) a statistical sense of the likelihood of that deviation—one that is rooted in an operational image of the distribution of a collection of sample percents relative to a population percent.

Though the diversity among students' responses would certainly be instructive to elaborate, I restrict my attention to a particularly telling commonality among them: they all focused largely on the first element, and they all made no connection to the sampling distribution shown in the table in Question *3*. This observation is consistent with the evidence that emerged during the discussion in Activity 10 of students focusing almost exclusively on the "error" part of margin of error. It is also supported by evidence from the post-interviews. Let us consider two representative responses to Question *3e* and corresponding discussion excerpts from the post-interviews more closely.

Perhaps the most articulate of the "somewhat consistent" responses to Question *3e* was given by Sarah:

> "Not necessarily. This statement means that in the past they have done surveys of 500 *people, and they usually are 4 percentage points below to 4 percentage points above the actual population percentage. They do not know if this specific poll was exactly accurate, but they do know that this percentage they came up with is close to the population percent that [inserted: they know] lies in between 32%-40%"*

For the purpose of analysis, I parse Sarah's response into a sequence of statements:

I. *Not necessarily.*

II*a. This statement means that in the past they have done surveys of 500 people,*

II*b. and they usually are 4 percentage points below to 4 percentage points above the actual population percentage*

III*a. They do not know if this specific poll was exactly accurate,*

[37] I should clarify that students were not expected to produce a response to Question *3e* of a form and scope comparable to the model response. Rather, their responses were judged on the basis of whether they broached the substantive ideas expressed in the model response.

III*b*. *but they know that this percentage they came up with is close to the population percent that [they know] lies in between 32%-40%.*

Sarah's response appears to address two questions (see Figure 8.39):

1) Is Stan's interpretation of margin of error a good one? (i.e.,"Margin of error being 4% means that between 32% and 40% of TN voters believe an income tax is necessary"), and

2) What should be a good interpretation of margin of error?

Statement I of Sarah's response presumably refers to the first question; it asserts that Stan's interpretation is not necessarily a good one. Statements II through III refer to the second question; they describe what a good interpretation of margin of error should be. Given her description in these later statements, it is reasonable to interpret her first statement as referring also to Stan's definition. In other words, the "not necessarily" means that Sarah thought Stan was saying "it is *definitely* the case that between 32% and 40% of TN voters believe an income tax is necessary" and that she took issue with the certainty that Stan attached to his prediction. Her statement "not necessarily", then, is also a counterclaim to Stan's claim of certainty: she is asserting that there is, in fact, no guarantee that the true value of the population percent is between 32% and 40%.

Now let us examine Sarah's description of a "good" interpretation of margin of error, in statements II through III. In statements II*a* and II*b* Sarah expressed the idea that the values of the sample percent for samples of size 500 selected in the past have usually been within 4 percentage points of the population percent's true value. Sarah was presumably referring to the patterns she had observed in the instructional activities of Phase 4. What is unclear, however, is her intended reference in the word "they". She seems to have been referring to the Gallup company mentioned in the question scenario. But polling companies do not typically know a population percent's true value. Presuming that she was thinking coherently, by past "surveys" Sarah may have meant sampling simulations from populations having parameter values that are known (like those employed in the instructional activities).

In statement III*a*, Sarah acknowledges that the particular poll (result) in the scenario is not known to be "exactly accurate", by which I take her to mean that the particular sample percent's

value is not guaranteed to be highly representative of the population percent's true value.[38] This suggests that Sarah had a sense that the individual sampling result's representativeness cannot be assessed without reference to other information. In statement III*b*, however, Sarah expresses confidence that the particular percentage value obtained in the poll is close to the population percent's true value. This confidence is, again, presumably based on the distributions of sample percents that she had observed in the instructional activities. She concludes statement III*b* with the assertion that the population percent's true value is known to be in the interval (32%, 40%). This last assertion is a bit perplexing, as the population percent's true value is not known.[39]

Sarah's explanation touches on some of the key ideas entailed in the intended conception of margin of error. There are two specific values prominent in her thinking: the particular sample's percent and the true population percent. Also prominent in her thinking is an expectation that these two values might deviate from each other by up to ± 4 percentage points, as determined by the way prior sampling results "usually" turned out. However, as suggested by Sarah's vague sense of "usualness" and by her focus on the particular sample percent, she evidently was not thinking of a distribution of sample percents in relation to the population percent in an operational way. This assessment is further supported by the fact that Sarah made no evident connection to the table in Question 3. Had she been so inclined, Sarah could have used the sampling distributions shown in the table to express her sense of expectation or "usualness" quantitatively—that is, in terms of the fraction of sample percents, for samples of size 500, that are contained within 4 percentage points of the population percent. It is in this last point that Sarah's understanding differs most significantly from the intended conception of margin of error.

During her post-interview, Sarah was again asked to respond to Question *3e*, and was provided with all of the same information. At first she simply re-read her written response. Then, when prompted for elaboration by the interviewer, she made several attempts at elaboration that

[38] This interpretation of "accurate" is consistent with the one that students seemed to gravitate toward in Phase 3 of instruction.

[39] I would speculate that in making this claim, Sarah was perhaps reasoning in reverse from what she had observed in the instructional activities of Phase 4. She may have remembered that large proportions of sample percents were contained within 4 percentage points of the sampled population percentages, and then used this as a basis for inferring that the population percent in this case (i.e., percent of TN voters favoring a State income tax) is definitely contained in the interval (32%, 40%), thereby fudging over the fact that high statistical likelihood is not the same as certainty.

were consistent with her written response. Consider, for instance, the following excerpts from the start of the interview:

> *"Yeah, basically they've done, uh what I got from doing this in class was that uh from what they know in the past of doing uh polls that are about this size or surveys about that size, they usually run between 4 percentage below to 4 percentage page--, points above the population percent"*
> *[…]*
> *"It [margin of error] means that they could be, I mean the actual, uh, reality, not just the poll they took but like the actual, uh, fact of what they, you know they're, the actual thing could be actually, uh, 32% to, or 40--, or uh even up to 40%"*

Here is an excerpt from the end of Sarah's comments on Question *3e*:

1. Sarah: […] Well, they know that the population percent is somewhere, is around 36, either 4 up or 4 down, 4 percentage points higher or 4 percentage points lower, usually. Like that's the general case [short pause], like this is [?] how it happened.
2. I: Hmm hmm. About what, about what percent of the time might that happen?
3. Sarah: Uh [short pause] about 99% of the time.
4. I: Really? Is there anything on that page that you could use to answer the question I just asked?
5. Sarah: Uh [sighs, medium pause], uh [sighs], hmm no I don't think
6. I: Ok
7. Sarah: I don't know.

In the last sentence of line 1 of this excerpt, Sarah's use of the definite articles "that", "this", and "it" presumably referred to her recollection (from Activities 8 and 9) that large proportions of sample percents are expected to fall within a 4-percentage-point range on either side of the population percent. But Sarah appears to have reasoned in reverse from this observation to infer that the population percent in this case will be within 4 percentage points of the particular sample percent's value.

In line 2, when the interviewer asked "… what percent of the time might that happen?", he was referring to the event "sample percents fall within 4 percentage points of the population percent". Sarah's response in line 3 ("99% of the time") seems like an arbitrary estimate, perhaps chosen to reflect her sense of almost certainty that the true population percent lies between 32% and 40%. As indicated by the rest of the excerpt, this estimate was evidently not based on an operational sense of the distribution of sample percents for samples of size 500, for Sarah clearly made no connection to the table in Question 3 containing such information (lines 4 and 6).

Peter's response to Question *3e,* together with his post-interview comments, suggests that he and Sarah had very similar understandings of margin of error. Below is his written response:

> "*No the interpretation is not a good one. The margin of error did not come from this sample. It came from other ones like it with the same sample size number. The margin of error of 4% does mean that between 32% and 40% MAY believe an income tax is necessary"*

The idea that margin of error does not come from the particular sample (percent) obtained, but is instead derived from information about other samples of the same size, was evidently salient for Peter. However, there is no evidence in his response to suggest that he was mindful of the distribution of a collection of sample percents.

Peter's response also implies that he understood Stan to be saying that the reported margin of error being 4 percentage points means that the inferred population percent is definitely in the range (0.32, 0.40). In Peter's thinking, a "good" interpretation entails the realization that this interval, inferred to contain the population percent's true value, is at best a possibility rather than a certainty. This is suggested by his capitalization of the word "MAY", which he offered as a qualification of Stan's presumed claim. Thus, several items seem prominent in Peter's thinking: the particular sample percent's value, the population percent's true value, and the possibility of a 4-percentage-point deviation between these two.

Notice that Peter, like Sarah, also made no connection to the table in Question 3. Had he been inclined to do so, he could have used the information in this table to estimate what fraction of the time sample percents, for samples of size 500, are expected to lie within 4 percentage points of the population percent. In other words, Peter had the necessary information at his disposal to quantify his sense of the possibility that the true population value was contained in (32%, 40%) by appealing to the distribution of a collection of sample percents relative to the population percent. But he evidently did not do so.

Peter's post-interview comments are consistent with his written response, and support the above analysis. Here is a brief illustrative excerpt from the end of his interview, in which he attempted to elaborate the difference between his and Stan's interpretation of margin of error:

8. I: Ok, tell me about that difference, that qualification that it "may be" between 32 and 40 percent.
9. Peter: All right, well it doesn't really—
10. I: What do you mean by that?
11. Peter: It doesn't really have anything—that they didn't really get this 4% from

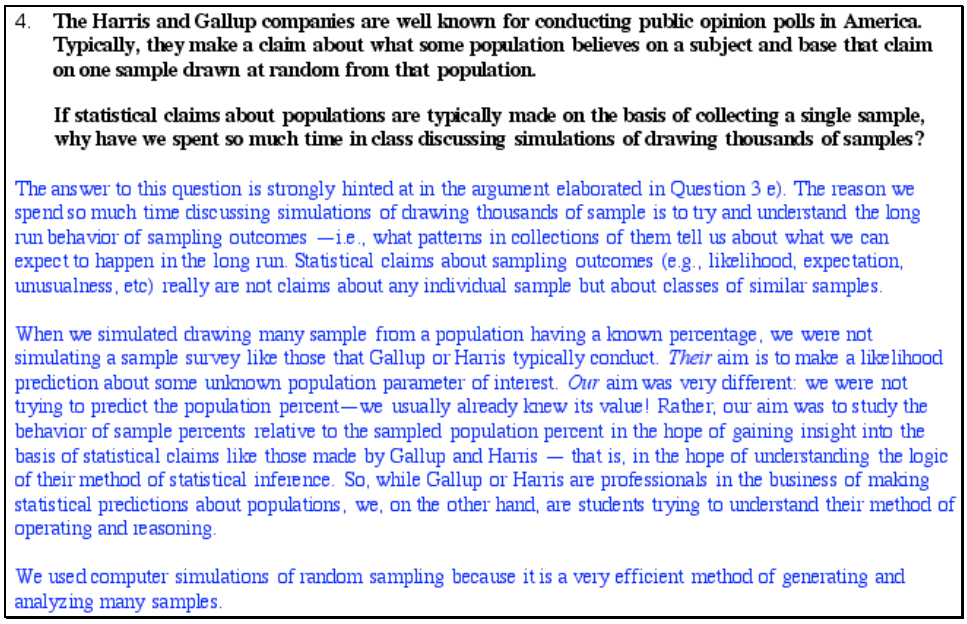
this problem.

12. I: Where did they get it from?
13. Peter: They got it from other, other samples taken.
14. I: Any old samples taken?
15. Peter: Samples taken with 500 sample size.
16. I: Ok, and where were they taken from?
17. Peter: (whistles) whatever! it probably doesn't even matter, I guess, I don't know, um (short pause) It's just saying that like it "may". Like, he's just saying that it actually "is", that like the actually 32% say that income tax is necessary, but—
18. I: Between 32 and 40.
19. Peter: Yeah, between 32 and 40, it is.
20. I: Ok, now you're saying that the better interpretation is that it "may be" between...
21. Peter: That "may be", yep.
22. I: And when you say "it may" is that...I'm not sure what you mean in terms of percentages.
23. Peter: I don't know...it may be between these percents but it doesn't have to be.

To summarize, evidence from students' written responses to Question *3e*, together with that from the post-interviews[40] and Activity 10 discussions, indicates that their conceptions of margin of error focused largely on the possible deviation between a particular sample percent's value and the population percent's true value. Theirs was, thus, a non-statistical conception; it was not rooted in an operational sense of how sample percents might be distributed around a population percent.

[40] 5 students were queried about Question *3e* during the post-interviews. The reasoning exhibited by Sarah and Peter is generally consistent with that of all 5 students.

Assessment Question *4*



4. **The Harris and Gallup companies are well known for conducting public opinion polls in America. Typically, they make a claim about what some population believes on a subject and base that claim on one sample drawn at random from that population.**

If statistical claims about populations are typically made on the basis of collecting a single sample, why have we spent so much time in class discussing simulations of drawing thousands of samples?

The answer to this question is strongly hinted at in the argument elaborated in Question 3 e). The reason we spend so much time discussing simulations of drawing thousands of sample is to try and understand the long run behavior of sampling outcomes —i.e., what patterns in collections of them tell us about what we can expect to happen in the long run. Statistical claims about sampling outcomes (e.g., likelihood, expectation, unusualness, etc) really are not claims about any individual sample but about classes of similar samples.

When we simulated drawing many sample from a population having a known percentage, we were not simulating a sample survey like those that Gallup or Harris typically conduct. *Their* aim is to make a likelihood prediction about some unknown population parameter of interest. *Our* aim was very different: we were not trying to predict the population percent—we usually already knew its value! Rather, our aim was to study the behavior of sample percents relative to the sampled population percent in the hope of gaining insight into the basis of statistical claims like those made by Gallup and Harris — that is, in the hope of understanding the logic of their method of statistical inference. So, while Gallup or Harris are professionals in the business of making statistical predictions about populations, we, on the other hand, are students trying to understand their method of operating and reasoning.

We used computer simulations of random sampling because it is a very efficient method of generating and analyzing many samples.

Figure 8.40. Question *4* queried students' sense of the overarching activity of re-sampling.

Table 8.5. The distribution of students' written responses to Question *4*.

Question	*Highly consistent*	*Somewhat consistent*	*Inconsistent*	*Ambiguous*
4	(0)	Sarah, Chelsea, David, Luke (4)	Nicole, Peter, Sue, Kit (4)	(0)

Question *4*—the final assessment item—is a meta-question designed to query what sense students made of the overarching activity of simulating drawing thousands of samples in light of the fact that actual polls make claims about a population on the basis of collecting a single sample.[41]

Table 8.5 displays the distribution of students' responses to Question *4*, with respect to their consistency with the model response. The responses are evenly split between "somewhat inconsistent" and "inconsistent". A striking commonality among them, however, is that they all refer explicitly to the idea of a sample's accuracy or precision, though not all students had in

[41] As I mentioned in the preamble to Chapter V, in a previous teaching experiment (Saldanha & Thompson, 2002) students' inabilities to reconcile these two ideas appeared to have been deeply implicated in their difficulties in understanding sampling distributions.

mind the same conception of accuracy or precision. It is instructive to consider an illustrative response more closely.

Nicole's written response, when taken together with post-interview discussions, is particularly instructive because it illustrates two senses of accuracy that are fairly representative of what a number of other students also had in mind:

> *"Although these claims are made on the basis of one sample, if you want accuracy you should take more. One sample cannot show the definite results of an entire population. So if you only take one, your results could be way off. Just because the sample you took was out of the ordinary. Even though a definite answer may never be achieved through lots of samples, it would be easier to draw a more accurate* conclusion"

Nicole's response obviously suggests that the idea of accuracy figured prominently in her thinking. As indicated in the first line, Nicole believed that selecting multiple samples somehow leads to greater accuracy, though it is unclear what she meant by "accuracy". The third and fourth lines indicate Nicole's acknowledging the idea that any individual sample might be unusual and therefore its result might depart significantly from the population parameter's true value. The response's final sentence would seem to suggest that the accuracy Nicole had in mind was something akin to the proximity of the inferred population parameter value (i.e., "the conclusion") and the true value (i.e., "definite answer"). Moreover, this last sentence reiterates the idea that this proximity can somehow be improved (i.e., made "more accurate") by selecting many samples.

This response is reminiscent of a proximity-based conception of accuracy that emerged among students in Phase 3 of instruction. It would appear that Nicole had assimilated the activity of simulating drawing thousands of samples into an image of sampling as the activity of attempting to infer the population parameter value with maximal accuracy, in the sense of coming as close as possible to the true value.

The post-interview around Question 4 reveals that there was, however, more to Nicole's thinking than her written response would suggest. She made several revealing attempts to elaborate her sense of "accuracy". These are illustrated in the following ordered interview segments.

Nicole's post-interview:

Segment 1

230. Nicole: […] All right, uh when I say "accurate results" I just mean like [short pause] results that reflect the true [short pause] population, from which we took the samples
231. L: Ok
232. Nicole: 'cause 1 sample could be off a whole lot from the rest of them
233. L: hmm hmm
234. Nicole: and so if they take one, if these companies take one sample they don't know whether or not it's gonna go along with the rest of them or if it's out of the ordinary
235. L: Right
236. Nicole: So, I think that we're taking a lot so we can make sure, when we have all the samples, that [short pause] our results are [short pause] what's gonna continually happen. Like, what, the, our results are, what we're gonna get over and over and over again.

As indicated in line 230 of the first segment, Nicole's sense of accuracy is consistent with the proximity-based conception that she expressed in her written response—it focused on the deviation between the particular sample result and the true population parameter. Lines 234-236 of the first segment indicate that Nicole believed the purpose of drawing many samples was to learn what to expect about the consistency of their results, thus providing a basis for judging whether individual results are unusual. This idea was reiterated in line 274 of the next segment, in which Nicole elaborated further on her sense of accuracy.

Nicole's post-interview:

Segment 2

263. L: We drew, we simulated drawing samples from a population that we knew the percentage of
264. Nicole: right
265. L: Now [why] are we doing that?
266. Nicole: I guess to see, I guess just to show, like us, how your samples aren't always gonna be accurate. And that's why you have to take more and more and more
267. L: Ok.
268. Nicole: because you can take that 1 sample and maybe it won't be, uhh accurate's not a good word, maybe it won't correlate with the sample, with the population percent
269. L: Ok.
270. Nicole: and so, the more you take it'll show you that [laughs]
271. L: What happens with the more you take? What does that do?
272. Nicole: [short pause] Well, it just

273. L: Does it make any one of those samples more accurate?
274. Nicole: No, but it, it uh [short pause], it, like if you take 1 sample it could be totally like inaccurate, and you wouldn't know because you don't have any others to judge it by. So if you take a lot of samples, you can like judge them against each other and see which ones occur more. So the ones that occur <u>more</u> would be the most accurate ones.

As lines 266-268 of the second segment suggest, Nicole believed, further, that the point of drawing many samples from known populations was to illustrate that not all samples will reflect or "correlate with" those populations. She took this expected variability among samples' representativeness as a rationale for selecting multiple samples, reasoning that the more samples selected, the better is one's basis for judging the "accuracy" of individual sampling results.

In lines 273-274, Nicole clarified that selecting multiple samples does not increase individual samples' accuracy. Rather, "more accurate" samples are those that occur most often. This may have been Nicole's way of expressing patterns that she recollected from the histograms employed in the instructional activities, in which large proportions of sample percents aggregated relatively close to the sampled population percents.

It would thus seem from these interview excerpts, that Nicole had two senses of accuracy in mind: accuracy as "proximity to the population percent" and accuracy as "frequency of occurrence". In addition, the idea of judging or anticipating the fit of individual sampling results on the basis of the consistency of multiple results seems to have been a central part of her thinking.

Nicole's ideas around Question *4* are consistent with those expressed by a number of other students. In addition, none of the students' responses to Question *4* expressed a clear sense of the distinction between simulating sampling from known populations and conducting a real poll, nor did they suggest the two were irreconcilable for them.

Chapter Summary

The activities in the final phase of instruction moved toward having students develop a sense of sampling distributions rooted in quantified variability—that is, the proportional measure of sample percents' dispersion with respect to intervals of various sizes around the sampled population percents. Activity 8 engaged students in investigating the relationship between sampling variability and underlying population percent, developing the generalization that the

two are relatively independent of one another. Activity 9 engaged students in investigating the relationship between sampling variability and sample size, culminating in the generalized inverse dependence relation between them—distributions of sample statistics for smaller samples tend to exhibit greater variability than those for larger samples.

Students' experiences in these activities together with the tripartite network of dependence relations that emerged from them (Figure 8.41) were intended as a basis for making statistical inferences, and for developing a statistical interpretation of margin of error in Activity 10.

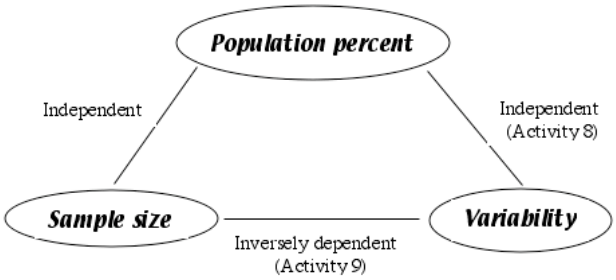


Figure 8.41. A three-part network of dependence relations among population percent, sample size, and variability.

Evidence from students' engagement in Activities 8 and 9 indicates that the emergence of this web of relations was anything but smooth and unproblematic for them. This evidence provides insight into students' thinking and the difficulties they experienced around two broad levels: 1) on the level of relatively fine details, in the intricate coordination and composition of ideas that underlay this network of relations, and 2) on a coarser level, in coming to see the significance of these relations for making statistical inferences.

With respect to the finer level, students experienced significant challenges navigating among the following ideas and quantities entailed in the sampling scenarios of Activities 8 and 9 (see Figures 8.4 through 8.19): populations of items, samples of items drawn from those populations, population percentages, ranges of sample percentages, groups of samples, numbers of samples comprising a group, numbers of items comprising a sample, patterns across scenarios, etc. Indeed, students experienced considerable difficulty not only in composing these ideas to give meaning to the quantity "percent of sample percents within x% of the population percent", but also in de-composing this quantity in terms of coherent relations among these ideas.

Students' difficulties with regard to this level were in the coordination of items and multiple actions, even once those seemed individually unproblematic for them. These difficulties are

consistent with the explanatory hypothesis advanced in Chapter VII. They suggest that students had not conceptualized the sampling scenarios as having a hierarchical structure that emerges from a stable scheme of conceptual operations centering around the images of repeatedly sampling from a large population, recording a statistic's value, and aggregating a collection of the statistic's values (see Figure 7.19).

In a number of discussion episodes during Activity 9 the distinction between people and samples (re)emerged as precarious for some students and appears to have been centrally implicated in their difficulties at this level. Those particular instances bolster this hypothesis.

Students' responses to post-instruction assessment questions (i.e., Questions *2a-2b*, *3a-3c*), however, suggest that the identification and elaboration of specific components entailed in sampling distributions may have become relatively unproblematic for them, in comparison to the difficulties they experienced during instruction.

With respect to the coarser level of detail, an assessment item administered at the end of Activity 9 (see Figure 8.29) indicates that a significant number of students appeared to have internalized the inverse relation between sample size and variability as salient. Results of the post-instruction assessment questions relevant to this level suggest some diversity among students' ideas about what factors affect the distribution of sample statistics. Specifically, in one task (Question *2c*) half of the students implied that they understood sample size to be the only factor having a significant effect on a distribution's variability. Other students considered the number of samples in a collection to be the relevant factor. In another task (Question *3d*) almost all students were oriented toward either population percent or number of samples, but not both, as unimportant for the distribution of sample percents.

In regard to the concept of margin of error, evidence from students' engagement in Activity 10 is consistent with that from the relevant post-instruction assessment responses. The evidence suggests that students had an essentially non-statistical interpretation of margin of error. Students understood that margin of error gives an estimate of the expected deviation between a particular sample percent's value and the population percent's true value. Some even understood this estimate to be derived not from the former, but rather from results of past samples of the same size. However, students were evidently not inclined to construe this estimate in statistical terms, by relating the particular sample percent's value to a distribution of a collection of such values in relation to a population percent's value.

CHAPTER IX

SUMMARY AND CONCLUSIONS

This concluding chapter begins with a broad overview of the teaching experiment, highlighting central aims of each phase of instruction and summarizing insights into students' thinking that emerged across them. The chapter then elaborates the study's contributions, implications, and limitations. It concludes with a post-analytic perspective on the nature of instructional interactions in the teaching experiment, considering them in terms of constructs drawn from complexity science.

Overview

This study has explored and characterized the reasoning that emerged among eight students as they engaged with instruction designed to support their developing coherent understandings of statistical inference. Instruction followed students, unfolding in concert with the research team's evolving sense of what students understood at different points in time. My analyses followed this unfolding, tracking how instruction evolved in tandem with the emergence of students' ideas across 17 lessons. I have characterized the flow of instructional interactions as giving rise to an emergent instructional trajectory unfolding in 4 phases, each distinguished by the focus of its aims and direction:

Phase 1: Orientation to statistical prediction and distributional reasoning

Phase 2: Move to conceptualize probabilistic situations and statistical unusualness

Phase 3: Move to conceptualize variability and distribution

Phase 4: Move to quantify variability and extend distribution

Figure 9.1 displays the unfolding of these phases across the experiment's duration.

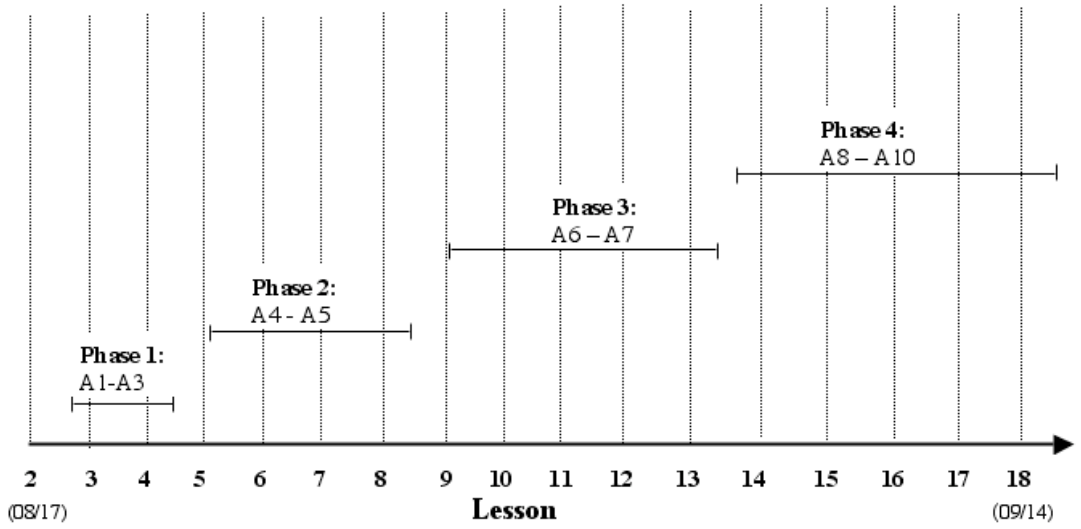


Figure 9.1. A time line of the instructional trajectory across the experiment's duration.

Taking a broad perspective, instruction in Phases 1, 3, and 4 engaged students in organizing collections of values of randomly generated sample statistics in increasingly sophisticated ways intended to support their abilities to construe such collections as distributions. Phase 2 addressed the probabilistic experiment. It engaged students in re-construing scenarios as idealized probabilistic situations, focusing on their repetitive structure, so that they might come to conceive an event's expectation as a statistical quantity.

Phase 1

Phase 1 began by engaging students in a concrete sampling activity (Activity 1), a central aim of which was to provide them with an experiential basis for understanding the simulation-based sampling explorations that came thereafter. Activity 1 highlighted the distinction between making a population inference on the basis of a single sampling outcome and on the basis of a collection of outcomes. The instructor invoked variability among outcomes as a rationale for orienting students toward collections of outcomes as more useful for estimating the population parameter's value.

Activity 2 followed by engaging students in comparing collections of sampling outcomes, generated by computer simulations. This prompted a need for students to structure such collections in ways that enabled their comparisons. Evidence from students' engagement in Activity 2 and from their work in an assessment activity (i.e., Activity 3) suggests that they were sensitized to collections of outcomes as a preferred basis for making population inferences.

Moreover, many students compared and constructed collections of sampling outcomes in ways compatible with structuring a collection in terms of the relative number of samples in it that contained a majority of one outcome (e.g., red candies). This structuring suggests that students were able to compose and coordinate two levels of imagery, each involving the quantification of a distinct attribute: one level involves individual sample compositions (e.g., the number of red candies out of 5 selected), while another level involves the relative weight of a collection's part satisfying some specific outcome criterion (e.g., the number of samples out of 10 selected that contain at least 3 red candies).

Finally, students' engagement in Activity 2 revealed little evidence of their having experienced difficulties making structural distinctions between various objects. They unproblematically followed instruction that shifted their focus of attention to increasingly complex objects: from individual sampling outcomes, to collections of outcomes, and then to a collection of similarity decisions that emerged from pair-wise comparisons of collections of outcomes.

Phase 2

Instruction in Phase 2 engaged students in designing simulations of repeated sampling experiments as a method for investigating whether an event can be considered statistically unusual. Activities 4 and 5 were intended as a context for occasioning the re-construal of scenarios as probabilistic situations and for moving students toward conceiving expectation as a statistical quantity. With regard to the former goal, students experienced significant difficulties construing given scenarios as probabilistic experiments. Their difficulties centered around making idealized assumptions about a situation so as to make it amenable to model as a repeatable sampling experiment. Structuring a situation in terms of idealized assumptions, entailing the construal of a population, a sample, and a method of selection was non-trivial for students. With regard to the latter goal, students experienced difficulty moving beyond an intuitive "gut feeling" sense that a particular event might occur and toward an operational notion of expectation: one that entails imagining the possibility of repeating a sampling process a large number of times and using the relative frequency of the observed event as a basis for quantifying expectation.

By the end of Phase 2 students showed evidence of being able to identify and construe a population, a sample, and a repeatable sampling process in a given scenario. However, many students also experienced difficulty coordinating these components to think of the scenario as a coherent and holistic probabilistic situation. In retrospect, this finding foreshadowed what emerged as a salient and robust conceptual problem for many students in Phases 3 and 4 of instruction: their profound difficulties in coordinating and composing components of an imagined re-sampling experiment into a stable and coherent network of ideas.

Phase 3

The ushering of Phase 3 was motivated by evidence, emerging at the end of Phase 2, that students' sense of variability was restricted to "differences between outcomes" and did not extend to ideas of distribution. Phase 3 emerged out of the research team's effort to address this problem through instruction. Activity 6 turned students' attention toward examining the deviations between collections of values of sample percents, generated by computer simulation, and the sampled population percent's value. The instructional aim was to support students' developing a sense of values' dispersion relative to the sampled population percent's value. The instructional interactions that unfolded around Activity 6 witnessed the emergence of the idea of accuracy as proximity and of proto-distributional images of a collection's dispersion within a continuum. However, Activity 6 turned out not to support the development of these nascent ideas into an operational sense of distribution.

Activity 7 engaged students in constructing and interpreting frequency histograms of collections of values of a sample statistic. The experience of organizing such collections into histograms was intended to support students' coming to interpret the inscription as depicting a sampling distribution. Classroom discussions around Activity 7, particularly those relating to students' interpretations of histograms, provide compelling evidence that many experienced profound difficulties in construing the frequency histograms as distributions of sample statistics. Their difficulties centered on distinguishing people from samples of people and understanding that the histograms depicted something about the latter rather than the former. These difficulties are consistent with students not having conceived the sampling scenarios that gave rise to the accompanying collections of values in terms of a hierarchical structure involving the composition of several levels of imagery: imagine selecting a number of individual items from a

large population to accumulate a sample of items; imagine recording the value of a statistic of interest for that sample; imagine repeating this process to accumulate a collection of values of a sample statistic.

Phase 4

The final phase of instruction aimed to move students toward developing an operational sense of distribution rooted in the quantification of sampling variability—that is, a proportional measure of the dispersion of a collection of sample statistic's values within various small intervals around the population parameter's value. Activities 8 and 9 formed a structured sequence that engaged students in systematically investigating the relationship between sampling variability and population percent, in the former, and sample size, in the latter. These activities culminated with the formulation of the tripartite network of generalized dependence relations depicted in Figure 9.2: sampling variability and underlying population percent are largely independent of

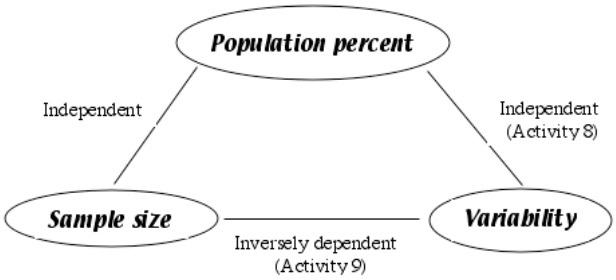


Figure 9.2. A three-part network of dependence relations among population percent, sample size, and sampling variability.

one another[1], as are sample size and population percent, while variability and sample size are generally inversely related—in the sense that distributions of sample statistics for smaller samples tend to be more densely dispersed around the population percent than those for larger samples.

Evidence from students' engagement in Activities 8 and 9, and from post-instruction assessment, provides insight into their thinking around two issues: 1) the intricate coordination

[1] I reiterate that the sampling distributions and population proportions are not really independent (see footnotes 1, 7 and 19, in chapter VIII). However, they are close enough to being independent to support the approach the research team took in developing the concept of distributions of sample proportions as a way to address the idea of margin of error.

and composition of ideas that underlay this network of relations, and 2) the salience and significance of these relations for statistical inference.

With respect to the first issue, students experienced significant difficulties navigating among the various items entailed in the sampling scenarios of Activities 8 and 9: populations of items, samples of items drawn from those populations, population percentages, ranges of sample percentages, groups of samples, numbers of samples comprising a group, numbers of items comprising a sample, patterns across scenarios, etc. Their difficulties were in coordinating and composing these items and actions across them into a coherent network of interrelated ideas. This suggests that students had not conceptualized the sampling scenarios of Activities 8 and 9 as having a hierarchical structure that emerges from a stable scheme of conceptual operations centering around the images of repeatedly sampling from a large population, recording a statistic's value, and accumulating a collection of the statistic's values (see Figure 46).

With respect to the second issue, a diversity of orientations and conceptions emerged. For instance, immediately after Activity 9 a significant number of students appeared to have internalized the inverse relation between sample size and variability. At post-instruction students revealed their beliefs about what factors affect the distribution of sample statistics. In one task (Question *2c*) half the students implied that they understood sample size to be the only factor having a significant effect on a distribution's variability. Other students considered the number of samples in a collection to be the relevant factor. In another task (Question *3d*) almost all students were oriented toward either population percent or number of samples, but not both, as unimportant for the distribution of sample percents.

Activities 8 and 9 provided a segue into the experiment's culminating instructional activity (Activity 10), which entailed discussions of the concept of margin of error in the context of an actual public opinion polling scenario. Evidence from these discussions and from post-instruction assessment strongly indicates that students generally interpreted margin of error as an estimate of the possible deviation between a particular sample percent's value and the intended population percent's value. Some students understood this estimate to be derived from results of other samples of the same size, rather than from the particular sample selected. However, students evidently did not construe this estimate in statistical terms, by relating the particular sample percent's value to a distribution of a collection of such values in relation to a population percent's value.

That students' conceptions of margin of error were evidently not rooted in the idea of sampling distributions is a bit surprising, given their apparent engagement in the culminating discussions of Phase 4 in which the connection between the two ideas was made rather explicit. The fact that students did not spontaneously make this connection when given the opportunity at post-instruction is reason to believe that they did not develop a strong disposition to make inferences by relating a particular sampling outcome to the behavior of a collection of such outcomes in operational terms.

Contributions and implications

An overarching and salient finding of this study is that students experienced significant difficulties coordinating and composing multiple objects and actions entailed in re-sampling scenarios into a coherent and stable scheme of interrelations that might underlie a powerful conception of sampling distributions, even when their envisioning of individual components seemed unproblematic. This suggests that the required coordinations and compositions are non-trivial. Moreover, since distributions of sample statistics provide the conceptual underpinning of statistical inference, it is stands to reason that developing such a scheme of interrelations is requisite for developing a deep understanding of inference. Indeed, an implication of this study is that a powerful understanding of inference is difficult to achieve because it entails more than thinking of making claims about a population on the basis of the assumption that an *individual* random sample mimics that population. Rather, the difficulty seems to lie in that it entails a scheme of interrelated ideas involving the coordination of images of sampling on multiple and hierarchically structured levels to arrive at judgments based on distributions of sample statistics.

If the research community develops principled and theoretically-based insights into what might be entailed in developing powerful understandings of statistical inference, in all of its necessary richness and detail, then it will be in a better position to begin addressing how to support the development of this important concept through principled instructional engagements. This study is a concerted effort toward developing such insights, and it already strongly suggests that supporting such developments may pose significant challenges for instructional design in the area of stochastic reasoning.

These last points touch on this study's generalizability. There are two types of interrelated generalizations that this study moves toward, beyond the context of the particular group of students who participated in this particular teaching experiment.

On one hand, this study contributes to the emergence of a cognitive-based *conceptual framework* for thinking about how to support the development of coherent stochastic reasoning. It does so by generating the above-mentioned insights. As an illustration of how the move toward such a framework might guide the design of principled instruction, let us consider the fact that the students in this study experienced significant difficulties coordinating and composing objects and actions into a coherent image of sampling distributions at late stages of the teaching experiment. This fact suggests that students' engagement in the early instructional activities did not provide sufficient support for their developing such a coherent image. It is thus natural to want to reconsider the design of those activities in retrospect.

As Chapter V describes, instruction in Phase 1 of the teaching experiment made some attempt to orient students to structural distinctions between objects and actions entailed in the re-sampling scenarios: a population; individual items drawn from the population; the accumulation of individual items into samples of items; individual samples drawn from the population; the value of a statistic recorded for individual samples; the accumulation of a collection of values of a statistic; discerning patterns in a collection of values of a statistic as a basis for estimating the population parameter's value. In retrospect, however, it seems that these distinctions were at best tacit and implicit in Activity 1. Given that these ideas turned out to be so problematic for students to coordinate and compose, it would be potentially useful to design an extension to Activity 1, for some future use, that would entail making distinctions and connections between these ideas an *explicit* topic of discussion and reflection. For instance, once Activity 1 is completed, as described in Chapter V, it might be useful to follow it with a meta-activity consisting, first, of a group discussion directed at having students 1) identify the various objects, processes, and values encountered and considered in Activity 1, and 2) elaborate and summarize the structural distinctions and connections among these. A second part of the activity might entail having students construct a written summary of these objects, distinctions, and connections. This could then serve as a basis for assessment and feedback for students. A third part of the meta-activity might make the issue of distribution more explicitly problematic, as in determining what proportion of the samples produced results within certain ranges.

Clearly, this type of meta-activity might constitute a useful follow-up to any of the activities that entail interpreting a re-sampling scenario. However it is done, the aim would be to provide an opportunity for students to step back and reflectively abstract from their experiences in Activity 1 essential components and relations that can serve as a conceptual anchor for making sense of future re-sampling scenarios and for conceptualizing sampling distributions in terms of a hierarchical structure.

The other generalizable aspect of this study is the research methodology employed to generate insights into conceptual problems entailed in understanding inference in relation to instructional engagements. This methodology, characterized in Chapter III and referred to as *epistemological analysis* (Thompson, 2002; Thompson & Saldanha, 2000), is replicable. Recall from the description in Chapter III that central to this methodology is a reiterative cycle of design, engagement, and interpretation driven by the research team's reflective interactions with students. This approach, taken together with the a posteriori retrospective analysis presented here, has proven to be a powerful one for developing insights into conceptual issues in understanding statistical inference. The application of this methodology to research learning and conceptual issues in other domain-specific content areas can, in principle, also be productive.

Although its small convenience sample does not support making claims about the prevalence of this study's central findings to a broader population of students, in a previous study conducted by the same research team (Saldanha & Thompson, 2002), there surfaced clear evidence of similar difficulties in coordinating and composing the above objects and actions, among a different group of students. I therefore strongly *suspect* that the difficulties our students experienced in that regard are not unusual. This points to a potentially relevant area for further research: investigating the conceptualization of hierarchically structured objects and processes. It would be feasible to conduct a study employing a similar methodology, but more narrowly focused on supporting the conceptualization of hierarchically structured processes and objects within diverse content areas. Given the difficulties our students experienced in spite of having participated in rather sustained instructional engagements, insight into what might be entailed in supporting such conceptualizations through instruction would be a welcome addition to the research base in mathematics learning and instruction.

This study entailed a combination of particular features that renders it distinct from prior relevant research discussed in Chapter I: the substantive mathematical ideas and connections

addressed in instruction, the sustained efforts to engage students with these ideas and connections at relatively intense levels, the method of following the current of students' ideas and allowing it to shape instruction, and the scope and grain size of the analyses detailed here. This distinctiveness has both constraining and potentially generative implications.

On the constraining side, because this study is largely incomparable with others that address a similar mathematical content matter, it is difficult to look back to the existing relevant research base for support or disconfirmation of its findings.[2] On the generative side, by characterizing the unfolding interplay between instructional design, instructional interactions, and student engagement and conceptions at the level of detail that it does, this study provides a well-documented case of a design experiment that did not shy away from having students explore difficult ideas and connections of statistical inference. As such, the study lays itself bare and invites substantive dialogue about issues of instructional design and learning with respect to the concept of statistical inference and its attendant schemes of interrelated ideas.

Limitations

Every empirical study is constrained, at its root, by the nature and quality of its data corpus. This study drew on three principle sources of data, listed here in their decreasing order of abundance: 1) whole-group classroom discussions, 2) students' written work, 3) individual interviews. Some less obvious limitations of this study are rooted in how the nature of source 1) and the scarcity of source 3) constrain the analyses developed.

The classroom discussions in this teaching experiment were generally orchestrated with the aim of having students confront and reflect on ideas addressed in the instructional activities. These discussions turned out to contain much evidence of students' thinking. However, on a local level these group discussions were not regimented and scripted affairs that unfolded in a consistently orderly manner. On the contrary, they were often locally messy, disorderly, and non-recurrent. The student-instructor interactions contained within these discussions were complex affairs that often developed out of unforeseen issues and events. Significantly, there did not always arise opportunities in these discussion for all students to extensively articulate and

[2] An exception to this claim is a previous study conducted by the same research team (Saldanha & Thompson, 2002).

elaborate their ideas and ways of thinking. Similarly, it was not always possible for the instructor to consistently follow-up individual students' ideas, to seek elaboration.

A consequence of these features is that students' ideas and thinking often emerged in snippets, and were articulated in incomplete, tentative, or halting forms. Analyzing such data for the purposes of this study sometimes necessitated making certain assumptions and imputations concerning students' foci of attention and objects of discourse, for instance. Consequently, analyses of some students' thinking, particularly in Chapter V, are somewhat speculative and interpretation-laden. I would argue that this could not helped; it is an unavoidable occupational hazard of analyzing data of this kind. With the exception of the overarching findings already highlighted, ample evidence of which emerged across much of the experiment, individual analyses should be taken as *viable* rather than *hard* claims about students' understandings and underlying images and conceptual operations.

Although the teaching experiment generated a rather voluminous data corpus, interview data of the kind collected at post-instruction (i.e., querying all students' sense of a common set of ideas, activities, and tasks addressed in instruction) is relatively scarce. This inhibited me from making substantive assertions about individual students' development. Collecting such interview data at the end of each phase of instruction might have provide a basis for making more systematic comparisons of individual students' thinking across time than is possible with the current data corpus.[3] More interview data might also have provided further basis for elaborating insights developed from analyses of classroom discussions and of students' written work.

Complexity and emergence: A post-analytic perspective

In comparison to prior relevant research studies—specifically, those studies addressing the teaching and learning of stochastic understandings of sampling and inference—this study opens up the usual "black box" of instructional interactions and takes an extended and sustained look inside. It views instruction not as a treatment or intervention, but rather as directed and sustained engagements and interactions among agents that are seen as dynamic, synergistic, and evolving. These interactions are not downplayed as intermediate to before-and-after states of knowing.

[3] Unfortunately, the research team lacked the necessary resources to conduct interviews this frequently during the experiment's progression.

Rather, they are seen to constitute the very medium within which students' ideas emerged and were or were not sustained. They are thus taken as worthy of relatively detailed examination.

In this closing section I borrow from an emerging theoretical perspective in educational research as a lens for elaborating the characterizations given in the above paragraph. Drawing on the work of Weaver (1948), Davis and Simmt (2003) characterize complexity science as "the science of learning systems, where *learning* is understood in terms of adaptive behaviors of phenomena that arise in the interactions of multiple agents" (p. 137). The authors (ibid.) define complex phenomena in terms of two key characteristics. First, complex systems are *adaptive*, meaning that they are self-modifying and can change their own structure. This feature, the authors rightfully point out, makes complex systems more amenable to evolutionary analyses than deterministic analyses. Second, complex phenomena are *emergent*, meaning that they are "composed of and arise in the co-implicated activities of individual agents. In effect, a complex system is not just the sum of its parts, but the product of the parts and their interactions." (ibid., p. 138).

In Davis and Simmt's (ibid.) view, complexity is not seen as just another category of phenomena. Rather, they advance complexity as a *perspective* that acknowledges that certain phenomena are perhaps not best understood through deterministic analyses alone. This perspective posits that "a different attitude is required for their study, one that makes it possible to attend to their ever-shifting characters and that enables researchers to regard such systems, all at once, as coherent unities, as collections of coherent unities, and (likely) as agents within grander unities" (ibid. p. 140).

The synergistic coupling of different levels of activity among the inter-actors in the teaching experiment in this study created a system that exhibited such features. For instance, the messiness and non-linearity of classroom interactions in the teaching experiment in this study were often recursively amplified by their coupling with an instructional method that followed the currents of students' thinking. This feedback system induced a classroom milieu exhibiting on-going self-modifications, a feature rendering the system poorly suited for deterministic analyses and more aptly characterized in the language of complexity science and emergence.

Indeed, the complexity perspective softens the traditional positivistic discourse of direct attribution, determination, and prediction. It moves away from making hard separations and relations between components of a system, such as assertions concerning specific cause and

effect relations between students' ideas' and particular things said or done in the course of particular interactions. Instead, this discourse posits agents' contributions, at different levels, to the constitution of conditions that might support, occasion, enable, and facilitate the co-emergence of ideas among participants in a collective and coherent macro-behaviors of such collectives.

As Davis and Simmt (2003) point out, some key features of complexity are prominently represented within such theoretical positions as radical constructivism, enactivism, situated learning, and some versions of social constructivism. In particular, the socio-cultural or emergent framework developed by Cobb and his colleagues (Cobb & Yackel, 1996; Cobb & Bauersfeld, 1995) foregrounds the idea of a reflexive and cyclical relationship between two perspectives: a psychological perspective that focuses on individual students' thinking as they participate in communal activities, and a social perspective focusing on the collective's behavior. The two phenomena are seen to be in mutual specification; as students reorganize their individual mathematical activity they are thought to contribute to the evolution of the collective's practices, activities, and orientations. The latter, in turn, feed back into individual students' activity and so forth. This idea has influenced analyses developed in various parts of this study, particularly in Chapter V of the dissertation.

Three characteristic properties

Davis and Simmt's (2003) elaboration of complexity entails the characterization of a set of interdependent qualities thought to be necessary for the emergence and viability of a complex system, each of which they see as simultaneously referring to a system's global properties and to local activities of agents within it. Here I elaborate three of these properties that aptly describe key features of the instructional interactions in the teaching experiment: internal diversity, redundancy, and organized randomness.

Internal diversity

Internal diversity is the potential for individual agents within a community to contribute in diverse ways to it. The instructional strategy employed in the teaching experiment was specifically geared to encourage participants to contribute their personal ideas and interpretations with regard to concepts fielded for discussion in particular instructional activities. This tended to

create a potential space not only for the "safe" sharing of personal ideas and understandings, but also for their public comparison and contrast among participants. This discursive space functioned as a medium in which students could develop and clarify theirs and others' ideas; differing points of view and tensions that might arise from them could provide a potential for the discussion of substantive issues.

This discursive space also served to expose a cross section of students' ideas to the instructor, who could then use this information in coordination with the team's specific instructional agenda to steer instruction in a bottom-up direction. In these ways the emergence and maintenance of such a discursive space served the goals of the research agenda.

At the same time there was diversity in the way students participated and engaged in instruction. Students' productions—their oral and written descriptions and explanations—and their observable efforts varied. Some students were able to give extended elaborations and seemed to readily engage in doing so with little prompting from the instructor, while others struggled to do so even when the instructor offered intended support. Moreover, such variation occurred both inter-personally and intra-personally. In these ways, internal diversity impeded the research team's efforts to develop coherent interpretations of students' understandings. This, in turned, added to the density and complexity of interactions, as the team's prompts for further elaboration from individual students often led discussions to branch off in unanticipated directions.

Redundancy

Davis and Simmt (ibid.) use the term *redundancy* to refer to "duplications and excesses of the sorts of features that are necessary to particular events" (ibid., p. 150). But they do not mean to evoke images of superfluousness and non-necessity. Rather, redundancy is the property that participants in a community share enough commonalities of experience, expectation, purpose, etc—that they be more similar than different—so that the system they constitute is able to maintain its coherence. Redundancy plays two key roles: 1) it enables interactions among agents; 2) it enables agents to compensate for others' failings, when necessary.

In the teaching experiment, redundancy provided a basis for meaningful conversations among participants. I might add, however, that redundancy in the form of shared expectations about engagement in activities and purpose of that engagement was not a given but instead

required negotiation. With respect to the second key role, redundancy is apparent in many of the classroom discussions in the teaching experiment. These discussions often entailed fragmented student participations that, under orchestration of the instructor, eventually coalesced into discussions having coherent directions and elaboration of issues. The fragmentation was due to different participants interjecting to fill in gaps when others were stumped or halted their nascent ideas in mid-thought. The end result of such discussions is taken to be greater than the sum of their parts. In the end, despite their syncopated nature, such discussions ended up forming the basis of significant instructional interactions.

In the teaching experiment there was also a *quest*, on the part of the research team, for redundancy of a kind that Davis and Simmt (2003) seem not to have in mind. I am alluding to the team's efforts to support students' developing conceptions and understandings compatible with those targeted in instruction. Indeed, instructional interactions were not prompted just for the sake of interacting, or just for the sake of having individual students table their ideas. Rather, this was all in the service of an overarching instructional agenda to help move students toward more or less coherent conceptions and lines of thought.

Organized randomness

Organized randomness is a structural feature that helps to determine a balance between diversity and redundancy among agents in a system (Davis & Simmt, 2003, p. 154). Complex systems are rule-bound, and while rules typically impose certain boundaries of activity they need not, in their totality, limit possibilities. Organized randomness, as I frame it here, speaks to productive combinations of an instructional situation or task's *pre*scritiveness and *pro*scriptiveness.[4] It has to do with setting activities for engagement so that they are neither too organized and redundancy-oriented nor too unfocused and diversity-oriented. Rather, activities should strike a "delicate balance between sufficient organization to orient agents' actions and sufficient randomness to allow for flexible and varied response" (ibid., p. 155).

In the teaching experiment discussed here, organized randomness often emerged as a salient feature of interactions and engagement, on both local and global levels. This emergence is traceable, on the one hand, to instructional design decisions made by the research team. On the

[4] In my interpretation, the randomness to which Davis and Simmt (2003) refer is not the same as statistical randomness. Instead, I presume they have in mind a feature akin to indeterminacy or unpredictability.

other hand, it also emerged partly as a consequence of the instructional methodology and practice of following the currents of students' ideas as they occurred in the moment.

For instance, in regard to the former, the research team strove to design instructional activities so that student engagement with them would create a productive combination of structuredness and unstructuredness. Indeed, the team's orientation was to design instruction with the aim of creating a particular "dynamical space that will be propitious for individual growth in some intended direction, but will also allow a variety of understandings that fit with where individual students are at that moment in time" (Thompson, 2002, p. 194). Organization was typically built into a particular activity in the form of explicit (written) task directives, questions, and prompts relative to idea(s) or issues that the activity sought to address. This structure was countered by "randomness" that often emerged from engagement with it in a whole-class discussion context, a component that was more diversity-oriented.

In regard to the latter, the class was subject to typical school constraints arising from the broader institutional culture within which it was embedded. These constraints imposed a manner of organizational structure on the class, such as rigid seating arrangements within the classroom and fixed duration of lessons. At the same time, norms of engagement and interaction that were negotiated by the research team served to counter these relatively rigid organizational constraints by attempting to open up interactions and expectations, thereby pushing the limits of possibilities for interaction and emergence of ideas among students than might otherwise have been possible.

To summarize, ideas of complexity science and emergence (Davis and Simmt, 2003) provide a post-analytic descriptive and interpretive frame to help the reader make sense of the perspectives and analyses of the previous chapters. In particular, this frame provides a point of reference for situating the often complex and messy classroom interactions and discussion excerpts drawn from the teaching experiment's data corpus.

APPENDIX A

TWO PROBLEMS FROM THE RESEARCH LITERATURE

Maternity Ward Problem

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day. In the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes it may be lower.

For a period of 1 year, each hospital recorded the days on which more/less than 60% of the babies born were boys. Which hospital do you think recorded more such days?

1) The larger hospital

2) The smaller hospital

3) About the same (that is, within 5% of each other)

Post Office Problem

When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 9 inches.

Every day for one year, 10 men registered at post office A and 25 men registered at post office B. At the end of each day, a clerk at each office computed and recorded the average height of the men registered there that day. Which would you expect to be true? (circle one)

1. *The number of days on which the average height was 6 feet or more was greater for post office A than for post office B.*
2. *The number of days on which the average height was 6 feet or more was greater for post office B than for post office A*
3. *There is no reason to believe that the expected number of days on which the average height was 6 feet or more was greater for one post office than the other.*

APPENDIX B

REVIEW OF COURSE IDEAS

Population

A collection of items having something in common that is of interest to a researcher.

Population parameter

A measurement of the entire population that you would like to know (e.g., the percent of the voting population who will vote for a political candidate; the population's yearly income per person; etc.)

Sample

A subset of a population.

Sample Statistic

A measurement of the sample that parallels the population parameter (usually calculated so that it is just like the population parameter of interest, except calculated for the sample). Typically, the population parameter you are after tells you what statistic to calculate using data gathered from your sample.

Common-sense Sort-of-Random Sample

A sample selected in such a way that the actual items selected are not predetermined. This is not the way statisticians think of "randomly selected." See "Random Sample", below.

Random Sample

Items selected from a population in a way that gives every member of the population an equal chance of being selected.

What does "equal chance of being selected" mean? If we were to repeat the selection procedure a huge number of times (huge relative to the size of the population), then each item in the population should be selected approximately the same fraction of the time.

Draw a sample from a population to gain information about it

Samples drawn truly at random (meaning that every member of the population is equally likely to be chosen) tend to reflect the characteristics of the population from which they are drawn. However … see "doing simulations to gain information. …" below.

Perform simulations to gain information about the process of sampling

The characteristics of samples drawn at random will not always accurately reflect the population's characteristics because characteristics of samples will vary from sample to sample. This is called sampling variability. At times a sample may even reflect the population's characteristics very poorly. By simulating the process of drawing samples at random from various populations with known characteristics can gain insight into *the process' likelihood* of producing estimates that are within certain percentage points of the population's actual percent.

Measuring the variability of a collection of sample percents

We can measure how variable are percents calculated from samples of a given size chosen at random from some population. One common measure is to determine the fraction of a collection's sample percents that arfe within certain ranges of the population percent.

Statistical unusualness

In statistics an event is said to be "unusual" (or "unlikely", or "rare") if *over the long run we expect to see it a small fraction of the time*. This way of thinkng about unusualness does not say anything about the event per se. Rather, it emphasizes our expectation that, for whatever reason and relative to certain circumstances, we expect to see it relatively infrequently. By convention statisticians have agreed that *a small fraction of the time* means 5% of the time or less. Thus, an unusual event is one that we expect will occur in 5% or less of a large number of times that it can occur.

Statistical pattern

A pattern that emerges only over the long run. It is impossible to predict the value of the next element in the sequence, only from the long term behavior can a pattern be discerned.

Example: The distribution of sample percents calculated from samples of a given size drawn randomly from a population. Because of sampling variability we cannot accurately predict what the outcome of any one sample will be. Only in the long run, after the random drawing of a large number of samples, does the distribution emerge.

Measurement error

TWO PERSPECTIVES.

1. The perspective of a carpenter who is considering A SPECIFIC item and is concerned that THIS PARTICULAR measurement is within a specific tolerance of the item's actual measure.
2. The perspective of an architect who is considering ALL MEASUREMENTS TAKEN AT A SITE and is concerned with WHAT PERCENT of all measurements taken by all carpenters are within a particular range of the items' actual measures.

Suppose an architect is asked about the accuracy of one specific measurement. He or she DOES NOT KNOW how accurate that measurement is. The best he or she can say is something like, "When we've studied this issue in the past, 99% of all carpenters' measurements were within 5% of the item's actual measure as determined by a much more accurate instrument, so I expect this one measurement to be pretty accurate."

Drawing one sample is like taking one measurement. The person paying for the sample is like the carpenter – he or she is interested in the accuracy of THAT ONE SAMPLE. But you, the statistician, are like the architect. You DON'T KNOW how accurate this sample is. You can only justify your conclusions by appealing to what happens over the long run.

Suppose a customer, who has paid a lot of money for you to conduct a survey of 1600 people, asks, "How do you know that these results are accurate?"

The best you can say is something like "We took great care to ensure that we used a truly random selection process in choosing these 1600 people. In our simulations of sampling 1600 people at random from a large population, 99% of the samples were within 2 percentage points of the population's actual percent. And this was true regardless of the actual population percent. So, it is unlikely that this sample's percent differs from the actual population percent by more than 2 percentage points.

What has “error” in it? The PROCESS of drawing a sample. It is not that any sample statistic is *wrong*. Rather, the idea of “error” is that a statistic computed from a sample will deviate from the actual population parameter.

In inferential statistics, we are fundamentally concerned with what happens over the long run if we were to repeat a process a large number of times. The reason for this is that insight into what happens over the long run is how we judge the trustworthiness of any individual result.

Margin of Error

This is a technical term meant to convey an idea of how variable are sample statistics calculated from randomly drawn samples of particular a particular size.
Here is a table that shows results of simulations of drawing samples of various sizes.

Sample Size	Number of Samples Drawn	Percent of Sample Percents within 1 percentage point of Population Percent	Percent of Sample Percents within 2 percentage points of Population Percent	Percent of Sample Percents within 3 percentage points of Population Percent	Percent of Sample Percents within 4 percentage points of Population Percent
100	2500	405/2500= 0.16	805/2500= 0.32	1172/2500= 0.47	1483/2500= 0.59
200	2500	612/2500 = 0.24	1147/2500 = 0.46	1590/2500 = 0.64	1914/2500 = 0.77
400	2500	840/2500 = 0.34	1544/2500 = 0.62	2002/2500 = 0.80	2284/2500 = 0.91
800	2500	1142/2500 = 0.46	1981/2500 = 0.79	2331/2500 = 0.93	2448/2500 = 0.98
1600	2500	1523/2500 = 0.61	2296/2500 = 0.92	2463/2500 = 0.99	2500/2500 = 1.00
3200	2500	1957/2500 = 0.78	2453/2500 = 0.98	2500/2500 = 1.00	2500/2500 = 1.00

The idea of margin of error comes from asking the question, “In what range of the actual population percent will we find at least *x*% of the samples when we draw samples of size N at random from the population? "Margin of error" is not about how accurate is any one sample. Rather, it is about how accurate, over the long run, do samples tend to be?
The idea of margin of error being about how accurate, over the long run, samples tend to be can be turned into a procedure. Suppose we surveyed 400 people on a subject. We want to determine a "plus or minus" range for samples of size 400 so that we expect at least 60% of all such samples to fall within that range of the actual population percent. The table below suggests that around 62% of all 400-item samples will be within 2 percentage points of the actual population percent.
So, for a sample of 400 that came up with a sample percent of, say, 45%, we would say

45% of this sample of 400 people (said such and such).
These results have a margin of error of ±2% with a
confidence level of 62%

Please note that "±2%" IS NOT ABOUT THIS PARTICULAR SAMPLE!!! Rather, it specifies the long-term accuracy that we have in mind—the fraction of samples we expect to be within two percentage points of the population percent. "62%" is that fraction of all samples of 400 people that we expect, over the long run, to be within ±2% of the population parameter.

It would be a **big** mistake to think that **this** sample of 400 people is within ±2% of the population percent. We cannot make that claim. We have **no idea** how accurate **this** sample is.

We are like the architect who cannot judge the accuracy of a particular measurement by a particular carpenter who is at a site many miles away. Instead, all he can vouch for is that his carpenters are within a certain range of actual measurements a high percent of the time.

Simulation as grounds for justification

We justify our claims about samples' accuracy by assuming that the future will resemble the past. In our simulations we drew *lots* of 500-item random samples from populations having various splits. It turned out that 95% of all 500-person samples in any of the simulations fell within ±4 percentage points of the actual population percent. The actual population's actual percent did not matter. So, we expect that in the future, 95% of the time we draw a 500 items at random and calculate a percent, it will be within ±4 percentage points of the actual population percent.

In our simulations, ninety-five percent of all 800-item samples were within ±3 percentage points of the actual population percent. So, we expect that, in the future, 95% of the time that we draw 800 items at random and calculate a percent, it will be within ±3 percentage points of the actual population percent.

REFERENCES

American Statistical Association. (1998). *What is margin of error?* Alexandria, VA: Author.

Bettoni, M. C. (1998). *The "Attentional Quantum" model of concepts and objects*, [website]. Bettoni, M. C. Available: http://www.fhbb.ch/weknow/ini/front.htm [2000, January 15].

Ceccato, S. (1949). Il teocono o " della via che porta alla verita" ('Theocogno' or of the path that leads to truth). *Methodos, 1*(1), 34-54.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 547-589). Mahwah, NJ: Lawrence Erlbaum Associates.

Cobb, P. (1998). Theorizing about mathematical conversations and learning from practice. *For the Learning of Mathematics*, 18(1), 46-48.

Cobb, P. (2000). Conducting teaching experiments in collaboration with teachers. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 307-333). Mahwah, NJ: Lawrence Erlbaum Associates.

Cobb, P., & Bauersfeld, H. (1995). The coordination of psychological and sociological perspectives in mathematics education. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning: Interaction in classroom cultures* (pp. 1-16). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cobb, P., & Whitenack, J. (1996). A method for conducting longitudinal analyses of classroom videorecordings and transcripts. *Education Studies in Mathematics*, 30(3), 213-228.

Cobb, P., & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational Psychologist*(31), 175-190.

Cobb, P., Yackel, E., & McClain, K. (Eds.). (2000). *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cortina, J. L., Saldanha, L. A., & Thompson, P. W. (1999). Multiplicative conceptions of arithmetic mean. In F. Hitt & M. Santos (Eds.), *Proceedings of the Twenty First Annual Meeting of the International Group for the Psychology of Mathematics Education - North American Chapter* (Vol. 2, pp. 466-472). Cuernavaca, Mexico: ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, OH

CTGV. (1992). The Jasper series as an example of anchored instruction: Theory, program description and assessment data. *Educational psychologist, 27*, 291-315.

Davis, B., & Simmt, E. (2003). Understanding learning systems: Mathematics education and complexity science. *Journal for Research in Mathematics Education, 34*(2), 137-167.

Data Description (1999). *Data Desk* Computer program. Author

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

diSessa, A. (1988). Knowledge in pieces. In G. Forman (Ed.), *Constructivism in the computer age*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Freudenthal, H. (1972). The empirical law of large numbers, or the stability of frequencies. *Educational studies in mathematics*(4), 484-490.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129-161). New York: Wiley

Gigerenzer, G. (1998). Ecological Intelligence: An adaptation for frequencies. In D. D. Cummins & C. Allen (Eds.), *The evolution of mind* (pp. 9-29). Oxford: Oxford University Press

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine.

Glasersfeld, E. v. (1978). Radical constructivism and Piaget's concept of knowledge. In F. B. Murray (Ed.), *Impact of Piagetian theory* (pp. 109-122). Baltimore, MD: University Park Press

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

Gravemeijer, K. (1994). Educational development and developmental research in mathematics education. *Journal for Research in Mathematics Education, 25*(5), 443-471.

Gravemeijer, K., Cobb, P., Bowers, J., & Whitenack, J. (2000). Symbolizing, modeling, and instructional design. In P. Cobb , E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design* (pp. 225-274). Mahwah, NJ: Lawrence Erlbaum Associates

Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisted: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12*, 275-305.

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. Chicago: The University of Chicago Press.

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

Kahneman, D., & Tversky, A. (1982b). Variants of uncertainty. *Cognition, 11*, 143-157.

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.

Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.

Lunneborg, C. E. (1994). *Modeling experimental and observational data*. Belmont, CA: Duxbury Press.

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

Piaget, J. (1968). *Six psychological studies*. New York: Vintage Books.

Piaget, J. (1971). *Genetic Epistemology*. New York, NY: W. W. Norton & Company.

Piaget, J. (1974). *La prise de conscience (The grasp of consciousness)*. Paris: Presses Universitaires de France.

Piaget, J. (1977). *Psychology and epistemology: Towards a theory of knowledge*. New York, NY: Penguin Books.

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

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific concept: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227.

Public Agenda. (2003). Best estimate: A guide to sample size and margin of error. Retrieved Sept 5, 2003, from http://www.publicagenda.org/aboutpubopinion/aboutpubop4.htm.

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

Saldanha, L., & Thompson, P. (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 Associates

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 Proceedings of the Research Presession Symposium of the 77th Annual NCTM Conference*. San Fransisco, CA.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114-145.

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 Associates

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage Publications.

Thompson, P. W. (1994). The development of the concept of speed and its relationship to concepts of rate. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 179–234). Albany, NY: SUNY Press

Thompson, P. W. (1996). Imagery and the development of mathematical reasoning. In L. P. Steffe , P. Nesher , P. Cobb , G. Goldin & B. Greer (Eds.), *Theories of mathematical learning* (pp. 267-283). Hillsdale, NJ: Erlbaum

Thompson, P.W. (2000). Radical constructivism: Reflections and directions. In L. P. Steffee & P. W. Thompson (Eds.), *Radical constructivism in action: Building on the pioneering work of Ernst von Glasersfeld* (pp. 291-315). London: RoutledgeFalmer.

Thompson, P. W. (2002). Didactic objects and didactic models in radical constructivism. In K. Gravemeijer , R. Lehrer , B. v. Oers & L. Verschaffel (Eds.), *Symbolizing, modeling and tool use in mathematics education* (pp. 191-212).

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 International Group for the Psychology of Mathematics Education - North American Chapter* (Vol. 2, pp. 403-408). Tucson, AZ: ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, OH

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. 95-114). Reston, VA: NCTM.

Velleman, P. (2000). *ActivStats*. Computer program. Addison-Wesley Longman.

von Mises, R. (1957). *Probability, statistics, and truth*. London: Allen & Unwin.

Vuyk, R. (1981). *Overview and critique of Piaget's genetic epistemology: 1965-1980, vol. I, II*. New York: Academic Press.

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

Weaver, W. (1948). Science and complexity. *American Scientist* 36, 536-544.

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.

Yates, D. S., Moore, D. S., & McCabe, G. P. (1999). *The practice of statistics*.