TEACHERS' UNDERSTANDINGS OF PROBABILITY AND STATISTICAL INFERENCE AND THEIR IMPLICATIONS FOR PROFESSIONAL DEVELOPMENT

By

Yan Liu

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements for

the degree of

DOCTOR OF PHILOSOPHY

in

Education and Human Development

August, 2005

Nashville, Tennessee

Approved:

Professor Patrick W. Thompson

Professor Richard Lehrer

Professor Paul A. Cobb

Professor Philip S. Crooke

TABLE OF CONTENT

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

STATEMENT OF PROBLEM

Teachers' understanding of significant mathematical ideas has profound influence on their capacity to teach mathematics effectively (Thompson 1984; Ball and McDiarmid 1990; Ball 1990; Borko, Eisenhart et al. 1992; Eisenhart, Borko et al. 1993; Simon 1994; Thompson and Thompson 1996; Sowder, Philipp et al. 1998; Ball and Bass 2000), and, in turn, on what students end up learning and how well they learn (Begle 1972; 1979). To elaborate, first, teachers' personal understanding of mathematical ideas constitutes the most direct source for what they intend students to learn, and what they know about ways these ideas can develop. Second, how well teachers understand the content they are teaching have critical influence on their pedagogical orientations and their ability to make instructional, curricular, and assessment decisions (Thompson 1984; McDiarmid, Ball et al. 1989; Borko, Eisenhart et al. 1992; Dooren, Verschaffel et al. 2002). This ensemble of teachers' knowledge (Shulman 1986), orientations (Thompson, Philipp et al. 1994), and beliefs (Grossman, Wilson et al. 1989)—of mathematical ideas, and of ways of supporting students' learning of these ideas, plays important roles in what students can learn and how well they learn in the instructional settings.

This has important implications for how teacher educators think about ways of supporting teachers' professional development. That is that, supporting transformation of teaching practices takes careful analysis of teachers' personal and pedagogical understanding. Such efforts increase the likelihood that what teachers teach and how they

teach have the potential of supporting students to develop coherent and deep understanding of mathematics.

Probability and statistical inference are among the most important and challenging ideas that we expect students to understand in high school. Probability and statistical inference have had an enormous impact on scientific and cultural development since its origin in the mid-seventeen century. The range of their applications spread from gambling problems to jurisprudence, data analysis, inductive inference, and insurance in eighteen century, to sociology, physics, biology and psychology in nineteenth, and on to agronomy, polling, medical testing, baseball and innumerable other practical matters in twentieth (Gigerenzer, Swijtink et al. 1989). Along with this expansion of applications as well as the concurrent modification of the theories themselves, probability and statistical inference have shaped modern science, transformed our ideas of nature, mind, and society, and altered our values and assumptions about matters as diverse as legal fairness to human intelligence. Given the extraordinary range and significance of these transformations and their influence on the structure of knowledge and power, and on issues of opportunity and equity in our society, the question of how to support the development of coherent understandings of probability and statistical inference takes on increased importance.

Since 1960s, there have been abundant research studies conducted to investigate ways people understand probability and statistical inference. Psychological and instructional studies consistently documented poor understanding or misconceptions of these ideas among different population across different settings (Kahneman and Tversky 1973; Nisbett, Krantz et al. 1983; Konold 1989; 1991; Konold, Pollatsek et al. 1993a;

Fischbein and Schnarch 1997). Contrary to the overwhelming evidences of people's difficulties in reasoning statistically, there is in general a lack of insight into what is going on in the transmission of this knowledge in classroom settings. Particularly, research on statistics education has attended to neither teachers' understanding of probability and statistics, nor to their thinking on how to teach these subjects (Truran 2001; Garfield and Ben-Zvi 2003).

The goal of this dissertation study is to explore teachers' personal and pedagogical understanding of probability and statistical inference. To this end, our research team designed and conducted a seminar[1] with eight high school mathematics teachers. This study is an early step of a bigger research program, which aims to understand ways of supporting teachers learning and their transformations of teaching practices into one that is propitious for students learning in the context of probability and statistics. As a precursor, this study is highly exploratory. The research team designed the seminar with the purpose of provoking the teachers to express and to reflect upon their instructional goals, objectives, and practices in teaching probability and statistics. The primary goal was to gain an insight into the issues, both conceptual and pedagogical, that teachers grapple with in order to teach probability and statistics effectively in the classroom.

This dissertation will present a retrospective analysis of this seminar. Specifically, the aims of this dissertation are:

[1] This study is part of a five-year, longitudinal research project "An investigation of multiplicative reasoning as a foundation for teaching and learning stochastic reasoning," designed and directed by Dr. Patrick Thompson, my dissertation advisor and professor of mathematics education at Vanderbilt University. Since I joined the research team 5 years ago, I have been integrally involved in all of its facets: instructional design, data collection, organization, and interpretation.

1) To construct an explanation of teachers' personal and pedagogical understanding of probability and statistical inference;
2) To create a theoretical framework for constructing such an explanation.

To explicate my research purposes, let me first explain what I mean by "understanding" and the method I use in developing descriptions of an understanding. By "understanding" I follow Thompson & Saldanha (2002) to mean that which "results from a person's interpreting signs, symbols, interchanges, or conversation—assigning meanings according to a web of connections the person builds over time through interactions with his or her own interpretations of settings and through interactions with other people as they attempt to do the same." Building on earlier definitions of understanding based on Piaget's notion of assimilation, e.g. "assimilating to an appropriate scheme" (Skemp 1979), Thompson & Saldanha (*ibid.*) extend its meaning to "assimilation to a scheme", which allows for addressing understanding people do have even though it could be judged as inappropriate or wrong. As a result, they suggested that a description of understanding require "addressing two sides of the assimilation—what we see as the thing a person is attempting to understanding and the scheme of operations that constitutes the person's actual understanding." (*ibid.*, p. 11)

To construct a description/explanation of a person's understanding, I adopt an analytical method that Glasersfeld called conceptual analysis (Glasersfeld 1995), the aim of which is "to describe conceptual operations that, were people to have them, might result in them thinking the way they evidently do." Engaging in conceptual analysis of a person's understanding means trying to think as the person does, to construct a conceptual structure that is isomorphic to that of the person. This coincides with the

notion of *emic* perspective in the tradition of ethnographic research, i.e., the "insider's" or "native's" interpretation of or reasons for his or her customs/beliefs, what things mean to the *members of a society*, as opposed to *etic* perspective: the external researcher's interpretation of the same customs/beliefs. In conducting conceptual analysis, a researcher builds models of a person' understanding by observing the person' actions in natural or designed contexts and asking himself, "What can this person be thinking so that his actions make sense from his perspective?" (Thompson 1982) In other words, the researcher/observer puts himself into the position of the observed and attempt to examine the operations that he (the observer) would need or the constraints he would have to operate under in order to (logically) behave as the observed did (Thompson 1982).

As a researcher engage in the activity of constructing description /model /explanation (henceforth explanation) of his subjects' understanding, he should in the mean time subject his very activity to examination, i.e., to reflectively abstract (Piaget 1977) the concepts and operations that he applies in constructing explanations. When the researcher becomes aware of these concepts and operations, and can relate one with another, he has an explanatory/theoretical framework, which usually opens new possibilities for the researcher who turns to using it for new purposes (Steffe and Thompson 2000). There is a dialectic relationship between these two kinds of analyses—constructing explanations of a person' understanding and creating a theoretical framework for constructing such explanations. The theoretical framework and the explanations exert a reciprocal influence upon each other as they are simultaneously constructed. Theoretical framework is used in constructing explanations of understandings. As one refines the understandings, the appearance of the framework

changes, as one refines the framework, the understandings may be modified (Thompson 1982).

It is important to note that a theoretical framework does not emerge entirely from the empirical work of trying to understand a person's actions and thinking. It could draw upon theoretical constructs established in an earlier conceptual analysis, or informed by others' work in the existing literature. And most often it is heavily constrained/enabled by the epistemology or background theories that the researcher embraces in his work (e.g. Thompson 1982). In the following chapters, I will first present a review of relevant literature with the purpose of highlighting the theoretical constructs that might potentially constitute part of the framework. This first part of this review presents a historical and conceptual analysis of probability and statistical inference. The second part reviews existing research on ways people/students understand probability and statistical inference, and the difficulties they experience as they learn these ideas. My goal of this review is to provide a vantage point for understanding teachers' knowledge and to highlight a way of understanding these ideas that are grounded in meanings and making connections amongst these ideas.

In Chapter 3, I will present the background theories and methodologies that guide the conceptualization of my research questions and the design and implementation of the study. Chapter 4 is a conceptual analysis of the probability, hypothesis testing, and margin of error. In Chapter 5, I will first provide an overview of the seminar. Following this, I will sketch the background of this seminar by summarizing the prior teaching experiments we conducted with high school students. Last, I will provide a detailed description of the seminar by summarizing the daily activities and interviews, as well as

the themes that we intended to emerge. Chapter 6, 7, and 8 are each devoted to a particular set of ideas: probability, hypothesis testing, variability and margin of error.

CHAPTER II

LITERATURE REVIEW

Understanding Probability and Statistical Inference: a Historical and Conceptual Perspective

My investigation of teachers' understanding in probability and statistical inference is motivated by the purpose of supporting the development of students' understanding by improving teacher education in this subject area. This study not only has to be built upon a knowledge of students and teachers' understanding from existing literature and prior research, but also an appreciation of the many ways probability and statistical inference are understood historically.

The development of the theories of probability and statistical inference has been riddled with controversy. For example, the concept of probability is often used to refer to two kinds of knowledge: *frequency-type probability* "concerning itself with stochastic laws of chance processes," and *belief-type probability* "dedicated to assessing reasonable degrees of belief in propositions quite devoid of statistical background" (Hacking 1975 p. 12; Hacking 2001 pp. 132-133). Since 1654, there was an explosion of conceptions in the mathematical community that were compatible with this dual concept of probability, for example, frequentist probability, subjective probability, axiomatic probability, and, probability as propensity (cf. Von Plato 1994; cf. Gillies 2000). Yet, until today, mathematicians and scientists continue to debate and negotiate meanings of probability both for its theoretical implication, and for its application in scientific research. There are subjectivists, e.g., de Finetti, who have said that frequentist or objective probability can

be made sense of only through personal probability. There are frequentists, e.g. von Mises, who contend that frequentist concepts are the only ones that are viable. According to Hacking (1975), although most people who use probability do not pay attention to such distinctions, extremists of these schools of theories "argue vigorously that the distinction is a sham, for there is only one kind of probability" (*ibid*, p. 15).

As noted by Nilsson (2003), the controversy surrounding the theories of probability and statistical inference presents a difficult question to educators: What do we teach? Instructional practices and research that sidesteps this question will likely to result in shortsighted design, which does not take into account of the consequence of students learning over the long run. It also renders the fact that researchers in psychological and instructional studies on probability and statistics tend to differ in their use of terminology. This makes it problematic both to communicate the research results to each other (Shaughnessy 1992), as well as to apply the research results to the classroom (Hawkins and Kapadia 1984). Against this background, I will first provide a brief overview of the theories of probability and statistical inference. Given the nature of my study, in my review I will highlight the conceptual complexities of probability and statistical inference, which I hope will help me in becoming sensitive to the subtleties of teachers' understanding and in anticipating their difficulties in making sense of these ideas in different ways.

Probability

There are many different views about the nature of probability and its associated concepts, such as randomness, chance, and likelihood. Fine (1973), von Plato (1994),

Gillies (2000), and Hendricks, et al. (2001) provide overviews of the debates that have been ongoing since the early 17th century, and Todhunter (1949), David (1962), and Hacking (1975) provide overviews of the development of probability prior to that. In what follows, I will sample a representative set of interpretations of probability that have profoundly influenced the research and curriculum design of probability thus far. The sequence of discussion roughly follows the chronological order of the work reviewed and attempts to give a sense of the historical development of the probability theory.

Laplace's classical probability

The essential characteristic of classical, or Laplacian, probability is "the conversion of either complete ignorance or partial symmetric knowledge concerning which of a set of alternatives is true, into a uniform probability distribution over the alternatives." (Fine 1973 p. 167) The core of this approach is the "principle of indifference"—alternatives are considered to be equally probable in the absence of known reasons to the contrary, or when there is a balance of evidence in favor of each alternative. For example, in this approach, all outcomes are equally probable in the toss of a die, or in the flip of a coin. Thus, the probability of the occurrence of any outcome is one out of the number of all possible outcomes. This approach to probability was the most prevalent method in the early development of probability theory, as the origins of the theory of probability were games of chance involving the notion of equal possibilities of the outcomes supposed to be known a priori (Todhunter 1949; David 1962).

However, classical probability builds on a number of troubling bases. First, it assumes an equal likelihood of alternative outcomes. Yet, "equal likelihood" is exactly synonymous with "equal probability." It is in this sense von Mises (1957) argued that,

"unless we consider the classical definition of probability to be a vicious circle, this definition means the reduction of all distribution to the simpler case of uniform distribution." (*ibid*, p. 68)

Even though one accepts such constraints of classical probability, objections still hold against making assumptions of equal likelihood of outcomes based on ignorance, lack of evidence, or partial symmetric knowledge. von Mises (1957) critiqued the reasoning of those who wish to maintain that "equally likely cases" in the game of dice can be logically deduced from geometrical symmetry or kinetic symmetry. He concluded that "at the present stage of scientific development we are not in a position to derive 'theoretically' all the conditions which must be satisfied so that the six possible results of the game of dice will occur with equal frequency in a long series of throws" (*ibid*, p. 74). Fine (1973) concurred that the present-day cubical, symmetrical die is evolved from many years of experimentation on ancient, irregular die (cf. David 1962), and that "it is this lengthy experience that may be elliptically invoked rather than the principle of indifference" (Fine 1973 p. 169). The attempt to justify the assumption of equally likely cases by having recourse to the principle of indifference leads to enormous inconsistencies and failures in the interpretations of problems concerning probability (Von Mises 1957). In sum, von Mises suggested two essential objections to the classical definition of probability—"On the one hand, the definition is much too narrow; it includes only a small part of the actual applications and omits those problems which are most important in practice, e.g., all those connected with insurance. On the other hand, the classical definition puts undue emphasis on the assumption of equally possible events in the initial collectives."(*ibid*, p. 79)

Von Mises' limiting relative frequency probability

von Mises' (1957) relative frequency definition of probability is based on two central constructs, namely, that of collective, and randomness. He limits probability to apply only to infinite sequences of uniform events or processes that differ by certain observable attributes, of which he labels "the collective." The definition of probability is concerned only with the probability of encountering a certain attribute in a given collective. Two hypotheses about collectives are essential in von Mises' definition of probability. The first is the existence of the limiting value of the relative frequency of the observed attribute. In other words, a collective appropriate for the application of the theory of probability must be "a mass phenomenon or a repetitive event, or simply, a long sequence of observations for which there are sufficient reasons to believe that the relative frequency of the observed attribute would tend to a fixed limit if the observations were indefinitely continued" (*ibid*, p. 15). The second hypothesis is a condition of randomness, called "the principle of the impossibility of a gambling system," in other words, "the impossibility of devising a method of *selecting the elements* so as to produce a fundamental change in the relative frequencies" (*ibid*, p. 24). von Mises requires that a collective (to which the theory of probability applies) also fulfils the conditions that the limiting value of the relative frequency of the attribute remains the same in all partial sequences which may be selected from the original one in an arbitrary way.

The strength of von Mises' limiting relative frequency theory of probability is that it offers both a physical interpretation of, and a way of measuring, probability (as opposed to mathematical probability, which I will discuss in the following section). It offers an operational definition of probability based on the observable concept of

frequency. This should be considered in concert with von Mises' background as a physicist and his close philosophical tie with Ernst Mach' Positivist tradition. Because of his main scientific interest in physics, von Mises is more concerned with the link between probability theory and natural phenomena (as opposed to, for example, a mathematician's interest in formalizing probability theory). His philosophical conviction, Positivism, holds that physical laws are merely summaries of sensory experience and the meaning of physical concepts is determined only by specifying how they are related to experience. It is in this sense that von Mises regards probability as "a *scientific theory* of the same kind as any other branch of the exact *natural science,*" which applies to long sequences of repeating occurrences or of mass phenomena (Von Mises 1951 p. 7).

Objections to von Mises' theory pinpoint its lack of connection between theory and observation by the use of limits in infinite sequences. It is well known that two sequences can agree at the first *n* places for any finite *n* however large and yet converge to quite different limits. Suppose a coin is tossed 1,000 times and the observed frequency of heads is approximately 1/2. This is "quite compatible with the limit being quite different from 1/2" (Gillies 2000 p. 101). To be more precise, the observation does not exclude the possibility that the probability (the limit of the relative frequency) is, say, 0.5007. Fine's (1973) position is in harmony with Gillies'. He suggested that "knowing the value of the limit without knowing how it is approached does not assist us in arriving at inferences," and radically concluded that a limit interpretation is "of value neither for the measurement of probability nor for the application of probability."

Kolmogorov's measure theoretical probability

Kolmogorov (1956) constructed the concept of probability on the basis of measure theory. A probability space (Ω, F, P) consists of a sample space, Ω; a σ-field F of selected subsets of Ω; and a probability measure or assignment, P. The elements of Ω are called "elementary events." The σ-field of subsets of Ω, F, has the following three properties.

1. $\Omega \in$ F.
2. If $F \in$ F, then $\overline{F} \in$ F (closure under complementation).
3. If for countably many i, $F_i \in$ F, then $\bigcup_i F_i \in$ F (closure under countable unions).

In lay terms, Kolmogorov formalized the notions that a probability space consists of (a) all the states (outcomes) in which an experiment can terminate, (b) a collection of events each of which is a collection of elementary outcomes, and (c) a way to assign numbers to events. It also has the properties that the sample space itself is an event, that an event not happening is itself an event, and that any combination of events is an event.

The probability measure P is a function from F to the interval [0, 1] that satisfies the following four axioms.

1. *Unit normalization* $P(\Omega) = 1$.
2. *Nonnegativity* $(\forall F \in \text{F})\ P(F) \geq 0$.
3. *Finite additivity* If $F_1,...,F_n \in$ F, and $F_i \cap F_j = \emptyset$ for all $i \neq j$, then $P\left(\bigcup_{i=1}^{n} F_i\right) = \sum_{i=1}^{n} P(F_i)$.
4. *Continuity* If $(\forall\, i)\ F_i \supseteq F_{i+1}$ and $\bigcap_{i=1}^{\infty} Fi = \phi$, then $\lim_{i \to \infty} P(F_i) = 0$.

These four axioms also capture basic intuitions: The first two capture the ideas that an experiment giving rise to outcomes always gives rise to one of its potential outcomes and that negative probabilities are impossible, The third says that the probability that any of a set of mutually exclusive events occurs is the sum of their individual probabilities. The fourth axiom is technical, in that it says that if an infinite sequence of nested events "vanishes", then probabilities of successive events approach 0. The reason for the fourth axiom is to ensure that *P* satisfies the law of large number. It is important to note that while Kolmogorov's axioms capture basic intuitions, they also capture ideas not normally associated with ideas of experimentation, such as the probability that an irrational number in the interval [0,1] is in the Cantor set. We cannot operationalize the process "pick an irrational number at random".

Kolmogorov's approach to probability has been regarded as a benchmark in the development of probability theory. It is considered to be almost universally applicable in situations dealing with chance and uncertainty. However, probabilists had a hard time accepting Kolmogorov's approach—"The idea that a mathematical random variable is simply a function, with no romantic connotation, seemed rather humiliating…" (Doob 1996 p. 593) The words "romantic", and "humiliating" indicates that Doob sensed a clear disconnection between Kolmogorov's deductive system and the inductive, experimental approach to probability and its applications. The tension between two approaches is even sharper in the work of Fine (1973)who argues that the probability scale in Kolmogorov's approach is "occasionally empirically meaningless and always embodies an arbitrary choice or convention." The very term "arbitrary choice or convention" betrays the author's belief in unspecified yet easily understood ontological premises. Fine continues,

"While conventions can be harmless, there is the danger that the apparent specificity of the Kolmogorov setup may obscure the absence of a substantial grip on the structure of random phenomena." Here Fine's statement is more specific: he yearns for the existence of a particular structure of random phenomena and is convinced that a probability theory should "grasp" this structure.

However, when analyzing Kolmogorov's measure theoretical probability and its implications to mathematical and scientific development, one has to keep in mind that Kolmogorov's approach follows very much Hilbert's formalist philosophy of mathematics. David Hilbert, one of the greatest mathematicians in history, set the tone for twentieth century mathematics with his advocacy of axiomatization. Hilbert's program aimed at establishment of a firm foundation of mathematics shaken in the crisis brought by paradoxes related to set theory (cf. Davis and Hersh 1981; Tiles 1991). Hilbert formalizes Geometry and Algebra by formulating them as formal systems of symbols and rules in which every theorem can be logically deduced from a set of axioms. The axioms captured the "essence" of a mathematical system. In brief, in the formalist approach, mathematics is a science of logical deduction and the meaning of the symbols and mathematical theorems are something "extra-mathematical"; Hilbert himself said once that his formal system of geometry can use the terms "tables, chairs, and beer mugs" instead of "dots, lines, and planes" (Reid 1970 p. 57). Kolmogorov apparently shares this intention as he tried to formalize probability theory in the same way. He wrote, "The theory of probability, as a *mathematical discipline*, can and should be developed from axioms in exactly the same way as Geometry and Algebra. This means that after we have defined the elements to be studied and their basic relations, and have stated the axioms by

which these relations are to be governed, all further exposition must be based exclusively on these axioms, *independent of the usual concrete meaning of these elements and their relations.* … The concept of a field of probabilities is defined as a system of sets which satisfies certain conditions. *What the elements of this set represent is of no importance in the purely mathematical development of the theory of probability*." (Kolmogorov 1956 p. 1, italics in original)

Kolmogorov's approach to probability makes a distinction between probability as a deductive structure and probability as a descriptive science. It is in this sense that when designing probability curriculum, one has to keep in mind what aspects of probability theory are of prominent importance to teach in the classroom. To use an analogy, teaching strictly Kolmogorov's axiomatic probability is like teaching the concept of circle as "a set of (x, y) that satisfies the condition $x^2 + y^2 = a(a \geq 0)$". If this definition precedes the introduction of the concept of distance and measurement of length, students may have considerable difficulties in visualizing the image of a circle as a set of points having the same distance from a fixed point in a two-dimensional surface. The implication is that although Kolmogorov's approach to probability does not lead to any contradiction, a probability curriculum only addressing the axiomatic approach to probability may prevent students from developing conceptual understanding of probability rooted in their experiences. It may also conceal the possible applications of probability theory.

De Finetti's subjective probability

Subjective probability is also known as Bayesian approach to probability. It is jointly attributed to de Finetti (1937), Ramsey (1931), and Savage (1954). All three authors

proposed essentially the same definition of probability, namely "the *degree of belief* in the occurrence of an event attributed by a given person at a given instant and with a given set of information" (de Finetti 1974 p. 3). It is often understood by many that the theory of subjective probability is based on the assumption that probability is a degree of belief or intensity of conviction that resides in human being's conscious mind, as opposed to an "objective probability" that exists irrespective of mind and logic (Good 1965 p. 6). de Finetti (1974) critically examined the hidden assumptions made by "objectivist" approaches to probability. The notion of "collectives" in von Mises' approach, for example, presupposes a certain degree of personal initiatives, meaning it is somewhat arbitrary to choose a collective against which to evaluate probability (Ayer 1972). As de Finetti wrote, "when one pretends to eliminate the subjective factors one succeeds only in hiding them" (de Finetti 1937; as quoted in Piccinato 1986 p. 16).

The defining property of subjective probability is the use of further experiences and evidences to change the initial opinions or assignment of probability. This is expressed symbolically as Bayes' theorem

$$P(B|A) = P(A \text{ and } B)/P(A)$$

Good (1965) argues that it is not a belief in Bayes' theorem that makes one a Bayesian, as the theorem itself is just a trivial consequence of the product axiom of probability. Rather, it is a readiness to incorporate intuitive probability into statistical theory and practices that makes one a subjectivist. However, the very "subjective" character of de Finetti's approach has been the most intensively discussed and criticized. In particular, his intention to view probability as subjective was considered as introducing arbitrariness in probability theory, which "invalidates the power of Bayesian theory" (Piccinato 1986

p. 15). For example, one may claim that one person with a 0.5 degree of belief actually had a stronger belief than another person who had a 0.55 degree of belief.

De Finetti justified subjective probability by introducing the idea of coherence. In gambling situation, for example, probability judgment is coherent if it does not expose one player to certain loss if his opponent is prudent/clever. According to coherence principle, it is perfectly possible that for a same uncertain event one person will have 0.5 degree of belief of its occurrence and that another have 0.55, but they will converge to the same final estimates of probability if faced with all available data/evidence. De Finetti proposed this thought experiment to illustrate the idea of coherence:

> You must set the price[2] of a promise to pay $1 if John Smith wins tomorrow's election, and $0 otherwise. You know that your opponent will be able to choose either to buy such a promise from you at the price you have set, or require you to buy such a promise from your opponent, still at the same price. In other words: you set the odds, but your opponent decides which side of the bet will be yours. The price you set is the "operational subjective probability" that you assign to the proposition on which you are betting.
>
> The rules do not forbid you to set a price higher than $1, but if you do, your prudent opponent may sell you that high-priced ticket, and then your opponent comes out ahead regardless of the outcome of the event on which you bet. Neither are you forbidden to set a negative price, but then your opponent may make you pay him to accept a promise from you to pay him later if a certain contingency eventuates. Either way, you lose. The bottom-line conclusion of this paragraph parallels the fact that a probability can neither exceed 1 nor be less than 0.
>
> Now suppose you set the price of a promise to pay $1 if the Boston Rex Sox win next year's World Series, and also the price of a promise to pay $1 if the New York Yankees win, and finally the price of a promise to pay $1 if *either* the Red Sox or the Yankees win. You may set the prices in such a way that
>
> $$\text{Price(Red Sox)} + \text{Price(Yankees)} \neq \text{Price(Red Sox or Yankees)}.$$
>
> But if you set the price of the third ticket too low, your prudent opponent will buy that ticket and sell you the other two tickets. By

[2] In de Finetti's theory, bets are for money, so your probability of an event is effectively the *price* that you are willing to pay for a lottery ticket that yields 1 unit of money if the event occurs and nothing otherwise. De Finetti used the notation 'Pr' to refer interchangeably to Probability, Price, and Prevision ('foresight'), and he treated them as alternative labels for a single concept.

considering the three possible outcomes (Red Sox, Yankees, some other team), you will see that regardless of which of the three outcomes eventuates, you lose. An analogous fate awaits you if you set the price of the third ticket too high relative to the other two prices. The bottom-line conclusion of this paragraph parallels the fact that probability is additive (see probability axioms).

Now imagine a more complicated scenario. You must set the prices of three promises:
* to pay $1 if the Red Sox win tomorrow's game; the purchaser of this promise loses his bet if the Red Sox do not win regardless of whether their failure is due to their loss of a completed game or cancellation of the game, and
* to pay $1 if the Red Sox win, and to refund the price of the promise if the game is cancelled, and
* to pay $1 if the game is completed, regardless of who wins.

Three outcomes are possible: The game is cancelled; the game is played and the Red Sox lose; the game is played and the Red Sox win. You may set the prices in such a way that

$$\text{Price}(\text{complete game}) \times \text{Price}(\text{Red Sox win} \mid \text{complete game}) \neq \text{Price}(\text{Red Sox win})$$

(where the second price above is that of the bet that includes the refund in case of cancellation). Your prudent opponent writes three linear inequalities in three variables. The variables are the amounts he will invest in each of the three promises; the value of one of these is negative if he will make you buy that promise and positive if he will buy it from you. Each inequality corresponds to one of the three possible outcomes. Each inequality states that your opponent's net gain is more than zero. A solution exists if and only if the determinant of the matrix is not zero. That determinant is:

$$\text{Price}(\text{complete game}) \times \text{Price}(\text{Red Sox win} \mid \text{complete game}) - \text{Price}(\text{Red Sox win}).$$

Thus your prudent opponent can make you a sure loser unless you set your prices in a way that parallels the simplest conventional characterization of conditional probability (de Finetti 1937).

Statistical inference

Statistical inference is “the theory, methods, and practice of forming judgments about the parameters of a population, usually on the basis of random sampling”(Collins English Dictionary 2000). The problem of statistical inference, as Hacking (1965) stated, is “to give a set of principles which validate those correct inferences which are peculiarly statistical”(*ibid*, p. 85). Because statistical inference is more concerned with logic and

conventions and thus involves less philosophical intricacy than does probability, there is generally less controversy as to both its nature and application. There are two important themes in statistical inference: hypothesis testing and parameter estimation. In general terms, the first is concerned with whether two (or more) sets of observations should be considered similar or different, while the second has to do with to decide how big is a difference (Simon 1998).

Hypothesis testing

The first published study on hypothesis testing was conducted by John Arbuthnot in 1710 (Hacking 1965). Arbuthnot studied the hypothesis that a new-born child has an equal chance of being male or female. He took 82 consecutive years of birth register in London as his data. On every year more boys were born than girls. Arbuthnot argued that if the hypothesis were true, that in fact there were an equal chance for male and female births, then there would be only a miniscule chance of getting more boys born in 82 consecutive years: $(1/2)^{82}$. Based on this result, Arbuthnot rejected the hypothesis. His reasoning was: an event had happened in London, as reported in the registers. If the hypothesis were true, the chance of that event happening would have been minute. So the hypothesis should be rejected (Hacking 1965) (Note that the idea of hypothesis testing is built upon the concept of probability from the very beginning). Fisher (1956) elaborated this reasoning into the logic of simple disjunction: either an exceptionally rare chance has occurred, or the hypothesis is not true. In other words, suppose a hypothesis is true, and according to which an event has a very small chance of occurrence if drawn at random. Now suppose the event does occur, then either we acknowledge that we encounter a small chance event, or we reject the hypothesis. But then, on what basis should we reject

a hypothesis? If we reject a hypothesis because what happens would happen rarely if the hypothesis were true, we might reject a true hypothesis because what would happen rarely could still happen. More over, if a hypothesis is judged to be not a viable explanation, what is a good explanation?

One of the solutions to this question was the use of rival hypotheses. It was very intuitive: Do not reject a hypothesis if what happens would happen rarely if the hypothesis were true. Reject it only if there is something better. Gossett wrote:

> A text doesn't in itself necessarily prove that the sample is not drawn randomly from the population even if the chance is very small, say .00001:what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05 (such as that it belongs to a different population or that the sample wasn't random or whatever will do the trick) you will be very much more inclined to consider that the original hypothesis is not true. (Hacking 1965 p. 83)

This leads to a theory of testing: a hypothesis should be rejected if and only if there is some rival hypothesis much better supported than it is. Gossett's theory played an important role in the work of Neyman and Pearson, who later invented the idea of significance level and developed the theory of hypothesis testing that is widely received today.

According to Neyman and Pearson, there should be very little chance of mistakenly rejecting a true hypothesis. Thus, the chance of an event occurring if the hypothesis were true has to be as small as possible for one to reject the hypothesis. This chance is called the significance level of the test. Introducing the idea of significance level to hypothesis testing was in essence adopting a convention as to when to accept or reject hypothesis. A hypothesis concerning the parameters of a population distribution will be rejected only if the probability of an observation/random sample or more extreme

samples from the given population (i.e., if the hypothesis were true) falls below a predetermined significance level.

However, rejection is not refutation, as Hacking (1965) put it. The fact that a hypothesis is rejected by some decision rule does not mean that it is necessarily false. Rather, it means that if we apply the same rule over and over again, more often than not, we will reject the false hypothesis. Neyman and Pearson wrote:

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we will ensure that, in the long run experience, we shall not be too often wrong. Hence, for example, would be such a rule of behavior: to decide whether a hypothesis H, of a given type, be rejected or not, calculate a specified character, x, of the observed facts; if $x>x_0$ reject H; if $x\leq x_0$, accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x\leq x_0$ or false when $x>x_0$. But it may often be proved that if we behave in such a way we shall reject when it is true not more, say than once in a hundred times, and it addition, we may have evidence that we shall reject H sufficiently often when it is false (Neyman & Pearson, 1933 as quoted in Hacking 1965, p. 104).

In other words, Neyman and Pearson proposed that we should not hope to find evidence about the truth of any particular hypothesis, but that we should consider the whole class of hypotheses that we shall ever test.

According to Hacking (1965), Fisher strongly objected to the Neyman-Pearson procedure because of its mechanical, automated nature. Use of a fixed significance level, say 0.05, promotes the seemingly nonsensical distinction between a significant finding if the P value is 0.049, and a non-significant finding if the P value is 0.051. Fisher insisted that although the Neyman Pearson theory worked for testing long sequences of hypotheses, as in industrial quality control, it was irrelevant to testing hypotheses of the sort important to scientific advance (Hacking 1965 p. 105). Nonetheless, despite Fisher and Pearson's long known feud against each other's work, their theories share a common

underlying logic that is related to Popperian inference, which seeks to develop and test hypotheses that can clearly be falsified (Popper 1959), because a falsified hypothesis provides greater advance in understanding than does a hypothesis that is supported (Johnson 1999).

Parameter estimation

While hypothesis testing tests the viability of hypotheses about a population characteristic, parameter estimation estimates the population characteristic through random sampling and quantifies the error in such estimation. Parameter estimation consists of point estimation and interval estimation. The first includes the idea of mean, median, variance, standard variation, etc. The second includes confidence interval and margin of error. These ideas, although often misunderstood (as I will address in the next chapter), are relatively straightforward in mathematics in terms of their meaning and interpretation.

However it is interesting to note that what is commonly referred to as confidence interval are generally regarded as a frequentist method, i.e., employed by those who interpret "90% probability" as "occurring in 90% of all cases". A “95% confidence interval” means that if the study were repeated an infinite number of times, 95% of the confidence intervals that resulted would contain the true population parameter. What is normally called “a confidence interval has a 95% chance of containing the population parameter” does not tell whether that particular confidence interval contains the population parameter, rather it says that if we were to repeat sampling and calculate the confidence interval at the like fashion, 95% of these confidence intervals will contain the true population parameter. This is consistent with what Neyman and Pearson wrote about

hypothesis testing in the sense that it is always a group attribute that is being considered and quantified. That frequentist probability is a foundation for statistical inference thus far talked about is because it was favored by some of the most influential statisticians in the first half of twenties century, including Fisher, Neyman, and Pearson.

Bayesian inference offers an alternative to the frequentist methods for hypothesis testing and estimation. In Bayesian inference, one starts with an initial set of beliefs about the relative plausibility of various hypotheses, collects new information (for example by conducting an experiment), and adjusts the original set of beliefs in the light of the new information to produce a more refined set of beliefs of the plausibility of the different hypotheses. In other words, Bayesian inference reduces statistical inference to Bayesian probability (see subjective probability). For example, sometimes the value of a parameter is predicted from theory, and it is more reasonable to test whether or not that value is consistent with the observed data than to calculate a confidence interval (Johnson 1999). For testing such hypotheses, what is usually desired is $P(H_0/\text{data})$. What is obtained, as pointed out earlier, is $P(\text{data}/H_0)$. Bayes' theorem offers a formula for converting between them:

$$P(H_0/\text{data}) = P(\text{data}/H_0)\, P(H_0)\, /P(\text{data})$$

This is an old (Bayes 1763) and well-known theorem in probability. Its use in the present situation does not follow from the frequentist view of statistics, which considers $P(H_0)$ as unknown, but either zero or 1. In the Bayesian approach, $P(H_0)$ is determined before data are gathered; it is therefore called the prior probability of H_0. There have been numerous debates on which methods of inference: frequentist or Bayesian, are more advantageous. However, it is not in the scope of this study to discuss the details.

A quantitative conceptual perspective

The perspective my research team and I (henceforth, I) took in designing the teachers seminar comes from Thompson' theory of quantitative reasoning (Thompson 1994). Briefly, Thompson's theory of quantitative reasoning is about people conceiving situations in terms of quantities (i.e., things having measures) and relationships among quantities. To conceive of physical quantities as measurable attributes of objects means one has come to conceive of a class of objects as having an attribute that can be seen as segmentable or as a relationship among segmentable attributes (Steffe 1991; Thompson 1994). From this perspective, understanding probability of an event entails first conceiving of something that is potentially measurable, such as conceiving an action that produces something having an "extent" or "intensity", and then conceiving a way to measure that extent or intensity. This points ultimately to understanding "the probability of an event" as a relationship between that event's extent and the extent of a universe of possibilities. In the realm of situations school students face, I would mean students understanding "the probability of E (event)" as meaning "the fraction of the time we expect E to happen". Note that I am not arguing for the relative frequentist' theory of probability, rather, I am talking about a conception of probability that fits with a quantitative perspective.

The advantage of a quantitative conception of probability is that it also supports thinking about an ontogenesis of conceptual schemes by which students can understand ideas of distribution and density of random variables, sampling (as a stochastic process), statistic as a measure of a group attribute, distributions of sample statistics, and statistical inference. Statistical inference is about inferring a population parameter by taking one

sample. One measures the accuracy of such an inference by making a probabilistic judgment of the sampling process' accuracy – the proportion of times a sample statistic would occur within a certain range were one to sample many times. By having students think of a probability as a statement of expectation of relative frequency – that to say an event has a probability of *.015* is to say that we *expect* an event to occur 1.5 percent of the time as we perform some process repeatedly, one builds a foundation for students to understand the idea of sampling distribution, margin or error, and confidence interval, etc.

Conversely, the context of sampling and statistical inference provides a natural environment for supporting the conception of probability as equivalent to mathematical expectation. Statistics instruction that aims to have students imagine distributions as emerging from repeatedly sampling a population and think of probability in relation to the distributions will likely divert students from thinking of probability as about a single event. It is in this sense that we follow Shaughnessy's (1992) use of the word "stochastics" to denote a combination of probability and statistics, i.e., we conceptualize probability and statistics in such way so that they are proposed to students as two expressions of a core scheme of operations.

Summary

In this chapter, I first reviewed the historical development of the ideas of and in probability and statistical inference. In my review, I attempted to highlight the conceptual essence of these ideas, the relations among them, and sometimes the pedagogical implications. In doing so, I hope to not only provide a glimpse of the controversies in the

development of these ideas, the complexities in understanding them, but also a vantage point for anticipating and making sense of teachers' understanding of these ideas. I then briefly described a way of conceiving of probability, statistical inference, and their relationships from a quantitative conceptual perspective. It is relevant to my study in an indirect but significant way. Although the teachers seminar intended to uncover teachers' understanding of probability and statistical inference, we, as designers of the seminar, could not, and did not, conduct the study without having in mind of what we hope the teachers would understand. Our rationale, in a nutshell, was: We develop a scheme of ideas that we hope students would understand so that they would develop coherent stochastic reasoning. By developing a scheme of ideas, I mean articulating ways of thinking about these ideas and their relationships. Having this scheme of ideas as an end-goal, we design instructions to probe students reasoning and support their learning as they engage in the instruction. In doing so, we construct knowledge of ways students operate on these ideas and the difficulties they experience as they attempt to assimilate the ideas we try to teach. This knowledge then becomes a resource for our work with teachers. Rather than coming to work with teachers without a clue of what we hope students and teachers would understand, we have a better idea of what we hope students to know and the difficulties they have in knowing, which allows us to narrow our focus and ask the question: given what we have known, what must teachers be grappling with in order to create a learning environment that is propitious for students learning? This zooming-in allows us to focus on probing teachers understanding on the big and difficult ideas pertaining students understanding and their teaching. This naturally leads me to the next

chapter, in which I will review literature concerning how people reason about probability and statistical inference in everyday and instructional settings.

Understanding Probability and Statistical Inference: a Review of Psychological and Instructional studies

Probability

Coming to understand probability: an epistemological perspective

Piaget and Inhelder defined chance as an essential characteristic of irreversible phenomena, as opposed to mechanical causality or determinism characterized by its conceptual reversibility. In their book, *The origin of the idea of chance in children* (Piaget and Inhelder 1975), Piaget and Inhelder described children's construction of the concepts of chance and probability in relation to the development of their conceptual operations. According to Piaget and Inhelder, children develop the concepts of chance and probability in three successive stages.

In the first stage (prelogical), generally characteristic of children under seven or eight years of age, children do not distinguish possible events from necessary events. "The discovery of indetermination which characterizes chance, by contrast with operative determination, entails the dissociation of two modalities, or planes of reality—the possible, and the necessary—while on an intuitive level they remain undifferentiated in reality or in being" (*ibid*, p. 226).

The second stage (concrete operation) starts when logical-arithmetical operations appears at around seven or eight years of age. Children start to differentiate between the necessary and the possible from the construction of the concrete logical-arithmetical

operations. At this level, the notion of chance acquires a meaning as a "noncomposable and irreversible reality" antithetical to operations which are reversible and composable in well-defined groups, and "the reality of chance is recognized as a fact and as not reducible to deductive operations" (*ibid*, p. 223). Piaget and Inhelder further hypothesized that beyond the recognition of the clear opposition between the operative and the chance, the concept of probability presupposes the existence of the sample space, that is, all the possible cases, so that "each isolated case acquires a probability expressed as a fraction of the whole" (*ibid*, p. 229). In other words, to get to this stage, children must 1) construct combinatoric operations, and, 2) understand proportionalities. They found out, however, that after distinguishing the possible from the necessary, children of the second stage failed to produce an exhaustive analysis of the possible. They argued that this was because an analysis of sample space (or all the possible cases) assumes operating on simple possibilities as hypotheses, yet children at this stage were only able to deal with the actual situations.

Finally, the third stage characterized by formal thought begins at eleven or twelve years of age. According to Piaget and Inhelder, during this period, children translate the unpredictable and incomprehensible chance into the form of a system of operations, which are incomplete and effected without order (in other words, according to chance). As such, chance becomes comparable to those very operations conducted systematically and in a complete manner. For example, once children have learned the operations of permutations, they can deduce all the possibilities if a chance situation to appreciate the fact that one particular outcome is " tiny" in comparison and is thus " unlikely" to occur. The judgment of probability thus becomes a synthesis between chance and operations.

The operations lead to the determination of all the possible cases, even though each of them remains indeterminate for its particular realization. Probability, being a fraction of determination, then consists in judging isolated cases by comparison with the whole.

The implication of Piaget's research is two fold. On one hand, it revealed the developmental constraints that children have in learning probability. On the other hand, it described the conceptual challenges one has to overcome in order to develop probabilistic reasoning, namely: 1) distinguishing uncertainty from deterministic situations and events, 2) developing a sense of the magnitude of possibilities of a chance event, and 3) understanding proportionalities. Recent research has found out that these conceptual challenges are not only functions of age, but also other variables. For example, Kahneman, Slovic, and Tversky (1982) and Konold (1989; 1991) found that people who have passed beyond the age levels identified by Piaget and Inhelder could still fail to distinguish between uncertain and necessary events due to a deterministic world view. That is, they often think that observable phenomena are connected to one another in cause-effect, perhaps complicated, ways.

Modeling the development of students' probabilistic reasoning

A number of studies (Shaughnessy 1992; Jones, Langrall et al. 1997; Horvath and Lehrer 1998) proposed models of students' development of probabilistic reasoning. Shaughnessy (1992) elaborated a model of stochastic conceptual development. According to this model, people's understanding of stochastics indicates various levels of conceptual sophistication, characterized by the following four types along an increasing advance scale:

1. *Non-statistical*. Indicators: responses based on beliefs, deterministic models, causality, or single outcome expectations; no attention to or awareness of chance or random events.
2. *Naïve-statistical*. Indicators: use of judgmental heuristics, such as representativeness, availability, anchoring, balancing; mostly experientially based and nonnormative responses; some understanding of chance and random events.
3. *Emergent-statistical*. Indicators: ability to apply normative models to simple problems; recognition that there is a difference between intuitive beliefs and a mathematized model, perhaps some training in probability and statistics, beginning to understand that there are multiple mathematical representations of chance, such as classical and frequentist.
4. *Pragmatic-statistical*. Indicators: an in-depth understanding of mathematical models of chance (i.e. frequentist, classical, Bayesian); ability to compare and contrast various models of chance, ability to select and apply a normative model when confronted with choices under uncertainty; considerable training in stochastics; recognition of the limitations of and assumptions of various models (*ibid*, p. 485).

While Shaughnessy modeled the stochastic understanding developmentally, Jones, Langrall, Thronton, and Mogill (1997) modeled students' probabilistic thinking along four conceptual constructs *sample space*, *probability of an event*, *probability comparison*, *and conditional probability*. In Jones' et al.'s framework, students' understanding of these constructs is demonstrated by their ability to exhibit certain behaviors when faced with uncertain situations. Specifically,

> An understanding of *sample space* is exhibited by the ability to identify the complete set of outcomes in a one-stage experiment (e.g., tossing one coin) or a two-stage experiment (e.g., tossing two coins [one at a time])… understanding of *probability of an event* is exhibited by the ability to identify and justify which of two or three events are most likely or least likely to occur… understanding of *probability comparisons* is measured by their ability to determine and justify: a) which probability situation is more likely to generate the target event in a random draw; or b) whether two probability situations offer the same chance for the target event…understanding of *conditional probability* is measured by their ability to recognize when the probability of an event is and is not changed by the occurrence of another event (*ibid*, p104-106).

According to Jones, et al., children exhibit different levels of thinking across these four constructs. These levels of thinking, from the least to the most sophisticated, are *subjective thinking*, *transitional between subjective and naïve quantitative thinking*, *informal quantitative thinking*, and *numerical reasoning*. Jones, et al. hypothesized typical behaviors associated within each level (attach chart).

Both Shaughnessy's and Jones, et al.'s models/frameworks are cast in term of behaviors or "observed learning outcomes" (Biggs and Collis 1982). Horvarth & Lehrer (1998) modeled probabilistic reasoning in terms of students' conceptions and imaginations. In their model, classical statistics has five distinct, yet related, components:

> 1) the distinction between certainty and uncertainty, 2) the nature of experimental trial, 3) the relationship between individual outcomes (events) and patterns of outcomes (distribution), 4) the structure of events (e.g., how the sample space relates to outcomes), and 5) the treatment of residuals (i.e., deviations between prediction and results, model and phenomenon.)

An "expert" model of statistics along these five components, Horvath and Lehrer suggested, are

> 1) understanding uncertainty as a conception that is situated in a context, instead of a fundamental property of a phenomenon; 2) understanding a trial as an instantiation of an experiment that yields a public outcome. This is a means of marking or classifying each event in order to combine sets of events that are required in the models of probability; 3) realizing that although individual events may be highly unpredictable, global patterns of outcomes are often predictable, which is often referred to as the "law of large number;" 4) having a means of systematically and exhaustively generating the sample space, and mapping the sample space onto the distributions of outcomes; 5) understanding that there will be residuals (or differences) between model (abstractions of key structures of relationship present in the phenomena) and the phenomena being modeled.

Note that while Shaughnessy's model described people's conceptual sophistication in probability and statistics in general, Jones, et al. and Horvath & Lehrer focused

specifically on students' probabilistic thinking, and they investigated students' thinking along several components. Although Horvath and Lehrer and Jones, et al. both studied children's reasoning along what they considered to be key constructs of classical probability, the latter's constructs seems to be more conceptual, while the former's constructs seemed to exhibit only task differences. Horvath & Lehrer's constructs are distinctive of each other with regard to their conceptual entailment, yet they also roughly form a natural progression in understanding chance and probability. For example, as Horvath and Lehrer suggested, understanding the relationship between simple events and distribution presupposes an understanding of the nature of experimental trial, and understanding the role of sample space in chance investigations presupposes an understanding of some relationship between simple events and distributions. On the contrary, the four constructs in Jones, et al.'s framework do not seem to form a progression in probabilistic thinking. However, these constructs seem to be specifically selected and employed to prescribe different tasks that were used to evaluate children's thinking. In sum, these three models provide a post-Piagetian interpretation of how probabilistic reasoning develops. They identified key constructs of probabilistic reasoning from different perspectives, which afford multiple ways we might make sense of aspects of teachers' understanding of probability.

Judgment heuristics

The research of Kahneman, Tversky and their colleagues (Kahneman and Tversky 1972; 1973; Tversky and Kahneman 1973; Kahneman, Slovic et al. 1982) documented the persistent "misconceptions" that people demonstrate when making judgment about situations involving uncertainty. Among these misconceptions are systematic mental

heuristics that do not conform to the mathematically-normative ways of reasoning under uncertain situations. According to the "representativeness heuristic", for example, people estimate the likelihood for events based on how well an outcome represents some aspect of its parent population, or reflects the process by which it is generated (Kahneman and Tversky 1972 p. 430). One problem often cited in the literature to illustrate the representativeness heuristic has been referred to as the "Linda Problem." Tversky and Kahneman (1983) presented the following personality sketch to a large number of statistically naïve undergraduates:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Subjects were subsequently asked which of two statements about Linda was more probable: 1) Linda is a bank teller, or 2) Linda is a bank teller who is active in the feminist movement. Tversky and Kahneman reported that 86% of subjects chose statement 2)—a choice which violates the conjunction rule of probability. They attributed this violation to the subjects' having based their choice on its resemblance to the sketch and concluded that the subjects did not have valid intuitions corresponding to the formal rules of probability when making judgment. Of course, it is also possible, as noted by Konold (1989), that subjects were not answering the question that Tversky and Kahneman intended. They may have been answering the question, "Which is the more accurate ("probable") description of Linda?"

Another example illustrating what was called "base-rate misconception" was documented in Kahneman and Tversky (1982):

> A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
> 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the nights of the accident and concluded that the witness correctly identified each one of the two colors 80%of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green?

Kahneman and Tversky anticipated that people would answer the question, "What is the probability that the cab is blue *given that the witness said it is blue*?" The correct answer to the question asked is 15%. The correct answer to the question they intended is about 41%, meaning that 41% of the time that the witness says it is blue, the cab is really blue. However, a typical answer from a large number of subjects is 80%, which answers the question, "What percent of the time does the witness correctly identify a cab's color?" (which is not the same as the percent of the time that the witness correctly identifies blue cabs). Kahneman and Tversky interpreted their results as indicating that people tend to ignore the base rate information because they see it as incidental rather than as a causal factor. Kahneman and Tversky did not, however, entertain the possibility that people understood their questions differently than they intended.

Kahneman and Tversky suggested that the systematic errors and misconceptions are disconcerting—either because the correct answer seems to be obvious in retrospect, or because "the error remains attractive although one knows it is an error"(Kahneman, Slovic et al. 1982). Such aspect of judgment-heuristics has been interpreted in two distinctive, yet interrelated directions. One interpretation concerns mathematics, and it says that some of the principles of mathematical probability are non-intuitive or counter-intuitive, which might account for the difficulties students have in assimilating the ideas

of probability. The second interpretation concerns psychology. It argues that human minds are simply not built to work by the rules of probability (Gould 1991 p. 469; Piatelli-Palmarini 1994).

Later research, such as by Konold and his colleagues (Konold 1989; Konold, Pollatsek et al. 1993) and by Gigerenzer (1994; 1996; 1998), Hertwig and Gigerenzer (1999) suggested that Kahneman and Tversky might have over-interpreted their data. While Kahneman and Tversky mainly focused on how one measures probability, Konold and Gigerenzer shifted the focus towards students' interpretation of probability questions. Konold (1989) suggested that, "Hidden in the heuristic account is the assumption that regardless of whether one uses a heuristic or the formal methods of probability theory, the individual perceives the goal as arriving at the probability of the event in question. While the derived probability value may be non-normative, the meaning of that probability is assumed to lie somewhere in the range of acceptable interpretation" (*ibid*, p. 146). In other words, Kahneman and Tversky seemed to assume that their subjects assumed a mathematical, or relative frequency meaning for "probability", while it may have been that many of them had a deterministic understanding of events, and that numerical probability simply reflected their degree of belief in the outcome.

In their experiments with the Linda problem, Hertwig and Gigerenzer (1999) found that many subjects had *nonmathematical* interpretations of "probability." For example, they may interpret the Linda problem as a task of looking for a plausible or accurate description of Linda. Such discrepancy in interpretations of probability comes from the fact that while subjects assume that the content of the Linda problem should be relevant to the answer, Kahneman and Tversky were actually testing a sound reasoning

according to which the content is irrelevant, to borrow Gigerenzer's phrase, "all that counts are the terms *probable* and *and"* (Gigerenzer 1996 p. 593, italics in original). Hertwig and Gigerenzer (1999) suggested that if students assumed a nonmathematical interpretation of probability, then their answers could not be taken as evidences of violation of probability theory "because mathematical probability is not being assessed" (*ibid*, p. 278).

Gigerenzer (1996) further argued that these judgment-heuristics were too vague to provide any meaningful explanations of people's reasoning. "The problem with these heuristics is that they at once explain too little and too much. Too little because we do not know when these heuristics work and how; too much, because, post hoc, one of them can be fitted to almost any experimental result"(*ibid*, p. 592). In other words, Gigerenzer believed that judgment-heuristics do not count as explanations as they are merely re-description and do not account for the underlying cognitive processes subjects undergo that make them choose a particular answer.

Outcome approach

As I will elaborate later, a number of studies have established an opposition between causal analysis and probabilistic reasoning, i.e., sound probabilistic reasoning precludes causal analysis of a probabilistic situation. Contrary to this traditional viewpoint, Konold (1989) argued that a formal probabilistic approach does not necessitate the denial of underlying causal mechanisms in the case of chance events. In practice, however, a causal description is often seen as impractical if not impossible (Von Mises 1957). Accepting a current state of knowledge, a probability approach adopts a "black-box" model according to which underlying causal mechanism, if not denied, are ignored

(Konold 1989). In his study on students' informal conceptions of probability, Konold claimed that the preference for causal over stochastic models has been linked to the preference for predicting outcomes of single trails rather than sample results. He then proposed that people's non-normative responses to probability questions might be due not only to their indiscriminate application of judgment-heuristics (Kahneman and Tversky 1972; 1973; Tversky and Kahneman 1973; Kahneman, Slovic et al. 1982), but also to their non-normative interpretation of probability and understanding of the goal in reasoning under uncertainty.

Konold investigated this hypothesis with a small sample of psychology undergraduate students. In individual interviews, students verbalized their thinking as they responded to questions about situations involving uncertainty. Konold then conducted both statistical and qualitative analysis on the interview protocol. On the basis of his analysis, Konold developed a model of students' reasoning that he called the *outcome approach* (Konold 1989; 1991; Konold, Pollatsek et al. 1993). Outcome oriented thinking is characterized by three salient features: 1) predicting outcomes of single trails, 2) interpreting probability as predictions and thus evaluating probabilities as either right or wrong after a single occurrence, 3) basing probability estimates on causal features rather than on distributional information. Individual employing an outcome approach do not interpret probability questions as having to do with a stochastic process. Instead of conceiving a single trial or event as embedded within a sample of many such trials, they view each one as a separate, individual phenomenon. Consequently, they tend to interpret their decision-making task as one of correctly predicting for certain, and on the basis of relevant causal factors, what the next outcome will be, rather than one of estimating what

is likely to occur in the long run on the basis of frequency data. Konold's finding suggested that causal analysis is tied with students' understanding of the goal of probability as predicting the outcome. He further claimed that if the outcome approach is a valid description of some novices' orientation to uncertainty, then the application of a causal rather than a black-box model to uncertainty seems the most profound difference between those novices and the probability expert and, therefore, perhaps the most important notion to address in instruction.

Konold (1995) also conjectured that students could hold multiple and often contradictory beliefs about a particular situation. For example, in one experiment, students were given the following problems:

> Part 1: Which of the following sequences is most likely to result from flipping a fair coin 5 times?
>
> (a) H H H T T,
> (b) T H H T H,
> (c) T H T T T,
> (d) H T H T H,
> (e) All four sequences are equally likely;
>
> Part 2: Which of the above sequences is least likely to result from flipping a fair coin 5 times?

Konold reported that while 70% of the subjects correctly responded the first part of the problem, that the sequences are equally likely, over half of these subjects did not choose (e) for the second part. Rather they indicated that one of the sequences is "least likely", which inadvertently contradicts their response to the first part. After interviewing these subjects for their reasoning, Konold concluded that this inconsistency resulted from the subjects' applying different perspectives to the two parts of the problem. In part 1, many subjects thought they were being asked, in accordance with outcome approach, to predict which sequence will occur. They chose (e) not because they understood that the

probability of each sequence occurring is the same, but because they couldn't rule out any of them. In part 2, many of these subjects applied the representative heuristics. For example, one might choose (c) as being least likely based on the fact that it contains an excess of T's.

Causal analysis

Although causal analysis was indicated in the above studies as associated with the important obstacles student must overcome in reasoning probabilistically, a number of studies explicitly discussed the implications and consequences of causal analysis. In the context of investigating students' difficulties in understanding sampling, Schwartz et al. (Schwartz and Goldman 1996; Schwartz, Goldman et al. 1998) suggested that one of the difficulties is that interpreting certain everyday situations, such as opinion polls, in terms of sampling requires the ability to manage the tensions between ideas of causality and of randomness. For instance, understanding a public opinion poll as a random sample involves giving up analysis of the causal factors behind people's opinions. Schwartz et al. referred to people's tendency to focus on causal association in chance situation as the covariance assumption, which describe specifically the phenomena that 1) people reason as though they assume everyday events should be explained causally, and 2) people search for co-occurrences or temporal associations between events and/or properties that can support this kind of explanation.

Biehler (1994) differentiated what he called two cultures of thinking: exploratory data analysis and probabilistic thinking. Unlike probabilistic thinking, which requires ruling out causal analysis, exploratory data analysis highly values seeking and interpreting connections among events. The inherent conflict between these two ways of

thinking raises the question, to borrow Biehler's words, "do we need a probabilistic revolution after we have taught data analysis?" (Biehler 1994) The term "probabilistic revolution" (Krüger 1987) broadly suggests a shift in world view, in the community of science in between 1800-1930, from a deterministic reality, where everything in the world is connected by necessity in the form of cause-effect, to one in which uncertainty and probability have become central and indispensable. While some researchers (Fischbein 1975; Moore 1990; Metz 1998; Falk and Konold 1999) claim that in learning probability, students must undergo a similar revolution in their thinking, Biehler (1994) argued for an epistemological irreducibility of chance, instead of an ontological indeterminism that the probabilistic revolution seems to suggest (e.g. quantum mechanics as the epitome of an inherently non-deterministic view of natural phenomena). He says,

> …this ontological indeterminism, the concept of the irreducibility of chance is a much stronger attitude… the essence of the probabilistic revolution was the recognition that in several cases probability models are useful types of models that represent kinds of knowledge that would still be useful *even when further previously hidden variables were known and insights about causal mechanisms are possible*. (*ibid,* p. 4, italics in original)

This point of view concurs with Konold (1989) who suggested that a probability approach adopts a "black-box" model, which ignores, if not denies, the underlying causal mechanism. A basic metaphor taken by the 19th century statisticians appeared to suggest the possibility of co-existence of causal analysis and probabilistic reasoning. Such is the idea of system of constant and variable causes that influence an event. The law of large numbers holds if the variable causes cancel each other out and the effect of the "constant" causes reveals itself only with large numbers (Biehler 1994). Biehler further suggested that the ontological debate of whether something is deterministic or not may not be

useful, rather, a situation can be described with deterministic and with probabilistic models and one has to decide what will be more adequate for a certain purpose.

In summary, an epistemological world view that one embraces as a general principle in guiding one's perception and actions is thought of as having to do with his or her development of probabilistic reasoning. One who regards the world as being intrinsically deterministic may naturally seek recourse in causal analysis when judging probabilities. Whereas one who views the world as being irreducibly non-deterministic will seek models and modeling in achieving maximum information on uncertain events. Yet, so far research disagreed on whether a change of deterministic world view is necessary in learning probability. On one side of the debate, probabilistic reasoning is considered to presuppose a "probabilistic revolution" in people's mind (Fischbein 1975; Moore 1990; Metz 1998; Falk and Konold 1999). On the other side, probabilistic reasoning does not necessarily conflict with a deterministic world view (Konold 1989; Biehler 1994). One can view the world as being cause-effect connected, yet intentionally ignore seeking causal factors. In such case, probability is considered as a model that one chooses over a certain situation, in approximating phenomena and quantifying information. However, research agrees on the fact that students having a deterministic view tend more to have a non-stochastic conception of events, e.g. thinking of events as being single and unique, as opposed to thinking of an event as being one of a class of similar events.

Proportional reasoning

Early developmental studies (Piaget and Inhelder 1975; Green 1979; 1983; 1987; 1989) have demonstrated that weak understanding of fraction and proportional reasoning

imposed limitations on children's ability to make probabilistic judgments, and that the concept of proportionality or ratio is prerequisite to an understanding of probability. One can infer that if a child understands probability, he must also understand the concept of proportionality. Fischbein and Gazit (1984) found evidence against this argument. They claimed that although probabilistic thinking and proportional reasoning share the same root, which they call the intuition of relative frequency, they are based on two distinct mental schemata, and progress obtained in one direction does not imply an improvement in the other. Fischbein and Gazit acknowledged, however, that probability computations may require ratio comparisons and calculation, and that it is the probability as a *specific mental attitude* that does not imply a formal understanding of proportion concepts. In this regard, Fischbein and Gazit's argument did not contradict Piaget and Inhelder and Green's thesis and the apparent conflict only resulted from the different uses of the term probability. Recent studies (Garfield and Ahlgren 1988; Ritson 1998) agreed that the ability to engage in probabilistic reasoning is highly dependent on the ability to think about fractional quantities and to think about ratios and proportions. Reciprocally, one may also use probability instruction as a context to teach the concept of fraction and ratio (Ritson 1998).

Evolutionary psychologists, such as Gigerenzer and his colleagues, have a different point of view. Gigerenzer (1998) argued that, "relative frequencies, probabilities, and percentages are to human reasoning algorithms like sodium vapor lamps to human color-constancy algorithms" (*ibid*, p. 13). In other words, Gigerenzer proposed that the natural and original way human reason about numerical information of uncertain situations is to use a format of natural frequencies (For example, saying an

event happens 3 out of 10 times is a format of natural frequency; saying an event happens 30% of the times is a format of relative frequency). When this reasoning system enters an environment in which statistics information is formatted in terms of proportion or relative frequencies, the reasoning will fail. He then argued that people are more likely to make probability judgments when the information is presented in natural frequencies than in a probability format. A second justification of Gigerenzer's proposal of natural frequency is what he suggested as a correspondence between representations of information and different meanings of probability. Gigerenzer (1994) considered single-event probabilities and frequencies to be two different representations of probability information. By framing probability information in the format of natural frequency, one avoids the confusion brought by multiple meanings of probability. Gigerenzer (1994) and Hertwig and Gigerenzer (Hertwig and Gigerenzer 1999) showed that students ceased to employ judgment-heuristics when they were engaged in activities that are formatted in natural frequencies. Sedlmeier (1999)demonstrated that natural frequencies were proven effective in training people how to make probability judgment and Bayesian inferences. Gigerenzer's suggestion of replacing relative frequency by natural frequency appeared to be an effort to eliminate the concept of proportionality. I argue that this is in fact a failed attempt. First, Gigerenzer (1998) suggested that a natural method of teaching is "to instruct people how to represent probability information in natural frequencies" (*ibid*, p. 25). However, such ability ostensibly entails a fractional understanding, as most probability information people encounter in their everyday lives is expressed as fractions, ratios, or percentages. Second, to deduce a probability judgment from information that is presented in natural frequency format, once again one needs to understand the

proportional relationship of the quantities that are involved. Consider the example given by Gigerenzer (1998):

> A scenario in probability format: The probability that a person has colon cancer is 0.3%. If a person has colon cancer, the probability that the test is positive is 50%; If a person does not have colon cancer, the probability that the test is positive is 3%. What is the probability that a person who tests positive actually has colon cancer?
> In natural frequency format: 30 out of every 10,000 people have colon cancer. Of these 30 people with colon cancer, 15 will test positive. Of these remaining 9,970 people without colon cancer, 300 will still test positive. Imagine a group of people who test positive. How many of these will actually have colon cancer? (*ibid*, p. 17)

To solve the problem in probability format, one uses Bayes' rule: (0.3%x50%)/(0.3%x50%+99.7%x3%). In the natural frequency format, 15/(300+15) will suffice. However, only one who has in mind that the quantitative information remains proportional will succeed in justifying why this is the case for any number of people one might choose. In fact, a strong case can be made that for one to choose this method spontaneously, it is with the felt assurance that the method will provide the same answer regardless of the number one picks.

In sum, proportional reasoning and its related conceptual operations appear to support probabilistic judgment once students conceptualize a probabilistic event as being a particular case in reference to a class of events. Indeed, a numerical probabilistic judgment is essentially a fraction, or a ratio. Yet, understanding the concepts of fraction, ratio, and percentages is no less complicated than understanding probability. An elaboration of fractional/proportional reasoning is beyond the scope of this paper. There has been extensive research in mathematics education on the development of students' understanding of fraction, ratio, and other conceptions involving relative comparison of quantities (Mack 1990; Behr, Harel et al. 1992; Kieren 1992; Thompson and Thompson

1992; Kieren 1993; Steffe 1993; Harel and Confrey 1994; Thompson and Thompson 1994; Pitkethly and Hunting 1996; Thompson and Saldanha 2002), but the role of proportional reasoning in the learning and teaching of probability has not been extensively researched.

Stochastic and non-stochastic conceptions

A *non-stochastic* conception of probability expresses itself when one imagines an event as unrepeatable or never to be repeated, whereas a *stochastic* conception of probability expresses itself when one conceives of an event as an expression of an underlying repeatable process (Thompson and Liu 2002). A non-stochastic conception disallows one to make sense of common probabilistic statements (e.g., "What is the probability it will rain on February 4, 2055?"). It reduces an event to the conceptual equivalent of a Bernoulli trial – the event will happen or it will not, and thus logically having a probability of 1 or 0 ("On February 4, 2055 it is either going to rain, or it is not"). Moreover, it leads people to act incoherently, e.g. talking about the chance of an event being x ($0<x<1$) while having in mind that it is never to be repeated. Analyses of school and college statistics texts found little attention being given to the problem of students, when asked to reason probabilistically, not conceiving events stochastically or of them not thinking that a specific event is but one outcome of some repeatable process (Thompson and Liu 2002). Textbook authors exhibit tendencies to state a probability question as if it were about a specific outcome, which impedes the recognition that it signifies an underlying stochastic process' long-term behavior.

The distinction between non-stochastic and stochastic conception ties closely with Kahneman & Tversky's (1982) *singular and distribution modes* of thinking, and

Konold's (1989) *outcome approach* to answering probabilistic questions. What distinguishes a stochastic conception from the latter is its basis in an image of a process that generates outcomes, instead of an image of a given class of elementary events as references for measuring probability (Von Mises 1957; de Finetti 1974). Kahneman and Tversky (1982) characterized this distinction as two different modes of judgment people adopt in attributing external uncertainty – 1) a distributional mode, where the case in question is seen as an instance of a class of similar cases, for which the relative frequencies of outcomes are known or can be estimated; 2) a singular mode, in which probabilities are assessed by the propensities of the particular case at hand. Mathematicians disagreed on which of these two modes are correct ways of evaluating probability, as I have discussed in the last section. While acknowledging that many questions can be approached in either singular or distributional mode, Kahnman and Tversky (1982) conjectured that people generally prefer the singular mode, in which they take an "inside view" of the causal system that most immediately produces the outcome, over an "outside view", which relates the case at hand to a sampling scheme.

The distinction between singular and distributional modes of reasoning is central to our thinking about the development of probabilistic reasoning. A singular mode and a distributional mode correspond to many of the issues addressed above. A singular mode might be a result of a deterministic worldview, or a result of an outcome approach. A student having a singular mode is likely to view an event as a single event. Consequently, he/she is inclined to interpret the task of probability judgment as predicting the outcome of this single event (outcome approach), is also inclined to look for causal factors in searching for justifications for his/her judgment (causal analysis), and is prone to use

judgment-heuristics in making probability evaluations. By contrast, a student adopting a distributional mode views an event as one of a kind, in other words, a special case of a class consisting of similar cases. As a consequence of such conception, it is possible for him/her to ignore the causal mechanism and instead, to adopt a frequency approach in evaluating probability. The question then becomes "how can we help students to adopt a distributional mode, while abandoning a singular mode?" In other words, "how might one conceive of an event as being a particular case of a given class?"

Thinking of an event as one case in a class of cases is reminiscent of von Mises' idea of "collective" (Von Mises 1957 p. 18). In von Mises' theory of probability, a "collective" is given as a priori, and is imposed on the probability model. This approach invited controversies on the justification of collectives, and the existence of limit, etc. Ayer (1972) demonstrated that there is arbitrariness in choosing the referent systems to evaluate the probability in von Mises' framework. Such arbitrariness poses serious challenges in instructional design if one were to accept the ontological assumptions inherent in von Mises' frequentist theory.

To avoid such difficulty, a particularly promising hypothesis is to shift focus from thinking about" collective" as given to thinking about random processes from which collectives are generated. Thompson and Liu (2002) hypothesized that to develop a stochastic conception one may have to go through a series of conceptions and operations as follows:

- Thinking of a situation (an "event") …
- Seeing it as the expression of some process
- Taking for granted that the process could be repeated under essentially similar conditions
- Taking for granted that the conditions and implementation of the process would differ among repetitions in small, yet perhaps important, ways

- Anticipating that repeating the process would produce a collection of outcomes
- And reciprocally, seeing collections as having been generated by a stochastic process

This hypothetical trajectory brings to the fore a conception of random process, at the joint of which the concepts of events, individual outcomes, and collections of outcomes are interconnected as a coherent scheme of ideas. It is important to note that this series of conceptualizations, were students to pass through it, finesses issues of deterministic versus non-deterministic views of reality. One could still hold a deterministic view of events and still imagine that the repetition of a process happens imperfectly.

Statistical inference

Hypothesis testing is one of the most difficult topics students encounter in an introductory statistics course. However, studies in students' understanding of hypothesis testing are scarce. Textbook authors typically incorporate a multi-step approach to present the logic and implementation of hypothesis testing. These approaches generally include: stating the null and alternative hypothesis, defining the critical value, calculating the test statistics, finding the *p-value*, deciding about the null hypothesis, and interpret the situation (Yates, Moore et al. 1998). Research has found that students have difficulties with almost every step of hypothesis testing. For example, Evangelista and Hemenway (2002) conjectured that students may have difficulty in distinguishing a test of hypothesis situation from other situations such as estimation or finding probabilities. Albert (1995) and Link (2002) found that students have difficulty recognize the population parameter to be tested. For example, they would fail to distinguish a population parameter from a sample statistics. This led to their difficulty to formulate the null and alternative

hypothesis. Albert (1995) argued that the idea of sampling distribution, fundamental to hypothesis testing, is too hard for students to learn. Bady (1979) and Moshman & Thompson (1981) found that people have a strong tendency to test hypothesis[3] by seeking information that would verify the hypothesis, instead of falsifying it.

Numerous studies have argued that the concept of confidence interval is non intuitive (eg. Howson and Urbach 1991; D'Agostini 2003). A common misconception of a 95% confidence interval is that 95% of the times the population parameter will fall within that interval. Fidler and Finch (2000) also found that some students think confidence interval is an estimate of sample mean, e.g. they would say things like, 95% confident that the sample mean will fall within the interval.

While most of these above studies focused on students' misconceptions and difficulties, they often take as unproblematic what it means to understanding hypothesis testing and confidence interval. For example, Hong and O'Neil (1992) studied ways of supporting students in building mental models (Streitz 1988) in hypothesis testing. They described experts' mental model as consisting of 1) formulas for computations, 2) steps to test a hypothesis, and 3) the following diagrammatic representation:

[3] Their studies were conducted in the context of scientific/logical hypothesis testing, instead of statistical hypothesis testing. However, in our teaching experiments with students, we found similar evidences—that students seek information to verify instead of reject the hypotheses.

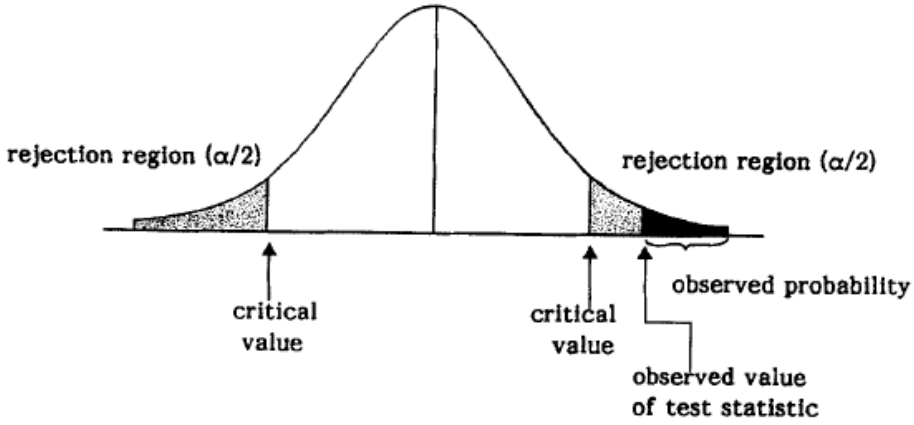
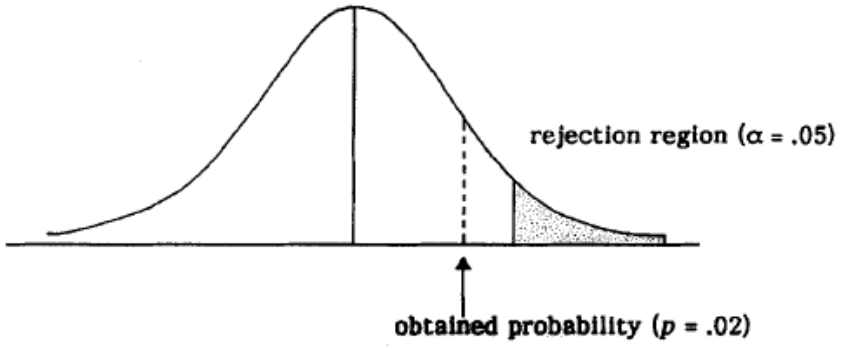


Figure 1. An example of relevant mental models in introductory hypothesis testing.

An intermediates' mental model:



The subject should have rejected the null hypothesis. The subject made a wrong decision because he or she applied the number concept (.02 < .05) on the number line instead of in the area under the curve.

Figure 2. An example of an incorrect mental model in introductory hypothesis testing.

However, these diagrams do not seem to constitute sufficient explanations about 1) what it means to understand hypothesis testing, and 2) what is the qualitative and conceptual difference between experts' and novices' understanding?

Lack of attention on what it means to have a coherent understanding also led researchers to be insensitive to the subtleties of students' thinking. For example, Fidler and Finch (2000) gave an example of an interpretation of confidence interval given by students that they deemed as correct: 95% confident that the population mean would lie

in this interval. However, this interpretation could, at best, leave us undecided about how the students think about confidence interval. They could have meant something entirely different from a valid interpretation of confidence interval, i.e. they could have an image of 95% of all possible population means fall within one particular interval calculated from one sample statistics.

Summary

In this chapter, I reviewed psychological and instructional studies that are pertinent to students' understanding and learning of probability and statistical inference. In summarizing literature on understanding probability, I first presented Piaget's epistemological study on children's learning of chance. His study provided a genetic analysis of how the idea of chance and probability are constructed in combinatoric situations. I then juxtaposed and compared recent studies that attempted to model probabilistic understanding developmentally and conceptually. Following that, I discussed common misconceptions of probability, and factors that contributed to development of probabilistic understanding such as worldview/orientations (towards phenomena, uncertain situations), and proportional reasoning, and stochastic reasoning. This ensemble of studies contribute to our knowledge of 1) how probability is understood or misunderstood, 2) what might have contributed to particular ways of understanding probability, and 3) what are the important ideas that need to be addressed in order to support students' development of probabilistic reasoning.

Literature related to understanding statistical inference is scarce. Systematic investigation on how students understand statistical inference is largely absent. However,

among the reviewed studies, there is a general consensus that the ideas of hypothesis testing and confidence intervals are very hard for student to understand. Students have difficulties comprehending the purpose and logic of hypothesis testing, and they often misinterpret what confidence interval is and what it does.

This synthesis of literature lays the groundwork for my study. First, although the literature does not explicitly explore teachers' understanding, it provides a rich body of knowledge on learners' understanding of probability and statistical inference, which would subsequently afford me the opportunity to make sense of and to be sensitive to various ways teachers might understand these ideas. I do so by assuming that without empirical evidence I have no reason to believe teachers understand these ideas differently from other learners. Second, the synthesis also highlights ideas that are central to understanding probability and statistical inference, as well as difficulties students experience in forming these ideas. These issues contributed significantly to our design of the seminar, especially, ideas that we wanted to raise with the teachers, and ways of supporting reflective conversations surrounding them. Last, a few selected studies also suggest important methodological considerations for the design of the seminar, as well as the retrospective analysis of data. For example, researchers could miss important information about their subjects' understanding as a result of their own lack of differentiation among various ways an idea could be understood. Subsequently, they could miss the opportunities of further probing their subjects' understanding, and misinterpret that understanding in retrospective analysis. What this means for the design of the seminar as well as my analysis is, as researchers, we have to understand the complexities and subtleties in understanding and learning probability and statistical

inference in order to see these issues in teachers' thinking. This understanding affords a vantage point in understanding teachers' knowledge in that it constitutes an initial framework for explaining their understanding. In this proposal, I attempted to broaden this understanding through three channels: 1) the historical and conceptual analysis of probability and statistical inference in the previous chapter, 2) the review of psychological and instructional studies on probability and statistical inference in the current chapter, and 3) the results from our previous teaching experiments with high school students on understanding probability and statistical inference which I will discuss in chapter 5.

This synthesis also points to directions in which my study with teachers will contribute to this body of literature. My observation is consistent with Garfield and Ben-Zvi (2003) in that research on statistics education has attended to neither teachers' understanding of probability and statistics nor their thinking on how to teach these subjects. Thus, this study's potential contribution to the research in statistics education by adding a new dimension is conspicuous. Understanding teachers' conceptual and pedagogical understanding in probability and statistical inference will not only add a new dimension in the body of literature, but also have significant implications in the practices of statistics education as well as teacher education. I will discuss this in detail in later chapters.

CHAPTER III

BACKGROUND THEORIES AND METHODOLOGY

Background Theories

The background theories that underlie this study coordinate a radical constructivist perspective on individual learning with a symbolic interactionist perspective on human communication. Radical constructivism is a theory that describes a particular way of looking at knowledge and knowing (von Glasersfeld 1995; Steffe and Thompson 2000). It emerged from von Glasersfeld's elaboration of Piaget's genetic epistemology (Piaget 1971; Piaget 1977)and the cybernetic concept of viability. Symbolic interactionism highlights the reflexiveness nature of human action (Mead 1910; Blumer 1969) and of mathematics classroom conversation (Bauersfeld 1980; Bauersfeld, Krummheuer et al. 1988).

Radical constructivism

A central assumption in radical constructivism is that individual cognizing agents construct what they know on the basis of their own experiences (von Glasersfeld 1995). The fundamental principles of this epistemological position are:

1. Knowledge is not passively received either through the senses or by way of communication;
   Knowledge is actively built up by the cognizing subject.
2. The function of cognition is adaptive, in the biological sense of the term, tending towards fit or viability;
   Cognition serves to organize one's experiential reality, not to discover ontological reality (*ibid*, p. 51).

In traditional western philosophy, knowledge is seen as a representation of an ontological reality. Both knowledge and reality are considered as independent of the knower. Von Glasersfeld opposed with this realist view of knowledge by proposing knowledge as a conceptual means to make sense of experience. Knowledge is understood as the concepts and relations in terms of which cognizing agents perceive and conceive the experiential world in which they live, and it is generated and proven viable on the basis of their experience (von Glasersfeld 1992) . This view of knowledge and knowing is fundamentally instrumentalist (Dewey 1981). "Cognitive structures—action schemes, concepts, rules, theories, and laws—are evaluated primarily by the criterion of success, and success must be ultimately understood in terms of the organisms' efforts to gain, maintain, and extend its internal equilibrium in the face of perturbation" (von Glasersfeld 1995 p.74).

The purpose of the proposed study is to develop an insight into teachers' personal and pedagogical understanding of probability and statistical inference. What this means within a constructivist framework is to develop viable models of teachers' understanding. By "model" I follow Thompson (1982) to mean "a conceptual system held by the modeler which provides an explanation of the phenomenon of interest", in this case, teachers' understanding. This is different from talking about a system of knowledge that teachers need to have (Shulman 1986; Simon 1995; 2000). If we agree on the assumption that to ensure successful intervention in teacher education we have to start where the teachers are and build from their prior knowledge, then it follows that we must be able to think as if we were the teachers. Building viable models of teachers' understanding of a particular mathematical idea means that we construct a conceptual system that

approximates teachers' conceptual constructs, which allows us to make sense of the coherence or a lack of coherence in teachers' understanding of the idea. The rationale underlying this line of work assumes that, from a constructivist perspective, cognizing agents have no direct access to each other's experiential reality. If this in fact is the case, how do I go about investigating teachers' understanding? Or, what is the rationale of any study that purports to understand others' knowledge? To answer these questions requires that we take the researcher (the modeler, the investigator) into the picture. We have to keep in mind that "the others we experience are the others we construct" (von Glasersfeld 1995 p. 191). If we take this idea seriously, it follows that "whenever they prove incompatible with our model of them, this generates a perturbation of the ideas we used to build up the model. These ideas are *our* ideas, and when they are perturbed by constraints, we may be driven to an accommodation" (von Glasersfeld 1995 p. 191).

Symbolic interactionism

One of the most important implications of constructivism is that "it provides a continual reminder that, regardless of our subjective connection with our personal environment, humans are biological organisms whose only way to exert mutual influence, aside from physical harm or pleasure, is through mutual interpretation" (Thompson 2000 p. 296). This orients us to look for explanations and descriptions of human interactions that highlight the process of mutual adaptations wherein individuals negotiate meanings by continually modifying their interpretations(Bauersfeld 1980; Bauersfeld, Krummheuer et al. 1988). From a symbolic interactionist point of view, humans are pragmatic actors who continually adjust their behavior to their interpretations of the actions of other actors. We

can adjust to these actions because we are able to interpret them (Blumer 1969). This points to a view of communication as reflexively constituted by the participants as they interpret other peoples' expressions with varying degrees of reflectiveness. Drawing from work on symbolic interactionism and information theory (MacKay 1955; 1964; 1965; Bauersfeld 1988; Richards 1991), Thompson and Thompson (1994) described this model of communication as

> Each party in a conversation is living with his or her own world of ideas, but takes into account the sense he or she makes of other people's expressions, and attributes, knowingly or unknowingly, intentions and motivations to others in the conversation. Conversations are quite like non-linear, chaotic systems, in that the possible directions they might take at any moment is a function of the participants' current understandings and intentions, and those understandings and intentions are influenced by the directions taken within the conversation(Thompson and Thompson 1994 p. 2).

Adopting this model of communication has two major implications for the current study. First, it orients me to looking at two ways of interpreting a conversation, as a participant and an observer. People who are conversing with each other can create an illusion of mutual understanding, or at least an impression that they are talking about the same thing (from a participant's perspective), when they don't understand each other fully or in fact have different points of views (from an observer's perspective) (Thompson and Thompson 1994). Occasions of miscommunication of this type can serve my purpose in reveal the underlying assumptions and thinking of the participants. As Dewey (1910) said,

> If two persons can converse intelligently with each other, it is because a common experience supplies a background of mutual understanding upon which their respective remarks are projected. To dig up and to formulate this common background would be imbecile; it is "understood"; that is, it is silently sup-plied and im-plied as the taken-for-granted medium of intelligent exchange of ideas. If, however, the two persons find themselves

> at cross-purposes, it is necessary to dig up and compare the presuppositions, the implied context, on the basis of which each is speaking. The im-plied is made ex-plicit; what was unconsciously assumed is exposed to the light of conscious day (*ibid*, p. 214).

Second, an interactionist perspective highlights people's agency in communication and action in general. Thus, although we attempt to bring out reflective discourse during the seminar discussion, I will not assume that each teacher reflects and reorganizes his/her mathematical understanding in the discourse. Cobb et al. (Cobb, Boufi et al. 1997) investigated the nature of reflective discourse and its relationship to participants' cognitive development. They contended that reflective discourse constitutes conditions for the possibility of mathematical learning, but does not inevitably result in the learning of each participant. This will prevent me from taking for granted that teachers' understanding are in line with the intended instructional goals, or what appears to be collectively understood.

Methodology

Constructivist teaching experiment

In conducting this study, my research team adopted a modification of constructivist teaching experiment methodology (Thompson 1979; Steffe and Richards 1980; Cobb and Steffe 1983; Hunting 1983; Steffe 1991; Steffe and Thompson 2000). The constructivist teaching experiment methodology was adapted from the Soviet-style teaching experiment (Kantowski 1977) to serve the purpose of developing conceptual models of students' mathematical knowledge in the context of mathematics instruction.

As I have mentioned earlier, radical constructivism entails the stance that any cognizing organism builds its own reality out of the items that register against its experiential interface. As such, it is necessary, in a teaching experiment, to attribute mathematical realities to subjects that are independent of the researchers' mathematical realities. While acknowledging the inaccessibility of the subjects environment as seen from their points of view, constructivists also believe that the roots of mathematical knowledge can be found in general coordination of the actor's actions (Piaget 1971). These assumptions then frame the specific research goals of a teaching experiment as being constructing models of subjects' mathematical realities while inducing changes to their understanding and probe their conceptual adaptability and endurance.

A distinguishing characteristic of the constructivist teaching experiment methodology is that the researchers induce changes in the subjects whose knowledge are to be investigated, whether the teaching episodes are realized with the researcher acting as teacher (Cobb and Steffe 1983; Steffe 1991), or in collaboration with teachers (Cobb 2000). The primary purpose of doing so is for the researchers to experience, firsthand, subjects' mathematical learning and reasoning. Without such experiences, there would be no basis for coming to understand the mathematical concepts and operations they construct or even for suspecting that these concepts and operations may be distinctly different from those of researchers/teachers (Steffe and Thompson 2000). In a teaching episode, the teacher-researcher has the initial tasks of 1) interpreting what he or she sees the subjects doing; 2) attempting to perform the act of de-centering by trying to understand the *mathematics of the children [other]* (Steffe 1991). After gaining experiential acquaintance with and making essential distinctions in subjects' ways and

means of operating in domains of mathematical concepts and operations that are of interest, in the next phase, the major goal moves to that of generating and testing research hypotheses. The primary purpose of doing so is to formulate the boundaries of subjects' ways and means of operating by teaching them with the goal of promoting the greatest progress possible in all participating subjects. In the actual instruction, however, the teacher-researcher may be too immersed in interaction to be able to step out of it, reflect on it and take action on that basis. It is, therefore, critical to appeal to an observer of the teaching episode for an alternative interpretation of events, which may help the teacher-researcher both to understand the subjects and to posit further actions (Steffe and Richards 1980; Steffe and Thompson 2000).

As a matter of course, the interactive mathematical communication in a teaching experiment are audio- or video- recorded. Retrospective analysis of the records is a critical part of the methodology. Careful analysis of the audio- or video- tapes offers the researchers the opportunity to activate the records of their past experience with the subjects and to bring them into conscious awareness. It also provides a chance for the researchers to make a novel or an alternative interpretation in terms of their evolving concept of the subjects' mathematics. Through retrospective analysis, the activity of model building is brought to the fore. This effort is equivalent to that of proposing answer to the question "What mental operations must be carried out to see the presented situation in the particular way one is seeing it?"(von Glasersfeld 1995 p. 78) Steffe and Thompson (2000) illustrated the nature of modeling subjects' understanding by comparing it to modeling in science: "As scientists, we want to provide explanations for the phenomena we observe. That is we want to propose conceptual or concrete systems that can be

deemed intentionally isomorphic to the systems that generate the observed phenomena"(Maturana 1978).

Multitiered teaching experiment

Although constructivist teaching experiment methodology was developed in the context of working with students, it describes a general methodology that provides an emphasis on conceptual analysis of mathematical ideas and on researchers constructing models of subjects' mathematical realities as they induce changes to subjects' understanding and probe their conceptual adaptability and endurance. Thus, it also applies to working with teachers. Building on this idea, the design of the teachers seminar draws on the multitiered teaching experiment methodology developed by Lesh and Kelly (2000) which highlights the aspects of investigating teachers' conceptual, psychological, and pedagogical understanding using what one knows about students' learning experiences.

Lesh and Kelly (2000) gave an example of a three tiered teaching experiment: tier 1 aimed at investigating the nature of students' developing knowledge and abilities; tier 2 focus on teachers' developing assumptions about the nature of students' mathematical knowledge and abilities; and tier 3 concentrated on researchers' developing conceptions about the nature of students' and teachers' developing knowledge and abilities. Design of a multitiered teaching experiment was driven by the assumption that "none of these adapting, and self-regulating systems develops in isolation from one another"(*ibid*, p. 209). Consequently, as Lesh and Kelly observed, the kinds of research design that have proven to be most productive for investigating the nature of students and teachers' understanding tend to focus on both individual longitudinal development and their

interactions. For example, one type of three-tiered teaching experiment that they have found to be especially effective in investigating teachers' mathematical and pedagogical understanding is to first engage students in activities designed to reveal the nature of their mathematical understandings, and then engage teachers in activities that involve: (1) designing similar activities to reveal students' understanding; (2) assessing the strengths and weakness of such activities; (3) assessing the strengths and weakness of the results that students produce in response to these activities; (4) making insightful observations of students' engagement in the activities; (5) developing a classification scheme that teachers can use, during students' presentation of their results, to recognize the alternative ways of thinking that students can be expected to use (*ibid*, pp. 216-217).

Didactic objects and didactic models

While constructivist and multitiered teaching experiment methodology provide a rationale that guides to the design and implementation of the teachers seminar, the design of instruction (or specific topics of conversations) was organized around the ideas of didactic objects and didactic models (Thompson 2002). A *didactic object* refers to "'a thing to talk about' that is designed with the intention of supporting reflective mathematical discourse" (*ibid*, p. 198). As Thompson noted, "objects cannot be didactic in and of themselves. Rather, they are didactic because of the conversations that are enabled by someone having conceptualized them as such" (*ibid*, p. 198). Thus, a didactic object is a tool for teachers to engage students in a classroom conversation that purports to support students' learning of a particular idea. To create a didactic object requires teachers having a framework for thinking about the purpose and aims of the didactic

object. This framework is what Thompson called didactic model – "a scheme of meanings, actions, and interpretations that constitute the instructor's or instructional designer's 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. 211). It is important to note that Thompson's theory of didactic objects and didactic models applies to any setting that is designed to elicit reflective discourse. Thus, it applies equally well to settings that involve students and settings that involve teachers.

In the context of our study, we designed "didactic objects" by drawing artifacts and data from previous teaching experiments with students. The rationale for the design of these didactic objects was grounded in our desire to engineer situations that would engage teachers in activities and conversations that will support building psychological models of their understandings. The design of these didactic objects was guided by didactic models and the models of students' understanding constructed from previous teaching experiments and existing literature. Didactic models were our image of a web of conceptual relations that we hope teachers will understand and how that understanding might develop. Models of students' understanding informed us of important or difficult conceptual and pedagogical issues that teachers needed to address in order to support students learning. Together, they guided our decisions in the use of didactic objects and in orchestrating the reflective discourse around them: what conversations to have around the didactic objects, and what issues to raise in those conversation.

CHAPTER IV

RESEARCH DESIGN

This study is part of a NSF sponsored research project *"Investigating the role of multiplicative reasoning in the learning and teaching of stochastic reasoning"* designed and conducted by professor Patrick Thompson and his research team at Vanderbilt University. This project entailed a total of 5 studies conducted over a 40-month period and involved three different groups of participants. This study is the last in this project. The prior four studies are teaching experiments investigating high school students' thinking as they participated in classroom instruction designed to support their learning of sampling, probability, and statistical inference as a scheme of interrelated ideas. The aim was to develop epistemological analyses of these ideas (Glasersfeld 1995; Thompson and 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. Using the products and insights we obtained from these previous teaching experiments (Saldanha and Thompson 2002; Thompson and Liu 2002; Saldanha 2003), we engaged a group of high school teachers in the last experiment in the format of a seminar. Our purpose was to have teachers rethink what they hope students learn from statistics instruction and reflect on ways of affecting students' learning.

Below I first summarize the design and implementation of the seminar. Then I summarize the prior teaching experiments with students and their findings as they bear on

the seminar. Last, I describe the daily activities and themes of the seminar and their purposes.

Design and Implementation

Eight high school mathematics teachers—six female and two male teachers—participated in the seminar. The following table presents demographic information on the eight selected teachers. None of the teachers had extensive coursework in statistics. All had at least a BA in mathematics or mathematics education. Statistics backgrounds varied between self-study (statistics and probability through regression analysis) to an undergraduate sequence in mathematical statistics. Two teachers (Linda and Betty) had experience in statistics applications. Linda taught operations research at a Navy Nuclear Power school and Betty was trained in and taught the Ford Motor Company FAMS statistical quality control high school curriculum.

Table 1. Demographic information on seminar participants.

Teacher	**Years Teaching**	**Degree**	**Stat Background**	**Taught**
John	3	MS Applied Math	2 courses math stat	AP Calc, AP Stat
Nicole	24	MAT Math	Regression anal (self study)	AP Calc, Units in stat
Sarah	28	BA Math Ed	Ed research, test & measure	Pre-calc, Units in stat
Betty	9	BA Math Ed	Ed research, FAMS training	Alg 2, Prob & Stat
Lucy	2	BA Math, BA Ed	Intro stat, AP stat training	Alg 2, Units in stat
Linda	9	MS Math	2 courses math stat	Calc, Units in stat
Henry	7	BS Math Ed, M.Ed.	1 course stat, AP stat training	AP Calc, AP Stat
Alice	21	BA Math	1 sem math stat, bus stat	Calc hon, Units in stat

The research team consisted of the PI, a collaborating school math teacher (Terry), and three graduate students. The team designed the seminar activities and artifacts during the year prior to the seminar. We implemented the seminar using the

constructivist teaching experiment methodology (Thompson 1979; Cobb 2000; Steffe and Thompson 2000). The collaborating teacher hosted most of the seminar activities and conversations. The PI served as an observer of the seminar and occasionally hosted the seminar or participated in the conversation. One graduate student took field notes and managed miscellaneous logistic work. Another graduate student and I recorded the seminar sessions with front and back cameras and made observations and notes during the seminar.

The seminar discussion progressed over eight sessions in two weeks. The seminar began at 9am each day and concluded at 3pm each day, with a 30- minute lunch break. At the end of each day, the research team met briefly to discuss our observations and suggestions on modification of next days' activities. At the end of the seminar sessions, we also made photocopies of teachers' notes. Each teacher was interviewed 3 times for about 45 to 60 minutes each time. Interviews were conducted once before the seminar and once at the end of each week. We video recorded all interviews and kept record of teachers' work during the interviews.

As I will summarize later, in prior teaching experiments with students, the research team explored what it means for students to create coherent understanding of probability and statistical inference and the conceptual obstacles they meet in doing so (Saldanha and Thompson 2002; Thompson and Liu 2002; Saldanha 2003).Video data and students' work from these studies were employed as points of discussion with the teachers to support our attempt to have them become aware of and refine their understanding of ideas, objectives, and practices in teaching probability and statistics. This was guided by the multi-tiered teaching experiment methodology (Lesh and Kelly

1996). In each session, the teachers engaged in activities and discussions in which they continually reflected, critiqued, and refined their understanding of probability and statistical inference, their understanding of students' conceptions and learning difficulties, and their ideas of teaching probability and statistics. The type of activities the teachers engaged in included: working on conceptually challenging and pedagogically problematic tasks as first-order participants; watching videotapes of students engaged in classroom discussions or in problem solving activities and examining students' work chosen deliberately by the research team that highlighted particular problematic aspects of students' conceptions; and analyzing textbook excerpts and proposing suggestions on refinement of ideas, reading and discussing scholarly writing on probability and statistics teaching and learning.

We engineered the discussions so that teachers first worked on and discussed the problems as first-order participants. We used these occasions to construct models of teachers' personal understanding of the ideas of probability and statistical inference. We then initiated pedagogical conversations about these ideas—given these ways of understanding these ideas, what are the implications for teaching them? We intended to elicit reflective conversations in the sense that what was previously discussed became objects of thoughts and conversation (Cobb, Boufi et al. 1997). The interviews were designed to include general questions concerning teachers' stochastic reasoning, as well as specific questions that were tailored to each teacher according to our observation and conjectures about his or her knowledge and beliefs.

Background

Teaching experiment one (TE1)

Twenty-seven students from Grades 11 to 12 participated in TE1 conducted in a non-AP semester-long statistics course during winter 1999 at a suburban high school in the Southeastern U.S. TE1 focused on ideas of sample, inference, sampling distributions, margins of error, and interrelations among them. It stressed two overarching and related themes: 1) the process of randomly selecting samples from a population can be repeated under similar conditions, and 2) judgments about a sample's outcome can be made on the basis of relative frequency patterns that emerge in collections of outcomes of similar samples. These themes were intended to support students' developing a distributional interpretation of sampling and likelihood (Von Mises 1957; Kahneman and Tversky 1982; Konold 1989).

TE1 progressed over 9-consecutive lessons and unfolded in three interrelated phases. It began with directed discussions centered on news reports of data about sampled populations and news reports about populations, raising the issue of sampling variability. It then progressed to activities that led to questions of "what fraction of the time would you expect a sample result 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, but are affected significantly by sample size.

A preliminary report of the teaching experiment (Saldanha and 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*, 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 variability among sampling outcomes that supports reasoning about distributions of outcomes.

TE1 also found that students had difficulty differentiating hypothesis testing and parameter estimation. For example, when the instruction had students use a computer simulation to test a null hypothesis regarding a population parameter, students believed that the purpose of the computer simulation was to find out the population parameter and thus did not understand the reason for assuming a population parameter (null hypothesis). Students also seemed to take the meaning of probability as unproblematic. They would answer the question "what is the probability of event A?" without being able to articulate what the question meant. Thus, they found questions such as, "What does the statement, 'The probability of [some event] is 0.37.' mean?" to be nonsensical.

Teaching experiment two (TE2)

The second teaching experiment was conducted in fall 1999 within a yearlong non-AP statistics and probability course given at the same high school in which the first experiment was conducted. Eight liberal-arts-bound students in Grades 10 through 12 participated in the teaching experiment. TE2 addressed the same ideas and used a similar instructional approach as that of TE1. In addition, the point of departure for TE2 was shaped by conjectures that the research team formulated about students' difficulties from the result of TE1.

The teaching experiment unfolded in a sequence of 17 consecutive classroom lessons over a period of 28 days. Instruction started by engaging students in a concrete sampling activity, a central aim of which was to provide them with an experiential basis for understanding the simulation-based sampling explorations that came thereafter. It then engaged students in designing simulations of repeated sampling experiments as a method for investigating whether an event can be considered statistically unusual. Next, students examined 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 variability as related to the ideas of distribution. Finally, the instruction moved 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 intervals around the population parameter's value (Saldanha 2003).

Saldanha (2003) reported that an overarching and salient finding of TE2 was that students experienced significant difficulties coordinating and composing multiple objects

(e.g., individual, sample, population, and measure of each) and actions (e.g., looking at collection of individuals, or collection of samples) 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 coordination and compositions are non-trivial.

Teaching experiment three (TE3)

TE3 was unrelated to the issues of probability and statistical inference. It focused on issues of data analysis and the use of multiple regression in generating predictive relationship from a subset of a population to the population itself.

Teaching experiment four (TE4)

TE4 focused on the idea of probability. It was conducted during spring semester, 2000. The study involved eight junior and senior high school students enrolled in a non-AP, yearlong statistics course. This classroom instruction progressed over 25 consecutive sessions within the course of 6 weeks.

The classroom instruction began with lessons that were designed with two aims. The first was to have students understand that a specific event can be considered as an outcome of some repeatable process. The second was to have them understand that saying an event's probability being *x* indicated "an *expectation* that the process producing this event will end with an outcome like this one 100*x* percent of the time as we perform the process repeatedly." The instruction also intended that students be able to see specific

events in two ways, stochastically and non-stochastically. The goal was that they be able to explain how conceiving of the situation as a stochastic process supported reasoning probabilistically about it, while conceiving it as a "one-shot deal" made the situation non-probabilistic. For example, they could interpret "What is the probability that George Bush's Texas house is white?" stochastically or non-stochastically. A stochastic interpretation would be like, "Pick a Texan at random (and imagine that this will be repeated a large number of times). What fraction of the time will you pick someone living in a white home?" Interpreted non-stochastically, George Bush's home in Texas is either white or it is not. If white, the probability that it is white is one. If not white, the probability that it is white is zero.

In the next phase, instructional focus was placed on conditional probability with the aim of having students investigate the relationship between probability and sampling. Specifically, students were directed to look at contingency tables from a sampling perspective and to make connections between long-term behavior and sample space. Activities engaged students in partitioning a population into sample spaces and anticipating the expected outcome if they were to sample many times from various subparts of the population.

In the final phase, the symbolic representations of probabilities were introduced and students made connections between notational representations of probabilities and the meanings they express by folding back to the ideas in the previous phases. For example, P[X]—the probability of event X happening—was to be interpreted as “the fraction of the time that we expect an event X to occur in a long series of trails”. P[A and B] and P[A/B] were to be differentiated in terms of the collectives that the underlying

stochastic processes apply in generating the expected outcomes. The notations of probability was placed at the end of the instruction sequence because the research team believed that, an early focus on symbolic operation of probability, as in the traditional probability instruction, would divert students' attention from building a conceptual image of probability. While if students develop a concept image of probability as essentially "the relative frequency of an expected outcome of a stochastic process over the long run", they would be less inclined to think they can answer the questions by looking at the superficial aspects of the scenarios.

TE4 found that students displayed a strong tendency to interpret situations non-stochastically until instruction raised the issue explicitly. The relative frequency of such interpretations (relative to opportunities for such) dropped steadily over the first five lessons, with the caveat that as situations became more complex they often required explicit conversation to orient students to reinterpret them. The teaching experiment's early focus on conceiving situations stochastically had a salutary effect on students' abilities to control their interpretations. All but one student showed the ability to interpret events in either way and all but two eventually came to interpret situations stochastically spontaneously.

The teaching experiment's early focus on interpreting situations stochastically had another interesting effect. It helped clarify complexities inherent to probabilistic reasoning per se. It did this by removing the confusions introduced by students failing to interpret situations stochastically when asked probabilistic questions about them. Complexities revealed in instructional conversations and interviews, included:

1. Students' difficulties imagining that a stochastic process produces a collective (in the sense of Von Mises 1957), When present, this itself had two consequences:
    a. it obstructed students' ability to reason about probabilities as if they were fractions of a population.
    b. it obstructed their abilities to connect ideas of expected long-term behavior with ideas of sample space.
2. Students' difficulties in reasoning proportionally. When present, this revealed itself in their not drawing connections between population frequencies and events' relative weights.
3. Students' difficulties envisioning complex, multi-leveled processes (e.g., those leading to contingency tables). This revealed itself in students loosing track of what they were talking about — an outcome corresponding to a cell, an outcome corresponding to a margin, or an outcome corresponding to a cell relative to a margin.
4. Difficulties that emerged because of students' focusing first on the question being asked instead of on the situation that gave rise to the question.

Summary of Seminar Activities and Interviews

The following table encapsulates the sequence of activities and interview questions we designed for the teacher seminar.

Table 2: Overview of seminar activities and interview questions

Events	Abbrevi-ation[4]	Date (2000)	Day of seminar	Activity title	Duration (minutes)
Orientation		5/13		Orientation	
Pre-Interview	I1-1	5/29		General questions	
Pre-Interview	I1-2	5/29		Variability of investment	
Pre-Interview	I1-3	5/29		Interpreting histogram	
Pre-Interview	I1-4	5/29		Accuracy of measurements	
Pre-Interview	I1-5	5/29		Sampling distribution	
Pre-Interview	I1-6	5/29		Interpreting statements	
Pre-Interview	I1-7	5/29		Law of large numbers	
Week One	A1-1	6/11	1	Data, sample, and polls	160
Week One	A1-2	6/11	1	Chance and Likelihood	29
Week One	A1-3	6/11	1&2	Pepsi	180
Week One	A1-4	6/12	2	Hand-sampling	46
Week One	A1-5	6/12	2	Jelly Beans	84
Week One	A1-6	6/13	3	Movie theatre	106
Week One	A1-7	6/13	3	Fathom investigation	104
Week One	A1-8	6/14	4	Stan's interpretation	120
Week One	A1-9	6/14	4	Musician	67
Week One	A1-10	6/14	4	Textbook analysis	95
Mid-Interview	I2-1	6/15		Alumni association	
Mid-Interview	I2-2	6/15		Harris poll	
Mid-Interview	I2-3	6/15		Horness scale	
Mid-Interview	I2-4	6/15		Purpose of simulation	
Mid-Interview	I2-5	6/15		Fundamental idea	
Week Two	A2-1	6/18	5	Textbook analysis	110
Week Two	A2-2	6/18	5	PowerPoint presentation	67
Week Two	A2-3	6/18	5	Rodney King	104
Week Two	A2-4	6/19	6	Clown & Cards	138
Week Two	A2-5	6/19	6	Vanderbilt population	125
Week Two	A2-6	6/20	7	US Census	165
Week Two	A2-7	6/20	7	Drug testing	115
Week Two	A2-8	6/21	8	Data analysis	130
Post-Interview	I3-1	5/22		Five probability situations	
Post-Interview	I3-2	5/22		Three Prisoners	
Post-Interview	I3-3	5/22		Blue Cab	
Post-Interview	I3-4	5/22		Gambling	
Post-Interview	I3-5	5/22		Drug testing	
Post-Interview	I3-6	5/22		Vanderbilt population	

Below I will provide a chronological account of these activities and their rationales.

[4] Abbreviations: Pre-Interview is abbreviated as Interview 1 or I1, Mid-Interview I2, Post-Interview I3. I2-3 means "Mid-Interview question number 2". A2-5 means "Week 2, Activity number 5".

Orientation meeting

We held an orientation meeting & information session for the participants approximately a month prior to the seminar. During this meeting, we briefed the participants on seminar and handed out a rough calendar of seminar sequence (Table 3).

Table 3: Calendar of the seminar given to the teachers prior to the seminar

Week	Monday	Tuesday	Wednesday	Thursday	Friday
June 11- June 15	Data, samples, and polls Chance & Likelihood	Likelihood Hypothesis testing Putting it all together	Statistical unusualness Distribution of sample statistics Margin of error	Distribution of sample statistics Margin of error Confidence interval	Mid-interview
June 18- June 22	Probabilistic situations Stochastic conception of sample	Single outcome vs. long-run behavior Conditional probability	Formalization of probability Conditional probability	Exploratory data analysis	Post-interview

We gave the teachers nine articles related to understanding and teaching probability and statistics (Shaughnessy 1993; Thompson, Philipp et al. 1994; Konold 1994b; Konold 1995; Newport, Saad et al. 1997; Cortina, Saldanha et al. 1999; Thompson and Saldanha 2000; Best 2001; Saldanha and Thompson 2001), some of which we anticipated to discuss during the seminar. We also scheduled pre-interview with the teachers. Teachers were told to read a textbook excerpt from Moore's (1995) *Basic practice of Statistics* which the interview would center on. The excerpt was the section 4.5 *sample means* from Chapter 4: *sampling distribution and probability* (pp. 292-303), focusing on the idea of sampling distribution, central limit theorem, and the law of large number.

Pre-Interview

Pre-Interviews were conducted about two weeks prior to the seminar. The purpose of the pre-interviews was to develop a sense of teachers' understandings of sampling as a stochastic process and of sampling variability. We asked the teachers two types of questions regarding the textbook excerpt, first, general questions concerning their impressions and understanding of the excerpts, what they thoughts are the important ideas and problematic parts for students, and second, questions that ask for teachers' interpretations of selected statements about the ideas of variability, sampling distribution, probability, law of large number.

Week one

We started the seminar by having the teachers discuss the ideas of sample, population and random sampling (A1-1). We intended that the discussion center on 1) descriptive statistics, inferential statistics, and the difference between them; 2) random sampling: how is a sample selected, and how does the selection affect how well a sample represents its population? 3) What is the implied population a poll tries to represent? The purpose of activity was for the teachers to clarify the relationships between sample and population, the distinction factual data and sample estimate, all of which are important for understanding statistical inference.

In the next activity, we asked teachers to interpret *the meaning* of two probability statements[5] (A1-2):

[5] In this paper, the phrase "probability statement" means a statement that involves probability where the word "probability" conveys only the linguistic aspect of the statement, and its meaning is loosely defined and subject to many interpretations. For example, a situation that involves chance, risks, etc., would be called a probability situation. The word "probabilistic" has a

1. Today there is a 45% chance of rain.
2. A pollster asked 30 people about which they liked better, Pepsi or Coca-Cola. 18 people said Pepsi. How likely is this result?

We asked the question "What does the statement/question mean?" This question was intended to reveal the teachers' informal conceptions of probability. Although probability is the focus of the second week, we brought it up here because it is a fundamental idea in statistical inference.

Next, we then gave teachers a follow-up activity on the Pepsi poll of the second question (A1-3). This activity was designed as a structured investigation of the question "how likely is this result?" and as a means of helping students make sense of the idea of statistical likelihood and of how the idea relates to making statistical claims and inferences. By engaging the teachers in this activity, we wanted to understand the extent to which the teachers understood and employed the method of hypothesis testing, and the kinds of potential difficulties they might experience in trying to do so.

On the second day, in A1-4 we gave teachers a handout of a sampling activity that was designed with high school students in mind. The activity asked students to sample various kinds of objects by hand from populations whose compositions are unknown to them. We asked the teachers to examine the activity and discuss its possible instructional intent and merits or shortcomings. We then, in A1-5, presented the teachers with a homework assignment given to high school students after they had participated in the hand sampling activity, and samples of students work. The students work documented how students, in going through hand sampling and subsequent investigation, observed the variations among samples, and learned to make claims about population compositions

---

cognitive meaning. When a probability situation is conceived of stochastically by a person A, then to A this situation is probabilistic or stochastic.

from collections of samples, instead of individual samples. We asked teachers to raise any issues they had about the students' thinking and to discussion further the instructional intent and merits of engaging the students in sampling activities. At the end of second day, we asked the teachers to reflect upon their discussion during the past two days. We had them summarize and connect the big ideas that had been discussed.

On the third day, we turned to the idea of the statistical unusualness (A1-6). In statistics an event is said to be "unusual" if over the long run we expect to see it a small fraction of the time. This is a stochastic conception of unusualness and it supports thinking about statistical inference. We observed that when asked to investigate the unusualness of a probability event, students often take an outcome approach, i.e., thinking that they are predicting whether the event will occur next time. We engaged the teachers in this activity to learn about the ways in which teachers understand the concept of unusualness.

The rest of the first week focused on the concepts of variability and margin of error. We first had teachers work on generating distributions of sample statistics using a statistics program Fathom (A1-7). The teachers repeatedly drew many random samples of the different sizes from a population, and then explore how such collections of samples are distributed, and particularly, how the distributions vary in relation to the sample size. The intent of the activity was to have the teachers explore the relationship between sample size and variability.

Next, in A1-8, we had teachers investigate the relationship among variability, population parameter, and the number of samples. We did so by fixing sample size and making the population parameter and the number of samples vary. We then had teachers

critique a common misunderstanding of margin of error, followed by a discussion on margin of error and confidence interval as two ways of expressing the same idea.

The next activity (A1-9) centers on the relationship between sample size and variability amongst samples. We presented the teachers the results of computer simulations of various samples of different size taken from one population. We asked the question, "how *accurately* samples of various sizes reflect the population's composition", or "how large does a randomly chosen sample have to be so that we feel assured it is a fair *representation* of the population?"

Finally, we engaged the teachers in a textbook analysis (A1-10), in which we asked the teachers to interpret the given definition of sampling distribution and to discuss what they thought students would need to know in order to understand it. Our purpose was to evaluate teachers' conceptual coherency of the idea of sampling distribution and their pedagogical knowledge on what it takes for one to understand this idea.

Mid-Interview

The Mid-Interview questions were designed on the basis of our observation of salient issues emerged in the teachers' participation of the first week's activities. The questions that we designed further probed the ideas that we found problematic for the teachers. These ideas included statistical inference, margin of error, distribution of sample statistics, and the purpose of computer simulations, etc.

Week two

The second week focused on the idea of probability. We started by engaging the teachers in a critique of a textbook excerpt on probability with the intention that the discussion would reveal teachers' understanding of probability and how they thought it should be taught (A2-1). We then presented a series of probability situations in the format of a PowerPoint presentation (A2-2) and asked the teachers to interpret their meanings. Our purpose was to explore the extent to which the teachers interpreted these situations stochastically. We orchestrated the conversation so that two big ideas would emerge: 1) probability is not about any one particular instance of an event or outcome, and 2) A situation as it states might not determine whether or not it is stochastic. Rather, it is the way of *conceiving of* the situations that makes stochastic. In the seminar, a probability situation that is conceived of stochastically is referred to as *probabilistic situation.*

Next, we introduced the Rodney King scenario (A2-3) in which the point of discussion was about whether the dispatch of 15 all-white police officers in the Rodney King event was a random occurrence. There are two important issues that we intended to probe. First, will the teachers conceive the situation stochastically, i.e. will they conceive of a random deployment process of which this police dispatch was a single case? What decision rule will they design to determine whether an event is a random occurrence?

On the second day of the second week, we engaged the teachers in the discussion of Clown and Cards scenario (A2-4). The big idea that we wished to highlight in this activity was that "A situation per se is not probabilistic. It is how you conceive the situation that makes it probabilistic or not." A situation can be conceived in different ways, stochastically or non-stochastically. Furthermore, even when interpreted

stochastically, there could still be more than one interpretation. The Clown and Cards scenario presented such a situation. We anticipated that teachers would have split opinions on ways of interpreting this scenario. If not, we would bring out an alternative interpretation so as to explore how the teachers would respond to this type of situation. In addition, we also wanted to inquire about teachers' pedagogical decision when in their classroom students have multiple and conflicting interpretations.

Our next activity Vanderbilt population (A2-5) intended to broach the idea of conditional events & probability through the use of contingency tables. At the onset, the term "conditional probability" was not mentioned. We wanted to see what teachers made of this activity/question and whether they recognized it as addressing conditional events. After teachers' own ideas had been made evident, we steered the discussions so that it addressed a productive way for students to think about conditional probability: it involves restricting one's attention to a subset of the population and asking what proportion of that subset has some characteristic of interest. This line of reasoning entails compound proportional reasoning: thinking about a fraction of a fraction of a population. It thus involves *structuring* a set of data/sample space/population into distinct parts that can be considered hierarchically. Later, we introduced the issues of notation use in the learning & teaching of probability. We designed a problem context involves contingency tables and conditional events (A2-6). Teachers discussed the issues of decoding and interpreting probabilistic statements/ideas across different representational registers, and developing student understanding of connections between these. We intended that discussions bring out teachers' ideas on the use/purpose of symbols in probability, on the interpretations of

notation that they view as desirable for students to develop, and on the instructional strategies that might support students' development of these meanings & interpretations.

The last activity (A2-7) was designed with the intention of connecting the ideas of probability and sampling distribution. We first presented a probabilistic situation in the context of mammography testing that other studies had found to be very problematic for medical professionals and students. We intended to make evident what teachers experience in trying to make sense of this scenario, e.g., ways of reasoning they employ and difficulties they encounter. Then we shared with teachers a sampling activity designed to help students make sense of information like that in mammogram scenario. We employed a different situation and used sampling to have students conceptualize the population and its composition (its sub-populations and their relative proportions). However, we did not reveal this intent to teachers directly, hoping that the discussion would reveal their own ideas about its intent and usefulness. Finally, we solicited reactions from teachers as to 1) the relevance & importance of the scenario, 2) the need for instruction that aims to help students develop coherent understandings of such scenarios. In sum, we hoped to find out whether teachers see that a lack of sound probabilistic reasoning can have real and potentially serious ramifications, and that sampling activities can help students develop coherent understandings of probability that will support their making sound decisions in critical situations?

Post-Interview

Post-Interview questions were designed to further probe teachers' conceptions of probability, and the extent to which they could flexibly interpret a probability situation

both stochastically or non-stochastically. We also presented hypothetical instructional situations and asked what teachers might react to them. We were interested in finding out what teachers might do when, in their classroom, students were split in their interpretations of probability situations. We also further explored teachers' pedagogical understanding concerning the use of different representational forms of probability.

Data Analysis

The data for analysis include video recordings of all seminar sessions made with two cameras (36.5 hours) and individual interviews (approximately 1x8x3=24 hours), the teachers' written work, and documents made during the planning of the seminar. The analytical approach I will employ in generating descriptions and explanations is consistent with Cobb and Whitenack's (1996) method for conducting longitudinal analyses of qualitative data and Glaser and Strauss' (1967) grounded theory, which highlights an iterative process of generating and modifying hypotheses in light of the data. Analyses generated by iterating this process are aimed to develop increasingly stable and viable hypotheses and models of teachers' understanding. The specific procedure that I adopt in my analysis consists of the following steps.

The first level summary

I begin by first reviewing the entire collection of videotaped seminar sessions and interviews. My primary goal in this process is to develop a rough description of what had transpired in the seminar, and a sense of ways of organizing the data. To this end, in reviewing the tapes, I partition the sessions into video segments: a segment of video is

defined by a) a chronological beginning and end; b) a task/artifact; c) the rationale of task design, or the issues that the task was designed to raise; d) discussion/verbal exchange around the task and the issues that arise. For each session, I write a summary of all video segments chronologically. Descriptions of video segments consist of the above four components. For interview data, I organize the summary around the questions being asked, i.e., for each interview question, I juxtapose teachers' answers and explanations so that it facilitates easy comparison as well as developing an overall sense of teachers' understanding. In general, my interaction with data in this phase can be characterized by my openness in documenting the data. In other words, instead of looking for answers to specific questions from the data, I consistently document what happened without differentiating certain segments from the others in terms of how well they might reveal teachers' understanding. My purpose, again, is to develop an overall sense of what have transpired in the seminar sessions and interviews, and to create an organizational structure and brief summaries that will facilitate further analysis of the data.

The second level summary

In this next phase, I review the videotaped seminar sessions and interviews again. But this time, I start to identify video segments that seem potentially useful for gaining insight into one or more teachers' personal and pedagogical understandings of probability and statistical inference. Segments containing direct evidence of teachers' thinking, miscommunication in discussions of problems or ideas, or controversy about mathematical meanings or pedagogical practices are especially significant. I then enrich the first level summary by providing thick descriptions of these segments. Along with

these descriptions, I also make observations and initial hypotheses of what these segments seem to have revealed about teachers' understanding.

Transcription and transcript analysis

Differentiated video segments in the last phase are then transcribed. A transcriber working with us produces transcripts of the conversations, and I complete the transcripts by including in them all textual information employed during the discussion, e.g. activities handouts, video screen shots, sketches, etc.

The unit of transcript analysis is conversation around activity. For each activity, I will capture its global structure by parsing the conversation into hierarchies of episodes. The first level episode will be defined as conversations around the organizing questions in the activity, and thus can be conveniently named as the organizing question. The second level episodes depict the significant themes of the conversation in the first level episodes. I then take each of these second level episodes as primary unit of annotation—local interpretation.

My primary goal in the annotation is to clarify the meanings of teachers' utterances, i.e., discern from their utterances what they had in mind. For any utterance *x*, the questions I ask include:

- What motivated a teacher to say *x*? What was the point the teacher tried to make?
- How does it build on this teachers' interpretation of conversation preceding *x*?
- How did the other teachers interpret *x*?

The guiding question for making sense of a teacher' utterance is: *What might he have been thinking (or seeing the situation, or interpreting the previous conversation) so that*

what he said made sense to himself? An interpretation of teachers' thinking is a conjecture that subjects to further confirmation, refutation, or modification in light of further evidences. Thus, the viability of the interpretations is achieved by triangulation of evidences across the data.

After annotating an episode, the next step is to synthesize the conjectures developed during the annotation. These conjectures are then subject to constant comparison and modification with those of the rest of the data. More substantiated conjectures will emerge from this process and serve as potential answers to the research questions in this dissertation.

Analyses of analyses

In this phase, the hierarchical structure of conjectures made during transcript analyses becomes data that will be further analyzed (Cobb and Whitenack 1996) and re-organized along three categories of themes: 1) Teachers' conceptions of probability; 2) conceptions of hypothesis testing; 3) understanding of variability and margin of error.

Within each of these categories, there are sub categories of theoretical constructs that describe/capture the teachers' different conceptions or understanding. As the collection of these constructs emerges and their relationship among one another becomes clearer, they form a theoretical framework that will serve as a basis for constructing narratives of teachers' personal and pedagogical understanding of probability and statistical inference.

Narrative construction

In this phase, I will first provide a synthesis of the results from the last phase, i.e., a system of theoretical constructs that are used to describe teachers' understanding of probability and statistical inference (Chapter 5). Next, I will construct narratives of teachers' understanding using the theoretical framework described in Chapter 5. I will first organize the narratives in three chapters.

Chapter 6: Teachers' understanding of probability;

Chapter 7: Teachers' understanding of hypothesis testing;

Chapter 8: Teacher's understanding of variability and margin of error.

OVERVIEW OF CHAPTERS V TO VIII

Chapter 5 presents a conceptual analysis of probability and statistical inference. In it I presented conceptual frameworks that I have developed through the analysis of the data. It is important to understand that these conceptual frameworks emerged out of the analysis, and they constitute the end product of this study. The theoretical constructs and framework that I elaborate in this Chapter are the tools with which I describe teachers' understanding of probability and statistical inference in Chapter 6 to 8. Chapter 6, 7, and 8 each focuses on one or one set of interrelated ideas—Chapter 6 on probability, Chapter 7 on hypothesis testing, and Chapter 8 on variability and margin of error.

For ease of reference, Table 4 highlighted listed the activities and interview questions that will be discussed in the chapters 6 to 8.

The discussion of activities and interview will not follow a chronological order, but will instead be organized around particular themes. For example, I2-4 was designed to further probe a conjecture concerning the purpose of re-sampling simulation that emerged out of the discussion on the Part I of A1-8. Thus, the discussion of I2-4 will follow immediately after that of A1-8 Part I. Each Chapter consists of discussions of a number of activities and interviews that helps to address the research question.

Table 4: Activities and interview questions appeared in later chapters

Events	Abbrevi-ation[6]	Date (2000)	Day of seminar	Activity title	Duration (minutes)
Orientation		5/13		Orientation	
Pre-Interview	I1-1	5/29		General questions	
Pre-Interview	I1-2	5/29		Variability of investment	
Pre-Interview	I1-3	5/29		Interpreting histogram	
Pre-Interview	I1-4	5/29		Accuracy of measurements	
Pre-Interview	I1-5	5/29		Sampling distribution	
Pre-Interview	I1-6	5/29		Interpreting statements	
Pre-Interview	I1-7	5/29		Law of large numbers	
Week One	A1-1	6/11	1	Data, sample, and polls	160
Week One	A1-2	6/11	1	Chance and Likelihood	29
Week One	A1-3	6/11	1&2	Pepsi	180
Week One	A1-4	6/12	2	Hand-sampling	46
Week One	A1-5	6/12	2	Jelly Beans	84
Week One	A1-6	6/13	3	Movie theatre	106
Week One	A1-7	6/13	3	Fathom investigation	104
Week One	A1-8	6/14	4	Stan's interpretation	120
Week One	A1-9	6/14	4	Musician	67
Week One	A1-10	6/14	4	Textbook analysis	95
Mid-Interview	I2-1	6/15		Alumni association	
Mid-Interview	I2-2	6/15		Harris poll	
Mid-Interview	I2-3	6/15		Horness scale	
Mid-Interview	I2-4	6/15		Purpose of simulation	
Mid-Interview	I2-5	6/15		Fundamental idea	
Week Two	A2-1	6/18	5	Textbook analysis	110
Week Two	A2-2	6/18	5	PowerPoint presentation	67
Week Two	A2-3	6/18	5	Rodney King	104
Week Two	A2-4	6/19	6	Clown & Cards	138
Week Two	A2-5	6/19	6	Vanderbilt population	125
Week Two	A2-6	6/20	7	US Census	165
Week Two	A2-7	6/20	7	Drug testing	115
Week Two	A2-8	6/21	8	Data analysis	130
Post-Interview	I3-1	5/22		Five probability situations	
Post-Interview	I3-2	5/22		Three Prisoners	
Post-Interview	I3-3	5/22		Blue Cab	
Post-Interview	I3-4	5/22		Gambling	
Post-Interview	I3-5	5/22		Drug testing	
Post-Interview	I3-6	5/22		Vanderbilt population	

[6] Abbreviations: Pre-Interview is abbreviated as Interview 1 or I1, Mid-Interview I2, Post-Interview I3. I2-3 means "Mid-Interview question number 2". A2-5 means "Week 2, Activity number 5".

CHAPTER V

CONCEPTUAL ANALYSIS OF PROBABILITY AND STATISTICAL INFERENCE: THEORETICAL FRAMEWORKS

This chapter is an explication of theoretical constructs and frameworks for understanding the teachers' understanding of probability and statistical inference. These theoretical frameworks, partially built on existing literature, should be understood as an end product of this study. That is, they emerged and were developed from the analysis of teachers' understanding of probability and statistical inference.

Statistical Inference

Statistical inference is the theory and methods of forming inferences about the parameters of a population on the basis of random sampling. There are two important themes in statistical inference: hypothesis testing and parameter estimation. Hypothesis testing tests the viability of hypotheses about a population parameter. Parameter estimation estimates the population parameter through random sampling and quantifies the error in such estimation.

Understanding statistical inference entails, first of all, understanding the goals of making statistical inferences: Why do we make statistical inference? What work does it do for us? Example:

> Question 1: Are there more non-smokers than smokers in the department of teaching and learning at Vanderbilt University?

To answer this question, we can conduct a survey of all the faculty and staff in the department, and calculate the number of non-smokers and the number of smokers, and compare the size of them. The question can be answer with YES or NO. When answered with YES, it means that we established *a fact with certainty*: there are more non-smokers than smokers in the department of teaching and learning at Vanderbilt University.

Now let's look at the question:

Question 2: Are there more non-smokers than smokers in the US?

To obtain an answer of the same kind of certainty as above, we have to ask every single person in the US whether or not he or she is a smoker. This is theoretically possible yet completely impractical. However, by collecting random sample(s), we can make inferences about whether, at a particular moment in time, there are more non-smokers than smokers in the nation (hypothesis testing), or estimate the percent of smokers in the nation (parameter estimation). Thus, we make statistical inference because *there are important questions to be answered, and statistical inference is the most economical and sometimes the only sensible way to answer these questions*.

Understanding statistical inference entails an understanding that the information we obtain from statistical inference *carry less certainty* than those in descriptive statistics (as in Question 1). A process of a hypothesis testing does not end with a result that can be expressed as *a fact*, but it ends with *a viable hypothesis*. The viability of the hypothesis (i.e. the certainty about the result) is measured in the context of all possible hypotheses about the population parameter (I will elaborate on this later). In a similar vain, the process of parameter estimation does not end with a *population parameter*, but an *estimate of the parameter* and *an estimate of accuracy*. The certainty we have about this

estimate is expressed in terms of measurement errors (I will elaborate on this later). Thus, the *best* answer we can expect from making statistical inference is:

1. Hypothesis testing: We reject, or do not reject, a hypothesis, such as "There are more non-smokers than smokers in the US". Rejecting it does not establish a fact that there are equal number of non-smokers and smokers, or there are more smokers than non-smokers in the US. It only means that it is our best judgment based on an observed sample.
2. Parameter estimation: We obtain an estimate about the population parameter, say, 20% of the US population are smokers. This 20% does not necessarily equate to the actual proportion of smokers. The best we can say is that, if we were to obtain more estimates using the same method, a very large percent of all the estimates will be within a certain range of the actual population proportion.

I will unpack the meanings of these answers later. The key point is that drawing statistical inference is a particular form of induction. Unlike logical/mathematical deduction or descriptive statistics, the conclusions of statistical inference/induction are not results of necessity or statements of facts, where the degree of certainty is 100%. In statistical inference, the degrees of certainty are less than absolute, and the quantification of the degrees of certainty is an essential part of the practice. The following chart captures the differences between descriptive and inferential statistics:

Table 5: Differences between descriptive and inferential statistics

	Descriptive statistics	**Hypothesis testing**	**Parameter estimation**
Involves sampling	No	Yes	Yes
Conclu-sion	Fact/Population parameter	Whether or not reject a null hypothesis	Estimate of population parameter
Degrees of certainty	100%	We control the certainty of our conclusion by control the number of times that we wrongly reject a viable hypothesis (Type I error).	We obtain an estimate of measurement error: some percent of all the estimates generated from the same method will be within a certain range of the population parameter (margin of error).

To summarize, statistical inference applies to situations where random sampling is an essential part of solving the problem. Since we make inferences about populations from particular random samples, statistical inference is essentially inductive. The purpose of statistical inference is not to establish facts about a population, but to test hypotheses and to obtain estimates of population parameter. We do not talk about the truthfulness or falsehood of such hypotheses or estimates. Rather, we talk about their viability, plausibility, or degrees of certainty. These are essential components of statistical inference.

Conceptual Analysis of Hypothesis Testing

Table 6 depicts a standard view of hypothesis testing.

Table 6: Standard view of hypothesis testing

Testing hypothesis of a population parameter
a. Establish the null and alternative hypothesis about the population parameter; b. Randomly take a sample from the population and calculate the sample statistics α; c. Find out the *p-value*: probability of obtaining a sample as extreme as or more extreme than α if the null hypothesis were true; d. If the *p-value* is less than 5%, reject the null hypothesis in favor of the alternative hypothesis. If *p-value* is bigger than 5%, do not reject the null hypothesis.

The myth of null and alternative hypothesis

There are many heuristics for setting up null and alternative hypothesis. A brief search in traditional and hypermedia statistical textbooks reveals that not only are the ideas of null and alternative hypothesis defined by heuristics or rules, without explanation, the rules are often incomplete, sometimes false, and sometimes contradictory.

1. The statement being tested in a test of significance is called the null hypothesis. The test of significance is designed to test the strength against the null hypothesis. Usually the null hypothesis is a statement of "no effect" or "no difference" (p. 538)…The effect we [the researcher] suspect is true, the alternative to "no effect" or "no change", is described by the alternative hypothesis." (Yates, Moore, and MaCabe 1999, p. 533)

2. The null hypothesis is a statement about the value of a population parameter, and it must contain the condition of equality and must be written with the symbol =, ≤, or ≥. (When actually conducting the test, we operation under the assumption that the parameter equal to some specific value.)…The alternative hypothesis is the statement that must be true if the null hypothesis is false. (Triola 1997, p. 349)

3. The null hypothesis represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. The alternative hypothesis is a statement of what a statistical hypothesis test is set up to establish. (http://www.cas.lancs.ac.uk/glossary_v1.1/hyptest.html#h0)

4. The null hypothesis is often the reverse of what the experimenter actually believes; it is put forward to allow the data to contradict it. (http://davidmlane.com/hyperstat/A29337.html)

These texts communicate two heuristics for setting up null and alternative hypothesis.

Heuristic 1: *Usually, a null hypothesis is a statement of "no effect". It must contain the condition of equality. An alternative hypothesis is the opposite of the null hypothesis.*

Heuristic 2: *An alternative hypothesis is sometimes called a research hypothesis. A common rule of thumb about setting up null and alternative hypothesis is that we take what we tend to believe (a claim that we wish to be supported) to be the alternative*

hypothesis, and the null hypothesis is the opposite of what we believe. We set up hypotheses in this way so that if the conclusion is rejection of the null hypothesis then the research hypothesis is supported.

Although both heuristics communicate aspects of the logic of forming null and alternative hypotheses, and they do apply to most of the hypothesis testing situations, the connections between them are not apparent. They even suggested a different sequence in setting up hypotheses: in Heuristic 1 the null hypothesis is set up first by the stated rule; in Heuristic 2 the alternative hypothesis is established first while the null is set up for the sake of argument.

In addition, the highlighted portions of the quotes 3 and 4, taken literally, communicate contradictory messages about whether a null hypothesis is what we believe to be true, or the reverse of it. The highlighted portions of quotes 2 and 3 also convey the idea that null and alternative hypotheses are to be proven true or false, which is incompatible with the logic of hypothesis testing.

In short, the concepts of null and alternative hypothesis are not well communicated by statistics curriculum writers. Null and alternative hypothesis must be understood in the context of the entire process of hypothesis testing. A null hypothesis is always set up in a way so that it specifies a population parameter, regardless of whether it is believed to be true. This is because a null hypothesis has to point to a distribution of sample statistics that we use to gauge the rarity of an observed sample. For example, suppose we suspect that there are more adult non-smokers than smokers in the US:

> Are there more adult non-smokers than smokers in the US?
> h_0: there are equal number of non-smokers and smokers.
> h_1: there are more non-smokers than smokers.

To test the null hypothesis, we collect a sample of size *n* from the population of adults in the US, and calculate an appropriate sample statistic (e.g., proportion of non-smokers, and the sample statistic is 60%). We ask the question: With what probability would a sample percent of 60% or more occur by chance if in fact smokers and non-smokers occur equally frequently in the population?[7] We then look at the distribution of sample statistics of random samples of size *n* from the hypothesized population (the parameter being 50% as specified by the null hypothesis), and compare the observed sample with the distribution. If we judge that samples having 60% non-smokers are sufficiently rare, then we conclude that the null hypothesis is not plausible. In other words, a theoretical distribution of sample statistics that is partially defined by the null hypothesis (the other defining factors being the assumed population distribution and the size of the observed sample) is central to the logic of hypothesis testing. This distribution, sometimes called null distribution, shows what is going to happen by chance if the null hypothesis about the population is true.

The relationship between alternative hypothesis and null hypothesis can be clarified by comparing hypothesis testing with proof by contradiction.[8] In proof by contradiction, we assume, along with the hypotheses, the logical negation of the result we wish to prove, and then reach a contradiction. That is, if we want to prove "If p is true, then q is true", we assume the truth of p and $\sim q$. From these assumptions, we deduce that $\sim p$ is also true or we derive $\sim r$, where r is a statement already taken to be true, e.g. an axiom or theorem. This contradiction leads us to conclude that the original statement q must be true when p is true.

---

[7] We chose 50% because it is the most conservative alternative to “more non-smokers than smokers”. We could just as well have chosen 40% as our hypothesized population parameter.
[8] Also known as *reductio ad absurdum*.

The similarity between proof by contradiction and hypothesis testing lies in the juxtaposition of two competing statements, and the method of proving/supporting one by way of falsifying/rejecting the other. In proof by contradiction, we deduce the truth of a statement by assuming its logical negation in conjunction with assuming some initial condition, and subsequently bring the negation into question by demonstrating that assuming it leads to unacceptable conclusions. Either the negation or the assumed initial condition must be false. In hypothesis testing, we test the plausibility of h_1 (what we suspect is true) by assuming a rival hypothesis, h_0, and testing its plausibility in terms of the likelihood of the factual data to have occurred given h_0 is true.

Proof by contradiction (along with the rest of classical mathematics) is built upon the law of excluded middle[9]*:* For any proposition p, it is either true or its negation is true. Falsification of $\sim p$ is a proof to the truth of p. Hypotheses about population parameters, however, are evaluated on the basis of samples that provide different degrees of evidences for or against them. A hypothesis is viable when data provides strong evidence in support of it. In this framework, rejection of a null hypothesis does not mean that the null hypothesis is false or wrong. It only means that it is not viable. Consequently, rejection of a null hypothesis does not mean that the alternative hypothesis is true. It means the alternative hypothesis is more plausible than the null. The same line of argument can be constructed for failure to reject a null hypothesis.

[9] Also known as *tertium non datur*.

Probability, unusualness, and distribution of sample statistics

The logic of hypothesis testing is that one rejects a null hypothesis whenever an observed sample is judged to be sufficiently unusual in light of it. This idea builds on a scheme of interrelated concepts including probability, unusualness, random sampling, distribution of sample statistics, and relative density of samples within intervals of the distribution.

The process of hypothesis testing involves the following objects:

Table 7: Objects involved in hypothesis testing

	Population in question	**Hypothesized population**
Population parameter	An unknown parameter	A parameter specified in the h_0
Sample(s)	A random sample of size n	A collection of samples of size n
Sample statistics	x	A distribution of sample statistics of the above collection

In hypothesis testing, we take a random sample of size n [it is an actual sample] from the population whose parameter we are testing; we then compare the sample statistic x against a theoretical *distribution of sample statistics* of samples of size n from the hypothesized population. When we assume null hypothesis is true, it follows that the sample statistic x is a result of chance occurrence, since the sample is taken randomly from the population. If the *probability* of samples like x is highly *unusual*, this would conflict with the supposition that the sample x is a result of chance occurrence, and subsequently lead to the rejection of the null hypothesis.

In hypothesis testing, probability is a mathematical expectation. It is a measure of our expectation of the proportion of samples like the one in question over a large number of repetitions. A sample is *unusual* if, over the long run, we expect it to occur a small fraction of the time.

Conceptualizing unusualness quantitatively is nontrivial. We observed in prior teaching experiments that students had a robust intuitive sense of "unusual" as meaning simply that an observed sample outcome is surprising, where "surprising" meant "differing substantially from what one anticipates". By this meaning, if one had no prior expectation about what the outcome should be like, then no outcome would be unusual.

Conceptualizing unusualness quantitatively entails an image of a distribution of sample statistics. The unusualness/probability of a sample is evaluated against the distribution of sample statistics of samples of the same size. We observed that students rarely made theoretical assumptions about the distributions of outcomes, and their attempt at applying the logic of hypothesis testing often became a meaningless exercise.

In hypothesis testing, unusualness of samples as extreme as or more extreme than the observed sample translates into a small *p-value*. Note that *p-value* is a probability of a composite event. It is the relative density of a region of samples whose values are as extreme as or more extreme than the observed sample. The collection of sample statistics in this region, rather than only the samples having the same statistics as the observed sample, constitutes our criterion for "evidence against the null hypothesis".

A small *p-value* is evidence against the null hypothesis. It is not evidence against an observed sample. An observed sample happened. It is not to be challenged. Instead, the null hypothesis is to be challenged/tested in light of the observed sample.

Logic of hypothesis testing

Although I have touched upon the logic of hypothesis testing earlier, I will further expand on this logic to a more general conceptual framework in order to incorporate ways of

thinking people might have that are incompatible with the logic of hypothesis testing. The logic of hypothesis testing that I have talked about so far takes the following form:

> We randomly collect a sample and calculate a statistical value x. If by assuming h_0 we derive that values like or more extreme than x are highly unlikely, then we conclude that x is highly unlikely. This conflicts with the fact that x actually occurred, which lead us to reject h_0. (HT1)

When facing this conflict, there are four possible choices:

(1) maintaining that h_0 is true and reject h_1;

(2) maintaining that h_0 is true and arguing that x is not a random sample, in other words, what is highly unlikely to happen given the null distribution actually happened because of bias in the sampling process;

(3) maintaining that h_0 is true and that what is highly unlikely to happen actually happened because of variability of the sampling process, and arguing that more evidences are needed in order to reject that h_0, and,

(4) rejecting h_0, in other words, denying that x is highly unlikely.

These choices are reflected in Figure 1.

Table 8: Theoretical constructs in hypothesis testing framework

Q		If the outcome is unusual in light of h_0, do they reject h_0?
1	R h_1	Rejecting h_1
2	*s* biased	Asserting that outcome is biased
3	LOE	Reluctant to reject of h_0 for lack of overwhelming evidence
4	R h_0	Rejecting h_0
5	C h_0	Committing to h_0
6	C h_1	Committing to h_1

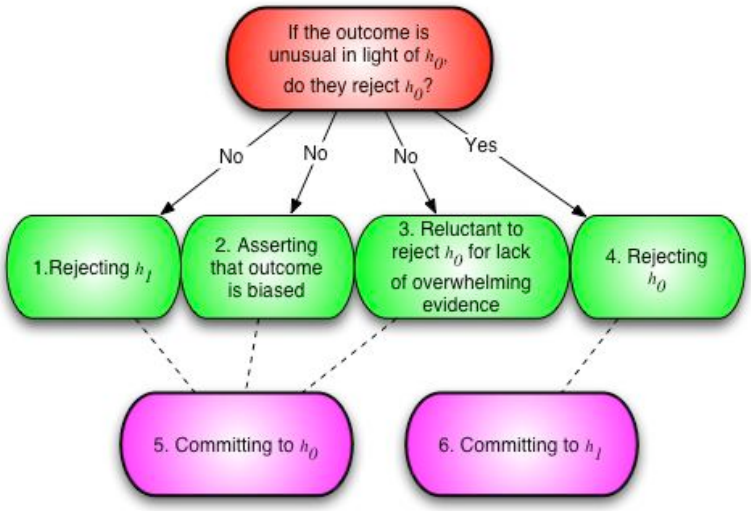


Figure 1: Theoretical framework for the logic of hypothesis testing

This conceptual framework, particularly choices (2) and (4), provides a way of conceptualizing hypothesis testing and random sampling each as an expression of a core conceptual scheme. The particular form of the logic of hypothesis testing (HT1) is an instance in this general framework when the randomness assumption (i.e., when *x* is a value of a random sample) stands, and when the last choice (rejecting h_0) is made to reconcile the tension.

> If h_0 is the actual population parameter, then making the second choice becomes:
>
> We collect a sample from a population with known parameter and calculate a statistical value *x*. If we derive that values like *or more extreme than* x are highly unlikely, then we conclude that *x* is highly unlikely. This conflicts with the fact that "*x* happened", thus it leads to conclusion that *x* is not a random occurrence, in other words, what is highly unlikely to happen happened because of bias in the sampling process. (HT2)

HT2 illuminates in essence the accepted practice in evaluating whether there is a bias in any sampling process: If a sample is highly unlikely, it will raise questions about the randomness of the sampling process.

Choices (1) and (3) point to potential ways of thinking people might have that leads to failure to employ the logic of hypothesis testing. In choice (1), a person rejects the alternative hypothesis on the basis of a small *p-value*. In choice (3), a person is reluctant to reject h_0 on the basis of a small *p-value* because what's highly unlikely to happen could still happen because of sampling variability, and he/she argues that more evidence is needed in order to reject the null hypothesis. (This misconception is a result of not understanding hypothesis testing as policy making embedded in the ideas of significance level and Type I error, which I will elaborate later). Both interpretations exhibit teachers' commitment to the null hypothesis, which is incompatible with the logic of hypothesis testing in which commitment is made to the alternative hypothesis.

Significance level

Embedded in the concept of significance level is the idea that hypotheses are neither true nor false. Decisions about particular hypotheses are results of applying decision rules. Such rules are justified in the broader context of a whole class of hypotheses that we shall never test. Neyman and Pearson (1933) addressed this matter by stating:

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we will ensure that, in the long run experience, we shall not be too often wrong. Hence, for example, would be such a rule of behavior: to decide whether a hypothesis H, of a given type, be rejected or not, calculate a specified character, x, of the observed facts; if $x>x_0$ reject H; if $x\leq x_0$, accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x\leq x_0$ or false when $x>x_0$. But it may often be proved that if we behave in such a way we shall reject when it is true not more, say than once in a hundred times, and in addition, we may have evidence that we shall reject H sufficiently often when it is false (Neyman & Pearson, 1933 as quoted in Hacking 1965, p. 104).

Note that the idea that we should not try to seek the truth of hypothesis is a result of the inductive nature of statistical inference, as I talked about earlier. The ideas of significance level and decision rules tells us that although we may not know the truth of hypotheses, we do know that if we consistently apply a decision rule, we can control the error rate (Type I error) within a reasonably low level. This measurement of error is an expression of the degree of certainty associated to our conclusion.

Conceptual Analysis of Margin of Error

What is margin of error?

Parameter estimation is about estimating a population parameter from taking a sample. Typically, the *accuracy* of an estimate is defined as the difference between the sample statistic and the population parameter. The smaller the difference, the more accurate the estimate is. Accuracy is about a *specific measurement* of an object: How far off is this measurement from the actual measurement of the object? Since in parameter estimation, the actual population parameter is unknown, it follows that the accuracy of individual estimates is unknown. The idea of margin of error tells us that, although we do not know how accurate a particular sample is, we do know that were we to repeatedly take samples of the same size, a certain percentage of the sample statistics will fall within a given range of the population parameter.

Although margin of error is the signature index of measurement error in poll results that appear in non-technical publications such as newspapers and magazines, it is understood poorly by the public. There is abundant confusion in both the lay and

technical literature about margin of error (Saldanha 2003). For example, the writings of ASA (1998) and Public Agenda (2003) misinterpreted margin of error as "95% of the time the entire population is surveyed the population parameter will be within the confidence interval calculated from the original sample".

The conventional definition of margin of error is based on the idea of sampling distribution and its meaning is expressed through the idea of confidence interval. A *Level c* confidence interval for a sample statistic is an interval centered on the sample statistics, and whose length (*2x margin of error*) is calculated from the standard deviation of the sampling distribution (when the population mean and standard deviation are known), or estimated from the sample standard deviation (when they are unknown). Level *c*, which also affects the margin of error, is the confidence level. Suppose *c* is 95%. This means that we expect 95% of the confidence intervals calculated from all samples of the same size will contain the population parameter.

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 and sampling distributions. First, we limited the discussion of margin of error to situations with populations of known parameters, thus excluding the scenario where margins of error have different values for samples of the same size. This allowed us to talk about the meaning of margin of error independently from confidence interval. A margin of error with 95% confidence level then means that 95% of all sample statistics will fall within the interval center on the population parameter and whose length is *2x margin of error*. Next, we focused on the idea of distribution of sample statistics instead of on the idea of sampling distribution. In this

approach, a sampling distribution is a special case of a distribution of sample statistics, so *distribution of sample statistics* is the more general idea. This approach provides the instructional benefit of easy simulation and demonstration of sampling process and results without compromising the path for understanding margin of error.

Two perspectives on measurement error

A prevailing misconceptions about margin of error is that margin of error is about a single sample statistic. However, margin of error is not about *specific measurements*, but about *collections of measurements*, or the *method* that generates the collections of measurements. How one relates margin of error reflects different perspectives on measurement error. Thompson (Teaching Experiments 1 and 2) clarified the distinction between two perspectives.

> 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 behavior 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 as determined by a more accurate device. 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:

> When we've studied this issue in the past, 99% 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 99% of this carpenter's measurements were within $\pm$1 millimeter, I have great confidence that this measurement is very accurate. (Thompson, Teaching Experiment 1)

Understanding the idea of margin of error entails that one adopts a contractor's perspective. Margin of error relates to a particular sampling result only to the extent that it is a measurement of the confidence that the *sampling process* that produced that result will produce results of which we expect a certain percent are within a given range of the actual parameter.

Margin of error, confidence level, and sample size

In this seminar, to say that a particular sampling method with confidence level $x\%$ and a margin of error r means that we anticipate that the interval $(p-r, p+r)$ captures $x\%$ of the sample statistics generated by it. The accuracy of a sampling method is simultaneously measured by both margin of error and confidence level. When the margin of error remains the same, a higher confidence level means more sample statistics fall within that range of the true population parameter, and thus conveys a higher confidence in the accuracy of results from the given sampling method. When the confidence level remains the same, a smaller margin of error means that sample statistics are clustered closer to the true population parameter, and thus conveys a higher accuracy of the sample method.

Comparisons of the accuracy of two (or more) sampling methods bring sample size into the picture. The relationship between margin of error, confidence level and sample size is:

When sample size is fixed, an increase in the confidence level will increase the margin of error. That is, if we want more sample statistics to fall within an interval centered on the population parameter, then we must increase the interval's width to capture a greater percent of sample statistics. For example, in a distribution of sample statistics obtained from random samples of size 512 from a binomial population with p=0.5, 95% of sample statistics are within 4 percentage points of the mean of the distribution, while 99% of sample statistics are within 5 percentage points of the mean.

If we fix the confidence level, then an increase in the sample size will decrease the margin of error. For example, in a distribution of sample statistics obtained from random samples of size 512 from a binomial population having p=0.5, 95% of sample statistics are within 4 percentage of the mean. In a distribution of sample statistics obtained from random samples of size 1024, 95% of sample statistics are within 3 percentage of the mean. This tells us that larger samples tend to be more accurate estimates because they are clustered closer around the mean of the distribution. And it means the phrase "*x*% sample statistics lie within a certain range of the true population parameter *(p-r, p+r)*" is another way of characterizing the variability of a distribution of sample statistics.

The above described relationships among margin of error, confidence level, and sample size is represented symbolically as

$$\text{Margin of error} = z * \frac{\sigma}{\sqrt{n}}$$

Where σ is the population's standard deviation, n is the sample size, and z* is the upper (1-C)/2 critical value (determined by confidence level C).

Margin of error and confidence interval

This study's definition of margin of error (the interval around the population parameter that captures *c%* of sample statistics) and use of populations with known parameters makes the use of confidence intervals inessential. Since a margin of error for a given confidence level and a given sample size is dependent only upon the population standard deviation, all confidence intervals will have the same width (*2r*, if margin of error is *±r*). Thus, for *c%* of the confidence intervals to contain *p*, the population parameter, *c%* of the sample statistics must be within *±r* of the population parameter. That is, the interval *(p-r, p+r)* will contain *c%* of the sample statistics.

A theoretical framework: a synthesis

Interpretations of margin of error involve some or all of these ideas: margin of error *±r* (*0<r<1*), confidence level *y%*, a population parameter *p*, a sample statistic *s* from a sample of size *n* (an estimate of *p*), a sampling distribution (distribution of all samples of size *n*) or in the context of the seminar, a distribution of a collection of samples of size *n*: s_i.

Margin of error, when centered around a population parameter, yields an interval that captures a certain percentage of sample statistics collected from repeatedly taking samples of a given size. Expressed symbolically, this interpretation is:

$$\textit{The interval } p \pm r \textit{ captures } x\% \textit{ of } s_i. \qquad x \in [0,100] \tag{1}$$

Reciprocally, when margin of error centered around the sample statistics, it yields confidence intervals *x%* of which contain the population parameter.

$$x\% \text{ of intervals } s_i \pm r \text{ contain } p. \quad (2)$$

Although typically, report of margin of error follows a sample estimate of an unknown population, margin of error in fact does not communicate to us how far off that sample statistic is from the population parameter. Rather it tells us that if we were to repeat the same sampling method, a certain percentage of all sample statistics will be within a given range of the population parameter. Therefore, with respect to one particular confidence interval, the best we can say is

$$\text{We don't know if the interval } p \pm r \text{ captures } s. \quad (3)$$

or

$$\text{We don't know whether the interval } s \pm r \text{ contains } p$$
$$(\text{but we do know that } x\% \text{ of intervals } s_i \pm r \text{ contain } p). \quad (4)$$

However, understanding of margin of error is not complete until one also understands that

$$x = y, \text{ i.e., } x\% \text{ is the confidence level} \quad (5)$$

In other words, the percentage of sample statistics captured by $p \pm r$ is the confidence level of a sampling method.

The combination of interpretations 1&3&5 conveys the definition/ways of thinking about margin of error as the research team had created. The combination 2&4&5 conveys a conventional interpretation/understanding of confidence interval.

Analysis of literature as well as data from the teachers seminar and prior teaching experiments found interpretations or ways of thinking that are incompatible with understanding margin of error. A classic misunderstanding of margin of error is:

The interval s±r contains p (6)

This interpretation is completely devoid of the idea of confidence level and a distribution of sample statistics. It exhibits the Carpenter's perspective that focuses on the accuracy of one individual sample statistic, and takes the margin of error as a measure of the distance between the sample statistic and the population parameter. Note that (6) is the direct opposite of the idea expressed in (4).

There are three other interpretations that indicate either a lack of or an erroneous understanding of margin of error. One interpretation is:

There is an x% probability that the interval p±r will contain s. (7)

This interpretation is not necessarily wrong, but certainly vague. Since we know that people hold various different meanings of probability, we cannot say for sure what they could mean by "*x*% probability". It could mean *x*% of sample statistics, in which case (7) is the same as (1), or it could simply denote a subjective belief, which means they do not have in mind a distribution of sample statistics. In other words, we do not have evidence, from (7), to claim that a person who says it is thinking that interval *p±r* would capture a portion of a collection of sample statistics. In the framework that I use to describe teachers' understanding, I will remove the ambiguity by assigning a subjective meaning to the word, "probability". That is, if a teacher says (7) but I have evidence that she is thinking (1), and I would assign (1) to her thinking.

The second interpretation is

The interval s±r captures x% of s_i, (8)

The interpretation conveys a distribution of sample statistics. However it says that *x*% of the sample statistics would be captured by the confidence interval constructed from the

sample statistics, instead of the confidence interval centered on the population parameter. The difference between (8) and (1) is the center of confidence interval constructed from the margin of error.

The third interpretation is

$$\textit{The interval } p \pm r \textit{ contains } x\% \textit{ of the intervals } s_i \pm r \qquad (9)$$

This interpretation is incoherent because all confidence intervals are of the same width ($2r$). It doesn't make sense to think that one interval will contain other intervals. Note that the interpretations 1, 2, 8, and 9 are all interpretations of margin of error that contains an image of distribution of sample statistics.

The above interpretations, taken together, constitute a theoretical framework/coding scheme (Table 9 and Figure 2) for understanding teachers' conceptions and interpretations of margin of error.

Table 9: Theoretical constructs in margin of error framework

1	The interval $p \pm r$ contains $x\%$ of s_i;
2	$x\%$ of the intervals $s_i \pm r$ contains p;
3	The interval $p \pm r$ either contains or does not contain s;
4	The interval $s \pm r$ either contains or does not contain p;
5	$x\%$ is the confidence level;
6	The interval $s \pm r$ contains p;
7	There is an $x\%$ probability that the interval $p \pm r$ contains s;
8	The interval $s \pm r$ contains $x\%$ of s_i;
9	The interval $p \pm r$ contains $x\%$ of the intervals $s_i \pm r$.
1or2or8or9	Interpretations that include distribution of sample statistics;
1&3&5	Understanding of margin of error;
2&4&5	Understanding of confidence interval.

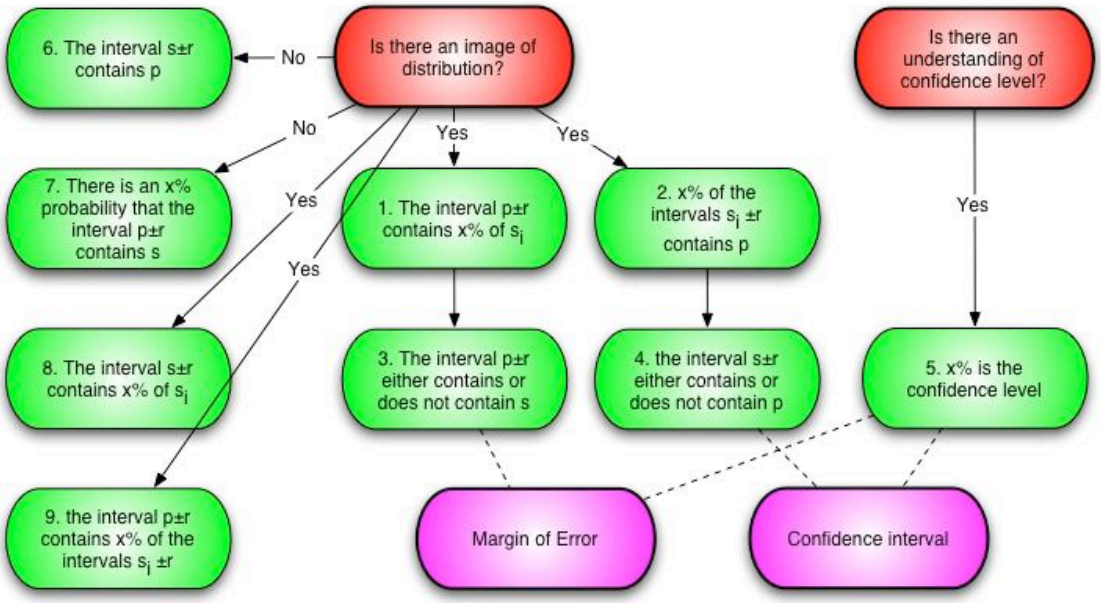


Figure 2: Theoretical framework for understandings of margin of error

Conceptual Analysis of Probability

Stochastic conception of probability

As I have elaborated earlier, statistical inference builds on the concept of probability. Statistical inference is about inferring a population parameter by taking one sample. In hypothesis testing, an inference is made on the basis of a probabilistic statement about *the relative frequency* of the observed sample over *a large number of repetitions*. In parameter estimation, the probability of a confidence interval containing the true population parameter is the *relative proportion* of all confidence intervals that contains the true population parameter. Hence, understanding statistical inference entails a stochastic conception of probability.

In a stochastic conception, an outcome A's probability being *x* means "an expectation that the long run repetition of the process that produces the outcome A will

end with an outcome like A *x* percent of the time." To say an outcome has a probability of *.015* is to say that we *expect* the outcome to occur 1.5 percent of the time as we perform some process repeatedly a large number of times.

A person having a stochastic conception of an event conceives of an observed outcome as but one expression of an underlying repeatable process, which over the long run will produce a stable distribution of outcomes. The conceptual operations entailed in this conception are:

1. Conceiving of a probability situation as the expression of a stochastic process;
2. Taking for granted that the process could be repeated under essentially similar conditions;
3. Taking for granted that the conditions and implementation of the process would differ among repetitions in small, yet perhaps important, ways;
4. Anticipating that repeating the process would produce a collection of outcomes;
5. Anticipating that the relative frequency of outcomes will have a stable distribution in the long run. (Thompson and Liu 2002)

For example,

What is the probability that 18 out of 30 people favor Pepsi over Coca Cola?

To conceive of the underlying situation stochastically entails

1. Conceiving of a random sampling process: selecting a number of people from a population, and asking each person whether he or she favors Pepsi or Coca;
2. Imagining repeatedly taking samples of size 30, and recording the number of people in each sample that favor Pepsi;
3. Understanding that this repeatable process will produce a collection of outcomes;

4. Understanding that because of the *random* selection process there exists variability in the collection of outcomes, but over the long run, the distribution of outcomes will become stable.

Probability of "18 out of 30 people favor Pepsi" is the relative frequency of this outcome within the distribution of outcomes.

A theoretical framework of understandings of probability

A stochastic conception, as I described above, is a coherent and powerful conception of probability that supports understanding of statistical inference. It is what one might take as an instructional objective when designing teaching of probability. In this study I proposed to develop a theoretical framework that, when applied to teachers, will result in descriptions of teachers' actual understandings of probability.

Below is the theoretical framework that I developed from the literature and from the analyses of teachers' interpretations of probability statements.

Table 10: Theoretical constructs in probability framework

1	Q1	Is there an image of a repeatable process?
2	Q2	Are the conditions of the process specified?
3	Q3	Is there an image of a distribution of outcomes?
4	OA	Outcome approach
5	ANA	Outcome is A or non A, prob.=1 or 0
6	PH	Proportionality heuristic
7	ANA	Outcome is A or non A, prob.=50%
8	APV	Probability as relative proportion: all possible values of random variables
9	APO	Probability as relative proportion: all possible outcomes
10	RF	Probability as relative frequency: distribution of all outcomes

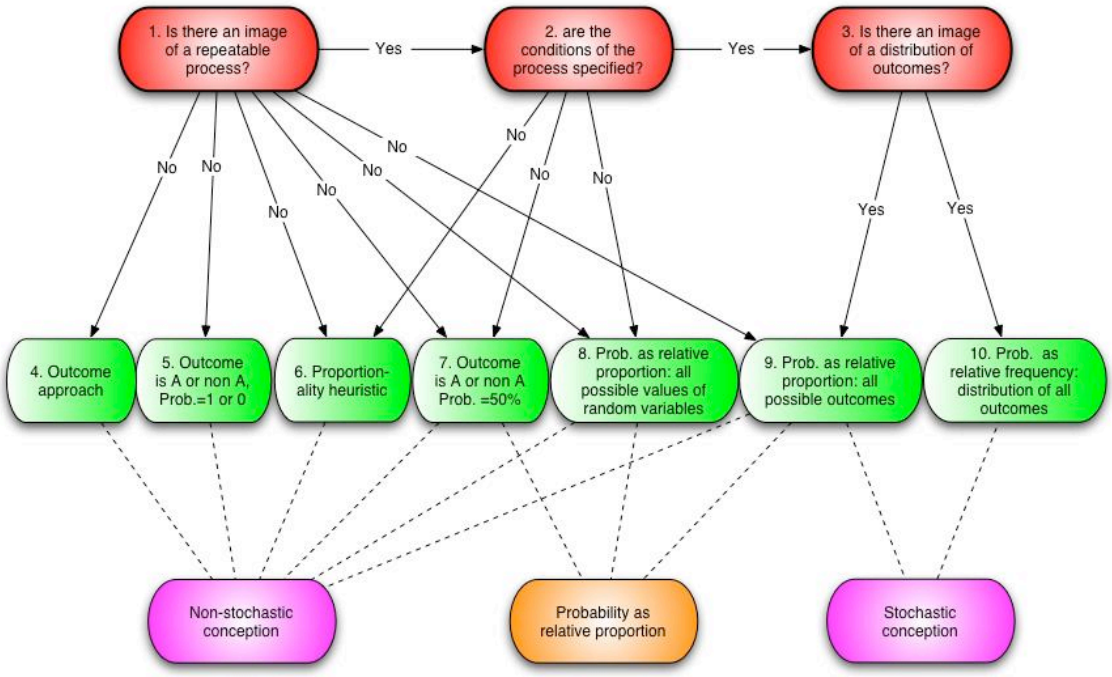


Figure 3: Theoretical framework for probabilistic understanding

Table 11: Explanation of Figure 3

Color	Meaning	Corresponding to
Red	Ways of thinking about a probability situation	Box 1, 2, 3.
Green	Interpretations of probability	Box 4 to 10.
Purple	Conceptions of probability	Path (See below)
Orange	A standard method of computing probability	Box 7, 8, 9.

Table 12: Explication of paths in Figure 3

Conceptions of probability	Path
Non-stochastic conception	1-4
	1-5
	1-6
	1-7
	1-8
	1-9
	1-2-6
	1-2-7
	1-2-8
Stochastic conception	1-2-3-9
	1-2-3-10

Interpretation 4 (outcome approach) is a subjective conception of probability. One who holds an outcome approach thinks that probability is a subjective judgment based on personal experiences. Interpretations 5 and 7 both adopt an approach that reduces the sample space, for probability of outcome A, to [A, not A]. Interpretation 7 further applies the principle of indifference—the probability of each outcome is 1/n where n is the number of outcomes. Note that people who hold such interpretations of probability will not be able to make sense of common probabilistic statements, such as, the chance of rain tonight is 40%. They will think instead that the chance of rain is either 1 or 0, or the chance of rain is always 50%. Interpretations 8 and 9 reveal a relative proportion conception of probability. Within this conception, the probability of an event is the relative proportion of the outcomes making that event compared to all possible outcomes. As we can see from the Example I of Table 13, the results of the interpretations 8 and 9 conflict with each other. This is because interpretation 8 confounded the values of the random variable (sum of the dots on the uppermost faces) with the process' sample space (the set of all possible states in which the process can terminate). Notice that interpretation 9 came from a standard method of computing probability. The limit of this method is that it can be applied only to experiments that have a sample space that can be defined so that all outcomes are equally likely. From this framework, we can see that a person who thinks probability is relative proportion of results might or might not think stochastically, depending on whether or not he/she has the set of conceptual operations listed in 1 to 3.

Table 13: Examples of paths & interpretations of probability

Path	**Interpretation of Probability**	**Example I:** *What is the probability that I will get a sum of 11 when I throw two dice?*	**Example II:** *What is the probability that 18 out of 30 people favor Pepsi?*
1-4	Outcome approach	11 is my favorite number, so I believe the probability is going to be 80%.	I'm a coke drinker. I think it is unlikely that 18 out of 30 people favor Pepsi.
1-5	Outcome is A or non A, Prob.=1 or 0.	I will either get an 11 or not. The probability is 1 if I do get an 11, 0 if I don't.	The probability is 1 if 18 out of 30 favor Pepsi, is 0 if not.
1-2-6	Proportionality heuristic	N/A	The likelihood of 18/30 is high if a larger sample reflects similar or the same proportion.
1-6	Proportionality heuristic	N/A	A person either likes or doesn't like Pepsi (implying that population parameter is 50%), thus the likelihood of 18/30 is high.
1-7 or 1-2-7	Outcome is A or non A, Prob.=50%	I will either get an 11 or not. The probability of getting 11 is 50%.	The outcome is either 18 or not 18, so the probability is 50%
1-8 or 1-2-8	8 All possible values of random variables	There are 11 possible outcomes: 2, 3, … 11, and 12. The probability of getting 11 is 1/11.	N/A
1-9 or 1-2-3-9	9 All possible outcomes	There are 36 possible outcomes: (1, 1), (1, 2)…, and (6, 6). Two of these outcomes (5, 6) and (6, 5) will give the sum of 11. The probability of getting 11 is 2/36.	There are 31 possible outcomes. The probability is 1/31.
1-2-3-10	Probability as relative frequency	If I repeatedly throw two dice, what fraction of the time will I get a sum of 11?	If we repeatedly take samples of 30 people, what fraction of time will get results of 18 people favoring Pepsi?

This framework allows me to make sense of data in two ways: 1) developing quantitative measures of teachers' interpretations of probability, 2) understanding teachers' conceptions of probability, and the extent to which their conceptions are close developmentally from the stochastic conception.

Teachers' understanding of probability

In the above section, I proposed a theoretical framework for describing a person's understandings of probability. Below I will elaborate two big ideas that have particular significance in pedagogy of probability.

From the above framework, it is conspicuous that people could have different conceptions of probability and interpretations of probability situations. Thus, we believe that it is important for the teachers to not only develop a coherent and powerful understanding of probability, but also to have an understanding of how probability statements and situations might be interpreted differently. In other words, we wanted the teachers to understand the idea that *a situation is not stochastic in and of itself. It is how one conceives of a situation that makes it stochastic or non-stochastic*. We believe that this idea would be essential for the teachers to become sensitive to alternative understandings their students might have.

Furthermore, as I have shown how a non-stochastic conception can have many different expressions, a stochastic conception, too, could lead to different interpretations of a probability situation. In other words, *a particular event could be seen as outcomes from different stochastic processes, and thus the probability of this event differs depending on how one conceives of the stochastic process*. For example, suppose we have urns A, B, and C, each containing a number of red and white marbles (see Table 14). Question: What is the chance that if you draw a red marble, it is from urn C?

Table 14: Number of marbles in urns

Urns	# of red marbles	# of white marbles
A	2	5
B	5	4
C	3	9

There are two ways of conceiving this situation stochastically. The first is to imagine that all the marbles are dumped into one container and each is labeled by what urn it came from. Of all 10 red marbles, 3 came from Urn C, thus we would expect that, over the long run, 30% of the time that we select a red marble, it will have come from Urn C.

A second way is to imagine a repeated process that we will first pick an urn at random, and then select a marble from that urn. In the long run, each urn will be chosen 1/3 of the *n* times we repeat the process. Each marble in urn A will be chosen 1/7 of the time that Urn A is chosen, so we will select a red marble from Urn A 2/7 of the time Urn A is chosen, or $\frac{1}{3}\times\frac{2}{7}$ of the *n* times we repeat this process. By the same token, we will draw a red marble from Urn B $\frac{1}{3}\times\frac{5}{9}$ of the *n* times we repeat this process, and we will draw a red marble from Urn C $\frac{1}{3}\times\frac{3}{12}$ of the *n* times we repeat this process. Therefore, of the times we select a red marble ($\frac{1}{3}\times\frac{2}{7}+\frac{1}{3}\times\frac{5}{9}+\frac{1}{3}\times\frac{3}{12}$ of the *n* times we repeat the process), we will have selected it from Urn C $\frac{1}{3}\times\frac{3}{12}$ of the *n* times that we repeat this process. Therefore, we will select a red marble from Urn C $\frac{\frac{1}{3}\times\frac{3}{12}}{\frac{1}{3}\times\frac{2}{7}+\frac{1}{3}\times\frac{5}{9}+\frac{1}{3}\times\frac{3}{12}}$ of the times we select a red marble (about 23% of the time).

This example illustrates that a probability situation can be conceived of from different stochastic perspectives, and that an answer to a probability question is valid as long as it is consistent with the underlying situation as one has conceived it.

CHAPTER VI

TEACHERS' UNDERSTANDINGS OF PROBABILITY

This chapter describes two sets of activities and interview questions in which we investigated teachers' conceptions of probability.

Table 15: Overview of the activities and interviews in Chapter 6

Chapter 6 Teachers' understanding of probability			
Section	**Activity (A) and Interview (I)**	**Day**	**Duration**
6.1 Stochastic and non-stochastic conception	A1-2 Chance and likelihood	1	29 m.
	A2-2 PowerPoint presentation	5	67 m.
	I3-1 Five probability situations		
	I3-4 Gambling		
6.2 Multiple interpretations of probabilistic situation	A2-4 Clown and Cards	6	138 m.
	I3-2 Three Prisoners		

Section 6.1 focuses on teachers' interpretations of probability situations, particularly on whether they conceived of the situations stochastically or non-stochastically. Section 6.2 introduced a probability situation that, if conceived stochastically, may subject to multiple interpretations, and investigated the ways with which the teachers responded to this type of situations.

Activity 1-2: Chance and likelihood

1. What does "today there is a 45% chance of rain" mean? 2. A pollster asked 30 people about which they liked better, Pepsi or Coca-Cola. 18 said Pepsi How likely is this result? (What does this mean?)

The key to interpret chance and likelihood stochastically is to conceive of a stochastic process, a process that generates a collection of outcomes of which the particular phenomenon in question is but one of those outcomes. For example, in the first situation, "a 45% chance of rain" means "of all those days having the similar conditions like today, 45% of them rain." The stochastic process was to examine the weather conditions of all the past days having the similar conditions as today. Interpreting the second situation stochastically means conceiving of an underlying population having a relatively stable proportion of people who favor Pepsi and a process of taking random samples of 30 people out of this population. Conceiving of the situation as such allows one to think of the result "18 people out of 30 favor Pepsi" as one (or one kind) of the possible outcomes of the sampling process. The likelihood of this outcome can then be quantified as the relative frequency of outcomes like this one against the total number of times the sampling process is repeated.

Activity 1-2, Episode 1: Question 1

What does "today there is a 45% chance of rain" mean?

The discussion around this question lasted for 17 minutes. Results show that three out of eight teachers interpreted the situation non-stochastically; two teachers interpreted the

situation stochastically. Available information was not sufficient to discern the other three teachers' conceptions.

Three teachers interpreted the situation non-stochastically. They were concerned, essentially, about *today's* chance of rain. Their conceptions of probability were non-stochastic because they were concerned about a single outcome, as opposed a collection of outcomes or a stochastic process (that generates these outcomes). There are two variations in these non-stochastic conceptions.

Two teachers, Betty and Linda, interpreted the meaning of "today there is a 45% chance of rain" by answering the question, "what would I do if today there is 45% chance of rain". See Excerpt 1[10]

Excerpt 1

2. Betty [It] means you'll probably wanna take your umbrella.

16. Linda I would do it based on what you were going to do that day. If you had an outdoor wedding planned, I'd say there's a good chance… you need to plan something else. But if I was planning for some outdoor sports activity, then I'd probably go ahead and do it. It is a relative—I don't know, It is not a number that really means 45 of anything. It is just…forty-five percent of anything, like if it were a hundred percent chance of rain then you know it is going to rain for sure, if it is eighty-five percent then you're almost sure it is going to rain, ten percent chance, well it probably won't rain so go ahead and do whatever you want that day.

Betty and Linda's answers suggested that they had a subjective conception of probability. A 45% probability of rain conveys the strength of belief about the likelihood of rain that lies somewhere between "definitely going to rain" to "no rain". The highlighted portion of Linda's utterance suggested that she did not conceive of a collection of days having

[10] The transcripts in the excerpts are numbered sequentially within one activity. For example, an utterance numbered 16 in Activity 2-4 means that it is the 16th utterances in the discussion around Activity 2-4.

similar weather conditions like today, and 45% as the relative frequency of days that rain. Rather, 45% has meaning only when placed in a scale from 0 to 1, which measures the strength of belief about the likelihood of rain on any particular day.

John was also concerned about *today's* chance of rain, although he had a different interpretation. His conception of probability closely resembled the "principle of indifference"— "conversion of either complete ignorance or partial symmetric knowledge concerning which of a set of alternatives is true, into a uniform probability distribution over the alternatives" (Fine 1973 p. 167).

Excerpt 2

26.	John	There's always a 50% chance of rain, it just so happens that on that day there's a little bit less than a fifty percent chance. It is either it will rain or it will not.
…		
112.	Terry	I want to go back to what John said about a fifty percent chance every day. Fifty percent chance of no rain or rain everyday, is that what you said?
113.	John	It is just that it is an event, I'm just saying that, if I don't watch the TV or anything, if I don't look outside, I know that outside either it is raining or it is not raining. It is fifty percent chance. What I would do is, say for instance the weatherman says forty-five percent chance of rain, I'll take my TI-83 (laughter)… I'll set up a random integer, and I go one to—or zero to ninety-nine, and if I get forty-four, I'm going to take my umbrella… or forty –five. (laughter)
114.	Henry	If all events are equal, you'd have a fifty percent chance of rain=
115.	John	Right.
116.	Henry	But if=
117.	Terry	But on any given day, is, do you have the same chance of having rain and not having rain??
118.	John	You don't—that's why they have meteorologists! They can go better predicting…

In John's conception, the probability of rain on any given day is 50% if we remain completely ignorant of the possible conditions or information that might provide evidences for or against any one of the two outcomes [rain] and [no rain]. As far as how 45% is determined, he left that solely to the expert, i.e., the meteorologists. From what he

said, he did not have a framework for making sense of the 45% probability, i.e., knowing where the 45% came from.

Only two teachers, Henry and Nicole, exhibited an orientation in thinking about where the number "45%" came from.

Excerpt 3

11. Nicole If you had 100 days just like today, that in forty-five days it is going to rain in the immediate area.

19. Henry I think today they run a lot of models, meteorological models…if you had a hundred days just like today, the model says it is going to rain forty-five of those a hundred days.

Excerpt 3 showed that both Nicole and Henry had an image of a collective in their thinking. They understood that the probability of rain on one particular day was calculated on the basis of a collection of days having similar conditions. Although they did not speak of a stochastic process, we believe that the process of "looking at past data on weather conditions over a period of time" was inherent in their thinking.

Activity 1-2, Episode 2: Question 2

A pollster asked 30 people about which they liked better, Pepsi or Coca-Cola. 18 said Pepsi. How likely is this result? (Question: What does "How likely is this result?" mean?)

The discussion around this scenario lasted for 12 minutes in two separate segments: 7 minutes in the afternoon of day 1, and 5 minutes in the morning of day 2.

Segment 1

Initial discussion revealed that only three teachers understood the question, "What does 'how likely is this result' mean?" Most of the other teachers answered a different

question. Henry answered the question: "What does the result '18 out of 30 people preferred Pepsi' mean?" Other teachers answered the question "how likely is this result?"

Excerpt 4

139. John — Yes, it is highly likely that it is 18. I would expect this number to be between 10 and 20, I doubt that it would be 0 and I doubt that it would be 29. But, I'm not trying to be funny, I'm being serious, it is either, either you like Pepsi or you don't like Pepsi…. So the other choice would be "not Pepsi", which would be Coca Cola. So it is either Pepsi or it is Coca Cola. And uh…how-- what does that tell us? Eighteen out of thirty, what does that tell me? Well, I can't say that 60% of the nation likes Pepsi because that's not enough data. That's just one sample of 30. But I would expect to see one sample of 30, say, 10 people liking Pepsi, or 20 people liking Pepsi, I would expect to see a lot of things, but I wouldn't expect to see 0 and I wouldn't expect to see 30.

148. Linda — It is about half way, so I would say it is very likely. Because 15 is half, so … I mean Pepsi and Coca Cola are, are always competitive with each other.

150. Betty — As a Coke drinker I would think more people would choose Coke, because I like Coke better and I can't see how anybody can drink Pepsi.

Only three teachers interpreted the meaning of "how likely is this result?"

Excerpt 5

130. Alice — Is that what you would have expected? ... that 18 would have preferred Pepsi.

144. Lucy — I would say they are asking, "Is that a valid result?", that you wouldn't expect to see that.

145. Terry — And what would make it valid? You said "valid result."

146. Lucy — Like if you were to sample a larger number of people, would you get the same result?

153. Sarah — Would we consistently get that result over and over?

The three teachers, John, Betty, and Linda, who tried to answer the question "how likely is this result", interpreted the situation non-stochastically. They evaluated the likelihood of "18 out of 30 people favor Pepsi" based on their beliefs of how the

population was distributed with respect to their preference. Betty, a Coke drinker, believed more of the population favored Coke, while Linda and John assumed that the population was evenly split. John's justification was, "either you like Pepsi or you don't like Pepsi… it is either Pepsi or it is Coca Cola." The logic of this reasoning is: "Because any particular person either likes Pepsi or doesn't like Pepsi, about half of the population like Pepsi and half don't." The absurdity of this logic can be easily seen by an analogy: "Because you either have AIDS or you don't, about half of the population have AIDS." The fault of this reasoning lies in the fact that one is thinking about individual cases while making statement about a collection. It is a coerced application of the "principle of indifference", always acting in a state of complete ignorance. That is, for any event that has *n* multiple outcomes, the chance of any particular outcome's occurrence in a number of repetitions of this event is always a fixed number *1/n*.

The defining characteristic of these three teachers' reasoning is that they tried to answer the question "how likely is this result?" on the basis of 1) the *particular* result, and 2) their beliefs of how the population was distributed—"I believe the population is distributed in this way. Given what I believe, the likelihood that 18 out of 30 people favor Pepsi is…" This way of reasoning is what I called a proportionality heuristic—evaluating the likelihood of a sample statistic by comparing it against the population proportion, or statistic of a larger sample, as we will see below in Lucy's thinking.

Lucy also applied this heuristic. By saying "If you were to sample a larger number of people, would you get the same result?" what she meant was something like this—in this scenario, 18 out of 30 people favor Pepsi. If we take a larger sample, say 90 people, and around 54 out of 90 favor Pepsi, then we would deem the result of "18 out of

30 favor Pepsi" as highly likely. The logic behind this reasoning could be that: "A larger sample is more representative of the population. So if you get the same result in a larger sample, that confirms that the result of the small sample is very likely." What distinguished Lucy's heuristic from John, Betty, and Linda's was that Lucy conceived of a repeatable sampling process.

The problem of proportionality heuristic lies in the question: What does it mean for two results to be *the same*? In Lucy's case, for example, would the two results be the same only when the larger sample has precisely 60% of people preferring Pepsi? What if there is 61% preferring Pepsi? Would that count as the same? If 61% could be counted as same/similar, where does one draw the line? In other words, the difference between proportionality heuristic and a stochastic reasoning is that a stochastic conception of likelihood is quantifiable, but likelihood in a proportionality heuristic is not. The answers to the question "how likely is this result" have to be expressed in qualitative terms, such as, "very likely" or "not likely".

Sarah's answer "would you get the result over and over" implies that she had conceived of a repeatable process and hinted on the idea of likelihood as long run expectation. However, since she did not elaborate on her thinking, we could not know what she had in mind about the process and how it related to the likelihood of the result.

Segment 2

This segment occurred during the follow-up activity (A1-3 on hypothesis testing, see Chapter 7) where the teachers were given a collection of 135 simulated samples of size 30 from a population that was evenly split in preference, and a series of questions that intended to help them make the connection between likelihood and long run

expectation— the idea that likelihood of a particular sample is measured by the relative frequency of samples like this over a large number of random samples of the same size. When the teachers tried to make this connection, some of them expressed their resistance. They argued that the word *likelihood* did not call for a quantitative interpretation.

Excerpt 6

341.Nicole	But "likely" feels like a weaker word than "what percent". "What percent" means to me that I need to do some floundering around to figure out a "mathie"-type answer and um, that's what "what percent" means. "How likely" just was (shrugs)..I mean in that case I wasn't making the assumption that it was 50-50 so it seemed quite likely 'cause I'd been basing it on my um…on a general knowledge… that's, you know, they're sort of equally distra-dist-distributing in Kroger, but I've never stood there that long and counted them up. So, sure!
342.Sarah	"Likely" is less definitive than percent=
343.Nicole	=That's right! That's how I felt and so=
344.Sarah	=and I don't disagree with that, but I'm like Henry, I couldn't come up with a better phrase! So..
345.Terry	When you think of the word "likely" , what question do you think, when you talk about likelihood, just in general=
346.Nicole	=what could've happened!?
347.Terry	And how would you measure "Could've happened"?
348.Nicole	Well, I wasn't measuring it because I thought that that's what the slide, you know=
349.Terry	=Ok=
350.Nicole	=I mean if, if you would ask me, you know, "three percent favor Pepsi, how likely is that result?" Just sort of based on my common knowledge I would've said, "I don't think it is likely, no."
351.Terry	OK.
352.Lucy	"Likely" sounds more like it is asking for a question like, for an answer like "fairly-likely" or "not very likely" or something.

In a sense, the teachers were wrapped around in the colloquial meaning of the word *likelihood*. They did not equate likelihood to the technical meaning of probability. Likelihood to probability is like, for example, *warmth* to *temperature*. When we ask a question, "How warm is the water?" we expect answers such as, "Yes, very warm", "Not very", or "It is not." When we ask, "What is the temperature of the water?" we expect a

measurement of the temperature expressed in quantitative terms, such as "70 degrees". By the same token, teachers thought that the question of "how likely" called for answers such as "fairly likely" or "not very likely". This non-quantitative conception of likelihood is consistent with, and perhaps could account for, the teachers' non-stochastic interpretations of the Pepsi situation.

Summary of Activity 1-2

The overarching finding that I draw from the discussion of these two scenarios is that, with exception of few instances, most teachers had a non-stochastic conception of chance and likelihood. Their interpretations to chance and likelihood situations were subjective, expressed either as purely personal beliefs or as results of a coerced application of the "principle of indifference". They held a non-quantitative meaning of likelihood, in which case likelihood expresses strengths of beliefs about the possibility of chance occurrence, as opposed a mathematical expectations.

A more detailed depiction of each teacher' conception of probability is provided below in Table 17 by coding teachers' interpretations using the conceptual framework elaborated in Chapter 5 (See Table 16 and Figure 4 below).

Table 16: Theoretical constructs in probability framework

1	Q1	Is there an image of a repeatable process?
2	Q2	Are the conditions of the process specified?
3	Q3	Is there an image of a distribution of outcomes?
4	OA	Outcome approach
5	ANA	Outcome is A or non A, prob.=1 or 0
6	PH	Proportionality heuristic
7	ANA	Outcome is A or non A, prob.=50%
8	APV	Probability as relative proportion: all possible values of random variables
9	APO	Probability as relative proportion: all possible outcomes
10	RF	Probability as relative frequency: distribution of all outcomes

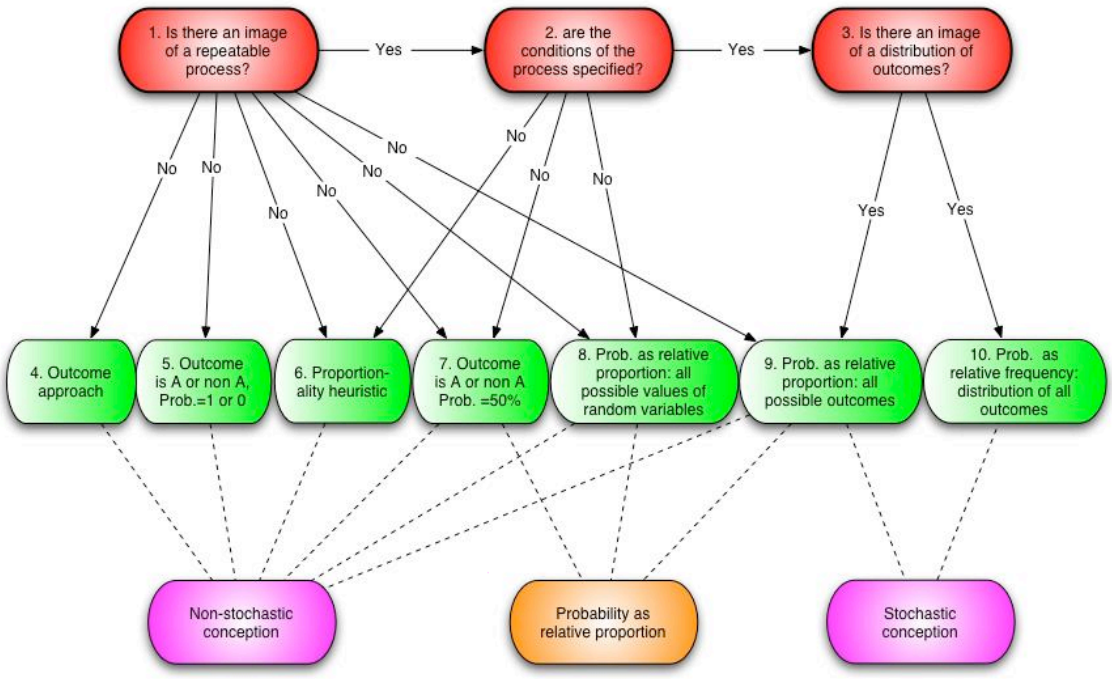


Figure 4: Theoretical framework for probabilistic understanding

Table 17: Teachers' conceptions of probability situations in Activity 1-2

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D1A1-2Q1	John	N						Y			
	Nicole	Y	Y	Y							Y
	Betty	N									
	Linda	N			Y						
	Henry	Y	Y	Y							Y
D1A1-2Q2	John	N					Y	Y			
	Nicole	N			Y						
	Sarah	Y	N								
	Lucy	Y	N				Y				
	Betty	N					Y				
	Linda	N					Y	Y			

Instruction for reading the table—The first column "locator" tells where an interpretation is located. For example, "D1A1-2Q2" means "Day 1, Activity 1-2, Question 2". Columns numbered 1-10 are theoretical constructs of the probability framework. Detailed description is provided in Chapter 5. "Y" and "N" means "Yes" and "No". Examples: 1) the "Y" in column 1 and row 4 means "For Activity 1 Question 1, Nicole conceived of a repeatable process." 2) the "Y" in

column 7 and row 3 means "John's interpretation of probability, as shown in Activity 1 Question 1, is "probability of an event A is 50% because there can be only two outcomes, A or non A." Codes across a row provide the information about a path. For example, row 3 denotes John's non-stochastic conception, path 1-5. Codes across a column tells the number of instances in which teachers exhibits a particular way of thinking. For example, column 1 tells the number of instances the teachers did or did not conceive of a repeatable process.

Table 17 showed that only two teachers, Nicole and Henry, clearly conceptualized chance stochastically when interpreting the first situation. For the second situation, none of the teachers conceived of it stochastically, i.e., conceptualizing a repeated sampling process that produces a collection of sample like the one in question, and evaluating the likelihood of the sample in question against this collection. Note that neither Nicole nor Henry interpreted the second situation stochastically. These inconsistency of interpretations across situations suggested that their stochastic conception was conditional to specific situations. They had not formed an orientation in interpreting probability situations stochastically.

Activity 2-2: PowerPoint presentation

Activity 2-2 was conducted in the beginning of the second week in which we focused intensively on teachers' conceptions of probability. In this activity, we presented the teachers with the following PowerPoint slides:

Table 18: The PowerPoint presentation on probability

#	Title	Slides	
1	Risky Rides	Your risk of being killed on an amusement park ride? One in 250 million.	What does that mean? How do you suppose they determined these values?
		Your risk of being killed driving home from an amusement park ride? One in 7,000.	
2	Vitamin Use	In a study of over 88,000 women, total folate (vitamin Bc) intake was not associated with the overall risk of breast cancer. However, higher folate intake (or multivitamin use) was associated with a lower risk of breast cancer among women who regularly consumed alcohol.	What does this mean to you? How do you suppose they determined this?
3	Gustav's Bad Luck	Gustav read in Newsweek magazine that drivers with three or more speeding tickets are twice as likely to be in a fatal accident as are drivers with fewer than three tickets.	What does this mean?
		The very next day he received his third speeding ticket.	What does this imply?
4	Car and weather situations	What is the probability that my car is red? What's the probability that the temperature tonight is below 40? What does that mean?	
5		Probability is what we anticipate will happen in the long run.	
6	Rishad's situation	Rishad's sister, Betty, rolled ten sixes in a row while playing a board game.	Rishad: "That is impossible!
			Rishad: "If a billion people rolled a die 10 times, what fraction of them would roll all sixes?"
7		The situation per se is not probabilistic. It is how you conceive the situation that makes it probabilistic.	
8	Rephrase the situations probabilis-tically	What's the probability that the next U.S. Attorney General is a woman? What's the probability that it will snow tomorrow? …the BHS gymnasium ceiling is 30' high? …you are off by no more than 2" when you measure the BHS gymnasium's height? …the Titans go to Super Bowl XXXXV? …you are dealt one pair in a 5-card hand from a standard deck?	
9		Two ways of think about probability: empirical and theoretical. You can determine that either empirically or theoretically. Experiment is the process that's being repeated. Outcome is a state an experiment could end. Event is a collection of outcomes.	

Slides 1-3 consisted of a number of probability statements, and we asked the teachers for their interpretations and their understanding of how the measures of probability (or risk) were determined. Slides 4 presented two probability situations and we asked the teacher

for their interpretations. Slide 5 was a statement that connects the idea of probability to long run expectation. Slide 6 presented a situation and its two different interpretations – one stochastic and one non-stochastic. We used this example to illustrate the idea, stated in slide 7, that *a situation per se is not probabilistic; it is how one conceives of the situation that makes it probabilistic*. In slide 8 we presented common questions about probability that are typically interpreted non-stochastically, and we asked the teachers to rephrase them from a stochastic perspective. Finally, the last slide connects the empirical and theoretical probability, and introduced the standard terminologies associated to the concept of probability.

The discussion around the slides lasted for 67 minutes. Discussion around slides 1 to 4 and slide 6 are more revealing of teachers' conceptions of probability.

Activity 2-2, Episode 1: Slide 1-3

1	Risky Rides	Your risk of being killed on an amusement park ride? One in 250 million. Your risk of being killed driving home from an amusement park ride? One in 7,000.	What does that mean? How do you suppose they determined these values?

Discussion on the first slide lasted for approximately four minutes. The following excerpts from the discussion suggested that Linda understood risk/probability as relative proportion.

Excerpt 7: Question 1

2. Terry All right, the risk of being killed on an amusement park ride is one in 250 million. What does this mean to you?
3. Sarah I don't want to be that one.
4. Terry You don't want to be that one.
5. John Well, the student's going to say it is not likely, but we can't say that.
6. Terry Ok.
7. Linda If you took, it took 250 million park rides for one person to get killed.

Excerpt 8: Question 2

12. Terry Your risk of being killed driving home from an amusement park ride was 1 in 7000. Ok? How do you suppose they figured that statistic out?

13. Linda From looking at the actual statistics.
14. Terry Specifically…
15. Linda The actual road deaths and the number of cars that they would estimate going from point A to point B.

Careful examination of Linda's responses suggested that there existed inconsistencies in the way she conceived of both situations. For the first situation, Linda conceived of a repeatable process of people taking park rides. However, the conditions of this process were not specified. Because while the risk in the situation intended to mean, "For every 250 million *people who take park rides*, one person dies", Linda's interpretation was, "For every 250 million *park rides*, one person dies." The same was true for the second situation. Linda conceived of two incompatible collections: actual road deaths and the number of cars. She did not conceive of a stochastic process of sampling from a collection of *people* (who drive home from an amusement park ride) and recording the number of those who were killed in auto accident.

2	Vitamin Use	In a study of over 88,000 women, total folate (vitamin B_c) intake was not associated with the overall risk of breast cancer. However, higher folate intake (or multivitamin use) was associated with a lower risk of breast cancer among women who regularly consumed alcohol.	What does this mean to you? How do you suppose they determined this?

Discussion around this question digressed as the teachers were concerned about the research design that led to the conclusion stated in the slide. The only evidence about their understanding of probability (or risk) was provided towards the end by Linda:

Excerpt 9

65. Terry But apart from the design of the study. Apart from whether there was a control group and all this stuff, just talking about when they talk about overall risk or lower risk, what does that mean?
66. Linda Less occurrence of breast cancer out of the total number.

Linda's interpretation of risk again revealed her conception of probability as relative proportions.

3	Gustav's Bad Luck	Gustav read in Newsweek magazine that drivers with three or more speeding tickets are twice as likely to be in a fatal accident as are drivers with fewer than three tickets.	What does this mean?
		The very next day he received his third speeding ticket.	What does this imply?

Excerpt 10

68.	Terry	Ok, this one I think we've already talked about. This is the one that Pat brought up last week. **'**Gustav reads in the Newsweek magazine that drivers with three or more speeding tickets are twice as likely to be in a fatal accident…Ok. What does that mean? What does that statement mean, that people being killed in wrecks have lots of speeding tickets?
69.	Sarah	But that's not true, but that's an answer you would get from a kid.
70.	Linda	The same number of drivers that had 3 or more speeding tickets, compared to the same number of drivers that had fewer than 3, there was a greater number of those that were killed, in the first group.
71.	Terry	Ok, when you say 'greater number'?
72.	Linda	Out of the same proportion, there's a greater proportion, but if you look at the same number, like 10,000 drivers of each group, there would be more that were killed in the top group than the bottom group.
73.	Terry	How many more?
74.	Linda	Um…Twice as much.
75.	Terry	Twice as many, yeah. That's one of those things where you can kind of read that and get a feeling of what that's telling you, but not really think about what it means—Does everybody agree with Linda's interpretation?
76.	John	Mmhmm Yeah.
77.	Terry	Ok. "The very next day he received his third speeding ticket. What does this imply?" Lucy, you're shaking your head, what does that imply?
78.	Lucy	[Inaudible]
79.	Terry	Ok. Does it say anything about…Gustav in particular?
80.	Henry	It says now he was a member of that second population.
81.	Terry	OK. Yeah, now he's in that other, as Linda, she said 10,000, now he's one of that=
82.	Linda	Well actually the number just became 10,001, but…His being in that group doesn't really tell us=
83.	Terry	Right. So now he's in the other group where twice as many of those are going to die, as the group he used to be in.

The highlighted answer to the first question in this slide given by Linda suggested that she conceptualized a collection of people (having *x* number of tickets) and a portion of that collection (of people who were in fatal accidents), and she interpreted likelihood (of

expected outcome) as the relative proportion of outcomes like this one in a group of all possible outcomes. Henry and Linda's answers to the second question also suggested that they understood the idea that likelihood/probability is not about a particular person, but rather it is a characteristic of a group.

Activity 2-2, Episode 2: Slide 4

4	Car and weather situations	What is the probability that my car is red? What's the probability that the temperature tonight is below 40? What does that mean?

The discussion around these two questions lasted for about 17 minutes. It is parsed into three segments. During the first and second segments, the teachers discussed their interpretations of the two situations respectively (hereinafter will be addressed as car situation and weather situation). During the last segment, the teachers argued about whether both situations could be interpreted in multiple ways, i.e., both stochastically and non-stochastically.

Segment 1: Car situation

Excerpt 11

87. Terry What is the probability that my car is red? [Stutters] let's say we're talking about *my* car.
88. Sarah What color is it?
89. Terry I'm not going to tell you.
90. Linda It is either 0 or 1.
91. Terry It is either 0 or 1, how come?
92. Linda Because there's only two outcomes.
93. Nicole That's a good example. That's a great question!
94. Terry Pardon?
95. Linda There's only two outcomes.
96. Terry Right.
97. Linda It either is or it isn't.
98. Terry Right. Either my car is red=
99. John But that's not what it is asking. But what's the probability of it being red, that's what we're asking. We know there're two outcomes, red or not red, but the probability of it being red is a different question.
100. Terry Yeah. So?

101.John So it can't be what I said, about the rain. Either it rains or it doesn't. So we don't know how it is=

102.Linda But this is a sure thing, though. This isn't dependent on any external occurrant things at all. You buy a car, it is either… The probability's 1 if you bought a red one and the probability's 0 if you bought a blue one.

103.Terry Right. This is, let me see what the next little thing is—

104.Sarah Well, wouldn't the probability of buying a red car have to do with how many red cars out of how many cars were on the lot?

105.Terry But is that the same thing, though? Is that the same thing to say what's the probability of buying a red car? Is that the same thing as saying what's the probability of *my* car that I own right now is red?

106.Henry Not quite.

Linda believed that the probability is either 1 or 0. In her thinking, there was one particular event (color of my car) with two possible outcomes (red and not red). Thus, the probability that "my car is red" is 1 when the outcome is red, and 0 when the outcome is not red.

John's response (lines 99 and 101) to Linda was both surprising and confusing. It was surprising because although he agreed with Linda that there were two outcomes, red or not red, he did not think the probability was "either 1 or 0", or 50%. He saw a distinction between the car situation and the one about the chance of rain where he argued that the chance was 50% because "either it rains or it doesn't". It was confusing because I do not know what distinction he saw, and what he meant when he said, "probability of it being red is a different question".

Sarah's comments in line 104 suggested that she attempted to interpret the situation in terms of relative proportion. That is, the probability of buying a red car is the relative proportion of the red cars out of all cars [that are for purchase]. Terry, the seminar host, dismissed this interpretation on the ground that the situation under discussion was about *my car*. Terry took a strong stance during the ensuing discussion.

She emphasized the phrase *my car* and insisted that the car situation was about a single event/object, and argued strongly that it did not make sense to talk about probability in this situation.

Excerpt 12

107.	Terry	Does it even many sense to talk about that as a probability?
108.	John	No. It doesn't make sense.
109.	Terry	No, it just doesn't even really make sense to talk about it because think about probability as repeating a process, and one of the characteristics being that we don't necessarily know what's going to happen, but once you go out there and see what color my car is, you know. It is not until I buy, you know, if I don't buy a new car, it is going to be the same color. There's no unknown there.
110.	Henry	There's a predictive sense, if you want to refine and play with it. If someone was trying to guess if you car was red, and you gave him information like well, my car is a sports car and if you knew that how many percents of all sports cars made were red, then that'd be a prediction. [mumbles something] Whether it is inclusive in that group=
111.	Terry	But would that even…But does that change what your answer would be?
112.	Henry	No, I'm not arguing with this, I'm just saying "Is there any point in asking this question?" I said there could be a point, if you wanted to try to make some sort of a prediction. Again, going back to what Sarah said about how many red cars are made out of the total population of cars, or, specifically, if you have a sports car…You see, I like to use that word 'likely' I'm trying to find another word…
113.	Henry	But if you have a sports car, there's a better chance of it being red=
114.	Nicole	Chance is on the can't do list. You can't say chance=
115.	Henry	Right. I can't use that either.
116.	Sarah	How about aptness.
117.	Various	[laugh]
118.	Terry	You can find all the synonyms you want. But the notion, if you think about this as a probability situation where, you know, if you were going to investigate this, what would you do?
119.	John	Randomly sample cars=
120.	Linda	You would just go out and look at the car.
121.	Terry	You'd go out and look at my car.
122.	Henry	[laughs]
123.	Terry	And once you've looked at it you could just say, well, my car is, in fact, not red. You'd say, well, her car wasn't red. But it doesn't really make sense because then, does it mean anything then to talk about probability? Well, what's the probability that my car is red? Well, it is=

124.Sarah Have you ever owned a red car?

125.Terry Yeah. I have had a red car, but my car right now is not red. So the probability that my car is red is 0, and then it doesn't really even make sense to talk about that, because as long as I don't buy a new car or get my car painted; if we work under those assumptions and everyday you go out and look at my car, you know it is not going to be red. You know, if we're working under the assumptions that I'm not buying a new car and I'm not going to paint it. So it doesn't even make sense in terms of thinking of it as a probabilistic…

Henry objected to Terry's claim that it did not make sense to talk about probability in this situation. He argued that there could be a way to see the situation so that talking about probability would make sense. The situation he conceived of was that of a person predicting the car color using the statistical information about the population of all the cars (or sports cars). Taken into account of John's comment "randomly sample cars", Henry's way of thinking in fact was a stochastic interpretation to the car situation. However, Linda and Terry opposed to this interpretation. For them, the repeatable process was "to go out and look at my car". Because this process did not produce variable outcomes (i.e., the outcome was already known, being either red or not red), Terry insisted that situation was not probabilistic.

Segment 2: Weather situation

The second question, "What's the probability that the low temperature tonight will be below 40 degrees?" was very similar to the probabilistic statement the teachers had discussed earlier, "Today there is a 40% chance of rain." (A1-2 in Day 1). Nicole, who had interpreted that statement stochastically, had a similar interpretation to the weather situation. However, as we will see in the following excerpt, when Sarah proposed a different stochastic interpretation (different in the sense that Sarah specified a particular

collection where Nicole vaguely expressed as "over a long period of time"), Nicole questioned her own interpretation.

Excerpt 13

131.Terry	What's the probability that the low temperature tonight will be below 40 degrees? What does that mean?
132.Nicole	Well, don't you think that it means that if we had conditions just like today, and the weather conditions generally around here over a long period of time, what's the probability that we'll get a temperature below 40 degrees?
133.Sarah	Or does it mean that on June the 18th, of all the June 18ths how many have we had that were below 40 degrees?
134.Nicole	That's right, is it tonight or is it all of these nights?
…	
139.Nicole	Anyway, it is zero for tonight, folks.

Sarah's mention of "June the 18th"(the day of the discussion), despite the fact that she was thinking of many years of June 18th, made Nicole realize that there could be two ways of interpreting this question. One is non-stochastic, i.e. thinking of the situation as being about one particular event: "The temperature *tonight* is below 40". The probability in this case is, as she said, "It is zero for tonight." The other way of interpreting the question is stochastic, i.e., imaging looking at the night temperature over an extended number of June 18ths.

Segment 3: The debate

The topic of discussion in this segment stemmed from Henry's observation. That is, the weather situation was interpreted in two ways. One way was to say: tonight temperature is or isn't going to be below 40 degrees; another way was to look at historical data about temperatures of days like today. But the only acceptable interpretation for the car situation was "it is either red or isn't red." Henry raised the question, i.e., essentially, "If there were two ways of interpreting the weather situation (stochastically and non-stochastically), why couldn't the car situation be interpreted stochastically?" (Excerpt 14)

Excerpt 14

147.Henry I was looking at those last two examples…I felt really good about both of those, but then I started sitting here and thinking in circles about what we're doing, but those were supposed to be two different examples, right?

148.Terry Uh huh.

149.Henry Because the one at the bottom I was looking at and saying 'Well, it either is or it isn't going to be less than 40 degrees tonight' and if you start making a prediction based on historical facts, then why can't, in the car example, you say 'Well, you've owned 5 cars in your life and four of them have been red, so… the probability of you owning a red car now is 4/5'.

150.Sarah [Because you can't divide red cars?].

151.Nicole Yeah, I mean I agree with him that the two questions are ambiguous because in one situation we're talking about *my* car and in the other situation we're downplaying the word *tonight.*

152.Terry Mmhmm.

153.Linda One has a large number of nights to base your---your statement about tonight, you can base it on a large number of previous nights, but if you talk about your car—if I said something about my car, I wouldn't base it on anything that I've ever owned before. I mean…

Nicole offered an explanation to Henry's observation: for the car situation the teachers fixated on *my* car, therefore the situation was about a particular event and thus non-stochastic; for the weather situation the teachers "downplayed" the word "tonight", therefore the situation could be about "a collection of nights" and thus was seen as stochastic. However, instead of questioning the criterion with which they chose to fixate or downplay particular words that led to different conceptualization of the situations (by asking, e.g. "Why did we only talk about *my* car in the first situation, but downplay the word *tonight* in the second? What was our rationale? Why did we interpret differently two situations that seemed very similar?"), she commented that the questions were ambiguous, and thus suggested that it was the ambiguity of the questions as they were stated that resulted in the differences in the interpretations. This suggested that she might have believed that a representation of a situation should *dictate* how a situation should be

interpreted. In her mind, there was no place for the person who is doing the interpretation. There was no distinction between a situation *as it is represented* and a situation *as it is interpreted*. An unambiguous situation or problem must have one and only one meaning (no matter who is interpreting it).

Linda argued against Henry and Nicole's claims. She did so by talking about what these two situations were about (to her), or how (she believed) they should be interpreted. She claimed that the questions were not ambiguous. The two situations were clearly different: The weather situation was stochastic and the car situation was non-stochastic. She did not, however, give any justification to her claims, as if they were entirely obvious and unproblematic. She did exactly what Nicole said they did: fixating on *my car*, and downplaying the word *tonight*, without much reflection of what she did. This could have meant that Linda, like Nicole, did not have in mind the person (in this case, herself) who was doing the interpretation. In other words, she was so wrapped around her own ways of thinking that she believed they were the only possible ways to think about the situations. She did not reflect on her own thinking: "how do *I* think about it, and why?" Rather, she believed the group was talking about "how to think about it" (implying that there was only one way to think about it) and she insisted that her way of thinking to be the only acceptable one.

The discussion continued as more teachers participated and took side in the debate. John and Nicole supported Henry that the car situation could also be interpreted stochastically and that these two situations were the same in the sense that they could both be interpreted in two different ways. Linda and Terry insisted that the car situation had to be non-stochastic.

Excerpt 15

154.	Henry	I've got a friend who's owned 50 cars. What's the probability=
155.	Betty	But I guess you could look at it, well…[microphone interference] the buying of the next car or the present car, is it based upon your actions from [above?].
156.	Henry	I've just=
157.	Betty	But I understand too.
158.	Henry	[inaudible] separate those two question for a little more clarity=
159.	John	I see. You've made a good point, Henry. Those are exactly the same question. Because, if the weatherman had to make a forecast about what kind of color your car would be, the weatherman would have the model go and look at the previous cars, just like he goes and looks at previous weather situations, so predict what tonight is going to be like.
160.	Linda	But there's a question about tonight. There's no question about what kind of car she has.
161.	John	I think those are the exact same question=
162.	Linda	There is no, no question. She either has a red car, or she doesn't=
163.	Terry	Ok, the language is—If you take, yeah, I see what Linda said—Do we know what the low temperature's going to be tonight.
164.	Linda	No.
165.	John	No, not until we get to tonight!
166.	Terry	Ok, do we know what it is going to be. So, and what we might think the low temperature is going to be right now, might be different from what we think the low temperature will be 6 hours from now.
167.	Linda	Maybe the question should be is it determined now. Whether we know it or not is really not the issue, it is is it determined?
168.	Terry	Right. Is it determined what the low temp—that's a good way to put it. It is determined what color my car is. There's no chance there. My car is a certain color. It is not going to change color when I walk out there. Whereas, it is not predetermined, at this moment, what the low temperature's going to be. So there is some chance involved there. So there's an element of—I can't predict what the low temperature is going to be.
169.	Henry	So we're talking about outcomes, then. We're saying that there's the possibility for different outcomes.
170.	Terry	Exactly. There has to be a possibility—and we're going to get to that in a second, but there has to be an unknown outcome. There has to be a possibility of different outcomes. If there's no unknown there, then it doesn't make sense to talk about probability. I mean, you can, you can say it is 1 or 0, but it doesn't really make sense to do that.

In this excerpt, we continue to see how Linda took her own thinking as unproblematic, and did not consider the alternative ways of thinking suggested by Henry and John. Terry also continued to defend her belief that the car situation was not stochastic, and the

weather situation was stochastic. Terry's justification was that while the outcome in the weather situation—the temperature tonight is below 40 degrees— was unknown, the outcome of the car situation—the color of my car—was known (it is either red or not red), and to Terry, a situation is probabilistic only when there are unknown conditions that will present possibilities of different outcomes. Linda modified Terry's justification from whether the outcome was *known* to whether it was *determined* (no matter if it was known). The outcome of the temperature tonight was yet to be determined, but the color of my car was determined. Therefore the car situation was not probabilistic. Terry and Linda's justification revealed two conceptions of probability: 1) a situation is probabilistic only when there are unknown outcomes, and 2) a situation is probabilistic only when the outcomes are yet to be determined.

What really accounted for the difference in the two groups' opinions, from my perspective, was not whether the outcomes were unknown or determined. It was, rather, hinged on the question: What was the outcome? To Terry and Linda, the outcomes for the weather situation were the possible temperatures for either tonight, or a collection of nights, but the outcomes for the car situation were "red" or "not red". The epistemological difference, it seems, is that in one case they anticipate variability (temperature) and in the other case they do not (color of car). In the former, they would be asking the same question about *different nights (or different time tonight)*. In the latter, they would be asking the same question about *the same car*. The other group, Henry and John, had a different point. The outcomes of weather situation could be either "below 40" or "above 40" when the situation was seen non-stochastically, which was parallel to the non-stochastic interpretation of car situation. By the same token, the car situation could

be seen from a stochastic perspective, in which case, the outcomes were colors of a collection of cars (previously owned by a person).

Linda and Terry's resistance to see the car situation from an alternative perspective indicated that they might not have seen a distinction between a representation of a situation and its interpretation. Linda and Terry conceived of car situation non-stochastically, stood firmly on their conviction that "the situation is not probabilistic." This conviction can be contrasted by "*my conceptualization of* the situation is not probabilistic." This suggested that for Linda and Terry, a situation either is stochastic/probabilistic or it is not. They did not realize that it was their conceptualization of the situation that made them see it as stochastic or non-stochastic. Once they were committed to their conceptualization, they resisted the alternative conceptualization (hence, interpretation) offered by Henry, John, and Nicole. I conjecture that there might be two possible explanations: 1) Linda and Terry implicitly assumed that any problem should have one correct answer and thus the situation it depicts must have one legitimate interpretation. As a result, the interpretation they embraced were the only interpretations and those who interpreted differently were wrong and thus did not deserve consideration, and 2) they might have accepted that situations could be interpreted differently but they were so ingrained in their own way of thinking that they were unable to entertain alternative ways of thinking. In either case, there was a lack of critical thinking or reflection on their own understanding, and a lack of orientation in trying to understand other people's understanding.

It was not until Nicole proposed another analogous situation that Terry began to differentiate a situation from its interpretation and realize that indeed the car situation could be thought of differently.

Excerpt 16

171.	Nicole	What about that guy that Pat was talking about, the doctor says to him you have 70% chance of surviving this cancer. And the guy says, no, it is either 0 or 1. Well maybe in the doctor's mind there really is some ambiguity in the unknown there. You see what I'm saying? Maybe there really is!
172.	Henry	So is that problem more like the car problem or more like the weather problem?
173.	Nicole	It is like the car problem=
174.	Terry	I think it is more like the car problem, but…Well, there is some unknown there, yeah, but it is not that there's not an unknown there. It is the way that he phrases the 70%. It is not that that person has a 70% chance of surviving, it is that of all people who have his type of cancer, 70% of those people survive.
175.	Terry	It is kind of like the Gustav problem. He's in this particular group where 70% of this group survive. It doesn't really talk about him in particular.
176.	Nicole	Right. I understand that, but you were trying to distinguish between the car problem and the=
177.	Terry	=Well I only have one car=
178.	John	=So it is like the weather problem!
179.	Terry	There's no group of cars. I have one car. Now, I guess, if I had several cars…
180.	Nicole	But if you only have one life, I presume you do, not being a cat, then it is like the car problem, you saying=
181.	Terry	=I don't understand why you're saying that those—What the doctor said…I'm not sure what you're saying is—What the doctor said was incorrect. You're right that the man living or dying is like the car problem.
182.	Nicole	Yeah.
183.	Terry	But then you're=
184.	Nicole	=But then you said something about=
185.	Terry	=But then if you=
186.	Group	[laughing and talking over one another]
187.	Terry	So if you conceive that guy as a member of a group of people, then you can talk about the proportion of that group of people. But if I have more than one car, you know, if I'm, like, a football player or a boxer or something and I had 10 cars, then we could talk about the probability that my car was red, because—What's the probability of the car that I

drove today…Now you turn it into something where there is some chance.

The situation Nicole proposed in line 171 was borrowed from an earlier discussion where Pat raised it to illustrate the idea that people could have different interpretations to a single situation. In this situation, a patient took a non-stochastic perspective to his "chance of survival", i.e., he either survive or he doesn't; the doctor, on the contrary, took a stochastic perspective, i.e. the chance of this patient's survival is the relative frequency of survivors out of all patients having the same diagnosis. The group had discussed this situation during the first week of the seminar, and it was understood that both interpretations were reasonable and that the difference in the interpretations was a result of different perspectives one held. By bringing back this situation, Nicole demonstrated its resemblances to both the car and weather situation, which made Terry finally understand the point that Henry, John, and Nicole were trying to get across—that the situations under discussion could be interpreted differently, depending on how one conceives of them.

The discussion during this segment established the idea that was coined in slide 7: The situation per se is not probabilistic. It is how you conceive the situation that makes it probabilistic.

Activity 2-2, Episode 3: Slide 6

6	Rishad's situation	Rishad's sister, Betty, rolled ten sixes in a row while playing a board game.	Rishad: "That is impossible! Rishad: "If a billion people rolled a die 10 times, what fraction of them would roll all sixes?"

In slide 6, Rishad's situation was presented together with its two interpretations. The first interpretation regarded the occurrence "Betty rolled ten sixes in a row" as a single

unrepeatable occurrence, while the second interpretation placed the occurrence in the context of a repeatable event "rolling a die 10 times". Our purpose of presenting these two interpretations was to clarify the distinction between two different perspectives in conceiving of a situation: a non-stochastic perspective and a stochastic perspective. Because of the discussion of slide 4 occurring prior to this slide, seeing this distinction was unproblematic for them here.

Excerpt 17

213.Terry Yeah, so in the first situation he may or may not be thinking about that in terms of probability, he's thinking about it in terms of a feeling, that he's never seen that before or…you know. Whereas in the second one, he's actually conceiving of it as a repeatable process and actually, you know, looking at what would happen in that circumstance.

…

225.Terry All right, so, the difference being that Rishad was thinking about the first one, just one—looking at one particular outcome. That's not really probabilistic, whereas you think about repeating a process and looking at what happens in the long-run. That's conceiving it as a probabilistic situation, ok? And it is not the situation *itself* that's probabilistic, this the way you think about it that makes it probabilistic, ok? So it is sort of getting back to what you said about the car thing and the weather thing. It is not the situation itself, necessarily, but it is how you *conceive* of it that either makes it probabilistic or not probabilistic. Ok?

In this excerpt, Terry interpreted Rishad's situation, connecting it back to the car and weather situation, and pointed out that big idea that "It is not the situation *itself* that's probabilistic, it is the way you think about it that makes it probabilistic". This idea bring to the forefront the imagery of "a person who is interpreting, thinking, conceiving" that was lacking in her and Linda's thinking earlier in the discussion of car situation.

Summary of Activity 2-2

To summarize teachers' conceptions of probability revealed in this activity, I coded the teachers' interpretations of the probability situations in the slides using the theoretical framework of probabilistic understanding. The codes are presented in the table below.

Table 19: Teachers' conceptions of probability situations in Activity 2-2

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D5A2-2S1	Linda	Y	N							Y	
D5A2-2S2	Linda	N								Y	
D5A2-2S3	Linda	N								Y	
D5A2-2S4Q1	John	Y	Y	Y							Y
	Nicole	Y	Y	Y							Y
	Linda	N				Y					
	Henry	Y	Y	Y							Y
	Terry	N				Y					
D5A2-2S4Q2	John	Y	Y	Y							Y
	Nicole	Y	Y	Y							Y
	Linda	Y	Y	Y							Y
	Henry	Y	Y	Y							Y
	Terry	Y	Y	Y							Y

As we can see from the table, the discussion around the first three slides revealed Linda's conception of probability as relative proportion. Linda did not conceive of well-defined stochastic processes for any of these situations. Therefore her conception of probability was non-stochastic.

Discussion around slide 4 showed that, while one group of teachers (Henry, John, and Nicole) realized that a probability situation could be interpreted in both ways, another group (Terry and Linda) insisted that the two situations in slide 4 must be interpreted differently. Terry and Linda were so committed to their own interpretations of the situations that they resisted alternative interpretations provided by the other group. They did not understand, initially, that there was a distinction between a situation and its interpretations, and that it was the interpretations that made the situation stochastic or non-stochastic, not the situation itself. Their resistance to entertaining alternative interpretations could be attributed either to their lack of orientation in understanding others' thinking or to their lack of reflection on their own thinking.

Eventually, all teachers came to understand that situations could be interpreted differently if one were to conceive of it differently. This result was achieved by clarifying the distinction between non-stochastic and stochastic perspectives through presenting situations (slide 6, and Nicole's example in segment 3 slide 4) with reasonable interpretations from both perspectives.

Interview 3-1: Five probability situations

The following Post-Interview question investigated 1) teachers' interpretations of probability situation, and 2) the extent to which teachers were aware of multiple interpretations of probability situations.

> Consider this list of statements:
>
> 1. What are the chances that it will snow in Billings, Montana on April 23, 2002?
> 2. What's the probability that the length of Dean Benbow's driveway is between 29' and 30'?
> 3. Your risk of being struck by lightning is 1 in 400,000.
> 4. How likely is it to be dealt one pair in a 5-card hand from a standard deck?
> 5. What's the probability that you are off by no more than 2" when you measure the length of your driveway?
>
> Tell me about the underlying situation that is being referred to in each statement.

Using the same method, I coded the teachers' interpretations of these probability situations (See Table 20).

Table 20: Teachers' conceptions of probability situations in Interview 3-1

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
I3-1Q1	John	N				Y					
	Nicole	Y	Y	Y							Y
	Sarah	Y									
	Lucy	N									
	Betty	Y	Y								
	Linda	Y	Y	Y							Y
	Henry	Y	Y	Y							
	Henry	N				Y					
	Alice	Y	Y	Y							
I3-1Q2	John	N				Y					
	Nicole	Y									
	Sarah	Y				Y					
	Lucy	N				Y					
	Betty	N				Y					
	Linda	N				Y					
	Henry	N				Y					
	Henry	Y	Y								
	Alice	Y	Y	Y							Y
I3-1Q3	John	Y	Y	Y							Y
	Nicole	N			Y						
	Sarah	N									
	Lucy	N									
	Lucy	Y									
	Betty	Y	Y	Y							Y
	Linda	Y	Y	Y							Y
	Henry	Y									
	Alice	Y	Y	Y							Y
I3-1Q4	John	Y									
	Nicole	Y	Y								
	Sarah	Y	Y	Y							Y
	Lucy	Y									
	Betty	Y	Y	Y							Y
	Linda	Y	Y	Y							Y
	Henry	Y	Y	Y							Y
	Henry	N				Y					
	Alice	Y	Y	Y							Y
I3-1Q5	John	Y									
	Nicole	Y	Y	N							
	Lucy	N									
	Lucy	Y									
	Betty	N				Y					
	Alice	Y	Y	Y							Y

If we look at the interpretations of each question, we find that the teachers consistently interpreted statement 4, the situation of "dealing 5 card" stochastically. The majority of them also interpreted statement 1 and 3 stochastically. Most of the teachers conceived of statement 2 non-stochastically. Only two teachers, Lucy and Henry, interpreted several situations in both ways (highlighted in red, Lucy 2 out of 5, Henry 3 out of 4).

Looking at interpretations across situations given by each teacher: Nicole, Linda, and Alice seemed to have an orientation in conceiving of probability situations stochastically. John, Sarah, Lucy, Betty, each conceived of two situations non-stochastically. Most teachers made statements about whether the situation was or was not probabilistic, as opposed to whether they conceived of the situation as probabilistic. Henry was the only one who consistently tried to interpret all situations in two ways. He acknowledged that he had difficulty seeing the third statement (risk of being struck by lightning) from a non-stochastic perspective. Perhaps it was because had he done so, then the probability would be either 0 or 1, which was in conflict with the given measure of probability1/400000.

Interview 3-4: Gambling

You must make a choice between:

a. Definitely receiving $225,
b. A 25 percent chance of winning $1,000 and a 75 percent chance of winning nothing.

1. Suppose this is a one-time choice. That is, you are presented with these options once and you will never be presented with these options again. Would you choose (a) or (b)? Why?

2. Suppose you are a gambler who will be presented with these same options many, many times. Would you choose (a) or (b)? Why?

This interview question resembles the slide 6 *Rishad's situation* in the PowerPoint presentation in that the distinction between non-stochastic and stochastic interpretations was made clear in the question. Teachers' answers to both questions revealed that they understood the distinction between making a one-time choice and making repeated choices, and they understood that consistently choosing *b* would yield more benefit over the long run. Four teachers made their choices based on this understanding (Table 21). The other four teachers, while understanding the distinction, nevertheless made their choices based on their personal preference of avoiding risk. It is worthwhile noting that Henry chose to accept risk even in the one-time situation. Perhaps he thought of this one-time situation as one of many one-time situations.

Table 21: The choices teachers made in Interview 3-4

Name	1	2
John	*a*	*b*
Nicole	*a*	*b*
Sarah	*a*	*a*
Lucy	*a*	*a*
Betty	*a*	*a*
Linda	*a*	*b*
Henry	*b*	*b*
Alice	*a*	*a*

Summary

Discussion of the two situations in Activity 1-2: *Chance and Likelihood* revealed that, on that occasion, the majority of the teachers had a non-stochastic conception of probability situations. Only two teachers held a stochastic conception when interpreting the first situation. There were three ways teachers interpreted probability non-stochastically: 1) Linda and Nicole's outcome approach, 2) John, Lucy, Betty and Linda's proportionality

heuristic, and 3) John and Linda's belief that probability of an outcome is 50% since an outcome either occurs or does not occur. My analyses of these interpretations in the context of the teachers' discussion revealed that while 1) and 2) both led to subjective judgment of probability, 3) was logically incoherent. Note that with exception of one instance (Nicole's outcome approach in interpreting one of the questions in the post-interview), none of these interpretations appeared in later activities and interviews.

Discussion on the first three slides of Activity 2-2: *PowerPoint presentation* revealed Linda's conception of probability as relative proportion. Evidence (e.g., consistently conceiving of incompatible collections for questions in slide 1) suggested she did not conceive of a clearly defined stochastic process, in which case, the relative proportion of observed outcomes out of all outcomes generated from this process. Therefore, I argued that Linda had a vague method of computing probability: probability is a "relative proportion" of quantity *x* over quantity *y*. Her conception of probability was non-stochastic.

Discussion on slide 4: *Car & weather situations* highlighted an idea that is pedagogically significant. That is, *a situation is what you conceive of it to be. It can be thought of differently*. Discussions showed that while one group of teacher consisting of Henry, Nicole, and John argued that the two situations could be interpreted both non-stochastically and stochastically, another group consisting of Terry and Linda insisted that one has to be non-stochastic and the other stochastic. Note that from this point on, teachers' conceptions of a probability situation hinged on Q1: Is there an image of a repeatable process? The interpretations that they offered were streamed into the following two types:

1) Non-stochastic (path 1-5): There is no repeatable process; an event either occurs or does not occur; probability is either 1 or 0.
2) Stochastic (path 1-2-3-10): There is an underlying repeatable process; probability is the relative frequency of the observed outcomes generated from the process.

At the end of the discussion of the PowerPoint presentation (end of discussion on slide 4: *Car & weather situations*, discussion on slide 6: *Rishad's situation*), there seemed to be a shared understanding in the group that a situation is not probabilistic in and of itself. It is how one conceives of it that makes it probabilistic.

In the post-interviews, teachers' interpretations to the five probability situations revealed that most of the teachers only interpreted the situations in one way. Only two teachers, Henry and Lucy, gave both non-stochastic and stochastic interpretations in a number of occasions (Henry 3 out of 4, Lucy 2 out of 5). We observed that while for some situations (e.g. question #2) most of the teachers interpreted it non-stochastically, for some (e.g. question #4) most of the teachers interpreted it stochastically, and yet for the rest, some teachers saw them non-stochastically, and others stochastically. Although we could explain some of these inconsistencies locally (e.g., the playing cards situation in question #4 is culturally understood to be repeatable), more evidence is needed to understand more about what drove the teachers to interpret probability situations non-stochastically or stochastically.

The second interview question *Gambling situation*, like the *Rishad's situation* in the slide 6 of the PowerPoint presentation, showed that the teachers did see the difference between the non-stochastic and stochastic conceptions whenever we made the distinction clear by juxtaposing two interpretations that embedded these two conceptions.

Activity 2-4: Clown & Cards scenario

> At the Cobb County fair a clown is sitting at a table with three cards in front of her. She shows you that the first card is red on both sides, the second is white on both sides, and the third is red on one side and white on the other. She picks them up, shuffles, hides them in a hat, then draws out a card at random and lays it on the table, in a manner such that both of you can see only one side of the card. She says: "This card is red on the side we see. So it is either the red/red card or the red/white card. I'll bet you one dollar that the other side is red."
> What is the probability that you would win this bet were you to take it?
> Part 1 - Discuss how you are thinking about this situation in order to formulate an answer to the question.
> Part 2 - We will watch video excerpts of students attempting to make sense of this situation. Discuss what they seem to struggle with.
> How might we help students reason about the situation in a way that is coherent and consistent with ideas of repeated sampling and probabilistic situations we have been discussing?

The last section revealed the teachers' stochastic and non-stochastic conceptions of probability situations across a number of activities and interview questions. In this section (activity), I will focus on multiple interpretations of probabilistic situations. [Note that I use the phrase "probabilistic situation" as opposed to "probability situation". It means that the section is about multiple interpretations of a situation when it is conceived of probabilistically.] As I have elaborated in Chapter 5, not only a non-stochastic conception has many different expressions, a stochastic conception, too, could lead to different interpretations of a probability situation. In other words, *a particular event could be seen as outcomes from different stochastic processes, and thus the probability of this event differs depending on how one conceives of the stochastic process*. The Clown & Card scenario presents such a situation that subject to many different interpretations. Our purpose in engaging the teachers in this activity was to understand the ways with which the teachers responded to this type of situations.

Discussion around this activity lasted for 138 minutes in total: 100 minutes before the break and 38 minutes after the break. Teachers' initial interpretations could be divided into three categories that led to three different answers to the original question: "1/2", "1/3", and "either 1 or 0".

Activity 2-4, Episode 1: Multiple interpretations

Interpretation 1

Betty, Sarah, Lucy, Henry, and Linda believed that the probability was 1/2.

Excerpt 18

20.	Betty	Okay, I took it from the two cards that were on the table, I mean, from the two cards that were red, that had red on them, not from the white-white because that seemed to be out of the picture. So I took those two possibilities, so it is either going to be on the other side, red or white, so that's, to me, one out of two.
…		
33.	Terry	What is the probabilistic situation? What is the process that is being repeated?
34.	Betty	Well it is— to me, as it is, maybe I messed up=
35.	Terry	Lucy?
36.	Lucy	Flipping the card that is right in front of you, flipping it up.
37.	Terry	Okay, so that's the repeatable process…
38.	Betty	Every time you drew a red card.
39.	Henry	That card only=
40.	Terry	=Just the one card that's=
41.	Henry	='Cause if it is the ah, white… well, I don't (inaudible)
42.	Sarah:	From either of the cards that would give you …
43.	Henry	I wasn't looking at it though from any of the three cards, I was just looking at it from the, the two that would be red or white.
44.	Sarah	From just that one card.
45.	Betty	To me that is the only thing that is being repeated – rather than drawing from the basket to start with or whatever…or the sack whatever.
46.	Terry	All right, Linda you were going to say something.
47.	Linda	It would be like drawing from a bag of cards that are red/red or red/white.
48.	Terry	So what are you saying there, are you agreeing that the probabilistic situation is flipping that card over?

49. Linda Yeah, it is the same thing, or, or having a bag of red and—red stuff and white stuff, because the issue is…that the issue is not the side facing you, you already know what that is.

50. Sarah Did you=

51. Linda =So you're down to picking between, picking things from a sample space of red vs. white.

…

631. Sarah Well, but I know, I mean, Betty was looking at this one card and we're looking at—and I'm thinking like Betty is – we've got two things going on here, you asked Betty what she *thought*, when I thought about it I thought about the way she did: you've got a card laying there that's red, it is been drawn out and to be red up it either has to be red or white on the bottom, there's only two outcomes, you flip it over, it is a one out of two possibilities on that particular part, but if you want it stated probabilistically…

There were two important assumptions made by this group of teachers. First, they believed that since a card was already drawn and a red side was up, the white/white card should be eliminated from their consideration, because had it been drawn, it could not have rendered the outcome of a red side facing up. The second assumption, more tacit than the first one, was that it did not matter which side of the card was visible. Only the color of the visible side mattered. As Linda said, "the issue is not the side facing you, you already know what that is."

Table 22: Outcomes conceived of by Betty, Lucy, Sarah, and Linda

	Up	Down
1st	R	R
2nd	R	W
	~~W~~	~~R~~
3rd	~~W~~	~~W~~

There were two ways the teachers conceived of the situation: One group of teachers, Betty, Lucy, and Linda, conceived of a stochastic process; Sarah conceived of the situation non-stochastically. The stochastic process that Betty and Lucy conceived was

"flipping the card whenever there is a red card comes up" (lines 36 and 38). Linda equated this process to that of repeatedly "drawing from a bag of cards that are red/red or red/white." In other words, were Linda to draw the R/R card, the other side would be red; were she to draw the R/W card, the other side would be white. Sarah saw the situation as: There was one card with red on top, the bottom was unknown but there were two possibilities, red and white (line 63). Therefore the probability (to Sarah) was the relative proportion of these two possible outcomes. In Sarah's conception of the situation, there was no image of a stochastic process. To summarize, here we saw the teachers giving the same interpretation to probability (as relative proportion of observed outcomes out of all possible outcomes) with Lucy, Betty, and Linda having a stochastic conception (path 1-2-3-9), and Sarah having a non-stochastic conception (path 1-9).

Interpretation 2

Nicole gave a different answer: The probability of winning of the bet was 1/3. Nicole had conceived of a two-stage sampling process. This process was "Pull a card from the bag, place one side on the table, and look at the other side of the card". Table 23 illustrates how she thought about the outcomes of this process:

Table 23: Outcomes conceived of by Nicole

Card	Side A	Side B
R/R	Up (R)	Down (R)
	Down (R)	Up (R)
~~W/W~~	~~Up (W)~~	~~Down (W)~~
	~~Down (W)~~	~~Up (W)~~
R/W	Up (R)	Down (W)
	~~Down (W)~~	~~Up (R)~~

Nicole evidently believed that picking a card randomly from the bag and placing it on the table produced 6 possible outcomes of the color of the upper side. There were two ways of placing the red/red and white/white cards: side A up or side B up. Since it was known that a red side faced up, three outcomes that had white on the upper side would be eliminated, hence the crossing out of WW, WW, and WR. Next, out of the three possible outcomes that had red on the upper side, only one had a white on the other side. Therefore, the probability was 1/3. Note that Nicole's conception of the situation was stochastic, and probability was the relative proportion of observed outcomes out of all possible outcomes (path 1-2-3-9). It is also important to note that Nicole focused on the outcomes being pairs of sides, one side up and one side down, and not simply on the color of the upper side.

Nicole's explanation in the following excerpt:

Excerpt 19

145 Nicole Why can't the clown, also say – it doesn't say this here – but why can't the clown just pull out a card and slap it down and let's assume it is white this time, and now the clown says, um… what's the probability the other side is white?

suggested that the situation she conceived of was " The clown plays the game repeatedly, drawing a different card, and making the bet with reference to whatever color is up." This conceptualization of the situation allowed her to accommodate to hypothetical scenarios where white cards were faced up.

Interpretation 3

John offered the third interpretation, that is, the situation was not probabilistic.

Excerpt 20

132.John I just wanted to say that, I think that we're wrong, I believe that the probability is one or zero, because we're talking about *that* card. You

want to repeat the process, why don't you flip it over one time you know the answer, so there's no repeating it=

133.Sarah That's a good call.

134.John This is not a probabilistic situation, would be my answer. This process cannot be repeated.

135.Terry So if you go, if you go to that booth at the fair, the clown is always going to have the same cards sitting there.

136.John No…no, no. See what this says, this says, we do=

137.Terry But you, in order for this to make sense you have to think about it, probabilistically.

138.John Yup, well, that's what I'm— the way I'm looking at it is that card's been sat down on the table, or, the card is there, it is that card, it is not asking what's the probability it pulls it out of the hat, it is after he's pulled it out of the hat and put it down, here's the card sitting on the table. The process being repeated is not pulling it—keep pulling it out of the hat=

John argued that both 1/2 and 1/3 were wrong and that the probability was either 1 or 0 because he believed that the situation was just a one-time event. The situation was about the clown and the one person who was betting the color of the down side of that one card, *once*. There were two possibilities: if the other side is red, the probability is 0, and you lose the bet; if the other side is white, then the probability is 1 and you win the bet (path 1-5).

Summary of Activity 2-4, Episode 1

Table 24 summarized the teachers' interpretation of the Clown and Card scenario. John and Sarah conceived of the situation non-stochastically, but they interpreted probability in different ways. John believed that the probability was either 1 or 0 depending on whether the bottom of the card was white or red. Sarah believed that the probability is ½, the relative proportion of white out of the two possible outcomes of the bottom color of the card. Lucy, Betty, Linda, and Nicole conceived of the situation stochastically, and they split in two groups in their interpretations of the situation. Lucy, Betty, and Linda conceived of a stochastic process that entailed randomly selecting one card out of two

cards, and thus deduced a probability of ½. Nicole conceived of a two-stage random process of selecting a card from the bag and showing the other side the side of the selected card, which led to the probability of 1/3.

Table 24: Teachers' conceptions and interpretations of probability situation in Clown & Cards scenario

Locator	Name	P.	1	2	3	4	5	6	7	8	9	10
			Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D6A7	John	1,0	N				Y					
	Nicole	1/3	Y	Y	Y						Y	
	Sarah	1/2	N								Y	
	Lucy	1/2	Y	Y	Y						Y	
	Betty	1/2	Y	Y	Y						Y	
	Linda	1/2	Y	Y	Y						Y	
	Henry	1/2										
	Alice											

Activity 2-4, Episode 2: Simulation

After the teachers laid out their different interpretations and answers to the question, they debated on which interpretation was correct. As shown from the excerpt below, the debate did not resolve the differences in teachers' interpretations as each teacher spoke from his or her own perspective.

Excerpt 21

143. John — But the bet, the bet is made after he pulls the card out not before, he's not stating, he's not stating, "Oh, what's the probability it is red given that this side is red, if I pull—" you know, he's not putting any constraint, he's pulling the card out and putting it down.

144. Betty — To me he's eliminated the white/white, totally, I mean that's not an option anymore.

145. Nicole — Why can't the clown, also say – it doesn't say this here – but why can't the clown just pull out a card and slap it down and let's assume it is white this time, and now the clown says, um… what's the probability the other side is white?

146. Sarah — (simultaneously) the probability that the other side is red… is white.

147.Nicole It is the same problem, so I'm arguing that the, what we're repeating is pulling the card out of the bag initially. And the clown doesn't give a flip what color…

…

153.Sarah I thought that the problem had to do with, once (inaudible) the card is on the table, now, you have to look at the probability over a length of time, you just, you realize in your mind that if you did this experiment over a length of time several opportunities, you would have either a red or a white down, and if you do it a lot it is going to work out to be about equal.

Each teacher's utterance fit with his or her own conception of the situation. To John, the situation was "The clown plays the game just once, with one person, on that one occasion". To Nicole, the situation was "The clown plays the game repeatedly, drawing a different card, and making the bet with reference to whatever color is up." To Betty and Sarah, the situation was "The clown plays the game repeatedly, drawing a different card until getting a red card up."

Pat commented that a situation could have multiple interpretations and that an answer would be valid as long as it was consistent with the way one conceived of the situation and the interpretation (that led to the answer) was consistent with the text. However, the teachers kept defending their own point of view and dismissed the others as being wrong or not viable. Some teachers proposed to use simulation to resolve the conflict. The discussion quickly fell back to the same debate "What is the situation?" as each teacher tried to set up the simulation in a way that fit with how he/she conceived of the situation.

Excerpt 22

237.Nicole Why can't we try to simulate this?

238.John Yeah I had a suggestion on that, like what we can do is set it up on our calculator to choose a zero or one, say zero represents red and one represents white, for instance, just a side of the card. So we'll say zero is red, so, if we set it up on our calculator and it gives us a red, that means the clown took it out of the hat and put a red down. So then we

know, that's our situation. If we do it and it gives us a one then we don't recognize that because that's not our situation, that's not—what we're trying to repeat this situation, so the situation has to be with a red card coming out. So we can go through, if it has a zero, then we do it again to see if we get zero to one, to see if it is red or white. That's, that's my suggestion.

239. Henry That's assuming that the same question is going to be asked every time.

240. John That's what I'm saying, if it is white you disregard that, only, if it is zero, that zero represents red so that's always going to be… -- I'm trying to get the same situation where there's a red card drawn.

241. Terry: I'm not sure that your doing that is going to be the same situation, though, because then you're not pulling from cards that are fixed in a certain way.

242. Sarah Can you do a calculation—calculator simulation where you do fix that first number. Where you always—where your first number's always…

243. Linda But that's the same as choosing from red or white on the other side.

244. Sarah Yeah…Not the same as the card situation because then the cards are not fixed in a certain way.

245. Terry Okay, well what are the, what are the— you *have to* start with the fact that you have three cards.

246. John Rrrright.

247. Terry So you, you=

248. John Half of them my red side, half of them my white.

249. Terry So you *gotta* know which card you got so we can do like Nic—you know Nicole had first card, second card, third card… okay, so that's part of the problem you've got to reach in and get a card. So that's—there's randomness there, "which card do you get?" so there's one level of the process. And then…

250. Sarah Then fix the wording of your question, but that's all.

From this exchange we see three types of simulations that the teachers conceived of. The first, proposed by John, was to simulate selecting between 0 (red) and 1 (white), discard 1 if selected, and then select again between 0 and 1. The second, made explicit by Linda, was to simulate selecting between red and white (the possible color of the other side once the red side is up). The third, proposed by Terry (who was supporting Nicole's interpretations), was to start from selecting one of the three cards and not discard any card in order to ensure randomness. The simulations these teachers wanted to set up fit precisely with the way they conceived of the situation.

The debate remained at the level of people trying to convince others that they were wrong, as opposed to reflecting on their differences. For example, Terry was convinced that Nicole's interpretation was the correct one. In communicating with teachers who did not share the same belief, she used the phrases such as, "you have to…", "you gotta…" Right after she said, "You have to start with the fact that you have three cards." John reiterated his own belief of what the outcomes would be, "Half of them my red side, half of them my white." (lines 246-249) Here is another example:

Excerpt 23

504.	John	I got the right answer the way I did it.
505.	Terry	Why don't we simulate it like this=
506.	John	=If you do it the way I did it you'll get the right answer.

These exchanges demonstrated that what Terry said made no impact on John's thinking. The same was generally true for the other teachers. The debate among the teachers did not resolve their differences because all the teachers were committed to their own conceptualization of the situation and none of them detached himself or herself to provide a criterion for judging which interpretation was valid.

Sarah's last comment—"Fix the wording of your question"— suggested that she believed the wording of the question should decide a correct way of interpreting the question, and thus a way of setting up the simulation. This supported my earlier conjectures that the teachers believed that 1) a situation has *one* correct interpretation, and 2) how a situation is worded/phrased dictates how it should be interpreted. If a situation is open to multiple interpretations, then it contains ambiguity and must be modified so that only one interpretation will be valid.

Discussion continued as teachers argued over how the simulation should be set up. The teachers did not realize that a simulation of a situation is essentially determined by one's conceptualization of that situation. This was revealed in the following excerpts.

Excerpt 24

499.	Pat	How could we settle this=
500.	Nicole	=I want to do a simulation.

Note that in line 237 Nicole had suggested simulation as a potential way of resolving the conflict. Nicole second proposal on simulation, after the long debate on how to set up the simulation, indicated that she did not understand that a simulation of a situation is tightly connected to conceptualizations of the situation. A simulation is a simulation of something. In the current context, this something is a well-defined random process of cards selection. If the question of what the process was were not settled, then the simulation could not be determined as a consequence. However, Nicole seemed to think of "doing simulation" not as an extension of one's conceptualization, but rather as a free agent (outside of the realm of thinking) that could decide which conceptualization was correct. The discussion continued as Pat confronted this misconception.

Excerpt 25

499.	Pat	How could we settle this=
500.	Nicole	=I want to do a simulation.
501.	Pat	besides by strength of argument?
502.	Nicole	I really want to do a simulation on the calculator.
503.	Pat	If you simulate it … if you simulate it so that in effect you are saying ... if you simulate it like John did, then you're going to get a probability of 1/2. If you simulate it like you did, then you'll get a probability of 1/3.
…		
509.	Nicole	Pat, I don't understand something you are saying.
510.	Pat	What's that?
511.	Nicole	I want to know if what you are saying to me when you gave your response is that I can't simulate it if I already know the answer.
512.	Pat	No. That wasn't it.
513.	Nicole	That's not what you said?

514.Pat If the process that you conceptualized is, pick a card, and it is either red or white on top, now… pick a card and it is red… with probability 1, that is you always get a red, now, its either red and white or red and red. If you simulate that, if that's the process that you simulate, then what do you see happening?

515.Nicole It would be 1/2, I guess.

516.Pat Yeah. See so, how you conceptualize the process is going to depict—is going to define your simulation.

517.Terry So if you conceptualize the process differently, then you're going to get different ...

518.Pat So, but I'm saying that running a simulation doesn't settle it.

519.Nicole Oh, I see what you are saying. So in a sense…

520.Pat The simulation=

521.Nicole The simulation… it depends on how I perceive the process as to how I set up the simulation.

522.Pat (simultaneously) as to what simulation you're going to run. (afterwards) That's right.

523.Nicole Well I'm struggling with setting it up *my* way. (laughs)

The idea Pat tried to convey during this excerpt was that one's simulation was based in one's conceptualization of the situation, and therefore a simulation could not, in principle, resolve the question of which conceptualization was correct. Lines 519 to 523 suggested that Nicole had understood this idea. Later discussion, however, revealed that this idea was not understood by some of the other teachers. Below I will present two excerpts taken from the ensuing discussion in which Henry and Terry continued to propose using simulation to decide a correct conception.

Excerpt 26

543.Pat But what people are doing is arguing for the correctness of their interpretations.

544.Nicole That's correct. I commend you.

545.Pat And I'm, I'm…the question I'm asking is, stand back from that. We've got different interpretations and everyone believes each of their interpretation firmly.

546.Henry I don't anymore. (laughter)

547.Betty Me either. Sorry I spoke up first.

548.Pat The question I'm asking is how to settle that argument other than everyone raising their voices saying, "My way of looking at it is right!"

…

564.Pat But please answer my question. How do you settle the matter?
565.Henry I think you could simulate it. Put the three cards=
566.Pat Which simulation do you run?
567.Henry Well, that's what I'm trying to tell you. I think you could take the three cards=
568.Pat But you're just restating the interpretations when you say, "This is the simulation to run." Because there are *two* simulations to run. How do you decide between them?
569.Henry I don't know which two simulations but I'm trying to tell you the simulation=
570.Linda (Simultaneous with Henry.) They're both right.
571.Pat No. One leads to a probability of 1/2 and one leads to a probability of 1/3. They can't both be right.
572.Henry No, see that's what I'm saying. I'm finally getting the fact that you're ... the probability is 1/2 is coming from the idea that there are only two *cards* in question. But the population that produced these two cards produces these three *outcomes* two of which are beneficial for the clown. So it is a ... I see that. I think you simulate by giving them the cards and letting them do it then they'll pay close attention to finishing the process. If they get a white card then they'll just stick it back in.
573.Pat Suppose somebody says, when you count, see the question is when you do the simulation, what are you going to count. So, if somebody reaches in and pulls out a card that is white on top=
574.Henry Then you stick it back in and you don't count it at all. Now if they have a red showing up now you're ready to count and then you flip it over if it is red then you mark it down as a win for the clown. Stick it back in the bag and pull out a card and if it is red on top and if it is white then you mark it down as a win for you. And 2/3 of the time that you do that the clown's going to win…you're going to win one…

This excerpt first showed that Henry and Betty did not firmly believed their interpretations that the chance of winning the bet was 50% (lines 546 and 547). By the time Henry said, "I'm finally getting the fact that…" (line 572), he had completely changed his interpretation. He now believed that the RR card should be counted twice and thus it gave a favorable edge to the Clown (line 572). He then described a process/simulation that fit with his current interpretation (lines 572 and 574). We did not know what had made Henry to change his interpretation, but we could infer from this excerpt that:

1) Once Henry changed his interpretation, he believed that the new interpretation was the only correct interpretation.

2) Henry believed that there was only one correct simulation to run, and this simulation would confirm the correctness of his interpretation.

I made such inferences based on the words that Henry chose to use when he explained his thinking.

> I don't know which two simulations but I'm trying to tell you *the simulation*= I'm finally getting *the fact* that…
> I think you simulate by giving them the cards and letting them do it then they'll pay close attention to finishing *the process*.

These phrases, as used in their context, revealed that Henry still firmly prescribed to the tacit assumption that a situation must have one correct interpretation. For this reason, when he finally came to understand the second interpretation, he dismissed his original interpretation as wrong (this was conveyed by his characterizing the second interpretation as a fact). When he described the simulation, he described it in a way as if it were the only possible simulation. As soon as he dismissed his original interpretation, the simulation that fit with that interpretation became completely absent from his mind.

To summarize, this excerpt illustrated that 1) Henry changed his interpretation of the situation, and 2) Henry believed that simulation could confirm that the validity of his interpretation. He did not know that the simulation he conceived of was a result of his changing conception of the situation.

The next excerpt occurred near the end of the morning discussion. Pat continued to push the discussion of "How to resolve the conflict". We would observe, in this excerpt, that both Terry and Henry still firmly believed that simulation could decide which of the two interpretations was correct.

Excerpt 27

641. Pat — Well the question I asked a long time ago, which *actually* I felt you did not want to address it, and that was "How, in principle, do you resolve="

642. Terry — =It is not that we don't want to answer it. I don't know *how* to answer it.

643. Betty — We're to that point where we don't have a clue how to answer it.

644. Pat — But we haven't even *talked* about it.

645. Linda — Could you give them some activities so we could give them a chance so they could go home and think about it. I mean that's what I do, because y'all can stand there and preach=

646. Terry — I think Henry's idea was a good idea. Give them three cards but don't tell them what to do with it. Give them a RR card, a RW card, and a WW card and you don't tell them what to do with it, you just say here are three cards.

647. Pat — Uh-huh.

648. Henry — Cause that's the situation.

649. Terry — And you don't tell them what to do with it. To me, somebody that's in the 50-50 camp, by trying to simulate—by trying to do that, some of those people ... maybe not all of them ... some of them are going to realize that what they are doing doesn't really make sense, because my process is that first I've got to reach in there and pick a red card, you know, and make sure I've got a red card and then flip it over, which ... it is not really the same process as what the clown is doing. Now, maybe they wouldn't realize that ... as they, or if somebody noticed and said, "Wait a minute. That isn't what the clown did. He didn't reach in and make sure he had one of the red ones. I didn't have one of the three cards face up and slide out one of the red ones and say this is red." Now, I don't know, maybe to the kids it wouldn't be any different because they know it is red, I don't know...

Henry's comment "Cause that's the situation" revealed, again, that he did not see the distinction between a situation and a situation as one sees it. He assumed, although without himself knowing it, that the way he saw the situation was the situation (or how this situation should be seen by anyone.) He insisted that his new interpretation was correct as if he had never thought otherwise. Terry, too, seemed so ingrained in her conception of the situation that she could not entertain the idea that alternative conceptions were both possible and reasonable. She still did not understand the idea that simulation was an expression of one's conception. Rather, she believed that simulation

results should confirm one's conception against competing conceptions. Thus she believed that to resolve the conflict meant to show the other camp's wrongness by simulation—the only simulation that she could envision.

To end the discussion of the episode, I want to add an observation that the teachers had become increasingly frustrated as their many attempts at resolving their differences (both by reiterating their own points of view and by proposals of simulation) went void. At this juncture, Pat decided to push the discussion to a different level. Here I turn to the next episode.

Activity 2-4, Episode 3: Pedagogical conversation

In this section, I will specifically focus on Pat's intervention—the direction to which he attempted to lead the discussion, how teachers interpreted him, and what they believed the discussion was about.

Pat observed that the way the teachers debated with each other was nothing more than reiterating their own points of view. He attempted to raise the pedagogical question, "How do you resolve the conflict without restating what you think is true?" His intention was to push the teachers to reflect the assumptions behind their thinking. Pat understood that the fundamental distinction between the two camps: those who believed the probability to be 1/2, and those who argued for 1/3, was that the first camp approached the problem with a fixation on the question that was asked, whereas the second camp focused on the underlying situation behind the question. Orientation to the underlying situation allows one to envision the space of all possibilities and to solve all the questions relevant to a situation, whereas an orientation to the question often leads one to choose only relevant information and thus prevents one from making sense of the entire situation

as a whole (Please recall the Urn & Marble example in the Chapter 5). Pat's intention, by raising the question of "How to resolve the conflict", was to invite the teachers to think about instructional design, that is, in designing probability instruction the teachers should avoid this type of conflicts by always orienting the students to think about the underlying situation as oppose to the question that is being asked. To engage the teachers in this reflective conversation required that the teachers take their debate as an object of reflection and consider the potential instructional action or design that would resolve the conflict. However, as we will observe in the following excerpts, the teachers did not engage in reflective conversation. Rather, they understood Pat as saying "How do we decide which interpretation is correct?" and thus kept arguing for their own interpretations. The following four excerpts illustrated Pat's persistent attempts at elevating the conversation and teachers' resistance to entertaining alternative interpretations, and engaging in pedagogical discussions.

Excerpt 28: First attempt (Starting time 35:30)

499.	Pat	How could we settle this=
500.	Nicole	=I want to do a simulation.
501.	Pat	besides by strength of argument?
502.	Nicole	I really want to do a simulation on the calculator.

Excerpt 29: Second attempt (Starting time 37:42)

548.	Pat	The question I'm asking is how to settle that argument other than everyone raising their voices saying, "My way of looking at it is right!"
…		
564.	Pat	But please answer my question. How do you settle the matter?
565.	Henry	I think you could simulate it. Put the three cards=
566.	Pat	Which simulation do you run?
567.	Henry	Well, that's what I'm trying to tell you. I think you could take the three cards=
568.	Pat	But you're just restating the interpretations when you say, "This is the simulation to run." Because there are *two* simulations to run. How do you decide between them?

569. Henry I don't know which two simulations but I'm trying to tell you the simulation=
570. Linda (Simultaneous with Henry.) They're both right.
571. Pat No. One leads to a probability of 1/2 and one leads to a probability of 1/3. They can't both be right.
572. Henry No, see that's what I'm saying. I'm finally getting the fact that you're ... the probability is 1/2 is coming from the idea that there are only two *cards* in question. But the population that produced these two cards produces these three *outcomes* two of which are beneficial for the clown. So it is a ... I see that. I think you simulate by giving them the cards and letting them do it then they'll pay close attention to finishing the process. If they get a white card then they'll just stick it back in.

Excerpt 30: Third attempt (Starting time 1:11:10)

617. Pat …My question is: How do you resolve that? [Pause]
618. Linda Could we take a break?
619. Pat Do you see, Henry, that the idea of running a simulation doesn't finalize it.
620. Henry I sure think it helps, though. Because if you have those three cards, and if you're running the simulation, then you can come back and ask them these sticky questions, and help to point in a clearer direction.
…
631. Henry See, if you run the simulation, though, you're not going to get 1/2.
632. Alice But you would run two different simulations.
633. Henry But there's only—you *think* you could run two different simulations. But there's *only* three cards. There's *only* one bag. And every time you run it, it is going to tally up.

Excerpt 31: Fourth attempt (Starting time 1:15:58)

641. Pat Well, the question I asked a long time ago, which *actually* I felt you did not want to address it, and that was "How, in principle, do you resolve-"
642. Terry =It is not that we don't want to answer it. I don't know *how* to answer it.
643. Betty We're to that point where we don't have a clue how to answer it.
644. Pat But we haven't even *talked* about it.
645. Linda Could you give them some activities so we could give them a chance so they could go home and think about it. I mean that's what I do, because y'all can stand there and preach=
646. Terry I think Henry's idea was a good idea. Give them three cards but don't tell them what to do with it. Give them a RR card, a RW card, and a WW card and you don't tell them what to do with it, you just say here are three cards.
647. Pat Uh-huh.
648. Henry Cause that's the situation.

649.Terry And you don't tell them what to do with it. To me, somebody that's in the 50-50 camp, by trying to simulate—by trying to do that, some of those people ... maybe not all of them ... some of them are going to realize that what they are doing doesn't really make sense, because my process is that first I've got to reach in there and pick a red card, you know, and make sure I've got a red card and then flip it over, which ... it is not really the same process as what the clown is doing. Now, maybe they wouldn't realize that ... as they, or if somebody noticed and said, "Wait a minute. That isn't what the clown did. He didn't reach in and make sure he had one of the red ones. I didn't have one of the three cards face up and slide out one of the red ones and say this is red." Now, I don't know, maybe to the kids it wouldn't be any different because they know it is red, I don't know...

The teachers kept responding Pat's attempts by restating what they thought was true/correct even though Pat had excluded it as a productive mode of argumentation. This was, in part, because the teachers themselves did not think they were simply restating what they thought was true. They were talking about simulations, which, in their mind, could decide which interpretation was correct. Nonetheless, teachers' responses revealed that they believed that discussion was about which interpretation was correct. They did not detach themselves from their own conceptions, and take the collection of, and the conflict of, different interpretations as the object of thought. Moreover, they seemed to completely forget that in their initial discussions of the problem, everyone except Nicole had proposed that the situation was like selecting two cards from a bag, one that is red on both sides and one that is red on one side and white on the other—which itself was a proposed simulation.

When Pat suggested that simulation could not be a means of settling the conflict, one of the teachers, John, said he would simply tell the other party that they were wrong.

Excerpt 32

575.Pat So, what do you say to the person who says that is the wrong simulation?

576.John You tell them that they're wrong when they say that (laughter) instead of taking two hours thinking that that's the right way to do it. If I had been told that that's incorrect, then I wouldn't be sitting here frustrated with this, trying to understand=

577.Nicole But I think Pat=

578.John -I don't mind someone telling me my thinking is wrong because I want to understand.

579.Pat I'm not telling you you're thinking is wrong

580.John It is wrong. I was wrong. And I just need to be told that.

John said that he would rather be told wrong than engaging in a discussion like the one they were having. This suggested his solution to the conflict: If one party cannot convince the other, then dictate it. Underlying this solution was John's firm belief in the necessity of having one correct interpretation. The following excerpts, appearing sporadically throughout the discussion, revealed that John was consistently looking for a correct answer.

Excerpt 33

305.John There's never a correct answer. That's why I get frustrated=

306.Terry I'm doing this to answer your question.

307.John I hate working on something if I'm not going to figure out the right answer.

Excerpt 34

504.John I got the right answer the way I did it.

505.Terry Why don't we simulate it like this=

506.John =If you do it the way I did it you'll get the right answer.

Excerpt 35

584.John I'm not asking for the answer. That's not what I'm asking for. I don't care about the answer. But I'm the kind of person that when someone explains something to me, then I learn it and I don't forget it. But when I don't understand how something works, and I'm trying to figure it out, I have no material here to teach myself, and take the—everybody's saying different things, then it just confuses me. I'm not being revealed the right answer. When I say "answer," I'm talking about the correct process to work the problem. If someone sits down and explains to me, "Here's where you are wrong. Here's what you needed to do. Oh, okay, I see. I'm okay." That's the type of person I am. I mean, maybe that's wrong for this. Maybe I shouldn't be like that. I don't know.

These excerpts indicated that the discussion to John was solely about deciding which interpretation was correct. John wanted to be told where he was wrong. He did not reflect on his own why he was wrong, nor did he attempt to understand alternative interpretations held by the other teachers. Had he been reflective, he would have been able to ask questions such as, "Could my thinking be wrong, and why?" or "What is it about, say, Nicole's interpretation, that made me think she was wrong?" etc. In other words, John demonstrated both a lack of reflection and critical thinking with respect to his own learning, and a lack of orientation in understanding others' thinking.

These excerpts also revealed John's difficulty in engaging in pedagogical conversations. He did not assume the identity of a teacher during the discussion. He thought and acted like a student, and his conception of learning was "to be told what is correct". For this reason, he believed that it was okay to simply tell students that they were wrong.

The teachers' belief that the discussion was about which interpretation was correct implied that they did not think of the discussion as an experience upon which they could reflect on both conceptual understanding of probability and the pedagogy of teaching probability (or any subject). John's frustration as well as the other teachers', for example,

Excerpt 36

638.Pat	Left to their own devices they would by and large, leave the classroom—in the classroom they would say, "um , okay, correct answer is…" – leave the classroom and they would be just as naïve as when they started the course. In other words, by not having them resolve the issue of "Why is it that the way I'm thinking about this is problematic", it never became problematic! And they kept thinking that way.
639.Lucy	But don't let it go on like this argument.

suggested that they did not think of the discussion as being productive and that they did not benefit from struggling with this conflict. They did not separate their identity of teacher from their identity of learner/problem solver and try to ask the question: "What can I learn from this in terms of teaching?" Rather, they seemed to have anticipated the seminar discussion to be a model that they would imitate in their classrooms, and their frustration was, in part, a result of the breakdown of that anticipation.

The pedagogical conversation culminated with Pat explaining his intention. He wanted the teachers to experience the dilemma of conflicting interpretations of a probability situation and to witness how irreconcilable it could be once it occurs in the classroom. He hoped that they came to see this dilemma as something to be avoided by cultivating in students an orientation to make sense of the problem situation before answering specific questions. Teachers' responses showed that some teachers appreciated this purpose, and some didn't.

Excerpt 37

690.	Sarah	So, maybe if you had asked us "How could you prevent or alleviate the dilemma, you would have gotten a much better answer than how do you reconcile=
691.	Pat	If you didn't see that trying to deal with that conflict at the moment it arose is something to be avoided, not something to be resolved=
692.	Sarah	I know. But once we got into the process when your question, your question to Henry or to whomever was about, "Well, how do you reconcile these two things", maybe at that point we're trying to answer that question instead of realizing that we shouldn't have allowed that circumstance to develop.
693.	Terry	But I think Pat's point was that unless you are actually in that position, where you are experiencing that ... I mean he wanted us to experience that conflict and to realize how=
694.	Nicole	It is going to happen in our classrooms=
695.	Terry	and how impossible it is to get out of it once it is happened, rather than saying to us, "Oh, if you do this". I mean we've actually=
696.	Sarah	But we were pretty far into that conflict when he asked us that question.

Sarah seemed to believe that the confusion and frustration in resolving the conflict of interpretations that they had experienced were something to be avoided. She did not see the purpose of the discussion as providing these experiences as an opportunity to reflect on pedagogy. Rather, It was about them giving good answers to the question, "How to reconcile the difference". Overall, the frustration these teachers experienced suggested that the amount of confusion they had experienced was over their comfort level, and this, in itself, may have caused the teachers to lose the perspective of what they were trying to do/learn.

Summary of Activity 2-4

Initial discussion around this activity revealed teachers' various conceptions of the situation and different interpretations of probability. John and Sarah conceived of the Clown and Cards situation non-stochastically, but they interpreted probability in different ways. John believed that the probability is either 1 or 0 depending on whether the bottom of the card was white or red. Sarah believed that the probability is ½, the relative proportion of white out of the two possible outcomes of the bottom color of the card. Lucy, Betty, Linda, and Nicole conceived of the situation stochastically, and they split in two groups in their interpretations of the situation. Lucy, Betty, and Linda conceived of a stochastic process that entailed randomly selecting one card out of two cards, and thus deduced a probability of ½. Nicole conceived of a two-stage random process of selecting a card from the bag and showing the other side of the selected card, which led to the probability of 1/3.

In the second episode, the teachers debated about which interpretation should be the correct one. Since each teacher was committed to his own interpretation, and thus

speaking form his own perspective, as a group they were not able to reach an agreement on which interpretation was correct. Nicole proposed to use computer simulation to decide the correctness of the interpretations. Pat commented that this would simply reaffirm each person's understanding of the situation. The conversation about the simulation fell back to the teachers arguing for the underlying situation the way they each saw. They did not understand that simulation was determined by ways of conceiving of the situation. Despite Pat's explanation, Henry and Terry kept proposing simulation as a means of proving the correctness of their own interpretations.

In the final episode, Pat attempted to engage the teachers in reflective conversations about their earlier debate. However, the teachers were not able to take the collection of interpretations as object of thoughts. Instead, the teachers kept arguing for what they believed was the correct interpretation, and proposing to use simulation to resolve the differences. When Pat suggested that simulation could not validate interpretation, John proposed that to resolve the difference they should simply tell the other party they were wrong.

Interview 3-2: Three prisoners

In the post-interviews, we presented a scenario (see below) similar to the Clown and Card scenario.

Students in a probability & statistics class were presented the following problem situation:

Three prisoners — John, Michael, and Quincy — share a cellblock. The Warden announced that, in celebration of the upcoming New Year, one of the three would be released, but he will not say in advance whom it will be.

Michael knew his chance of being released is 1/3. To improve his chances, he said to the Warden "I know you cannot tell me who will be released. But, will you tell me the name of one prisoner who will *not* be release?"

The Warden told him a name of someone other than Michael. Michael left, confident that his probability of being released had risen to 1/2. Is Michael right? Explain.

What is your understanding of this situation and problem?

Two students responded like this:

Student 1

> "Michael is right.
> Once the Warden tells him the name of who will not be released, then the Warden only has two prisoners (including Michael) from whom to choose. If the Warden will make his choice randomly, that means there's a 50%, or 1/2, chance that Michael will be the one released".

Student 2

> "Michael is wrong.
> His chance of being released is 1/3 because if the Warden chooses at random from the three prisoners, then each one is equally likely to be chosen! The Warden speaks to Michael after making his selection, so what he tells Michael has no effect on his chance for release".

1. Please comment on these two students' responses.
2. Suppose that you are the teacher in this class. How might you handle the situation of your students being split evenly between these two responses?

This interview question was similar to the Clown and Cards situation because it too could subject to multiple interpretations, as presented by the utterances of the students 1 and 2. Both interpretations took probability as relative proportions of outcomes, and the differences between the interpretations (and thus, conceptualizations of outcomes) could be accounted for by the question, "When did the warden make his decision?" If the warden had made his decision after telling Michael the name, the probability of Michael's release would be 1/2. If the warden had made his decision prior to telling

Michael the name, the probability of Michael's release would be 1/3. Both interpretations are viable because the statement of the situation does not exclude either one, and thus both probabilities should be accepted as correct in light of the students' explanations.

The summary of teachers' comments on the students' responses (Table 25), showed that almost all the teachers saw the similarity between this scenario and the Clown and Cards scenario (Nicole was the only teacher who did not state it).

Table 25: Teacher's answers to I3-3, Q1: comments on students' responses

John	It is very similar to the red card situation… Student 1 assumes warden makes the decision after talking to Michael. Student 2 assumes before. I'd have to say that both interpretations are incorrect. I definitely know that student 1 is not right. But I could be making a mistake about whether student 2 is wrong… Student 1 is coming in at the middle of the situation I think that's a mistake. If it works the way student 1 talks about, then he is right. So is student 2.
Nicole	I don't know if warden made the decision before or after talking with Michael.
Sarah	This is such the red card problem. I wouldn't comment on which one is right or wrong. My decision will be based on when the warden made the decision.
Lucy	It is kinda like that red card problem. Student 2 is correct, or closer to being correct. Because the warden made the decision before talking to Michael.
Betty	This is kinda like that red card situation. My first reaction is that Michael and student 1 is right. But when I see the student 2's reason… I'm easily swayed. I don't know (who is right). Depends on which one you are assuming.
Linda	This is kinda like the cards. These two interpretations are different. Student 1 thinks that the decision was made before, while student 2 thinks that the decision was made after talking to Michael.
Henry	This is very similar to the red card problem…I think they are both right depending on when you begin your process…Student 1 thinks like Michael, assuming that warden hasn't made the decision before talking to Michael. Student 2 thinks that just because the name was told, it didn't change the original odds. I'd say it is 1/3. I'd agree with both interpretations. But I think student 2 is more right.
Alice	It reminds me of the red card problem. I think Michael's chance of being released is 1/3, before and after talking to the warden. I know what student 1 was thinking. I've been there done that. But what he is missing is that if you're going to simulate the process, you wouldn't start there, you would go back to the 3 people. I agree with student 2. I think majority of people will think like student 1.

Table 25 shows that all teachers except Alice and Lucy thought that both interpretations were viable, depending upon when the warden made his decision, with Henry leaning towards student 2's interpretation. Lucy and Alice stood firmly behind student 2's interpretation, and believed that the warden had made his decision before talking to Michael.

The next question asked the teachers how they might handle a hypothetical situation of their students being split evenly between these two responses. Table 26 summarized the teachers' answers, and it shows that all teachers acknowledged the legitimacy of both interpretations.

Table 26: Teacher's answers to I3-3, Q2: How would you handle the situation?

John	To be honest, this is too tough a situation for me to see myself get out of.
Nicole	I think both student 1 and 2 could be correct. It is ok to have multiple interpretations to situations that are unclear.
Sarah	I won't necessarily persuade students to take either assumption. But I want them to justify their assumptions, because just from the original story we couldn't tell when the warden made the decision. I don't think it has to have a right or wrong answer, but I do think the justifications should be valid and stand up.
Lucy	I would say as long as you gave the assumption that's making the distinction, then that's ok.
Betty	In the past, before this past two week, I would have looked up the back of the book and chose which one is right. But since then, I probably would come to an assumption as a class, where are we going to decide on when the decision was made. And then let the class decide what they want to accept, or we could accept both problems given that they have information to back up their answer.
Linda	I will first demonstrate how each interpretation could be correct and then say "Ok, let's adopt a convention that it is either one or the other.
Henry	I really don't like this problem because it is subject to multiple subjective interpretations. This is a problematic design. … I think they both are right, depending on what their goals are. But that problem has a lot of room for argument. The kids are going to form an opinion and will be reluctant to change, unless they trust you.
Alice	Both are correct depending on how they view the problem. … But I think it is better to lay some groundwork to try to avoid it.

Henry's answer revealed his belief that a well-design problem should not be open to multiple interpretations. John, Henry, Linda, and Alice's answers suggested their commitment as a teacher to achieve a consensus in the classroom. Linda and Alice provided a solution to achieve this consensus—by either adopting a convention (Linda), or laying certain groundwork so as to avoid multiple interpretations and potential dispute (Alice).

To summarize, the teachers' responses to this interview question suggested that most of them had profoundly changed their ways of thinking with respect to interpretations of probability, as well as how to handle classroom situations involving multiple interpretations that are not excluded by the problem text. All teachers were open to the idea that a probability situation could be interpreted differently. Most of the teachers were comfortable with accepting students holding different interpretations as long as they provide sufficient justifications.

Summative Analysis of Teachers' Conceptions of Probability

This summative analysis examined each teacher's conceptions of probability across situations[11].

John

John's conception of probability, as we can see from Table 27, was situational. That is, his conceptions of probability change across different situations. Initially (in the first week), his conception of probability was non-stochastic. He interpreted probability in two

[11] In this summary, I included the data from Activity 1-6 and Interview 2-3 that were discussed in Chapter 7.

ways primarily. The first, a probability of an event A is 50%, because A either happens or does not happen. The second, the likelihood of a sample is evaluated against how the sample statistics resembles the parameter of population from which the sample is drawn. In the second week during the discussion of the PowerPoint slides (particularly the slides on car and temperature), John developed a stochastic conception of probability. However, his interpretation to the Clown and Cards situation as well as the post-seminar assessment suggested that his stochastic conception of probability was contingent upon situations. Out of the six situations, half of the time he interpreted them non-stochastically. Nonetheless, his non-stochastic interpretation of probability—probability of an unrepeatable event is either 1 or 0--was a coherent one.

Table 27: John's conceptions of probability situations

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D1A1-2Q1	John	N						Y			
D1A1-2Q2	John	N					Y	Y			
D3A1-6	John	N					Y				
I2-3	John	Y	Y	Y							Y
D5A2-2S4Q1	John	Y	Y	Y							Y
D5A2-2S4Q2	John	Y	Y	Y							Y
I3-1Q1	John	N				Y					
I3-1Q2	John	N				Y					
I3-1Q3	John	Y	Y	Y							Y
I3-1Q4	John	Y									
I3-1Q5	John	Y									
D6A2-4	John	N				Y					

Nicole

Nicole's conception of probability was predominantly stochastic. Only on two occasions, she held an outcome approach, and interpreted probability situations non-stochastically.

Table 28: Nicole's conceptions of probability situations

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D1A1-2Q1	Nicole	Y	Y	Y							Y
D1A1-2Q2	Nicole	N			Y						
I2-3	Nicole	Y									
D5A2-2S4Q1	Nicole	Y	Y	Y							Y
D5A2-2S4Q2	Nicole	Y	Y	Y							Y
I3-1Q1	Nicole	Y	Y	Y							Y
I3-1Q2	Nicole	Y									
I3-1Q3	Nicole	N			Y						
I3-1Q4	Nicole	Y	Y								
I3-1Q5	Nicole	Y	Y	N							
D6A2-4	Nicole	Y	Y	Y							Y

Sarah

Table 29 shows that only on one occasion Sarah interpreted a probability situation stochastically. This suggested that she had a predominantly non-stochastic conception of probability.

Table 29: Sarah's conceptions of probability situations

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D1A1-2Q2	Sarah	Y	N								
D3A1-6	Sarah	N			Y						
I2-3	Sarah	Y	Y	N			Y				
I3-1Q1	Sarah	Y									
I3-1Q2	Sarah	Y				Y					
I3-1Q3	Sarah	N									
I3-1Q4	Sarah	Y	Y	Y							Y
D6A2-4	Sarah	N									

Lucy

Table 30 shows that on many occasions we did not have complete information about how Lucy interpreted a probability situation. This was due to the fact that she rarely articulated her thinking during the seminar discussion. If we look at her responses to the

Post-Interview, we can see that she interpreted situations 1 and 2 non-stochastically, situation 4 stochastically, and situation 3 and 5 in both ways. This suggested that although she understood the distinction between non-stochastic and stochastic conceptions of probability, in practice her interpretations to probability situations were contingent upon how a situation was formulated.

Table 30: Lucy's conceptions of probability situations

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D1A1-2Q2	Lucy	Y	N				Y				
I2-3	Lucy	Y	Y	Y							Y
I3-1Q1	Lucy	N									
I3-1Q2	Lucy	N				Y					
I3-1Q3	Lucy	N									
I3-1Q3	Lucy	Y									
I3-1Q4	Lucy	Y									
I3-1Q5	Lucy	N									
I3-1Q5	Lucy	Y									
D6A2-4	Lucy	Y	Y	Y							Y

Betty

Table 31 shows that in the first week, Betty had a pro-dominantly non-stochastic conception of probability. During the Post-Interview, Betty interpreted situations 2 and 5 non-stochastically, and situations 1, 3, and 4 stochastically. This suggested that her conception of probability was situational.

Table 31: Betty's conceptions of probability situations

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D1A1-2Q1	Betty	N									
D1A1-2Q2	Betty	N					Y				
D3A1-6	Betty	N			Y						
I2-3	Betty	Y	Y	N			Y				
I3-1Q1	Betty	Y	Y								
I3-1Q2	Betty	N				Y					
I3-1Q3	Betty	Y	Y	Y							Y
I3-1Q4	Betty	Y	Y	Y							Y
I3-1Q5	Betty	N				Y					
D6A2-4	Betty	Y	Y	Y							Y

Linda

Linda had a mixture of conceptions of probability. In the first week, her interpretations of probability included that of outcome approach, proportionality heuristics, frequentist interpretation, as well as, thinking that probability of any event is 50%. In the second week, she interpreted most of the situations in the PowerPoint slides non-stochastically. Her interpretation was that of probability as relative proportions of expected outcomes. Her responses to the car and temperature slide suggested that her conception of probability was contingent upon situations. The rest of week she interpreted most of the probability situations stochastically.

Table 32: Linda's conceptions of probability situations

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D1A1-2Q1	Linda	N			Y						
D1A1-2Q2	Linda	N					Y	Y			
D3A1-6	Linda	N			Y						
I2-3	Linda	Y	Y	Y							Y
D5A1S1	Linda	Y	N							Y	
D5A1S2	Linda	N								Y	
D5A1S3	Linda	N								Y	
D5A2-2S4Q1	Linda	N				Y					
D5A2-2S4Q2	Linda	Y	Y	Y							Y
I3-1Q1	Linda	Y	Y	Y							Y
I3-1Q2	Linda	N				Y					
I3-1Q3	Linda	Y	Y	Y							Y
I3-1Q4	Linda	Y	Y	Y							Y
D6A2-4	Linda	Y	Y	Y							Y

Henry

Henry interpreted all the situations stochastically (for post-interview Question 1, 2, 4 where he gave both non-stochastic and stochastic interpretations). Therefore, we could claim that Henry's conception of probability was predominantly stochastic. In addition, his responses to post-interview questions also suggested that he was well aware of the distinction between non-stochastic and stochastic conceptions of probability.

Table 33: Henry's conceptions of probability situations

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D1A1-2Q1	Henry	Y	Y	Y							Y
D3A1-6	Henry	Y	N								
I2-3	Henry	Y	Y	Y							Y
D5A2-2S4Q1	Henry	Y	Y	Y							Y
D5A2-2S4Q2	Henry	Y	Y	Y							Y
I3-1Q1	Henry	Y	Y	Y							
I3-1Q1	Henry	N				Y					
I3-1Q2	Henry	N				Y					
I3-1Q2	Henry	Y	Y								
I3-1Q3	Henry	Y									
I3-1Q4	Henry	Y	Y	Y							Y
I3-1Q4	Henry	N				Y					

Alice

Table 34 shows that Alice's interpretations of probability situations were consistently stochastic.

Table 34: Alice's conceptions of probability situations

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D3A1-6	Alice	Y	Y	Y							Y
I2-3	Alice	Y	Y	Y							Y
I3-1Q1	Alice	Y	Y	Y							
I3-1Q2	Alice	Y	Y	Y							Y
I3-1Q3	Alice	Y	Y	Y							Y
I3-1Q4	Alice	Y	Y	Y							Y
I3-1Q5	Alice	Y	Y	Y							Y

Overall, this summative analysis revealed that five out of all eight teachers (John, Linda, Sarah, Betty, Lucy) had a situational conception of probability. Nicole, Henry, and Alice had a stochastic conception of probability. Henry also had a non-situational conception of probability. He was able to distinguish a situation from conceptions of the situation, and offer multiple interpretations to one situation.

Chapter Summary

This chapter provided a description of teachers' conceptions and interpretations of probability as they engaged in the seminar discussions and post-interviews. The teachers had various conceptions and interpretations of probability in the beginning of the seminar, most of which were non-stochastic (Table 17). Towards the end of the discussion on the PowerPoint presentation, the teachers' interpretations of probability had less variability: some interpreted stochastically (path 1-2-3-10), some non-stochastically (path 1-5) (Table 19). The idea that "a situation could be conceived of both stochastically

and non-stochastically" emerged during this discussion. However, in the Post-Interview, only two teachers, Henry and Lucy, exhibited an awareness of this idea. The other teachers interpreted the situations in either one way or another (Table 20).

The Clown and Cards situation was a prototype of situations that subject to many interpretations, i.e., even when interpreted stochastically, one could conceive of different underlying stochastic processes. Teachers' initial answers revealed differences in both the way they conceptualized the situation as well as their different interpretations of probability even when their conceptualization was the same. Teachers' discussion focused on validity of two particular types of interpretations, one leading to probability of ½, and one 1/3. They engaged in unreflective conversation, in the sense that they did not examine the assumptions behind each interpretation, but merely arguing for what they believed was true. They seemed to have an underlying assumption that there should be only one correct interpretation for any situation and they believed that computer simulation could decide which interpretation was correct. Despite Pat's persistent attempt to engage the teachers in conversations about pedagogy (thus entailing reflection on their thinking/interpretations as a collection), the teachers resisted and kept arguing over the correctness of their interpretations. In the post-interviews on a similar question, however, we saw signs of a significant change in teachers' responses. All teachers acknowledged the validity of multiple interpretations and stated that they would be conformable to accept their students holding different interpretations. However, half of the teachers also expressed a commitment to consensus, and claimed that they would avoid such situations in their classroom.

CHAPTER VII

TEACHERS' UNDERSTANDINGS OF HYPOTHESIS TESTING

This chapter describes three sets of activities and interview questions related to hypothesis testing as they unfolded over the two weeks. Broadly speaking, these activities and interview questions engaged the teachers in exploring aspects of hypothesis testing with the aim that their understanding of hypothesis testing be revealed in the discussion.

Table 35: Overview of the activities and interviews in Chapter 7

Chapter7 Teachers' understanding of hypothesis testing			
Section	**Activity (A) and Interview (I)**	**Day**	**Duration**
7.1 unusualness/*p-value*	A1-6 Movie theatre scenario	3	106 m.
	I2-3 Horness scale		
7.2 testing hypothesis of population parameter	A1-3 Pepsi scenario	1&2	180 m.
	I2-1 Alumni association		
7.3 testing hypothesis of randomness	A2-3 Rodney King scenario	5	104 m.

The chapter consists of three sections. The first section focuses on teachers' understanding of unusualness. As I have elaborated in chapter 4, a stochastic conception of unusualness (*p-value*) is a prerequisite for understanding the logic of hypothesis testing. Activity 1-6 and Interview question 2-3 are our attempt at understanding whether or not teachers had a stochastic conception of unusualness. Since unusualness is essentially a probabilistic conception, in this section I will adopt the theoretical framework for probabilistic understanding to describe teachers' understanding of unusualness. The second and third sections describe teachers' discussion of hypothesis testing scenarios. Section 2 focuses on the logic of hypothesis testing. Activity 1-3 and

Interview question 2-1 present two scenarios in which hypotheses about population parameters are tested. Activity 2-3 presents a scenario in which the randomness of sample is the hypothesis to be tested.

In each section, I begin by elaborating the rationale for the design and implementation of the activity. This is followed by a chronological recap of the discussions that unfolded around the activity, highlighting the interactions and teachers' thinking that emerged within them. Then, I highlight teachers' responses to the interview questions that provide additional insights. Finally a summary of major findings is provided in the end of each section.

Unusualness/*p-value*

Activity 1-6: Movie theatre scenario

Ephram works at a theater, taking tickets for one movie per night at a theater that holds 250 people. The town has 30 000 people. He estimates that he knows 300 of them by name.

Ephram noticed that he often saw at least two people he knew. Is it in fact unusual that at least two people Ephram knows attend the movie he shows, or could people be coming because he is there? (The theater holds 250 people.)

Assumptions for your investigation:
Method of Investigation:
Result:
Conclusion:

"Gut level" answer:

Overview

This activity was adapted from Konold (1994). The purpose of this activity was to investigate teachers' understanding of unusualness. This activity centered on investigating the question: Is it *unusual* that at least two people Ephram knows attend the

movie he shows? Ways of thinking about this question that indicates a stochastic conception of usualness would be:

1. Assuming that people go to the theatre randomly, i.e., people do not go to the theatre because Ephram is there;
2. Thinking of a collection of nights, when random samples of 250 people from the population of 30000 go to the theatre;
3. Recording the number of people Ephram knows each night;
4. Plot a distribution of these numbers, and calculate the density of "at least 2": the chance of at least two people Ephram knows attend the movie he shows;
5. If the proportion is smaller than 5% (a conventional significance level), then conclude that it would be unusual that at least two people Ephram knows attend the movie.

The discussion around this activity lasted about 106 minutes. I divide the discussion into four episodes:

Table 36: Overview of discussions around Activity 1-6 Movie theatre scenario

Episode	**Theme**
1	Gut level answer
2	Method of investigation
3	Simulation and the cut-off level
4	Sarah's confusion

Activity 1-6, Episode 1: Gut level answer

Initially, the teachers gave their gut-level answers. All teachers said it was not unusual that Ephram saw at least two people he knew. There was no discussion about what unusual meant. Nor did any teacher give a quantitative measure to unusualness. It was as if the question should be answered in one of the two ways: "Yes, it was unusual" or "No. It was not unusual."

Henry added another answer "Somewhat unusual" when Pat expressed his surprise at teachers' initial responses.

Excerpt 38: Henry's scale of unusualness

23. Pat I'm surprised that no one thinks it unusual.
24. Henry Well, it was just unusual here. I think it could be somewhat unusual.
25. Terry Okay, so you want to say somewhat=
26. Henry Somewhat.
27. Terry So we want to have a scale of unusualness?
28. Henry Yeah.

Teachers' "scale of unusualness", at this point, suggested that they did not have a quantitative conception of unusualness. Rather, they seemed to use unusualness to express a subjective feeling about chance occurrences. This scale of unusualness expressed in terms of "somewhat unusual" and "non unusual" paralleled with teachers' non-quantitative conception of likelihood: A question of "how likely is … " has answers such as "not likely", "highly likely", etc.

Activity 1-6, Episode 2: Method of investigation

This episode lasted for 48 minutes (transcript lines 78 to 391). The teachers set out to investigate whether the event "Ephram sees at least two people he knows" was in fact not unusual, as their gut instincts told them. John proposed the first method—comparing two relative proportions: 1) the proportion of people Ephram knows by name out of the entire population in town, and 2) the proportion of people Ephram knows out of all the people coming to the theatre, i.e. 2/250. Discussion of this method quickly turned into a discussion about what unusualness meant. In the following I will start with John's method of investigation, and then move on to discussions that revealed the teachers' conceptions of unusualness. I will organize the discussion by the types of reasoning that the teachers exhibit, yet also try to maintain a chronological order in which the discussion unfolded.

John's conception of unusualness

Excerpt 39: John's proportionality heuristic

78.	Terry	Let's move on and talk about the method of investigation. When it said method of investigation, it didn't necessarily say, what would be your way of going about figuring out whether this was unusual … what would you go about doing? How would you figure that out?
79.	Henry	Develop a proportion.
80.	Terry	Okay, and how would you develop a proportion?
81.	John	Well, he knows 300 and there's 30,000 people so there he developed a proportion from that. Out of the proportion of 30,000, how many does he know? So basically he knows 1 out of every 100 people.
82.	Terry	Okay
83.	John	So, that's what I did.
84.	John	I set up a proportion and basically got that he knows 1 out of every 100 people=
85.	Terry	Okay and then what did you do with that?
86.	John	So therefore I said he'd know approximately 2.5 people at the movie theater because if 250 come, out of the first 100, he should know 1 person out the next 100 he should know one and of the last 50 he should know half=
87.	Various	[chuckle]
88.	John	so of course you can't do that. So that's the way that I approached the problem.
89.	Terry	So you said he'd know … Just based on proportionality=
90.	John	Proportions, yeah.
91.	Terry	=he would know2-3 people?
92.	John	Right. Just based on proportions.

John's method was: *Since Ephram knows 300 people out of 30,000 people in his town, it means for every 100 people, he knows 1 person. On any given night he should know 2.5 people out of 250 people who come to the theatre, given that this 250 people is a random sample of 30,000 in his town. Therefore, it is not unusual that he saw in the theatre at least 2 people he knows*. This method employed the proportionality heuristic. Recall that John employed the same heuristic in the discussion of Pepsi scenario, where he proposed the method of evaluating the likelihood of "18 out of 30 people favor Pepsi" by comparing the proportion of 18 out of 30 to the underlying population proportion 50%.

John's method suggested his conception of unusualness was about *one particular event*: how likely is it that on one particular night Ephram sees at least two people he knows? This way of conceptualizing unusualness does not quantify unusualness. Rather, it leads to a subjective judgment on the likelihood of the outcome (of Ephram seeing at least two people he knows) in non-quantitative terms—"Yes, it is unusual," or "No, it is not". He did not employ a scheme of repeated sampling that would allow him to quantify unusualness, i.e., conceiving of "Ephram sees *x* people he knows" as a random event and evaluating the likelihood of outcomes "Ephram sees at least two people he knows" against the distribution of a large number of possible outcomes.

In the ensuing discussion, Terry pushed John to give a definition of unusualness:

Excerpt 40: John's definition of unusualness

119.	Terry	Can I come back to John, what you said. You said 2.5=
120.	John	Well I just did mathematically here=
121.	Terry	Okay but I guess what I'm saying is what are you using as your concept of unusual? Tell me what you're think=
122.	John	I've never felt it was unusual. I said no, it is not=
123.	Terry	But what definition are you using of unusual=
124.	John	My, okay, here's my definition. It is what I said earlier. If it was 1 out of 1000, if the proportion came to be 1 out of 1000, then I would say this would be unusual that he knows two people. Then I would change my first answer. It came out to be 1 out of 100, so 250 people as the first 100 come in, he knows one of them. But if that proportion had come out to be 1 out of 1000, then I would say 'hey! The odds of him knowing two people are not very likely because, you know, that theater would have to hold 2,500 people for me to believe that he could see 2.5.' Does everybody see where I'm …

Terry understood that it was *coincidental* that John's method—comparing the population parameter to the sample statistic—led to his conclusion (2.5 was fairly close to 2), and that his conception of unusualness was not based on a method that would provide a measure of unusualness. She thus asked John to clarify his definition of unusualness—to

provide a rationale for deciding whether an event is unusual regardless of the context.

John gave his definition of unusualness:

> If for every 1000 people Ephram knows 1 person, then for 250 people, he knows 0.25 person, which would make "Ephram sees at least 2 people he knows" unusual. But for every 100 people Ephram knows 1 person, so for 250 people, he knows 2.5 person. This makes "Ephram sees at least 2 people he knows" not unusual.

Ash shown in this excerpt, rather than giving a general definition of unusualness that provides a rationale for measuring the unusualness of an event, John merely presented two particular scenarios for which the proportionality heuristic works. This "contextual definition" of unusualness further confirmed that John's conception of unusualness was subjective and non-quantitative.

Sarah, Betty, and Linda's subjective conception of unusualness

The following excerpt revealed Sarah's conception of unusualness.

Excerpt 41: Sarah's subjective conception

99.	Terry	What would it mean to be unusual that he knows at least 2 people at the movie? What do you have to think about to think about whether it is unusual=
…		
114.	Sarah	I thought if he had a night where he saw 50 people that he knew, that would be unusual.
…		
131.	Terry	… I'm hearing people saying 'one night, just by proportions, 2-3 people that he knows are going to be there and what I'm saying is that that, to me, is not talking about whether that was unusual or not.
132.	Sarah	I think that would be your expectation. Something unusual would have to be something different than that, like a whole herd of people comes in that he knows.
133.	John	We'd have to go back to the assumption=
134.	Terry	Okay, okay stop! Let's not even talk about this situation. Just tell me what your definition of unusual is.
135.	Sarah	Something that does not meet the expected.
…		
164.	Terry	Let me come back to two things. Sarah, you said you don't expect it and you said you're doing it a certain … You have some cutoff that

		you're doing. When you say you're doing it and when you say expectation where does that come from?
165.	Sarah	In my mind, I don't have some kind of arbitrary number that makes it usual=
166.	Terry	That's fine
167.	Sarah	=Or unusual. It is just that there are certain things that I expect or don't expect to do.
168.	Terry	Where does your expectation come from?
169.	Sarah	Probably, if you're talking about me personally, from personal experience. For example, I don't expect to get on an airplane and fly to Chicago any time this summer, because that's not something that I regularly do. But I do expect to get on an airplane and fly to Raleigh, North Carolina because that's something I do 3 or 4 times a year because of family there. So do you see the difference in what I'm saying?

Sarah, in lines 114 and 132, claimed that the event "Ephram saw *a whole herd of* people he knew" was unusual. This was another example of teachers not conceptualizing "Ephram sees *x* people he knows" as a random repeatable event. While John was fixating at one particular night when *x*=2; Sarah was fixating at one particular night when *x=a big number*. Neither one of them had conceived of a collection of nights with varying outcomes. Lines 135, 167, and 169 revealed that Sarah's conception of unusualness was entirely subjective—something is unusual if it is unexpected, and expectations are made on the basis of personal experience.

The following excerpts showed that Linda and Betty's conceptions of unusualness were similar to Sarah's.

Excerpt 42: Linda's subjective conception

155.	Linda	Can I draw a different example?
156.	Terry	mmhmm
157.	Linda	Okay, in a college you expect that for every college algebra class, 30 people are going to sign up. But if one instruction has a 60 person waiting list, that's unusual.
158.	Terry	Okay
159.	Linda	Okay? Because you don't see that. There's something going on there. That's unusual.

Excerpt 43: Betty's subjective conception

177. Terry I think I came up with an example. Can I give my example? Okay. Let's say every day I drive to work I go down White Bridge Road, okay? And I say that it is unusual—The intersection of White Bridge Road and Harding is a big intersection—it is unusual for me not to get stopped at that light. It is unusual for me to not have to stop at that light=
178. Sarah Expect to stop
179. Terry =What is my criterion for me saying that's unusual, for me to not have a red light there?
180. Betty It is out of the ordinary. It is out of normal circumstances.

Excerpt 44: Betty's outcome approach

330. Pat All right let me ask you this. Tonight you are going to go to the movies and I'm going to ask you, would it be unusual for you to see … let's see, you're going to go to the movies in downtown Nashville. Would it be unusual for you to see somebody that you know?
331. Betty Yes
332. Pat And why do you say that?
333. Betty Because I don't go to downtown Nashville to the movies.

Linda's example in Excerpt 42 and Betty's comments in Excerpts 43 and 44 suggested that they both had a subjective conception of unusualness. More specifically, Betty's comment in Excerpt 44 exhibited a typical case of someone employing the outcome approach, i.e. an event/outcome is unusual if it is unanticipated to occur.

Henry and Alice's quantitative conception of unusualness

Only two teachers, Alice and Henry, had a quantitative conception of unusualness.

Excerpt 45: Alice's quantitative conception

142. Terry Is it unusual that people get struck by lightning?
143. Linda Yes.
144. Terry Why is it unusual that people get … I mean, why do you say yes?
145. Linda Because most people don't get struck by lightning.
146. Lucy In your experience. From what you've observed
147. Alice It happens to very few people.
148. Terry Okay when you say it doesn't happen very often, what do you mean?
149. Alice Think about the whole population. Compared to the number of times it happens within the whole population.

Excerpt 46: Henry's quantitative conception

161. Henry Also, again we're trying to remove ourselves from the problem, again, and define what unusual is. There may be a lot of people at the table

who have a mathematical definition of unusual, if you want to talk about numbers and problems and so forth, and then you may have a personal definition of unusual=

162. Terry Mmhmm

163. Henry =for instance, myself. I tend to think 'well, something's unusual if I'm doing it less that 50% of the time'. That's just sort of my rule of thumb for me as a person. If I'm not doing it 50% of the time or more, then it is unusual. It doesn't occur for the majority. It is less than the majority. It is the minority.

Line 147, 149 and 163 suggested both Alice and Henry's interpretations of unusualness involved some underlying repeatable process. To Alice, the unusualness of "people get struck by lightning" was measured by the low frequency of its occurrence over a large number of times. To Henry, an event was usual if it was repeated less than 50% of the time.

In this following dialogue, Terry directed the discussion back to the method of investigating the unusualness of "Ephram sees at least 2 people he knows".

Excerpt 47

296. Terry Well lets just say, you know, how would we investigate this. How are we going to figure out if it is unusual that—Let's say that we know the theater where Ephram works. How would we investigate if it was unusual that he saw 2 or more people? I mean you put it into a concrete context. You started to write down something. What is the first thing you said on your paper there?

297. Alice Each night record how many he knew out of the 250 and keep track of it over a long period of time.

298. Terry Okay, see Alice has the understanding that 'okay, for me to decide if it is unusual or not, I have to go down to the theater every night and somebody has to record how many people Ephram knew and if that came up a lot of times over how many nights we did it—a whole year or whatever—then we might say that it is not unusual.' And again what's 'a lot of times' would depend. But to get the students to understand that that's the question that's being asked, that what we're wanting to know what's unusual would be if a low percentage of the nights over a long period of time he saw 2 or more people that he knew, or didn't see a large proportion or—-

Alice's answer showed that she has conceived of "Ephram sees x people he knows" as a random repeatable event. This conception then led to the quantitative measurement of unusualness as the relative frequency of the event's occurrence.

Activity 1-6, Episode 3: Simulation and the cut off level

The teachers designed the simulation—simulating 100 times taking 250 random numbers from 1 to 30000, and count the number of time any number from 1 to 300 shows up. Before running the simulation, Alice asked the question:

Excerpt 48: Alice's question

545. Alice Okay I have a question. After you do all of this how do you then decide whether it is unusual or not?

Terry asked the group whether it would be unusual if "50 out of the 100 times he knew two or more people." Most of the teachers said No. But when she lowered the number to 10% and 5%, some teachers believed that 10% was low enough, but some others believed even 5% was not low enough. Pat commented at the end of this episode that 5% was a conventional cut-off level for unusualness, i.e., an event is considered unusual if it happens less than 5% of the time. At this point, Sarah raised two questions that revealed the inconsistency in her image of unusualness. I will examine her thinking in the next episode.

Activity 1-6, Episode 4: Sarah's confusion

Excerpt 49

589. Sarah I have two real quick yes/no questions, okay? One, can unusual be on the higher side. Could it occur more often than you would anticipate and also be classified as unusual? Because everything that we have talked about has been when it has not occurred as much as we anticipated. Can it occur more often than you anticipate and also fall into the category of unusual?

590. Pat Are you saying common … Can rare mean common?

591. Sarah Well I'm just saying if this is what you expect, and we've talked about it occurring less frequently being an unusual event, is that not what I've heard us say?

592. Terry I understand what you're saying.

593. Sarah Okay, so if you expected these people to show up and they showed up only 10 nights out of whatever you measured, you said that would be unusual. Well if they showed up every night, would you consider that unusual? Could it be interpreted that way?

594. Terry That, to me, is not the definition of unusual. Unusual, to me, means that over the long run, you repeated this process many, many times, a small percentage of the time would you see that particular event. A small percentage of the time makes it unusual.

595. Sarah Okay, so go back to our red and white candies. 50-50 in our bag. We would expect, over the long run, to extract half red and half white, over a large sampling over the long run. So if we extracted no reds, that would be unusual?

596. Terry No reds in one sample?

597. Sarah In a sample. Would that be an unusual sample?

598. Nicole How big were those samples?

599. Sarah I don't know. I mean, I'm just trying to get a concept here. And I guess my question goes back to if we extracted all reds, would we also classify that as unusual?

…

606. Terry … Let me ask a different question, what would make that sample unusual?

607. Sarah You get to define unusual, so.

608. Terry Okay, if we define it as being=

609. Sarah The low occurrence of something, then I'm just asking if a high occurrence of that same event could be classified as unusual, because it is on the other extreme from what you expect.

…

617. Sarah =let me say this. If you get none, if that's unusual, is getting all unusual? That all I want to know.

Excerpt 50

651. Sarah And then the other one [question] has to do with … We were limited with Ephram to two or more people. Can unusual have to do with quantity that occurs on any one given occurrence? Can it be unusual for 200 of those people that he knows to show up?

652. Terry Sure. You can ask that question. But how would you investigate that question?

653. Sarah Well, I don't know. I'm just asking for the use of the word "unusual".

654. Terry Okay, but when you say "unusual", what's the underlying implication when you say the word "unusual"?

655. Alice Over the long run.
656. Sarah Over the long run.
657. Terry If you went to the theatre 500 times, and saw how many times that he knew 200 people in the theatre=
658. Sarah Well if it even occurred once, it is going to be an unusual event because you don't expect it to occur, but … I'm not saying this the way you would say it.

Sarah's questions revealed that she left the discussion with an inconsistent understanding of unusualness. Earlier in Excerpt 41, we saw that her conception of unusualness was "Something is unusual if it is different from what you would expect". At this point, she seemed to have developed a proto-quantitative conception: "Something is unusual if it occurs less frequently than you'd expect" (line 591). She raised the question, "Is something unusual if it occurs more frequently than you'd expect?" (lines 589 and 593). The question sounded completely nonsensical— it was tantamount to asking, as Pat phrased, "Can rare mean common?" (line 590). However, her reformulation of the question as shown from lines 595 to 617 suggested that she had an entirely different image of unusualness. In her image of unusualness, there was a distribution (of number of red's in an evenly split population) and the mean of the distribution was the expectation (50 red's). She believed that the group had established that the events occurring less frequently from the expectation (e.g., 0 red) were unusual, and she wondered if the events in the right side of the distribution, i.e. occurring more frequently (e.g., 100 red's), would also be considered as unusual.

The inconsistency in Sarah's thinking was that she confused *a sample percent (relative proportion of some item in a sample S)* with the *relative frequency of samples like S over a large number of times*. In the context of Ephram's scenario where the random event was "Ephram sees *x* people he knows", *x/250* was the sample statistics.

Note that 2/250 is a small proportion, let's name it *y*. If the relative frequency of the event [x≥2], let's name it *z*, is small, then we would say the event [x≥2] is unusual. Sarah confused *y* with *z*. Her question was not "If *z* is big, is it unusual?" (Which is equivalent to "Can rare mean common?"). Instead, it was "If *y* is big, is it unusual?" (Which is equivalent to "Is it unusual that Ephram saw a whole herd of people?"). Sarah's confusion here suggested that she did not have in mind a distribution of sample statistics. Her conception of unusualness was "An event is unusual if it happens less frequently than you expect" where expectation continue to be subjective. This was also supported by her second question in Excerpt 50—Can unusualness be applied to a single instance? (line 615)

Summary of Activity 1-6

Table 37 revealed that, of the six teachers whose conceptions of unusualness were revealed in the discussion of the movie theatre scenario, all but one teacher had non-stochastic conceptions of unusualness. John applied a proportionality heuristic to "measure" the unusualness, while Sarah, Betty, and Linda had an outcome approach to unusualness. Only two teachers, Alice and Henry, conceived of unusualness as a statistical value.

Table 37: Teachers' conceptions of probability situation in Activity 1-6

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
D3A1-6	John	N					Y				
	Nicole										
	Sarah	N			Y						
	Lucy										
	Betty	N			Y						
	Linda	N			Y						
	Henry	Y	N								Y
	Alice	Y	Y	Y							Y

Alice's stochastic interpretation of the situation led the group to the simulation—simulating 100 times taking 250 random numbers from 1 to 30000, and count the number of time any number from 1 to 300 shows up. The group appeared to have achieved a consensus that "an event is unusual if it happens less than 5% of the time".

Sarah raised two further questions after the discussion on simulation and cut-off level. Her questions revealed that her conceptions of unusualness did change, from "something is unusual if it is different from what you'd expect" to "something is unusual if it occurs less frequently than you'd expect". Sarah's questions also suggested that the inconsistency in her thinking was a result of confounding *a sample percent (relative proportion of some item in a sample s)* with the *relative frequency of samples like s* over a large number of times.

Interview 2-3: Horness scale

Here is a partial data display of information gathered by the US News and World Report in 1997 on the country's top colleges.

TopColleges

	College	Reputati...	AcceptR...	Retention	GradRate	BrandVal	ClassesUnder20	ClassesOve
2	Allegheny U. (PA)	2.6	0.57	0.84		41	0.36	0.13
3	American U. (DC)	2.9	0.79	0.85	0.7	43	0.42	0.03
4	Andrews U. (MI)	1.8	0.65	0.66	0.47	39	0.68	0.04
5	Arizona State U.	3.3	0.79	0.71	0.48	19	0.28	0.18
6	Auburn U. (AL)	3.1	0.86	0.8	0.65	67	0.4	0.08
7	Ball State U. (IN)	2.5	0.92	0.7	0.54	32	0.35	0.09
8	Baylor U. (TX)	3.3		0.83	0.7	149	0.42	0.11
9	Biola U. (CA)	1.8	0.88	0.77	0.55	252		
10	Boston College	3.5	0.39	0.94	0.85	377	0.41	0.09
11	Boston U.1	3.4	0.55	0.84	0.7	125		
12	Bowling Green State U...	2.6	0.86	0.76	0.6	26	0.49	0.05
13	Brandeis U. (MA)	3.7	0.54	0.9	0.82	356	0.62	0.1

Different collegiate associations, such as NCAA conferences, were interested in developing a measure of overall association stature (you can probably guess which ones were for or against this!).

Dr. Robert Horness of Colgate University thought that the formula
mean(ReputationRating)× mean(BrandValueRating)
might be useful in this regard. A new association of 23 schools announced a score of 1300 on the Horness scale. Is that good?

Is 1300 good? The answer to this question, from a statistical point of view, entails a stochastic conception of the Horness scale. That is, if we look at a distribution of Horness scores from randomly collected samples of 23 schools, where does the 1300 lie? Table 38 summarized teachers' answers to this question. As we can see, with the exception of Nicole, all the teachers provided some rationales for evaluating whether 1300 was a good score. These rationales could be categorized into two types: 1) compare 1300 with one or a few more scores, and 2) compare 1300 in relation to a distribution of Horness scores.

Table 38: Summary of teachers' answers to Interview 2-3

John	Randomly pick 23 schools at a time, calculate the Horness scale for the sample, and repeat this. I get a distribution of Horness scales. If the distribution cluster around, say, 800, then 1300 is good because it is on the high end.
Nicole	I could take samples of size 23, and then compare them, but I still don't know, since I don't understand the underlying stuff here, why would make any sense to multiply … and basically wind up with unit squared … it just doesn't make sense to me.
Sarah	You can take some random sample of size 10, calculate the Horness measure, then you take a sample consisting of schools having high reputation and brandvalue ratings and calculate the measure. So you can compare 1300 with a random sample and a quality sample, if you take 10 good schools, they score 2000, and the random sample score 1000, then 1300 is probably okay … I'd take a random sample and a quality sample, and I'd do both and see where 1300 fall relative to those. Just do 4 or 5 random samples doesn't really tell me whether 1300 is good.
Lucy	I have nothing to compare to? I guess you can take some of those schools and calculate the Horness score and see if it gives you something else. I would probably take a random sample of size 23. I would do it 5 times to compare. All the scores centered around 900ish, so that would make 1300 pretty high.
Betty	Looks like we're trying compare different colleges. I would have see how it (1300) measured up to some other scores. Take random samples of size 23. (Luis simulated and got one sample result 1122.) According to this result, 1300 is good. (what is the criterion you use to judge is this is a good score?) If it is higher than other ones.
Linda	Well, I can take 23 schools, I can do some simulations … 23 at a time, plot a distribution.
Henry	One way is to compare with the Horness score of another set of schools with similar demographic and similar profiles. You can also try to determine what an average Horness score would be, try to develop a distribution of Horness score of all set of different colleges, and see at what point your 1300 fall in that distribution.
Alice	Consider this as a sample, and … we have to compare it to other samples of the same size. We could randomly select a sample.

Table 39 showed that majority of the teachers interpreted the situation stochastically, and compared the 1300 against a distribution of Horness scores. Two teachers, Sarah and Betty, did not have a distribution of scores in mind, and only compared the 1300 to a number of other scores. It is worth noting that Sarah's suggestion, while seemingly reasonable, in fact begs the question. First, it is a measure of conference standing, not individual school standing. Second, she proposes to select a handful of high-quality schools when quality is measured on the Horness scale. She would first need to know

what constitutes a high Horness score before she could examine conferences that score high on it.

Table 39: Teachers' conceptions of the situation in Interview 2-3

Locator	Name	1	2	3	4	5	6	7	8	9	10
		Q1	Q2	Q3	OA	ANA	PH	ANA	APV	APO	RF
I2-3	John	Y	Y	Y							Y
	Nicole	Y									
	Sarah	Y	Y	N			Y				
	Lucy	Y	Y	Y							Y
	Betty	Y	Y	N			Y				
	Linda	Y	Y	Y							Y
	Henry	Y	Y	Y							Y
	Alice	Y	Y	Y							Y

Testing Hypothesis of Population Parameter

Activity 1-3: Pepsi scenario

Overview

This activity was conducted after the discussion on Activity 1-2: The likelihood of 18 out of 30 people favoring Pepsi (see Chapter 6), in which we found that most of the teachers had non-stochastic conception of likelihood. We anticipated that this conception of likelihood would become obstacles for the teachers' understanding of hypothesis testing, As a result, in designing for the discussion of hypothesis testing, we provided the teachers with a handout that clarified a stochastic conception of likelihood. Our intention was to minimum the teachers' potential difficulty with the concept of likelihood, and to examine teachers' understanding of other conceptual issues in hypothesis testing, such as the logic of indirect argument and decision rules, etc.

The first part of the handout (Figure 5) accentuates a stochastic conception of likelihood.

?? A pollster asked 100 people which they like better, Pepsi or Coca Cola. 55 said "Pepsi". What can we conclude?

Let's consider "how likely" it is that we get results like these. To investigate "how likely … " we must assume some portion of the population actually prefers Pepsi to Coca Cola. But do not be misled—we will not make a *factual* assumption about who favors what drink. Rather, we will make *working* assumptions, such as "let's assume for the moment that people are evenly split between Pepsi and Coke."

Example: *Suppose the population of soft drink consumers is evenly split in their preferences. How likely is it that we get 55 people or more saying "Pepsi" as their choice?*

Put another way, asking "How likely is it that we get 55 people or more saying 'Pepsi' as their choice?" is like asking, "If we were to take a large number of 100-drinker samples (and take them without bias) from an evenly split population of drinkers, approximately what fraction of these samples would have 55 people or more saying 'Pepsi'?"

Figure 5: Handout of Activity 1-3 Pepsi scenario, part I

A likelihood of a sample statistics, presumes that the sample statistic of interest has some underlying distribution, for without assuming a distribution we have no way to gauge any sample's rarity. In other words, likelihood of a chance event, in the sampling context, builds on a stochastic conception of "take a sample". Thus, understanding likelihood of this particular event means to place event A in the context of repeated sampling. This requires

1. An image of a population having a certain parameter;
2. An image of repeatedly taking samples of 100 people from the population, asking their preferences of soft drink, and recording the number of people who prefer Pepsi;
3. Understanding that the likelihood of an event A is the relative proportion of A's occurrences.

The second part of the handout (Figure 6) presented the analogy of repeated sampling from an evenly split population to repeated coin tossing—a situation that we presume

teachers are more familiar with. We provided a list of 135 simulated samples of size 100 from repeated coin toss, i.e. from a population split 50-50 in preference (Figure 7).

This last way of thinking about the likelihood of finding 55 of 100 people preferring Pepsi when sampling from an evenly split population can be stated more generically. Since each person answers "Pepsi" or "Coke" the essential situation is like tossing a coin, where "heads" means Coke and "tails" means Pepsi. Collecting one sample of 100 persons' preferences is like tossing a coin 100 times, recording each toss. Collecting a large number of 100-person samples is like tossing a coin 100 times, and doing it over and over. So, if we want to collect information on how likely are samples like the one we have (55 people saying "yes") assuming an evenly split population, we can forget about buttonholing a large number of people and instead get busy tossing coins.

Here are results from a computer simulation made to generate 100 zeroes and ones randomly, repeated a large number of times. We can call "0" a head (Coke) and "1" a tail (Pepsi). Here are the results of running this simulation 135 times, each time having the program generate 100 "tosses" of a fair coin.

Figure 6: Handout of Activity 1-3 Pepsi scenario, part II

135 Samples, Each Having 100 Tosses of a fair coin

Sample	Heads	Sample	Heads	Sample	Heads	Sample	Heads
1	52	24	49	47	48	70	59
2	46	25	50	48	49	71	51
3	37	26	44	49	49	72	58
4	54	27	50	50	56	73	49
5	54	28	58	51	53	74	56
6	46	29	49	52	49	75	57
7	49	30	50	53	49	76	46
8	41	31	54	54	50	77	54
9	62	32	55	55	52	78	44
10	60	33	48	56	56	79	45
11	50	34	45	57	53	80	57
12	51	35	46	58	53	81	53
13	52	36	59	59	47	82	44
14	49	37	42	60	50	83	59
15	45	38	51	61	45	84	60
16	55	39	51	62	50	85	45
17	56	40	45	63	47	86	50
18	52	41	47	64	47	87	38
19	42	42	55	65	54	88	46
20	44	43	57	66	54	89	52
21	46	44	52	67	46	90	44
22	38	45	50	68	57	91	48
23	47	46	44	69	49	92	52

Sample	Heads	Sample	Heads	Sample	Heads	Sample	Heads
93	51	104	42	115	60	126	59
94	57	105	49	116	53	127	50
95	53	106	40	117	45	128	60
96	57	107	53	118	48	129	43
97	57	108	44	119	49	130	57
98	55	109	47	120	50	131	49
99	46	110	52	121	52	132	53
100	56	111	49	122	55	133	53
101	42	112	46	123	42	134	50
102	51	113	54	124	45	135	46
103	47	114	52	125	60		

Figure 7: Handout of Activity 1-3 Pepsi scenario, part III

The final part of the handout (Figure 8) gave instructions for interpreting the simulation results, and presented a list of questions and activities that we wanted the teachers to engage in.

The 9th row represents the 9th 100-toss sample and the fact that it had 62 heads. A 100-toss sample having 62 heads is like a 100-person sample having 62 people preferring Coca Cola. The 71st row represents the 71st 100-toss sample and the fact that it had 51 heads. This would be like a 100-person sample having 49 people preferring Pepsi.

Questions

1. What does row 37 of the above table mean in regard to coin tosses? What does it mean in regard to our investigation of people preferring Pepsi or Coca Cola?
2. Describe what the above table represents. For example, say something like "This table contains 135 entries. Each entry is made of a pair of numbers, and it represents ..."
3. Based on the simulated data given in the previous table, what fraction of the time can we expect a 100-person sample to have 55% *or more* of it favoring Pepsi when we draw from an evenly split population of drinkers?
4. Why is it important to look for "samples having 55% *or more* favoring Pepsi"? Why not just count the samples having 55% *of the sample* favoring Pepsi?
5. Assume that sampling procedures are acceptable and that a sample is collected having 60% favoring Pepsi. Argue for or against this conclusion: This sample suggests that there are more people in the sampled population who prefer Pepsi than prefer Coca Cola.
6. James argued this position: *If soft drink consumers really were evenly split, then it is not very likely that we would see 55 people or more preferring Pepsi. None of my friends can tell the difference between the two, so I think that 55 people saying "Pepsi" is really evidence that our procedure for selecting our sample and collecting our data somehow introduced a bias.* Does James have a point? Why? How should we respond?

Figure 8: Handout of Activity 1-3 Pepsi scenario, part IV

Questions1 and 2 aimed to have teachers interpret and understand the meaning of the simulation results. Question 3 related the simulation results back to the original question about the likelihood of "55 out of 100 people preferring Pepsi". Question 4 intended to plant the seed of the idea of the *p-value*—that in testing hypothesis about a population parameter, results more extreme than the observed sample statistics would count as counter evidence against the assumption about population. Question 5 was the target question of the activity. It presented a scenario that invited reasoning like this:

1. Establish the alternative hypothesis: there are more people in the sampled population who prefer Pepsi than prefer Coca Cola.
2. Establish the null hypothesis: the population is even split. Understanding that the null is the negation of alternative hypothesis.
3. Randomly take a sample from the population: in this scenario, the sample statistic is 60%.
4. Find out the *p-value*: probability of obtaining a sample as extreme as or more extreme than 60% favoring Pepsi if the null hypothesis were true.
5. If the *p-value* is less than 5%, reject the null hypothesis in favor of the alternative hypothesis. If *p-value* is bigger than 5%, do not reject the null hypothesis.

Against the background of this line of reasoning, we can talk about the kinds of support w built for the teachers through the instruction and the simulation results on the handout. That is that, it discusses what is conventionally considered a *p-value*. If the teachers were to conceptualize Question 5 as a hypothesis testing scenario, then they would know that *p-value* builds on the idea of likelihood, and that they could find the *p-value* by using the list of simulation results.

Questions 5 and 6 each called for a variation of the logic of hypothesis testing. For Question 5, the logic is: Given that 1) a *random* sample occurred, 2) the likelihood of the sample's occurrence is rare under a given assumption, we reject the given assumption. For Question 6, the logic is: Given that 1) a sample occurred, 2) the

likelihood of the sample's occurrence is rare under a given assumption, and 3) the given assumption is in fact true; we conclude that the sample is not drawn randomly.

The discussion around this activity unfolded over the course of two days, and it last 3 hours total. On day 1 the teachers worked on the problems and discussed their answers to the questions for approximately one hour. On day 2 the teachers spent two hours reflecting on the day 1 discussion.

Table 40: Overview of discussions around Activity 1-3 Pepsi scenario

Part & Episode	**Theme**
Part I	Initial discussion
Part I, Episode 1	Question 5
Part I, Episode 2	Question 6
Part I, Episode 3	Reflection on previous discussions
Part II	Teachers' reflection on the purpose of the activity
Part III	Further discussion
Part III, Episode 1	Purpose of the activity
Part III, Episode 2	Decision rule

Activity 1-3, Part I: Initial discussion

The teachers first worked in pairs, for 25 minutes, on the questions given in the handouts. They then spent 5 minutes discussing questions 1 to 4, which turned out to be fairly simple and straightforward to them. Question 5 was problematic for the teachers; the discussion around it lasted 23 minutes. Question 6 took 2 minutes. Below I will highlight the excerpts chronologically beginning with the discussion of questions 5 and 6, and embed the analysis of the teachers' thinking in the description.

Episode 1: Question 5

Assume that sampling procedures are acceptable and that a sample is collected having 60% favoring Pepsi. Argue for or against this conclusion: This sample suggests that there are more people in the sampled population who prefer Pepsi than prefer Coca Cola.

In the beginning of the discussion, three teachers, Lucy, John, and Henry, took the position that the argument *there were more people in the sampled population who prefer Pepsi than prefer Coca Cola* was false. They based this claim on the evidence that only 2.96% of the simulated samples had 60% or more favoring Pepsi (Excerpt 51 and Excerpt 52).

Excerpt 51

1.	Terry	All right, Question 5. "Assume that sampling procedures are acceptable" blah blah blah, "how likely is it that we have 60% favoring Pepsi?"
2.	Lucy	2.96%.
3.	Terry	Okay, and how'd you get two point nine six percent?
4.	Lucy	Counted. And there were 4 of them that are 40% and below.
5.	Terry	Okay and then, "Argue for this conclusion: This sample suggests that there are more people in the sample population who prefer Pepsi than prefer Coca-Cola"
6.	John	This is incorrect.
7.	Terry	Why?
8.	John	Because this is just one situation that, it just so happens that they got 60% favor Pepsi. That's why you have to repeat the samples several times, because that may not be the norm.
9.	Lucy	Because when you did it 135 times, only 2.9% of the time did you have it that high.

Excerpt 52

24.	Terry	Given, given—what's true? What was the initial assumption?
25.	Henry	If they have an average of fifty-fifty, if the mean was fifty, if the average of the population is fifty it means that according to the z score test, 2.7% of the times you would get a sample as extreme as this, meaning you get 60% favoring Pepsi or higher. That only occurs by chance 2.7% of the time.
26.	Terry	So two point seven percent, or—what? Two or three samples out of a hundred by random chance, would you get sixty or more people favoring Pepsi? Is that evidence more people favor Pepsi?
27.	Henry	No.
28.	Terry	Why not?
29.	Henry	'cause it is very small occurrence. Significantly small …

The teachers, as represented by Henry, Lucy, and John, saw the simulation as providing evidence that the population was *not unevenly* split. Their logic seems to have been: If the

population was indeed unevenly split, with more Pepsi drinkers than Coke drinkers, then you would expect to get samples like the one obtained (60% Pepsi drinkers) more frequently than 2.96% of the time. The rarity of such samples suggested that the population was *not* unevenly split. Henry's (line 29) and Lucy's (line 9) explanations indicated that they operated on the policy that "If the sample statistic is unlikely to occur, reject the argument that seemed be supported by the sample statistic". Figure 9 presents their collective logic.

1) Proposition 1: The population is evenly split.
2) Proposition 2: The population is not evenly split, as suggested by the observed sample statistic.
3) Find out probability of obtaining a sample as extreme as or more extreme than 60% favoring Pepsi if proposition 1 were true.
4) If the *p-value* is less than 5%, reject proposition 2.

Figure 9: John, Lucy, and Henry's line of reasoning for Question 5

The ensuing discussion illustrated Terry's attempt at pushing the teachers toward explicitly clarifying and elaborating their logic of hypothesis testing, and to make it the object of the discussion.

Excerpt 53

55.	Terry	Let me focus us a little bit. What we're talking about here is what we might call "unusualness". In statistics it is all based on, here is our assumption of what is going to happen, let's go out and see what is going to happen. If what actually happen in our sample is rare, unusual, occurs a small percentage of time, given our assumption, then we have to conclude there is something else going on there. The fact that of all the samples we could get, we got the one that's relatively rare to get, so what could be the reason that we got that? It could be chance.
56.	Henry	Most likely not. Most likely it is because of the process by which the data is generated.
57.	Terry	Or?
58.	Lucy	You were standing on the corner of the street … and Pepsi …
59.	Terry	Or?
60.	John	Or they are not evenly split.

61. Terry Or they are not evenly split. Everything we did was based on the assumption that they are evenly split. If I truly randomly sampled, if I come up with a sample that's very unlikely to occur, then I have to conclude that my original assumption was false, that I can't support my original assumption.

Terry, in line 55, essentially asked the question: how can we explain the tension between 1) a sample occurred, and 2) the likelihood of the sample's occurrence is rare under a given assumption[12]? Henry suggested one explanation, i.e., the sample was not randomly chosen (line 56). And John offered the other, i.e. the assumption was not valid (line 60). Together, this excerpt illuminated the logic of hypothesis testing. In the context of the Pepsi scenario, because the sample was randomly chosen, it should have implied to the teachers that the original assumption (the population was evenly split) was not valid (line 61).

This idea, however, was not understood by the remaining teachers. The following excerpt revealed that despite Terry's success in engaging Henry and John in a conversation about the logic of hypothesis testing, the other teachers thought differently.

Excerpt 54

62. Sarah As the size of your data set increases, you get closer to that evenly split thing.
63. Terry When you say "the size of your data set" …
64. Sarah I might have … anyway, what I did was I took the first column and average them and I got 49%, and then I add the second column to that, and I get 50.4%, and then as I increase the… you know the concept?
65. Henry Law of large numbers.
66. Lucy Law of large numbers.
67. Sarah Yeah, so as I increase the size of the data set, I get closer to the evenly split. Okay, now, I didn't finish. But if you take 135 and do an average, and come out with the evenly split, then your assumption is probably more accurate than if you just look at this average you got in one, kinda out of ordinary.
68. Terry The reason, why are you getting an average of 50% when you do that?

[12] The introduction of the phrase "assumption" is important to note. To place Terry's question in the context of hypothesis testing, the "given assumption" here means the null hypothesis.

69. Sarah What you are doing is that you increase the size of the sample, and so you're getting more and more close to …
70. Terry Why is it getting closer to 50%?
71. Sarah Because that's what your assumption is …
72. Nicole Validated.
73. Sarah I hope. I mean I guess, is that what I'm trying to say?
74. Terry That's where we started. I guess what I'm trying to say is, how does that answer the question of getting the sample of 60% or more. I'm not disagreeing with=
75. Sarah =It doesn't necessarily answer that. It is just saying that if you take one of those sub samples, and happen to get something that … you're going to get one of the other end upsetting.
76. Lucy Right.

This excerpt illustrated Sarah's reasoning: The sample of 60% favoring Pepsi might have been a result of small sample size. If you increase sample size, the proportion is going to be closer to 50%. Sarah seemed to believe that the sample of 60% favoring Pepsi was drawn from an evenly split population, and thus her response was an attempt to explain the question—how could it be possible that a rare sample (having 2.96% chance of occurring) has occurred? She proposed variability as the explanation. When Terry asked Sarah about the relevance of her comments to the question, Sarah acknowledged that she was not trying to answer the question (line 75). Line 75 further supported the conjecture that Sarah accepted the assumption of population being evenly split. Her conception of the situation was that of repeatedly sampling from an evenly split population, but the purpose of the repeated sampling was to approximate the population parameter (even if it was known).

Since Sarah was able to rationalize the rarity of any particular sample statistic by incorporating the idea of variability—what she phrased as the law of large numbers, it suggested that rejecting the initial assumption based on the rarity of one sample statistic

could be a shaky policy for the teachers. This conjecture became validated when Linda said,

Excerpt 55

81. Linda — I think the fact that you got one sample that's so rare to occur, you shouldn't conclude that your initial assumption is incorrect. I think it is just that we're going to show that you did get an unbiased sample, that you're not always going to get 50%.

Linda seemed to agree with Sarah that the population could in fact be 50%, and that getting a sample of 60% was simply because of variability. She also raised the theoretical question: Why reject the initial assumption based on the fact that the sample obtained was rare? She seemed to claim that one could not make any judgment based on the fact that a rare sample occurred, because the sample *could* occur however rare its chance of occurrence might be.

Pat gave an example in response to Linda's argument. He anticipated that an extremely unlikely event would probably convince the teachers to reject the initial assumption.

Excerpt 56

83. Pat — Suppose I tell you that while you're talking, I flipped out my pen and it landed on its tip and stayed there.
84. John — I will say do that 1000 more times, and I'll bet you it won't happen once.
85. Pat — Well. I'm not going to do that. But I'm asking you, do you believe it?
86. Henry — Do I believe you? If I know nothing about you, I would not believe you. But if I have a personal relationship with you, and I know that you have a tendency to tell the truth, and I know that it could happen, it'd be rare but it could happen, I might have a tendency to believe you. But if you have a equal likelihood of lying to me, then I would say that I don't believe you.
87. Pat — Why not?
88. Henry — Because it is very rare, very, very rare.

Henry's utterance (line 86) reveals that in an inference situation, he resorted to subjective judgment before using statistical information. When there was no basis for subjective

judgment, he would use a tacit decision rule—making inference about one instance on the basis of how frequently he would expect it to happen over the long run.

Henry and John eventually concurred with Pat that the data suggested that the chance of getting samples of 60% or more was sufficiently rare so as to reject the assumption that the population was evenly split.

Excerpt 57

102.	Pat	Now, suppose that you look at my pen, and it is landing on its tip, then what would you say?
103.	Henry	I would have to investigate the pen, the wire. I still would doubt it.
104.	Pat	Oh, no, you are looking at it.
105.	Henry	I have to investigate, seeing is not validity.
106.	John	We haven't been told, maybe some of the constraints of the experiment were left out.
107.	Pat	All right. In other word, what you assumed is the way it worked. You are saying it couldn't have worked the way it was assumed. Something is different.
108.	Henry	Something is different. My assumption was wrong.
109.	Pat	Yeah, so then what you are doing is that, saying that, "Gee, this happened. But I know the way these things work. And they in fact they work the way I assume they do, that it is so rare, and if it does happen, then probably it doesn't work the way I assume it works." See there is reverse logic to it.
110.	Henry	Right.

Excerpt 58

116.	Pat	Do you all see now that what that entails is hypothesis testing?
117.	John	Yeah.
118.	Pat	So we're deciding whether or not to reject the null hypothesis[13].
119.	John	Right.
120.	Henry	In which we would have.
121.	Terry	I probably would. 2.9%, that's pretty unlikely.

Pat, in Excerpt 57, highlighted again the logic of hypothesis testing: When a sample occurs, and the likelihood of the sample's occurrence is rare under a given assumption, we conclude that either 1) the assumption is right, but the sample is not randomly chosen, or 2) the sample is randomly chosen, but the given assumption is not warranted. Lines

[13] Please note that here Pat explicitly pointed out the equivalence of "the initial assumption" and "the null hypothesis".

107 and 109, expressed one variation of this logic: If 1) a sample occurred, 2) the likelihood of the sample's occurrence is rare under a given assumption, and 3) the sample is randomly chosen, then we conclude that the given assumption is not valid. Excerpt 57 and Excerpt 58 suggested that John and Henry might have understood this logic.

Episode 2: Question 6

> James argued this position: *If soft drink consumers really were evenly split, then it is not very likely that we would see 55 people or more preferring Pepsi. None of my friends can tell the difference between the two, so I think that 55 people saying "Pepsi" is really evidence that our procedure for selecting our sample and collecting our data somehow introduced a bias.* Does James have a point? Why? How should we respond?

Question 6, through James' argument, intended to highlight the other variation of the logic of hypothesis testing—Given that 1) a sample occurred, 2) the likelihood of the sample's occurrence is rare under a given assumption, and 3) the given assumption is in fact true; we conclude that the sample is not drawn randomly. The discussion around this question lasted 2 minutes.

Excerpt 59

124.	Lucy	That it shouldn't come out 55, somehow we made an error.
125.	Terry	We made an error, where?
126.	Lucy	Either select or …
127.	Terry	Okay. So somehow our data has bias in it for us to get 55, 'cause it is 50-50.
128.	Linda	You're not going to come out 50-50. I mean they can be 50-50, but you are not going to get 50-50 in your sample, unless you test everybody in the population. And then if you are correct, you're going to come out 50-50. That's the only, usually the only, I mean … just to realize that if you just take a sample, you're not going to get exactly the percentage.
129.	Terry	How could you help them to understand that?
130.	Linda	I would say that it is 55% in that sample, you are correct, but it is evenly split in the population. But since we're not, 'cause we're assuming that. Since we're not able to count everybody, we're only able to do samples, and do the best we can with our samples.
131.	Sarah	How about "we just have that 100 people here, James, we need to take a few hundred more?"
132.	Linda	Just keep doing, that's why we do so many over and over.

133.John Just give James a coin, and ask him to flip it 10 times, if it is correct, they you're supposed to get 5 heads and 5 tails, but what happens if he gets 3 heads and 7 tails, are you going to tell me that the coin is not 50-50? It just happens.

134.Terry The coin thing is a good example.

This excerpt demonstrated that while Lucy agreed with James' argument (lines 124 to 127) and was operating under the logic of hypothesis testing, Linda, Sarah, and John, were not (lines 128 to 133). They were using the concept of variability to explain the possibility of getting a sample statistic of 55% from an evenly split population. In other words, they cast doubts on James' position by arguing that the sample of 55% might have been just an odd case. This resembled the question that arose earlier from Linda's comments in Excerpt 55: Why rejecting the initial assumption (in this case, the randomness of the sample) based on the rarity of one sample? Sarah and Linda's suggestion to take more samples (lines 131 and 132) revealed their commitment to the initial assumption and their belief that in order to reject an assumption or to claim that a sample is biased they need to have more evidence than the rarity of one sample.

Episode 3: Reflection on previous discussions

The discussion on day 1 ended with a 7-minute discussion (Excerpt 60) where Terry attempted to push the teachers to reflect on their difficulties with the activity, and their understandings of the activity, both conceptually and pedagogically.

Excerpt 60

135.Terry What kinds of the things did you find problematic as you went through this activity?

136.Sarah No other than what I was overhearing. We heard "z score" (referring to Henry) and we didn't know what they were doing. We didn't know why they were doing that. I think I could come up with an acceptable answer without such words.

137.Lucy I thought so. I was questioning him the whole time.

138.Terry Do you see this activity as being useful in classroom? Where in a stat course would you use something like this?

139.Sarah It'd be really helpful EARLY ON to help them make a distinction between what they really mean and how they interpret the verbiage there.

140.Terry How does this tie into the sample we talk about earlier? Do you see that as being connected?

141.Sarah When you get to the question about the 60%, you would look particularly at one limited sample within the entire samples, which then if you'd define that, makes your population …

142.Terry That's a whole other issue, you got population and samples.

143.Sarah Right, right.

144.Terry How could you really figure out if students understand what's going on here?

145.Nicole You'd have to change the wording and doing the same types of, you have to give different, change it to more than Coke and Pepsi.

146.Henry The one thing that helps is when talking about hypothesis testing, they have to state a conclusion relevant to the hypothesis. That's really where a lot of it is crystallized for a lot of students, when they actually have to state it in a conclusion in a context of this problem. They've done the math, they've done the statistical calculation. Now explain what this means in reference to the sample you have, and what you'd expect … is complicated for them.

147.Terry What big ideas do you think the students will realize this activity is stressing? What important concepts do you hope the students would realize, pretty important concepts have to do with something like this?

148.Henry Averages, and how far things fluctuate from the average.

149.Sarah I think this is an example that you could revisit several times at your concept. Just take this set of data, and start out with what it is that you want to do, and then move to another concept, you know, you can start out with this data and do mean, you could do mean for these 135 samples, and then you can go on and write them and do your median, your mode, and then you can move on to standard deviation, and you know, just use the same data all the time. I think that's a pretty good size sample.

150.Terry Henry, you said averages, what do you mean by averages?

151.Henry The mean.

152.Terry The mean of?

153.Henry The mean of your sample in relation to the mean of the population.

154.Terry The mean of your sample?

155.Henry The mean of these samples.

156.Terry Oh, the mean of all the samples, is that what you mean, the mean of the distribution?

157.Henry Yeah. The mean of the distribution. Of course, when you take the mean of the means, which is the law of large numbers, taking samples, and finding averages, and then average the averages, it reduces your deviation. You get a more accurate distribution. This is the fundamental building block of all the statistical calculations and

> analysis that I have worked with, all build around normal curve distributions.

Terry formulated a number of questions to push the teachers towards reflecting on the idea of hypothesis testing, and discussing the pedagogical value of the activity. What was remarkable about the teachers' responses was that NONE of the teachers mentioned the logic of hypothesis testing. They did not raise it as a problematic issue (lines 135 to 137) despite their confusions. Nor did they think of it as an important concept (lines147-157) that this activity was addressing, despite evidence that suggested that John and Henry had understood the activity as being about hypothesis testing (Excerpt 57 and Excerpt 58). In particular, Sarah and Henry's answers to Terry's last question (line 147)—about the big ideas stressed by the activity—suggested that what they had conceived of the activity as being about was far from what we had intended. Henry, for example, was in the realm of parameter estimation (how to obtain an accurate estimate of the population by re-sampling), while the activity was about hypothesis testing.

Summary of Activity 1-3 Part I

Analysis of teachers' discussion in Part I leads me to develop the following framework for describing teachers' understanding of the logic of hypothesis testing.

Table 41: Theoretical constructs in hypothesis testing framework

Q		If the outcome is unusual in light of h_0, do they reject h_0?[14]
1	R h_1	Rejecting h_1
2	*s* biased	Asserting that outcome is biased
3	LOE	Reluctant to reject of h_0 for lack of overwhelming evidence
4	R h_0	Rejecting h_0
5	C h_0	Committing to h_0
6	C h_1	Committing to h_1

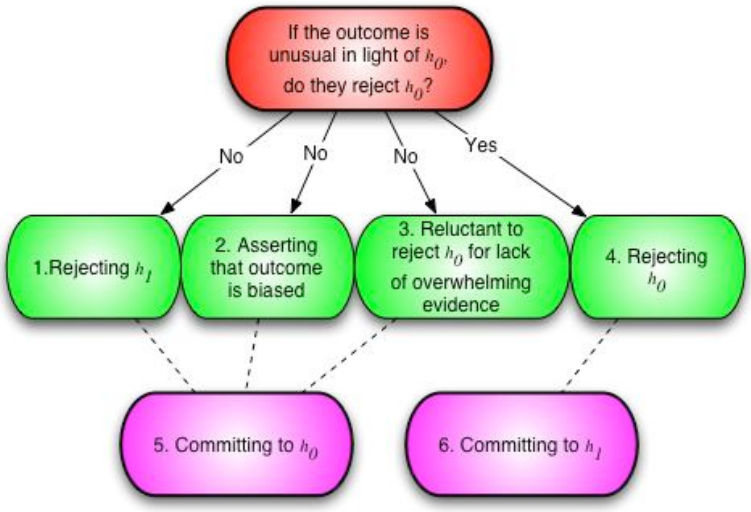


Figure 10: Theoretical framework for the logic of hypothesis testing

This framework captures the varieties of decisions (codes 1-4) people make when a small *p-value* is found. Decisions 1-3 are likely to be made by people who are committed to the null hypothesis, whereas people who are committed to the alternative hypothesis would reject the null on the basis of a small *p-value* (decision 4). Using this framework, I coded

[14] In this framework, I take the phrase "null hypothesis" or h_0 to mean "an assumption about the population parameter". That is, the reader should read "null hypothesis" with the awareness that the assumption that I refer to as the null hypothesis might not have been conceived of by the teachers as a null hypothesis in its standard sense, i.e. a hypothesis set up to be rejected in order to confirm a competing hypothesis. I chose to do so because the use of the phrase "null hypothesis" allows me to talk about the competing alternative hypothesis and relate the framework to the standard reasoning of the logic of hypothesis testing. Please keep this in mind when interpreting this framework and the descriptions following it.

the teachers' interpretations in Activity 1-3 Part I, see Table 42. The column under 1 to 4 shows the number of decisions the teachers made in each category.

Table 42: Teachers' logic of hypothesis testing

Locator	Name	1	2	3	4	5	6
		R h_1	*s* biased	LOE	R h_0	C h_0	C h_1
D1A2Q5	John	Y				Y	
	Lucy	Y				Y	
	Henry	Y				Y	
	John				Y		Y
	Henry		Y			Y	
	Sarah			Y		Y	
	Linda			Y		Y	
	John				Y		Y
	Henry				Y		Y
D1A2Q6	Lucy		Y				
	Linda			Y		Y	
	Sarah			Y		Y	
	John			Y		Y	
Counts		3	2	5	3	9	3

I would like to highlight the most distinctive feature in this table: The last two columns of the table shows that out of all 13 instances that I have coded, in 9 the teachers exhibited a commitment to the null hypothesis (the initial assumption that the population was evenly split), whereas in standard hypothesis testing, one's commitment is to the alternative hypothesis. That is, it is the alternative hypothesis that one suspects is true, and the logic of hypothesis testing provides a conservative method for confirming it.

In three cases, the teachers, John, Lucy, and Henry, tacitly accepted the truth of null hypothesis and rejected the alternative hypothesis on the basis of a small *p-value*. In one case, Henry suggested that the sample was not random, and thus no decision about null hypothesis could be made based on the *p-value*. In the other five cases, most prominently represented by Sarah and Linda, exhibited a clear commitment to null

hypothesis. They did not implicitly accept that the null hypothesis was true, but they believed that in order to reject the null hypothesis, they would need overwhelming evidence against it. A small *p-value* calculated on the basis of one sample does not constitute overwhelming evidence.

Activity 1-3, Part II: Teachers' reflection on the purpose of the activity

We began the next day by asking the teachers to write their perceptions of our intention in engaging them in the activity. In Table 43, I summarized the teachers' understandings of what the activity was about. It revealed that the teachers did not conceive of the activity as being about hypothesis testing.

Table 43: Summary of teachers' understandings of what Activity 1-3 was about

Henry	z-score process: the process of deducing the *p-value*—the likelihood of obtaining a sample as extreme or more extreme than the observed sample.
Sarah	Interpretations of data: critical thinking about data or assumptions about data.
Betty	Concepts of sampling, population, parameter estimation, and hypothesis testing.
John	The idea that assumption about a population is not always correct: if the chance of getting a sample is rare under this assumption, then we reject our assumption.
Linda	We make a claim about a population, say, "the population is evenly split", and then conduct experiment to see if the results of the experiment support the claim about the population.
Lucy	Setting up a need for hypothesis testing; modeling law of large number; simulation.
Nicole	Understanding probability, sampling, and sampling distribution.
Alice	The idea that when taking a sample from a population, the sample statistic may vary from the population parameter. It doesn't mean there is a bias in sample selection.

Although Betty and Lucy mentioned hypothesis testing, they did not think the purpose of the activity was to have them reflect on the logic of hypothesis testing.

Henry, John, and Linda's responses alluded to aspects of hypothesis testing. Henry took the activity to be about finding *p-value*. He missed the parts on developing

and testing hypotheses. John guessed that we wanted to hear that assumptions (null hypotheses) were not always right, i.e., they are to be tested in light of data/sample.

Linda's response was ambiguous: By "experiment" she could have meant either the simulation or "taking one sample from the population in question". It would make more sense if she meant the latter. If she meant the former, the results of simulating repeated samples from the hypothetical population would certainly support the assumptions about that population, which would make the activity meaningless. However, even if she did mean "taking a sample to support the claim about population parameter", she still did not understand the activity or hypothesis testing: The purpose of hypothesis testing is not to support the working assumptions one makes about a population, but to provide a conservative test of the viability of the claim one really has in mind. Linda's response suggests that her commitment was to the null hypothesis, or the working assumption.

Activity 1-3, Part III: Further discussion

Further discussion on day 2 revealed that the teachers continued to struggle with understanding 1) the purpose of the activity, and 2) the legitimacy of rejecting the null hypothesis (the assumption about the population parameter) on the basis of one sample.

Episode 1: Purpose of the activity

Excerpt 61

176.Nicole = I think the thing is that, I-I'll say that I felt uncomfortable enough … Of course I did dart out early and you—more might've happened before you all broke up but um (snickers from various teaches) … I was left a little confused about … um … where we were um … both number five and number six [questions 5 and 6 in Part I]=

…

191.Nicole =and see, where I was um … I think where I was totally misled was that … um … it never dawned on me that what I was supposed to do

was challenge the underlying assumption that is was 50-50 in the first place.

192.Terry mm hmm … Okay. All right yeah that's=

193.Nicole = Because I guess (shifting, exasperated voice) … I thought we were supposed to answer that question based on that … mm that the um that the population was 50-50=

194.Terry =Okay=

195.Nicole =I don't know if that makes any sense=

Excerpt 62

728.Terry Okay so we talked about, again, going back to what your understanding is of how doing something like this helps you to understand, would help students to understand. What other concepts or other ideas do you see this … Or do you have any understandings or any interpretations now that you didn't have at the beginning of today or that you didn't have yesterday when you left? That you could share.

729.John I think the main thing we talked about at the end of the day yesterday I'm just a little … I always assumed the assumptions are true. I think what you guys have been trying to do with these, at the end of the day yesterday and today is to show us is that the assumption could be false. I didn't ever think about it like that. That was an assumption that was *assumed* to be true. So …

The highlighted portion of these excerpts suggested that

1) Nicole learned that she was supposed to challenge the original assumption (in light of the data).
2) Nicole used to think that she was supposed to "answer the question" based on that assumption.
3) John learned that the original assumption about the population could be wrong.
4) John used to think the original assumption is always right.

Let's examine Nicole and John's thinking in 2) and 4) in light of the original question.

The original question stated that:

> Assume that sampling procedures are acceptable and that a sample is collected having 60% favoring Pepsi. Argue for or against this conclusion: This sample suggests that there are more people in the sampled population who prefer Pepsi than prefer Coca Cola.

The hypothesis to be tested in this situation (hereafter, alternative hypothesis) was that there were more people in the population preferring Pepsi. The underlying assumption that Nicole and John referred to was that the population was evenly split (i.e., the null hypothesis). If the assumption (null hypothesis) were always right, as presumed by Nicole and John, then the alternative hypothesis would be refuted logically by this assumption alone, as opposed to any investigation. This meant that Nicole and John did not conceive of the activity as being about hypothesis testing.

What Nicole and John claimed they had learned—1) and 3)—showed that Nicole and John still did not understand the purpose of the activity—testing the alternative hypothesis—and the logic of testing this hypothesis by assuming the opposite of it. They understood that a null hypothesis could be wrong and rejected, but they did not understand the purpose of the activity as of testing the alternative hypothesis, and of the relationship between these two hypotheses. In hypothesis testing, establishing a null hypothesis is a strategic move. But for Nicole and John, it was not.

Overall, this episode suggested that while Nicole and John had a commitment to null hypothesis during the discussion of Part I, they had learnt to give up that commitment. However, they still did not understand that 1) a null hypothesis is defined through an alternative hypothesis, and 2) the logic/process of hypothesis testing starts with a commitment to, or at least an interest in, (what is conventionally called) the alternative hypothesis.

Episode 2: Decision rule

Below I will analyze a 19-minute discussion centering on Linda's question: Why reject the null hypothesis (that the population was evenly split) based on the result of one sample?

Excerpt 63

730.	Linda	Are you saying that, I don't know if I heard you right, but did you say that based on that one experiment where you got a high percentage of Pepsi drinkers, or whatever it was, that we had to abandon our assumption that it was 50-50, just based on that one, that it was unlikely to occur so that you think something was wrong with your assumption, is that what you said?
731.	Henry	If you only had that one set of data, period=
732.	Linda	Is that what you said, though?
733.	Henry	Yeah
734.	Linda	Well see I have trouble with that, you know, because, I mean lightning could strike anybody but it is unlikely, it could still happen. I have trouble saying, just based on that one, let's abandon our hypothesis.
735.	Terry	That's true, you're right you could be wrong …
736.	Linda	So you wouldn't want to do anything based on that one experiment. I wouldn't think so.

In this excerpt, Linda expressed her resistance to the idea of rejecting null hypothesis based on one sample. She believed that the sample statistic, albeit rare, could have occurred by chance. Her argument—conveyed from the example of lightning (line 734)—was that even if an event was unlikely, it could still occur, therefore one should not reject the null hypothesis solely on this basis. This reluctance to reject the null, I conjecture, could be a result of her concern over wrongly rejecting, or in her words, abandoning the null hypothesis. Terry's comments in part supported this conjecture. She said, "You are right you could be wrong", meaning, Linda's reservation about rejecting null hypothesis was justified because we could wrongly reject the null hypothesis if what's highly unlikely did happen by chance.

Terry pursued Linda's thinking by asking what she would do alternatively if she was not willing to reject the null hypothesis.

Excerpt 64

747.	Terry	Okay, so … If you don't want … If changing your assumptions is going to come to, you know, make great difficulty for you as the person who has got this research or whatever. What could you do …
748.	Linda	Look at more cases.
…		
759.	Terry	When you said look at more cases, what do you mean?
760.	Linda	Do more experiments.
761.	Terry	Okay, so like instead of 100 … This graph represented 100 samples, do you mean do more samples?
762.	Linda	No. I'm saying don't make a decision based on that one that was so different. Why change your hypothesis based on the one that was so different?
763.	Terry	You mean sample another 100 people and see what you get?
764.	Linda	No, I'm just saying I wouldn't abandon the hypothesis based on the one that was so different!

This exchange further revealed Linda's thinking. When Terry asked her what she would do instead of rejecting null hypothesis, she responded, "Look at more cases/do more experiments" (lines 748 and 760). This supported an earlier conjecture: Linda was committed to the null hypothesis, and therefore she would reject a null hypothesis only if there were overwhelming evidences against it.

When Terry asked whether by that she meant collecting any other real sample, or simulating selecting more samples from the hypothesized population, she denied both, reiterating her point that she would not abandon the hypothesis based on one sample (lines 762 and 764). She added that " … one sample *that was so different*." What did she mean by that—from what was that sample different? In the context it seemed that she could have meant two things: 1) the difference between the sample statistic (60%) and the assumption about the population (50%), and 2) the difference between the sample statistic and the majority of the collection of sample statistics. In either case, what she

said added additional insights to my early conjecture. Earlier, I conjectured that Linda did not want to reject the null hypothesis based on ONE sample, because she was concerned about rejecting a true hypothesis. In this exchange Linda's stress on the phrase "one sample that was so different" indicated that she was not oriented to rejecting the null hypothesis. Because if she was so oriented, then the larger the difference (in a distributional sense, not additively), the more doubt would be cast on the viability of the null hypothesis, in which case her objection to "abandoning the null hypothesis based on one sample that was so different" would be nonsensical. Linda was oriented to "confirming the null hypothesis" (see also her written responses in Part II). Therefore the smaller the difference is, the more likely the null hypothesis is true. In this case, her objection to "abandoning the null hypothesis based on one sample that was so different" would be reasonable.

The following discussion (Excerpt 65) provides evidence for this conjecture about Linda's commitments, and also illuminates the other teachers' understanding of the issue under discussion.

Excerpt 65

771.Henry	[To Linda] You're looking for an absolute when there isn't one. Let me see if I understand what you're trying to say here.
772.Linda	I'm not. I'm saying that if you're going to abandon your hypothesis based on that one sample, that's an absolute=
773.Henry	Right, that's what you're saying=
774.Linda	=I'm saying don't do that. I'm saying you need more information.
775.Henry	But what you're saying is they haven't run 100 samples of 100 people. We're just talking about having gathered 100 people, and you got the one case of 100 people and you got this extreme data and you're saying 'okay, now this is saying that we should throw it out, but is our hypothesis really bad enough?' Well that's the dilemma that you're in. You have to figure out 'did I by chance alone get the chance occurrence, or is it more likely that the hypothesis is incorrect?'
776.John	It is kind of like, you know, when they test drugs and medical testing. You can't sit there and just keep using people and using people, you're

going to start killing everybody if you keep continuing sampling. So you have pick 10 or 20 and you do your test on them, you test the vaccine or whatever you're going to test, and then from there you have to kind of make … I mean, you can't just keep doing this test on people and then they, you know, you kill 10,000 people before you know it, so … I think that in that, if you were going to say where would you have to use it, I'd think somewhere along those lines, in the medical … Where human life or something precious is at risk. Here, so what? It is just Pepsi/Coke, we can just keep sampling it, right?

777. Linda No, I'm saying, based on one experiment, why throw out your data?

778. John I know, what I'm saying is … I'm saying in the position=

779. Henry =There is a choice to be made. You have to decide=

780. Linda =I thought we had decided.

781. Alice So there is no right or wrong?

782. Henry There is no absolute in statistics. There is more right, more wrong, usually, but there's no absolute.

783. Various (Laughter)

784. Henry Look, when you've got to make that choice and you've got to decide okay, your occurrence=

785. Linda =Well you don't have to decided based on the one, just wait and see=

786. Henry =this one that you have should've only occurred less than the four times out of 100. So, is it likely that you got the one that occurred less than four times out of 100 or is it more likely that there is something wrong with the hypothesis?

787. Linda Well, you would wait. You wouldn't kill more people, but you would wait until more people, wait and see what'd happen, you know, look at the history. You wouldn't have to do experiments=

788. John =What I'm saying is, with this Pepsi/Coke, so what? You test, you go, you get your 100. 60 people say they like it … No. What I'm trying to say is, you get your 100, 60 of them say they like Pepsi. All right you don't have to be satisfied with it because it is not going to hurt anything to go sample another 100. And then I would look at it they way you're saying. I would keep sampling, continue sampling and not be satisfied with that. But what I'm saying is there's probably some situation where you get your 100 and that's it! You're going to get the 100, you're going to get your proportion, and you have got to make a decision from that and I think that's what they're trying to get at with it. What do you do with that? If that's all you have, that one proportion, how do you make your decision? I think that's what's we're trying to talk about today.

789. Henry And if you could retest, if you could redesign the experiment and repeat the experiment, and if you could investigate all the variables and look for any confounding factors, then you'd have to ask yourself, your sample should only occur less than 4% of the time, so either that happened, or there is this larger possibility that your hypothesis is wrong and I would go with the hypothesis.

790.Linda	Right. I agree with that. But I'm saying that I thought she (referring to Terry) said that we'd needed to change our hypothesis. I though that's what she said.
791.Terry	That was my decision. That's my personal decision. And I made that decision based on the decision rule that was in my head. But you don't have to use the same decision that I used.
792.Linda	Oh. Okay!
793.Terry	If you're the Pepsi company, you may have a different decision rule. I have Coke stock. I have different decision rule than someone who really likes Pepsi may have … or somebody who likes that little curly-headed girl.

In this excerpt, Henry elaborated the logic of hypothesis testing (lines 775, 786, and 789). His idea was: Since an event that's highly unlikely to happen (under a null hypothesis) did happen, one has to decide whether it is more likely that this event happened by chance, or that the null hypothesis was in fact not true. In other words, rejecting a hypothesis would be a choice to be made, as opposed to a conclusion about the absolute truthfulness of the null hypothesis. Linda insisted that she would not abandon the null hypothesis (or in general, make a decision about it) based on one sample, and that she would need additional information (lines 772, 774, 777, 785, and 787). In lines 776 and 788, John tried to push Linda into making a decision from one sample by raising situations where no additional samples would likely be collected. John's utterances indicated that he also had a good understanding of the logic of hypothesis testing. But missing from both his and Henry's arguments was the idea of decision rule. The idea of decision rule is that if we say that we will reject the null hypothesis when the *p-value* is less than 5%, it means that if we could take sample repeatedly, over the long run, we would wrongly reject a true null hypothesis only 5% of the time. Linda was afraid that basing decisions on one sample would lead to FALSE decisions. But without the idea of

decision rule, she did not know how her fear of committing a Type I error would be eliminated or controlled, besides insisting on collecting more samples.

The discussion ended with Terry saying that it was her decision to reject the null hypothesis, and Linda said, "Okay." With what was she agreeing? It seems that what she understood was that rejecting the null hypothesis was Terry's *personal* decision and that she did not have to agree with it. In other words, Linda's "okay" indicated an agreement to disagree. The following two excerpts from later discussion supported this conjecture by revealing further evidences of Linda's reluctance to rejecting the null hypothesis.

Excerpt 66

799.Terry	Can I ask a question? Is there a point at which you would feel comfortable abandoning your initially assumptions, or saying that you disagree? Is there a point at which you would feel comfortable saying that my initial assumptions are false based on your one sample?
800.Linda	It would depend on what the application was. If it were human life at stake, I would say ehh, this person died when we use that, so, let's substitute=
801.Terry	What if you got a sample that occurred, um, 1 out of 10,000 times? Based on your assumptions if your result that you got from your sample occurred, by chance, 1 out of 10,000 samples? How would you feel about your initial assumptions? What would your conclusion be?
802.Linda	It would make me doubt it.
803.Sarah	If I am that patient, and the doctor tells me that this medication will probably cure what you have, but we have 1 out of 10,000 patients die, I believe I'm going for it.
804.Linda	You could still be that 1 of 10000, though. But that doesn't mean that the doctor was wrong.
805.Sarah	That's right. I mean, you know, but if he says well we have 4800 out of 10000 die, I'd say 'hmm I think I'll take my chances somewhere else'.

In this excerpt, Terry asked Linda at what point she would feel comfortable rejecting the null hypothesis. Linda said that even a chance as small as 1 out of 10,000 would only make her doubt the null hypothesis but not reject it (line 802). Her concern about whether or not the doctor was wrong (line 804) was analogous to her concern about whether or

not the null hypothesis was wrong. This again revealed that for Linda rejecting a null hypothesis meant making a conviction that the null hypothesis was wrong. A small *p-value* on the basis of one sample did not constitute as evidence strong enough for her to make that conviction.

Excerpt 67

806.Pat	Um, an observation about the conversation in total. It seems like that there are two different perspectives at play, and people aren't announcing which one they hold. One perspective is: suppose that we are trying to make a decision about the case that sits in front of us right now, okay? Like, should we give this person this medicine? Should we convict this person and assign him the death penalty? So, what you're trying to do is to make a decision about the specific case, then that's a non-probabilistic situation. You just investigate the heck out of it until you're satisfied that you're, you know, convinced of whatever you need to be convinced about. But if what you're saying is, 'Okay, I'm going to be making this decision over and over and over and over, I'm going to make this decision in a lot of cases, and I'm going to apply the same criteria to make my decision. Where should I set my boundaries as to where and when I will say yes and when I'll say no? So it is like a policy decision. You see the difference?
807.Pat	Now, Linda, which one do you think you were talking about when you were talking about the death penalty? The individual case or a policy decision?
808.Linda	I wasn't talking about an individual case. I was saying the proponents or opponents of death penalty in general. That they argue that, hey there might be … remember we found this one person who was innocent and got taken off death row, well the opponents of the death penalty could say, 'oh, that was one person out of 1 million. Our hypothesis is still valid, that it held … that … whatever the hypothesis is. In other words, the Supreme Court wouldn't overturn the death penalty based on one case. Like what we're saying here, that's the one case, where we had an unusual occurrence. That's what I was saying.
809.Pat	But what's the hypothesis in that case? That the death penalty exists?
810.Linda	No. That it is … that only people who deserve to be executed get executed.
811.Pat	That there are no errors.
812.Linda	mm hmm (confirms).
813.Pat	That's the hypothesis?!
814.Linda	I guess.
815.Sarah	Well if you find one error you can disprove your hypothesis. It is like you need direct proof for geometry.
816.Linda	Well, I guess it was a bad example.

817.Pat No, no, it is a good example. But I think it is an example of *what* is the question.

818.Linda Basing your decision on one occurrence.

In this excerpt, Pat raised the distinction between two perspectives: policy decision and individual case. He believed that Linda had a perspective of individual case, that is that, she was thinking about making a decision about that one particular null hypothesis, as opposed to designing a decision rule that would apply for situations like this over the long run. Linda did not seem to understand this distinction. She raised another analogy to express her concern over "basing your decision on one occurrence". Her exchange with Pat, in lines 811 and 812, constitute the most affirmative evidence for Linda's deterministic orientation, that is, she was looking for the truth of null hypothesis; she did not have the scheme of thinking that would justify the decision about a particular hypothesis.

Summary of Activity 1-3, Part III

In this Part, teachers continued to discuss Activity 1-3 and reflect on their earlier discussion in Part I. While Part I showed that majority of the teachers had a commitment to null hypothesis, in this Part of the discussion, at least two teachers, Nicole and John, had given up this commitment. However, this act alone did not constitute understanding of hypothesis testing because they still did not understand that 1) a null hypothesis is defined through an alternative hypothesis, and that 2) hypothesis testing starts with a commitment to the alternative hypothesis.

Part II and III further revealed Linda's thinking about the activity. Linda initially believed that the activity was about confirming the null hypothesis: We make an assumption/null hypothesis about a population, and then take a sample to see if this

assumption was accurate (see Part II). However, since the discussion was about whether or not to reject the null hypothesis, Linda vehemently opposed the idea of "rejecting an assumption based on one sample". Her argument was that no matter how rare the sample was, it could still occur, and thus it couldn't be used to reject an assumption. A mixture of beliefs and orientations helped to explain why she opposed to rejecting the null hypothesis. This beliefs and orientations include, as I have illustrated with many evidences,

1) Linda was committed to the null hypothesis. She would reject a null hypothesis only if there were overwhelming evidences against it. Therefore, she opposed to "rejecting the null on the basis of one sample" and proposed to take more samples to see if the null hypothesis was *right* or wrong (Note the "right" in this sentence means that she was not committed to reject the null, but confirming it).

2) Linda was looking for the truth of null hypothesis. Rejecting a null hypothesis, to her, means making a conviction that the null hypothesis was wrong. Because of this belief, she opposed to "reject the null hypothesis on the basis of one sample" because any rare sample could still occur theoretically.

Interview 2-1: Alumni association

The Metro Tech Alumni Association surveyed 20 randomly-selected graduates of Metro Tech, asking them if they were satisfied with the education that Metro gave them. Only 61% of the graduates said they were very satisfied. However, the administration claims that over 80% of all graduates are very satisfied. Do you believe the administration? Can you test their claim?

This interview question presents a typical hypothesis testing scenario—There was a stated claim about a population parameter: 80% of all graduates of Metro Tech were very

satisfied with the education that Metro gave them. A random sample of 20 graduates found that only 61% of them said they were satisfied. The implied question was, "Will the samples like or more extreme than 61% be rare enough for one to reject the claim of 80%?"

Almost all the teachers noticed the large difference between 61% and 80%, and they believed the small sample size was the reason why there was such a big difference. When asked whether they believed the administration's claim, the teachers had different opinions (Table 44). Two teachers said they did not believe the administration's claim. Four teachers said they did. Henry and Alice based their choice on the fact that 80% was possible, despite its difference to the sample result. Sarah, however, did not know that 80% was a claim. Rather, she thought it was a sample result. The other two teachers were hesitant in making a decision, with one of them, Lucy, leaning towards not believing the administration.

Table 44: Summary of teachers' answers to I2-1, Q1: Do you believe the administration?

John	I can't say either way.
Nicole	No.
Sarah	Yes. I have no reason to doubt administration's claim.
Lucy	I don't know. I would say the administration is a little high in their claim.
Betty	I think it needs more information to back that up. From this one sample, you can't believe them. You wouldn't believe it as well as if you have more samples to back up.
Linda	I'd say it is possible that 80% is true. First of all, I don't have reason not to believe the administration. The sample result 61% wouldn't lead me not to believe the administration. At least 80% is large, is bigger than half.
Henry	Yes. I would still believe the administration.
Alice	Yes. I think it is possible. They survey only 20 students, so it is possible that the 61% could be an extreme.

When asked how they would test the administration's claim, only Henry initially proposed to use hypothesis testing.

Table 45: Summary of teachers' answers to I2-1, Q2: Can you test their claim?

John	Take many samples of size 20, or take a larger sample.
Nicole	Repeatedly taking samples of 20, seeing how many times 17, 18, 19 people are very satisfied.
Sarah	Take many samples: take 5 samples of size 20, average the percentages, and see if that got you closer to 80%. Take 5 samples of size 20, if the percentages are close to 61%, then claim of 80% is too high. If the percentages are higher than 61%, then maybe 80% is okay.
Lucy	I will tell them to do it (take a sample) again. Compare the two. If this one is 61, and the other one is close, then I'd say 80% is pretty high.
Betty	Take another sample … if you take one more, and it gives you 61%, or somewhere close that, then it would give you more doubt about the 80%.
Linda	I could go out and do a survey of all the students on the campus.
Henry	We could test the ratio. 80% of 20 are supposed to be satisfied, but only 61% of 20 said they are. You could do a z test analysis to figure out how unusual it would be, or what percentage of the time you would get a percent as low as 61%.
Alice	By doing this repeatedly, or choosing a larger sample, which would be more representative of the population.

The methods the teachers proposed fall into the following categories:

1. Take many samples of size 20 (John, Nicole, Sarah, Alice)
2. Take a larger sample (Alice)
3. Take one or a few more samples of size 20 (Lucy, Betty)
4. Survey the entire population (Linda)

I conjecture that the reason that the teachers did not propose hypothesis testing was because they did not see this question was a typical one of the kinds of question that hypothesis testing was meant to solve.

Teachers' elaboration of these methods showed that they were not trying to design a policy, in the sense that their methods were not well defined so that other people who might use their methods would reach the same conclusion as they did. For example,

Excerpt 68

1. Luis how could you test that?

2. Betty Take another sample.
3. Luis You mean take one sample.
4. Betty Right. One sample can give you more information. Take more samples would be better.
5. Luis Tell me how one sample can give you more information.
6. Betty Well, if you take one more, and it gives you 61%, or somewhere close that, then it would give you more doubt about the 80%.
7. Luis So two samples are enough?
8. Betty Well, no … well, I mean, how much is enough? Maybe it will depend on what the next one would be. If the next one came out closer to 61%, then that's enough for me to question the 80%. If the next one came out close to 80%, I might take another one (laughing).
9. Luis Supposed you get a 62%, and then an 85%, what would you do?
10. Betty I might take another one.
11. Luis Say the fourth one gives you 77.
12. Betty If two of them are close to 80, then I might think that the 61was just on the lower end of the range there, or an odd sample possibly.
13. Luis What would lead you to believe or not believe the 80%? Will you stop at 3 (samples)?
14. Betty If there were two close to 80%, I will believe it. If two close to 61, then I would doubt it, but I don't know if I will take another or not.
15. Luis I got a sense that you're thinking about multiple samples, but I don't know where you will stop?
16. Betty I'm not sure. I think maybe I need to know what the population was … no I wouldn't … yes, I would.

This excerpt shows that Betty could not specify a decision rule. Similar to Betty, the teachers who proposed to take a larger sample did not specify a decision rule about how the result of the larger sample would determine the truthfulness of the claim.

The methods proposed by the teachers also indicated that they assumed that they would have access to the population. In response to this observation, the interviewers asked a follow-up question: Is there a way to test the claim without actual sampling? This question prompted some teachers to think about hypothesis testing. Out of the seven teachers who initially did not think in terms of hypothesis testing, three teachers, John, Nicole, and Lucy, proposed hypothesis testing. Henry, who proposed hypothesis testing initially and expressed it in terms of "z test analysis" was asked whether he could

investigate the claim without using z test. Below are the answers given by these four teachers.

Table 46: Summary of teachers' answers to I2-1 following-up question: Is there a way to test this claim without actually sampling?

John	Assuming the population is over 80% in favor. 100 integers, 1-80 satisfied, 81-100 not. Take 20 samples of size 20. Get a frequency distribution which center around 0.8. If most of samples center around 0.61, then the claim was false.
Nicole	Calculating the probability of getting 80% or more if in fact the population percent is 61%...100 marbles, 61 red, 39 blue. Repeatedly take samples of size 20. See how many times you get 80% or more.
Lucy	You could do a hypothesis testing. You could test, based on 61%, where this would fall … (drew a normal curve, 61% in the middle) … assuming 61% is correct, then you could see how many standard deviation that 80% would be away from it. If it is way out there, then I'd say they need to change that number. If it is only one standard deviation away, then I wouldn't feel so bad. I wouldn't argue with them too much.
Henry	I want to do a simulation, create a population where 80% of the population is satisfied, sample that population in the sizes of 20, sample that many times, for each sample calculate that percent, and compare those percents, and even average the percent of each sample and see if that approaches 80%.

As we can see, John and Henry had the same conception that Lucy and Henry exhibited during the discussion of Activity 1-3 Pepsi scenario. They started out with the assumption (null hypothesis) that 80% of the population was satisfied and created a distribution (John) or collection (Henry) of sample statistics. Then, John said, "If most of samples center around 0.61, then the claim was false". He did not realize that most of samples *will* center around 80% given that that was his assumption about the population parameter. Henry said, " … compare those percents, and even average the percent of each sample and see if that approaches 80%." He implied that if the average approached 80%, then the claim was true. Like John, he did not see that the average *will* approach 80%. In other words, both John and Henry failed to see the connection between their assumptions of the population parameter and the characteristics of the resulting distribution of sample

statistics. We can conjecture, from this instance, that seeing this connection is a necessary condition for understanding hypothesis testing.

The other four teachers, Sarah, Betty, Alice and Linda, insisted that they would need to have access to the population in order to investigate the administration's claim. The following except shows that even when the interviewer provided increasing support for her to conduct hypothesis testing, Linda could not succeed at it:

Excerpt 69

1.	Pat	Could you test them?
2.	Linda	I could if I know where it was coming from. Surely it is based on something.
3.	Pat	Whatever it is based on, is there any counter evidence to it?
4.	Linda	No, I don't (have any counter evidence)
5.	Pat	Could you use the sample result to test the claim?
6.	Linda	I don't know.
7.	Pat	Is there any way you can think of this as a hypothesis testing situation?
8.	Linda	Go and do the survey several times
9.	Pat	Could you use simulations?
10.	Linda	Yeah … you could use 61%, 20 people … see how many times that interval contains 61% … you could use the 20 … you draw a normal curve?
11.	Pat	Someone else started this: set up 100 integers, 1-80 satisfied, 80 to 100 not satisfied. Can you start from here?
12.	Linda	You could do several simulations. Assuming 80% satisfied. You can look at the frequency of percentages … I'm blank. I'd simply need to know the shape of that curve contains 80%.

The interview with Alice turned out to be one of the most confirmative cases for our earlier conjecture that the teachers were not aware of the effect of assuming a working/null hypothesis. The conversation showed that Alice understood that to conduct a simulation, one has to have a population whose parameter is well defined.

Excerpt 70

1.	Pat	Suppose you don't have access to the students in metro school, how would you test the claim?
2.	Alice	I don't think you can test it. You can only draw inference on it.
3.	Pat	Could you use simulation?

4. Alice — We would repeat taking samples of size 20, and look at the distribution of sample percents.
5. Pat — Would you have to make any assumption about the populations to do the simulation?
6. Alice — Each one has the equal chance of being selected.
7. Pat — That's about each item in the population. What about population as a whole?
8. Alice — Not sure how to answer that.
9. Pat — Would you have to assume that 80% are satisfied?
10. Alice — Based on their claim?
11. Pat — To test their claim.
12. Alice — That's what you would expect the outcome to be.
13. Pat — That their claim is true. But to do a simulation, what would you sample from?
14. Alice — All the graduates in metro.
15. Pat — When you collect samples, how do you know what you have?
16. Alice — Okay. We'd have to assign, I'm not sure, numbers from 1 to a certain value would represent yes, and from that value to the population would represent no.

Pat and Alice started to conduct a simulation. They agreed that they would assume that 80% of the population was satisfied, as the administration claimed. When Pat suggested that they were looking for the chance of obtaining a sample as extreme as or more extreme than 61%, Alice was surprised. She believed they were looking for the chance of getting samples of 80%. In other words, she believed that she was supposed to look for what she had assumed. It could be that she did not have an underlying scheme that connected the population parameter to the simulation results. What she was thinking was, "We want to see if 80% was correct, we conduct a simulation, if the chance of getting 80% is not very unlikely, then 80% is correct." A person who has made the connection between population parameter and the simulation knows that the chance of getting 80% from a population that split 80-20 is far more likely than those of getting any other sample. When Pat told Alice, "We're not looking for what we assume, we're looking for how many times we got the sample result," she said, "So we're not using the 80%, we're

using the 61%?" She appears to have meant, "So we should assume that the population parameter is 61%, instead of 80%?" Her conception of what they were supposed to do had not changed.

Excerpt 71

1.	Pat	What value would that be?
2.	Alice	I got caught up with that. I don't know.
3.	Pat	Could you do it that, from 1 to 1220 is yes?
4.	Alice	Where did you get 1220?
5.	Pat	61% of 2000.
6.	Alice	Okay. That's that one sample. If the administration is correct, then 80% of 2000, so 1600.
7.	Pat	So from 1 to 1600 would be yes, 1600 to 2000 would be no. (Pat conducted the simulation with ProbSim. 100 samples of size 20. assumption 80%)
8.	Pat	We're looking for 61% or fewer.
9.	Alice	We're looking for 80%, aren't we?
10.	Pat	No. That's their claim. We're not looking for what we assume, we're looking for how many times we got the sample result.
11.	Alice	So we're not using the 80%, we're using the 61%?

The following conversation confirmed our conjecture about Alice's conception. She argued that the administration's claim was accurate because the simulation result suggested so.

Excerpt 72

1.	Pat	We got 8 or more No's 5% of times. So 61% or less occur less than 5% of the times.
2.	Alice	12 or fewer say yes.
3.	Pat	So if the administration was telling the truth, how often would we see the result like 61%?
4.	Alice	8 or more say No, 12 or fewer say yes …
5.	Pat	What are we assuming?
6.	Alice	We assume 61% of the population says yes.
7.	Pat	No, we assume 80%.
8.	Alice	I think the alumni association's claim is invalid.
9.	Pat	No, they didn't make any claim. They only did the sample. The administration made the claim.
10.	Alice	Oh, I think the administration's claim of 80% is accurate.
11.	Pat	That's because the alumni association found 61%?
12.	Alice	No, it is because of the outcome of our sampling process.
13.	Pat	It is because 95% of the times we got more than 12 yes's?
14.	Alice	Yes.

In sum, teachers' responses on this interview question suggested they did not employ spontaneously the method of hypothesis testing for the situation. Instead, 7 out of 8 teachers proposed methods of investigation that presumed that they would have access to the population, and none of these methods were well-defined policies that would allow one to make consistent judgment. This led to our conjecture that even though the teachers might have understood the logic of hypothesis testing, they did not understand the functionality of it. In other words, they did not know the types (or a model) of questions that hypothesis testing was created for, and how hypothesis testing became a particularly useful tool for answering these types of questions. Hypothesis testing, as a tool for making statistical inference, was created because in situations involving statistics information, very often it is impossible or non-economical to attain the population parameter. Hypothesis testing ensures that we can make reasonable judgment on the viability of the hypothetical population parameter with a random sample. In general, hypothesis testing answers the prototypical question of whether two sets of observations are similar, observations could be a sample result or a true or conjectured population parameter. Essentially, to be able to employ hypothesis testing spontaneously entails that one connects the question at hand to this model of questions that hypothesis testing could answer.

With the interviewers' prompt, four teachers suggested the logic of hypothesis testing. But only one teacher's elaboration of hypothesis testing was correct. Nicole set up the sample statistic as the null hypothesis. Henry and John believed that they were supposed to look for the chance of obtaining a sample of 80% given the population parameter is 80%. The other four teachers initially responded that they could not

investigate the administration's claim without access to the population. The interviewers provided additional support for Linda and Alice. The conversation with Linda revealed that even with increasing amount of support, Linda still did not understand the logic of hypothesis testing. The conversation with Alice revealed that she shared Henry and John's belief about what to look for after establishing the null hypothesis.

Testing Hypothesis of Randomness

Activity 2-3: Rodney King scenario

In 1991 Rodney King was involved in an incident that led eventually to some of the worst riots in US history. The New York Times (March 18, 1991) reported, "at least 15 officers in patrol cars converged on King." The article broaches one of the issues that made this incident explosive: "In what other police officers called a chance deployment, all pursuing officers were white." The force, which numbers about 8,300, is 14% black.

Critics denied that this was a "chance deployment," claiming instead that all pursuing officers being white reflected an underlying prejudice within the LAPD dispatcher office.

1) Consider the statement "In what other police officers called a 'chance deployment', all pursuing officers were white."

- What does the statement itself mean?
- What is meant by 'chance deployment'?
- Is there a claim implied in this scenario? If so, what is the claim?

2) Do you think that "chance" is a reasonable explanation for all pursuing officers being white? Discuss how you might investigate this question.

3) Suppose you are a member of the California Supreme Court and that before you is the matter of what level of likelihood may be taken as "evidence of racial bias in police dispatching procedures." What level would you set for that and future cases?"

Overview

The nature of the dispute in the Rodney King scenario is about the randomness of the deployment of police officers. The police claimed that it was by chance that all 15

officers were white, whereas the critics argued that the deployment was not a random event. The solution to this dispute lies in the question: how likely is it that we get all white officers if we randomly select 15 police officers from a population of 8300 officers, 14% of whom are black?

In working with high school students on this problem, we found that they conceptualized this question in two ways. The first one, which we called a detective's perspective, was a non-stochastic conception with an outcome approach interpretation: It investigates that particular deployment—looking for additional information about what happened that night that would help explain why all 15 pursuing officers were white. This interpretation treats the situation in question as a single unrepeatable event. The second one was a stochastic conception. It asks the question: Of all the possible random deployments of 15 officers, in what percent of deployments do we see all white officers? If the relative frequency of all white officers is smaller than 5% (a pre-determined cut-off level), then the fact that all 15 officers were white in the Rodney King scenario could lead to the conclusion that the deployment was not random and that it indeed reflected racial prejudice.

Discussion around this activity lasted for 104 minutes. I will divide the discussion into 3 parts, each focusing on one set of questions asked in the handout.

Table 47: Overview of discussions around Activity 2-3 Rodney King scenario

Part & Episode	**Theme**
Part I	Question 1
Part II	Question 2
Part II, Episode 1	How to set up simulation
Part II, Episode 2	Assumptions about simulation
Part II, Episode 3	Decision rule
Part II, Episode 4	Simulation results
Part II, Episode 5	Interpretations of simulation results
Part III	Question 3

Activity 2-3, Part I: Question 1

Consider the statement "In what other police officers called a 'chance deployment', all pursuing officers were white."
What does the statement itself mean?
What is meant by 'chance deployment'?
Is there a claim implied in this scenario? If so, what is the claim?

Our intention in giving the first set of questions was to prompt the teachers to make sense of the scenario. We hoped that discussion around these questions would clarify the meanings of the arguments held by the two sides of the debate. Below I will elaborate what meanings we hoped the teachers have, and compare to the understandings that they actually had. As will be revealed later, such a comparison is important for us to make sense of the difficulties the teachers experienced later on.

The key to understand the statement in question is a stochastic conception of deployment. To claim a deployment a "chance deployment" means that it is a result of a *random* and *repeatable* process, i.e. the process of randomly selecting 15 officers from the population of police officers. A deployment is not a "chance deployment" if the relative frequency of deployments like this one occurs significantly rare. This is the meaning of "chance deployment" that we hoped the teachers would have. Discussion revealed that the teachers did not have a stochastic conception of the deployment.

Excerpt 73

2. Terry Okay, everybody had a chance to … to read the scenario, that I'm sure we're all familiar with. So what we want to do first is talk about, that statement, which I think is the last statement of the scenario … ah, and "what officers called a chance deployment all pursuing officers were white." First question is, "what is that statement, what does that statement mean?"
3. Linda The police officers are saying that it was … just a random occurrence … that there was no thought put into … race when they deployed the police officers that were sent, that it happened just by chance, it was just the luck of the draw.
4. Terry Okay, all right, so, sort of getting into the next question, "what is meant by chance deployment?" I guess that's what you're saying. (pause) Can we state that probabilistically? (pause)
5. Alice Of all officers … ah I don't know … (pause)
6. Linda You wouldn't be able to state what they said probabilistically, you could make a probabilistic statement from=
7. Terry =About chance,
8. Linda fourteen percent.
9. Terry specifically "chance deployment", I guess.
10. Terry You'd want to talk about the concept of "chance deployment" and talk about … what that means probabilistically.
11. Linda If you interpreted the paragraph up above I would say that, over a large number of deployments fourteen out of a hundred officers that responded, would be, or whatever ratio that is, would be black.
12. Lucy You mean fifteen at a time out of that eight thousand three hundred?
13. Terry Taking samples of fifteen?
14. Linda Yeah.
15. Nicole Repeated samples …
16. Betty Yeah.
17. Nicole of fifteen.
18. Terry So in repeated samples of fifteen officers where, any officer could have been chosen out of the eighty-three hundred I guess, that's just what happened. So when you said "the luck of—" I guess I was trying to get you to think, well you say, "luck of the draw" … specifically what are you saying? That … of all the groups of fifteen officers I could have picked that just happened to be the=
19. Linda =Happened to be white, yeah.
20. Terry So again, if you just, randomly generated sets of fifteen officers … assuming all officers all officers have the same chance of being chosen. Okay? Now is there a claim implied in this scenario?
21. Henry The claim would be … that you wouldn't get ah, zero black officers, out of the luck of the draw very often, in fact, *so* non-very often that this was, that the claim is that, that it was prejudice.
22. Terry Okay.
23. Henry That they were deployed it that way.

24. Terry Lucy, were you …?
25. Lucy That was pretty close to what I was going to say.
26. Terry Okay. So there's sort of an implied claim that, that=
27. Lucy =that there was an underlying prejudice.
28. Terry Right, that there was an underlying prejudice, that perhaps it wasn't, it didn't seem, it was unusual that we had no black officers in a group of fifteen. Okay? Um …

The discussion around these questions suggested that the teachers initially did not conceive of the deployment stochastically. In line 3 Linda answered the first question—the meaning of the statement—by replacing "chance" with "the luck of the draw". Terry pushed the group to think stochastically by asking, "Can we state that [chance deployment] probabilistically?" Linda (line 6) argued that it was impossible to state it probabilistically. The probabilistic statement she did make—over a large number of deployments fourteen out of a hundred officers that responded, would be, or whatever ratio that is, would be black (line11)—revealed that although she did conceive of the deployment as a repeatable process, she had a predisposition to focus on the central tendency. She did not have in mind a distribution of sample statistics that resulted from the repeated deployment. Therefore, she said what she said for the sake of "making a probabilistic statement". She did not conceive of chance deployment probabilistically. Terry further pushed her by asking what she meant by "the luck of the draw". Again, Linda paraphrased the statements (by using the phrase "happen to") as opposed to operationalizing "chance deployment".

In line 21 Henry's interpretation of the claim in question—"You wouldn't get zero black officers so *often*" suggested that Henry might have conceived of a repeatable process of selecting 15 officers. So did Lucy, perhaps (line 25). However, neither one of them articulated this repeatable process.

Activity 2-3, Part II: Question 2

2) Do you think that "chance" is a reasonable explanation for all pursuing officers being white? Discuss how you might investigate this question.

With respect to the question—Do you think that "chance" is a reasonable explanation for all pursuing officers being white?—all teachers, with the exception of Linda, thought that chance was *not* a reasonable explanation and that there existed racial prejudice in the deployment. Linda said that she was not inclined to make a judgment without any investigation.

Next I will focus on the teachers' discussion on how they would investigate this question. The discussion is parsed into five episodes.

Table 48: Overview of discussion in Part II of Activity 2-3 Rodney King scenario

Episode	**Theme**
1	How to set up simulation
2	Assumptions about simulation
3	Decision rule
4	Simulation results
5	Interpretations of simulation results

Episode 1: How to set up simulation

Teachers proposed that they investigate the question by conducting simulations in graphing calculator TI-83. The excerpt below illustrated how they set up the simulation and interpreted the simulation results.

Excerpt 74

82.	Terry	Okay, so how might we investigate this question?
83.	John	How 'bout=
84.	Linda	=Simulate.
85.	John	(laughs) Yes, simulate.
86.	Terry	All right, we could simulate what happened and *see* what would happen if we repeated the process again assuming this certain

circumstances remained constant, how often would we see a deployment with no black officers. How many of the deployments would we see. So if we're going to do a simulation what do we need to do.

87. Nicole Well, let's assign some numbers …
88. Henry Set up a population that has the same ratio.
89. Terry Okay, same ratio, right.
90. Henry And then sample that population, with fifteen officers each time,
91. Terry Okay.
92. Henry and then look for … you'd have a number of samples that had zero Black officers.
93. Terry Okay, so what proportion.
94. John Yeah, that's what I did. I just did one to eighty-three hundred … and I randomly took fifteen at a time, put it in list one, and then said sort list one, and every time I've done it, I've gotten some—I just said that one to eleven sixty-two is the black officers,
95. Terry Okay,
96. John and then eleven sixty-three to eighty-six hundred are the white.
97. Terry Okay.
98. John So every time I've done it I've gotten Black officers.
99. Terry Okay.
100. John And I've taken fifteen at a time.
101. Terry All right.
102. Henry You would expect to see two Black officers, because fourteen percent of fifteen is two point one, so you're expecting … well, it is an average of two.

Henry proposed to simulate randomly selecting samples of 15 officers from a population of 8300 of which 14% are black and 86% are white (lines 88 to 90), and suggested in line 92 that they then look for the frequency of samples of all white officers. Taking into account of what he had said earlier (line 21), it is reasonable to claim that Henry had conceived of the deployment stochastically, and that by simulation he was aiming to find out the relative frequency of all white officers.

John in the meantime had already conducted the simulation and observed that none of his samples in several repetitions consisted of all white officers (lines 94 to 100). Since he did not count how many black officers he got each time, it is reasonable to conjecture that he was concerned with the question *how often do I get a sample with no*

black officers? In other words, he was looking for the relative frequency of samples like the one in Rodney king scenario.

Henry's response to John's observation (line 102) had a different focus—*how many black officers would you expect to see in a typical sample?* It seems like that he was providing a rationale for John's observation (of the lack of non-black samples)—since you should expect to see, on average, two black officers in a sample, it means that you wouldn't get non-black samples very often. This reasoning revealed the conception behind the "proportionality heuristic": i.e. the composition of a sample drawn from a population should resemble that of the population (regardless of the sample size).

Episode 2: Assumptions about simulation

As the teachers started the simulation process in their calculators, Terry asked the teachers about the assumptions that they made. It was made clear that one of the assumptions was that "every officer has an equal chance of being deployed." (lines 144 and 145)

Excerpt 75

136.	Terry	=so what assumptions are we making …
137.	Nicole	That the proportion of …
138.	John	It is of …
139.	Nicole	black officers on duty at any given time, is=
140.	John	=yes.
141.	Terry	Fourteen percent?
142.	John	Yes. Yeah we have to make that. That's right.
143.	Terry	We have to make that assumption, and, what other assumption do we have to make about …
144.	Henry	That they all have an equal chance.
145.	Terry	Every officer has an equal chance of being deployed, at any given time … which may or may not be true if …
146.	Nicole	Right.

Episode 3: Decision rule

The teachers then moved on to the discussions about the decision rule—*How might one decide whether the deployment in the Rodney King scenario was a chance deployment using the simulation results?*

Before we look at the decision rules proposed by the teachers in this episode, let me clarify the decision rule for this situation from a conventional point of view. In this situation, there are two competing claims: (1) There is no bias in police dispatch, i.e., getting a sample of all white is random, and (2) There is a bias in police dispatch. The population consists of 8300 officers, 86% of which are white and14% are black. We randomly select 15 officers at a time and repeated this process a large number of times. Over this larger number of times, *x*% of all samples have all white officers. The decision rule is: If x<5, it means a very rare event occurred, in which case we will conclude that there was a bias in the dispatch (i.e., claim (2)), and reject the claim (1).

Excerpt 76 revealed three teachers, Nicole, Terry, and Henry's decision rules.

Excerpt 76

187.	Terry	Um … I guess one of the things, and I guess you said, but I just in terms of talking about, how would you, I guess what would you be looking for to decide, what would help you decide whether chance was a reasonable explanation? Once you did the simulation.
188.	Henry	Frequency.
189.	Linda	Variance …
190.	Henry	You've also got to set, what did we call it last week the baseline, or … what was the word we used?
191.	Nicole	Margin of error?
192.	Henry	Nah, that um, you gotta set (inaudible) a line for range, you gotta set. What did we call that last week? You guys set a line?
193.	Linda	Confidence level?
194.	Terry	Oh, the decision rule? Okay.
195.	Henry	A decision … and compare what you get with a certain, you've already run it several times, you haven't got it, theoretically you're only going to get it five point nine percent of the time, and that's *right*

	there in the very margin of error, whether it is chance or it isn't chance, so you gotta make … is that what you're talking about?
196.Terry	Mm-hm, specifically what would you be looking about, what would you be looking at? Let's say you've done a hundred simulations of fifteen officers?
197.Nicole	Well, ninety-five out of a hundred, let' say, – would that be a reasonable break point?
198.Terry	Okay.
199.Nicole	and simulations … ninety-five out of a hundred simulations have *only* white officers in them, then … it is not just *chance* deployment that caused this group to be only white.
200.Terry	If ninety-five of the hundred do have …
201.Nicole	… simulations have just white officers …
202.Terry	Then it *would* be just chance.
203.Henry	If only five, five of them had *just* white officers then it would …
204.Terry	Oh five of them just had …
205.Nicole	I need help if I screwed this up …

This excerpt shows that Henry understood Terry's question as asking for a significance level, or cut off level (although he had trouble recalling the terminology). Nicole proposed 95% to be the "breaking point" (line 197). However, the "decision rule" she described in line 199 showed that she was not thinking about the cut off level. Rather, she was talking about her interpretation of a hypothetical data that would be received about the actual deployment. In other words, she was addressing the question: What would happen if 95% of the deployments consisted of all white officers? Her answer was it would mean that there was a bias in the dispatch, and her rationale was likely to be: Given that the population composition was 86% white and14% black, getting 95% of all samples having all white would suggest that there was a bias in the dispatch. However, throughout the discussion Nicole did not provide a rationale for her argument. It was not made evident to the group why she thought the way she did. Conceptually, although what she said was a viable conjecture, it was not a decision rule. She talked about *one* hypothetical scenario (95% of the samples are all white) that would potentially confirm a

hypothesis (that there was a bias in the dispatch). She was not thinking of developing a distribution of sample statistics and comparing the relative frequency of sample of all white against a significant level, which would then lead to the decision of whether the deployment was a result of chance or bias.

Terry's response to Nicole (line 200-202) suggested that she understood Nicole to mean "If 95% of samples are all white, then getting one sample of all white must be a result of bias." This contradicted with her conception of bias: "If a rare event occurred, it must be a result of bias." Therefore, she modified (or negated) Nicole's argument into: If 95% of all samples are all white, then getting one sample of all white could be just by chance. The fundamental difference between Terry and Nicole was that Nicole was thinking about the actual deployments, and Terry was thinking about the simulation.

Henry attempted to "correct" Terry and Nicole's arguments. Combining his utterances in line 203 and 213, we can reconstruct his image of the situation and his main argument. For him, the purpose of the activity was to test the unusualness of the samples of all white. He imagined a distribution of sample statistics that had around 5% of samples of all white and 95% of samples having at least one black officer. The argument that he seemed to express, but did not articulate clearly, was the decision rule that if less than 5% of samples are all white (the rest of the samples have as least one black), that would suggest there was a bias in police dispatch in the Rodney King scenario. Table 49 summarizes the "decision rules" held by the teachers.

Table 49: Summary of Nicole, Terry, and Henry's decision rules

Nicole	If in multiple deployments of 15 officers, 95% of the deployments have all white officers, it means the dispatch (that produced these deployment) was biased. Because it was unlikely that "95% of the deployment have all white officers" since only 86% of the population are white.
Terry	If 95% of the simulated random deployments have all white officers, it means that it is NOT unlikely "to have one deployment that have all white officers". This means that the dispatch (in Rodney King scenario) is not biased.
Henry	If only 5% or less than 5% of samples are all white (the rest of the samples have as least one black), that would suggest there was a bias in police dispatch in the Rodney King scenario.

The following excerpts further revealed the similarities and differences among these three teachers' decision rules.

Excerpt 77

206. Henry	We're looking for black officers …
207. Nicole	What?
208. Henry	We're looking for, we're assuming there's going to be black officers … most of the time, at least one.
209. Terry	(talking to Nicole) If ninety-five percent of your samples had all white officers … ninety-five percent of your simulations where you're looking at chance deployment had all white officers, what would that suggest?
210. Nicole	I would argue then that it is not … chance simula—chance deployment that we wound up with only whites, in the car.
211. Henry	(pointing to Terry) Yeah, you just said backwards as well, you just said if ninety-five percent was all white,
212. Terry	I thought that's what she (pointing to Nicole), I thought that's what she said.
213. Henry	That is what you (pointing to Nicole) said and you (pointing to Terry) said it as well but what we're looking at is ninety-five percent of them would have at least one black officer, that's what we would expect, if and only five percent would be all white.
214. Terry	If I do a hundred, okay let me just get this straight, if I do a hundred simulations …
215. Henry	You would expect only five=
216. Terry	=I'm not talking about what I would expect. If I'm doing a hundred simulations, and ninety-five of those simulations they were all white …
217. Henry	Then there's not a problem (pointing to sheet).
218. Terry	Right …
219. Alice	Then it *is* chance.
220. Terry	It *could* be chance.

221. Henry Yes.
222. Terry (talking to Nicole) Does that make sense? If ninety-five out of my hundred simu=
223. Nicole =Oh, oh oh oh, okay.
224. Terry You see? I was a little confused.
225. Nicole Yeah, okay. I said it wrong.

The discussion revealed that the differences among these three teachers' thinking lie in a number of elements: 1) *actual deployment or simulations*? Nicole was thinking about actual deployments, while Terry and Henry were thinking about simulation. 2) *which dispatch are they talking about?* Nicole was talking about a hypothetical dispatch. It was neither the dispatch in Rodney King scenario, or a random dispatch (i.e. simulation). Terry was talking about the dispatch in Rodney King scenario. Henry was talking about simulation, i.e. a random dispatch.

The similarity among their thinking was that *none of them were concerned about "coming up with a decision rule"*. Nicole tried to answer the question "how do we know there is a bias in the dispatch?" She came up with a hypothetical scenario "95% of the deployments have all white" that would lead to the conclusion that the dispatch was biased. She did so without thinking how "the dispatch" in her hypothetical scenario related to the dispatch in the Rodney king scenario. She was not constraint by Henry's concern "the simulation was not likely to produce 95% of all white samples" because she was not thinking what was likely to happen in simulation. Nicole answered her question "how do we know there is a bias in the dispatch?" by proposing one scenario where an unlikely event happened. This unlikely event is "95% of all samples are all white." This is a very significant finding in that it is a case of *conceptualizing the situation stochastically, but not having an understanding of p-value*. Terry had the same implicit criterion for biasness, i.e. "a rare event happened." However the event she had in mind

was “one sample of all white officers”, i.e., the deployment in the Rodney King scenario. Her main objective during this part of the conversation was to correct Nicole’s statement (although without truly understanding what Nicole really meant). Henry tried to make sense of Nicole and Terry’s thinking from his own frame of mind. He was concerned about what was likely to happen in a simulation. He was more active in “correcting” Terry and Nicole than he was in developing a decision rule.

The end of this excerpt (lines 222 to 225) suggested that Terry had “convinced” Nicole that she was wrong. The beginning of the next excerpt was marked by a transition, in which Terry shifted from responding to Nicole’s statement to reiterating the purpose of activity, i.e., to test whether the deployment of all white in the Rodney King scenario was unusual.

Excerpt 78

226.	Terry	All right, so I get, one of the things this sort of ties into is kinda the thing with the Coke and Pepsi thing is that, basically for us to decide whether or not chance is reasonable, we’re looking to see if, you know, what, if this, if this sample of fifteen white officers is unusual, is this an unusual sample=
227.	Nicole	Right.
228.	Terry	and the only way we can judge if it is unusual is to do what?
229.	Henry	Sample.
230.	Terry	And … do what with the sample?
231.	Henry	Compare.
232.	Terry	Right, and compare that to what would happen in a whole bunch of other samples, so it is related to that idea of unusualness again.
233.	Nicole	Okay, you state what I should have said then.
234.	Terry	If ninety-five percent of your samples had all white officers, were all white, then that would be, that would suggest to me that chance is reasonable, because it happened, a lot.
235.	Nicole	Yeah. (Leans to the table and writes down notes)
236.	Terry	Okay? All right, let’s go ahead and look at a simulation of, or simulations, and we’ll … (turns on overhead) all right, now we did this on ProbSim, you could do it with Fathom …

In line 233 Nicole asked Terry to "state what I should have said". She was in fact asking for a decision rule, i.e., to answer her original question, "How do we know there was a bias in the dispatch?" However, Terry literally restated what she thought Nicole should have said (line 234)— If 95% of the samples had all white officers, then we would claim that the deployment was not unusual. Since the purpose of this discussion was to find a decision rule, and that this statement (line 234) marked the end of this discussion, it was likely that the group had taken it as the decision rule.

Episode 4: Simulation results

The group conducted the simulation and they found that in multiple simulations from 7% to 12% of the samples had all white officers.

Episode 5: Interpretations of simulation results

After the simulation, the teachers appeared to be at a loss as to what to make of the results. None of the teachers tried to compare the *p-value* (7-12%) against a significance level (e.g., the conventional cut off level for unusualness, 5%), and conclude that the investigation did not provide enough evidence to claim that there was a bias in the dispatch. The following four segments showed how the teachers did interpret the simulation results.

Episode 5, Segment 1: John

Based on the simulation results, John concluded that the deployment was not a chance event (transcript 297-348). He had a hard time articulating his rationale. Here is a summary of his reasoning: John had an image of a normal distribution of sample statistics in mind. The center of the distribution is the population mean, and 95% of the sample statistics fall within an interval centered on the mean. Imagine the two extreme ends

(each containing 2.5% of the sample statistics) are shaded. John's argument was since 7~12% fell in the un-shaded region it meant that it was not unusual, which led to the conclusion that the deployment was a chance occurrence. Essentially, John implicitly set a cut off level of 2.5% and compared the *p-value* against it to reach the conclusion.

Episode 5, Segment 2: Sarah

Sarah's comments in the following excerpt confirmed my earlier conjecture (Episode 3), that she (perhaps among others) had taken Terry's statement (line 234) into a decision rule: If the sample of all white officers occurred ninety-five percent of the time or more, then we would think it was by chance that all white officers showed up.

Excerpt 79

349.	Sarah	Let me see if I hear what you're saying, a while ago we said that, and this is when Henry interpreted what the two of you were saying about, if it occurred ninety-five percent of the time then we would think it *was* by chance that all white officers showed up, okay, now we look at this and see that over a large, very large sampling it is occurring ten percent of the time, do we now attribute it to chance?
350.	Terry	Around ten percent, yeah, around ten percent of the time we're getting.
351.	Sarah	And that's outside of your ninety-five percent, I mean you're not going to, you're not going to … that ninety-five percent, that arbitrary ninety-five percent of the time that you said that if white officers showed up that much that you have treated it not to be as chance, I'm not real comfortable, I'm not real comfortable, with, with that big a number, I mean, can it not be attributed to chance, like you said the very first day something about if something happened to you fifty percent of the time you considered it usual … well, this isn't even fifty=
352.	Henry	Well, in the field of psychology, they go with eighty percent.
353.	Sarah	Right. But I'm not, I'm not sure that I'm comfortable with something having to happen, something like this having to happen ninety-five percent of the time and then you're generating data that shows that basically *none* of the (inaudible) is occurring five percent of the time.
354.	Terry	Okay, so you're saying it looks to you like … based on our simulation seeing that we have between seven to twelve percent of our samples …
355.	Sarah	I guess my *feel*=
356.	Terry	=Uh-huh.=

357.Sarah =is that based on what I've seen there that this fifteen could have been all white by chance, but it not have to be that ninety-five percent, didn't have to be that I, you know …

According to this rule, anything that happen less than 95% of the time would be regarded as unusual. In lines 351 and 353 Sarah questioned this cut-off level as being too high. Yet, she did not propose an alternative cut-off level. She believed that occurrences that happen 7~12% of the time were not unusual, and this belief was based on a subjective "feel" (lines 355 and 357).

Episode 5, Segment 3: Nicole

Excerpt 80

375.Nicole Can we say just this that if ten percent of the time we get an all-white group then? Can't we leave it at that, that that means that it, that *it is possible that it is a chance*?

In this excerpt, Nicole asked two questions. In the first question, Nicole suggested the possibility of making a compromise, i.e., do not make a decision, instead, simply state the *p-value*: the percent of time we would get a deployment like this if we do it over and over again. This suggestion was perhaps due to a lack of clarity in the cut-off level. However, the second question suggested that Nicole did apply an implicit/subjective cut-off level, which led her to conclude that the *p-value* being 10% meant the deployment could have been a chance occurrence.

Episode 5, Segment 4: Betty

In this segment, Betty proposed that the group revisit the concept of chance deployment. In doing so, she pushed the group to reflect on their understanding of chance & randomness, and in turn, achieve an agreement on the decision rule.

Excerpt 81

378. Betty Well, can we revisit what we said was chance deployment? I'm getting confused on that.

379. Terry What did you say? What did you think?

380. Betty I didn't, I don't, I was … don't know.

381. Terry Who gave us our definition of chance deployment? Linda, was that you?

382. Linda Mm-hm.

383. Betty I didn't get all that down there, to revisit on my paper's the reason.

384. Nicole I wrote down if we repeatedly, assuming the random samples of fifteen officers, assuming the samples of fifteen officers was random, if we repeatedly did this, what we were looking at was … um, geez, I didn't write down a whole sentence. (laughter)

385. Terry Let's think about what would make you think that it was *not* chance deployment, based on our simulations that we just did? What would make you think that being chance deployment would be an unreasonable … what would make you suspect that it is *not* chance? Can you think about it that way? Given all our samples, we did two hundred samples of size fifteen, could there have been something about those samples that could have made you think, "well maybe this isn't chance?" What would have happened?

386. Betty Well I guess I get hung up on chance, is that, are we saying that, that it doesn't happen often, or it is rare that it would occur?

387. Henry Chance is reasonable, chance is, what percent can you reasonably expect … in our simulation we're saying roughly ten percent of the time you're going to get an all-white sample.

388. Terry I wouldn't even say chance is necessarily rea—in *this* context it is reasonable, but chance is what would happen, what percent of the time would you see, fifteen perc—fifteen all-white officers if the only way they were deployed was completely randomly.

389. Lucy If there were no other forces.

390. Terry If there were no other forces acting on it … how often would it occur, or how many times would you have repeated this would it happen that we would get all white officers in your fifteen officers that were deployed. All right, *that's*, that itself would be the measure of the probability of the chance … now, whether the chance is reasonable is not is whether you, you believe that that really … suppose that we – I think that I can answer your question by answering a different question, let me try this – suppose we run our simulations and we had seen two percent of the samples were all white, three percent of the samples were all white, one percent of the samples was all white, two percent of the samples was all white, four percent of the samples was all white, all right, that's what happened by doing this randomly … all right, let's pretend that's what happened … all right, given that, looking at, we see that, by randomly generating these samples we get things like, one percent, two percent, three percent of the time that's what happened, we got a set of all-white officers, would you think that

that's a reasonable—would you think that it would be reasonable, in this situation … that you had all white officers, given that it appears that that event only occurs one to two to three percent of the time?

391. Betty No, that doesn't seem reasonable.

Teachers' responses to the Betty's question "What is chance/chance deployment?" suggested that no one had an operational understanding of chance deployment, that is, if an event happens less than 5% of the time, we would call it *not* a chance occurrence. Betty (line 386) assumed the opposite meaning for "chance". She equated it with "unusualness". Henry (line 387) suggested that he took "chance" to mean probability—the likelihood of an occurrence. Lucy (line 389) said that chance meant "no other forces", i.e., free of biases in sampling process.

In line 390, Terry engaged the teachers in a thought experiment: Suppose that only 1-4% of samples had all white officers, would that lead to the conclusion that the deployment in the Rodney King scenario was biased? Betty's answer (line 391) indicated that she had an intuitive understanding that "if a rare event happens, it means that there is a bias". However, the idea of the cut-off level remained tacit. Question 3 of this activity, which I will turn to next, brought this idea to the foreground.

Activity 2-3, Part III: Question 3

Suppose you are a member of the California Supreme Court and that before you is the matter of what level of likelihood may be taken as "evidence of racial bias in police dispatching procedures." What level would you set for that and future cases?

Question 3 focuses the idea of the significance level: A significance level of *x*% (x is a small number) means that a sample that happens less than *x*% of the time is considered to be significantly rare, and that if it does happen, we will conclude that it is a result of bias in the sampling process. We make the conclusion with the knowledge that it might have

happened by chance (i.e. we might have made the wrong conclusion), but if we apply the same rule over and over again, in the long run we will not made the mistake more than *x*% of the time.

Excerpt 82

580.Terry	Okay … All right? … All right, suppose you are a member of the California Supreme Court, and this is being … that the matter is, what level of likelihood may be taken as evidence of bias in police dispatch procedure, what level would you set, for that, in future cases? Which I think is what your question asks (Alice), did most of you feel, somewhere Sarah you said that ten percent, somebody else, John you said that, something that occurs ten percent of the time by chance, you didn't feel like that was … enough evidence of racial bias or that there was something other than chance … what level would you chose? Do you see the point of that question?
581.Sarah	I do, and I don't think that I have any problem at all with ten percent being a nice clean cut-off. I still have trouble getting beyond a thousand other factors, and I think the thing John brought up, the fact that there in the curve somebody would show you an assignment of officers to specific areas, I, I think that those would be all things that if I were the person who was going to bring in the suit, I would know all of those things, I would've investigated those kinds of things instead of just a flat number of all the employed police officers in L.A. county.
582.Terry	Okay.
583.Sarah	Do you understand what I'm saying?
584.Terry	Mm-hm, and that's something that the students got kinda—but we're not, you know, but we're not talking about necessarily specifically this event, if we're thinking about, we're thinking about all events that would be like this.
585.Sarah	Right, but if you're asking me about as a Supreme Court member, relative to this Rodney King case, then what would be considered evidence and bias, racial bias and all that kind of stuff, then I think all that other does have to come in place=

Terry started by asking: What was the cut off level? Sarah's responses in lines 581 and 585 suggested that although she accepted a 10% cut off level, when it came to decision making she held a detective's (non-stochastic) perspective, i.e. she took the position of a detective who would investigate what might have caused the deployment to be all white officers. The question asked for a decision rule—what level of likelihood we would set as

significant so as to set a standard for convicting such cases. But for Sarah, the task was to decide whether or not to convict the police department in the Rodney King scenario. In the following excerpt, Pat tried to distinguish between these two different perspectives with the aim of pushing the teachers towards thinking about policy making.

Excerpt 83

587.Pat	Um … let me try putting it another way, and, and see if it changes the way you're thinking about, what they mean by the question. (pause) Suppose that … the Supreme Court Justices say, "what we're really deciding is what percent of innocent, of people falsely charged are we going to … convict?" (pause) Is that the question that they're addressing?
588.Alice	Say it again?
589.Pat	What percent of defendants charged falsely … are we going to allow to be convicted?
590.Nicole	Okay.
591.Alice	Wouldn't you want that to be zero? (laughter)
592.Nicole	No … can't be.
593.Terry	No it is probably …
594.Pat	Do you see—but is that, do you see that as the question that they are addressing?
595.Sarah	Do *I* see that as the question they are addressing?
596.Pat	Yes.
597.Sarah	No. (pause) Should I see that as the question—do *you* see it as=
598.Pat	= Well, I'm just wondering if=
599.Sarah	=do *you* see it as the question?
600.Pat	if you see the two as being related.
601.Terry	Linda's nodding her head earlier. You want to say what you're thinking, Linda?
602.Linda	Um … well what he's saying is that, would you except a level of—what percent probability that you got from there would you be willing to accept and still convict the police department of being wrong, of course, I would say something very close to zero, like one percent two percent.
603.Pat	Now, suppose that … eh, um, continue that line … suppose that what we're talking about, suppose that they have a thousand cases, like, the Rodney King … scenario … in which, um … yeah, suppose that we have a thousand cases and in fact all of them were falsely accused. If we said, "ten percent is good enough to convict them", the ten percent probabilities=
604.Nicole	That means you got a hundred of them out there in jail, I mean … right? A hundred times …
605.Pat	Mm-hm.

606.Nicole they've been in error.
607.Pat Yeah. So you got a hundred of them in jail, falsely.
(pause)
608.Nicole I think we have to set our standard much lower than twelve percent here, or whatever …
609.Terry Mm-hm.
610.Nicole I mean …
611.Sarah Now, are we talking about convicting somebody or the probability of all white officers showing up?

In this excerpt, Pat rephrased the original question (Question 1) into Question 2 and asked the teachers if they saw them as related.

> Question 1:
> Suppose you are a member of the California Supreme Court and that before you is the matter of what level of likelihood may be taken as "evidence of racial bias in police dispatching procedures." What level would you set for that and future cases?
>
> Question 2:
> What percent of defendants charged falsely are we going to allow to be convicted?

Underlying Question 2 was the idea of Type I error: the error of rejecting a true hypothesis. By asking the question, Pat hoped to see if the teachers saw the connection between significance level and Type I error. Sarah's response (line 597) suggested that she did not make such connection. Her question (line 611) further confirmed that she was holding a detective's perspective, concerning about the decision making in the Rodney King scenario. The exchange between Pat and Nicole in lines 603-608 indicated that Nicole had understood that the issue was about deciding a level of rejecting a true hypothesis. She concluded that the level should be much lower than 12%.

Excerpt 84

635.Terry =What percentage of samples that were all white, how low would that have to be for you to be persuaded that you're going to say that there was some racial bias there?
636.Sarah Now then *that's* a different question then the one you asked earlier, because you asked=

637.John Two point five percent.

638.Terry All right, John, very firmly says two point five percent, you (Linda) said one percent?

639.Linda I said "something close to one or two", yeah.

640.Terry Okay.

641.Henry (inaudible) you can make that decision based on these facts. A decision as large as that would have a lasting (inaudible) that that, you're talking about setting up a precedent, legal precedent, you should make a legal precedent based on one singular fact like that, alone, but the accepted one is five percent, if you're just talking about generically looking at things five percent is pretty darn fair, this is twice that, it is ten percent.

642.Terry Okay, so you're saying you would—five percent would be low enough for you?

643.Henry I'm not saying I would dare do that based on this, I wouldn't make any ruling period based on this fact even if it is flat zero.

644. …

645.Nicole I want to change the context from this situation to another number that's similar … most of the people in this room are women, and if breast cancer strikes people randomly, which we all know it doesn't, one out of eight women will get breast cancer. That's twelve point five percent of the population, and twelve point five percent of the population is enough to motivate fifty-two percent of the entire American population or whatever to, to think that, the women in the population to think that breast cancer is truly frightening—the NSF is putting a whole truckload of money and the National Cancer Society, etcetera, etcetera, etcetera – American Cancer Society (corrects herself) – so twelve percent is an alarming number, so therefore, we *can't* convict this, um – my argument is – we can't convict the LAPD for being racially biased, twelve percent is, is, well within the … I mean it can really happen.

646.Linda What if it were five percent?

647.Nicole I think that we gotta make it lower if you were doing this in the Supreme Court.

648.Linda It should be very close to zero.

649.Nicole Yeah I really wanna go lower.

650.Linda If not zero.

651.Sarah I just don't think as a Supreme Court decision you can arbitrarily say a percentage.

There are three observations that I would like to highlight. First, this excerpt showed that John, Linda, and Nicole were comfortable with a low level of significance. John fixed on 2.5%, Linda 1-2%, while Nicole wanted something closer to 0%. Second, Henry pointed

out that the conventional significance level was 5%. However, he also stated (in line 643) that he would not make any decision based on the simulation results even if the *p-value* was zero. This suggested that he had a similar conviction as Linda's reasoning emerged in earlier discussion, that is that, he would not reject the null hypothesis (that the dispatch was random) on the basis of the unusualness of one occurrence. Finally, Sarah's comment (line 667) was consistent with her detective's perspective. She did not connect the significance level with Type I error, and believed that a significance level was arbitrary.

Summary of Activity 2-3

What have we learnt about teachers' understanding of hypothesis testing from the discussion around Activity 2-3: Rodney King scenario? The most conspicuous result from this discussion was the difficulty the teachers experienced with the idea of decision rule. This difficulty expressed itself in at least three ways: 1) in understanding the question and how a decision rule relates to it; 2) in designing a decision rule before the simulation; and, 3) applying a decision rule to the simulation results.

Discussions in Part I and in Part II (episode 5 segment 4) suggested that the teachers did not have an operational understanding of "chance deployment", and therefore did not know how to answer the question of whether the police deployment in the Rodney King scenario was a chance deployment. Episode 1, 2, 5 of Part II showed that although the teachers conceived of the policeman dispatch stochastically and in turn designed the right simulation, they did not know how to use simulation to answer the question. Episode 3 focused on the difficulty the teachers experienced with coming up with a decision rule. The three active participants in this episode, Nicole, Terry, and

Henry, each held different images of the situation and of decision rules, yet no one clarified a decision rule that they could apply to the simulation results. Episode 4 summarized the teachers' attempts at interpreting the simulation results (of selecting 15 officers at random from the entire 8300 member force 84% of whom was white). Because of the lack of a decision rule, the teachers could not reach a conclusion about how the simulation results illuminated the question of whether an all-white group of 15 officers constituted evidence of bias, and the topic of discussion returned back to the meaning of "chance deployment".

Discussion in Part III showed that a number of teachers including Nicole, John, Linda, Sarah, and Henry, had understood the concept of decision rule. The discussion also revealed that Sarah and Henry, despite their understanding of decision rule, held a non-stochastic (detective's) perspective, and would not make a decision about the deployment in the Rodney King scenario on the basis of statistical investigation.

Overall, the discussion of the Rodney King scenario illustrated the complexity of understanding hypothesis testing. Knowing how to answer the question—Is the deployment in the Rodney King scenario a "chance deployment" – entails a scheme of thinking that includes a stochastic conceptualization of the situation, a stochastic conception of *p-value*, and understanding the interrelated concepts of randomness, decision rule, significance level, and Type I error. The discussion of Rodney King scenario has painted a picture of different teachers possessing different fragments of such a scheme, which helps to explain the difficulty they had with answering the question. At the same time, teachers' difficulty with understanding the question and how to answer the question led me to conjecture that this difficulty was a result of them not having such a

scheme of thinking. In other words, possessing such a scheme of thinking (as a tool) seemed to be a pre-requisite for one to understand the question (i.e., quickly recognize the task for which the tool could be utilized).

Chapter Summary

This chapter explored teachers' understanding of hypothesis testing. Section 7.1 focused on teachers' conception of unusualness or *p-value*. Discussion around Activity 1-6 Movie theatre scenario revealed that, of the six teachers whose conceptions of unusualness were identified, all but one teacher Alice had non-stochastic conceptions of unusualness. Only two teachers, Alice and Henry, conceived of unusualness as a statistical concept. We also found that Sarah's incoherence in her understanding of unusualness was a result of confounding *a sample percent (relative proportion of some item in a sample s)* with the *relative frequency of samples like s* over a large number of times. Results on Interview 2-3 Horness Scale showed that six out of eight teachers conceived of the situation stochastically, and compared a sample statistic against a distribution of sample statistics. Only two teachers, Sarah and Betty, did not have a distribution of sample statistics in mind when answering the question.

Section 7.2 explored the teachers' understanding of the logic of hypothesis testing. Discussion around Activity 1-3 Pepsi scenario revealed that the teachers did not conceive of the situation as entailing hypothesis testing. Out of 13 instances, in 9 the teachers exhibited a commitment to the null hypothesis, whereas in standard hypothesis testing, one's commitment is to the alternative hypothesis. That suggested that the teachers did not understand that a null hypothesis was typically set up to be tested for one

to conservatively confirm a competing hypothesis. In three cases, the teachers, John, Lucy, and Henry, tacitly accepted the truth of null hypothesis and rejected the alternative hypothesis on the basis of a small *p-value*. In one case, Henry suggested that the sample was not random, and thus no decision about null hypothesis could be made based on the *p-value*. In the other five cases, most prominently represented by Sarah and Linda, exhibited a clear commitment to null hypothesis. Linda did not implicitly accept that the null hypothesis was true, but she believed that in order to reject the null hypothesis, she would need overwhelming evidence against it. A small *p-value* calculated on the basis of one sample, to her, did not constitute overwhelming evidence.

Teachers' reflection on their initial discussion around Activity 1-3 revealed that although Nicole and John had given up their commitment to the null hypothesis, they still did not understand that 1) a null hypothesis is defined through an alternative hypothesis, and that 2) hypothesis testing starts with a commitment to the alternative hypothesis. This discussion further revealed Linda's understanding of hypothesis testing. Linda initially believed that the activity was about confirming the null hypothesis. When she understood that the discussion was about whether or not to reject the null hypothesis, she vehemently opposed the idea of "rejecting an assumption based on one sample". Her argument was that no matter how rare the sample was, it could still occur, and thus it could not be used to reject an assumption. A mixture of beliefs and orientations helped to explain why she opposed to rejecting the null hypothesis, and these include:

1) A commitment to the null hypothesis. She would reject a null hypothesis only if there were overwhelming evidences against it. Therefore, she opposed to "rejecting the null on

the basis of one sample" and proposed to take more samples to see if the null hypothesis was *right* or wrong.

2) A concern for the truth of null hypothesis. Rejecting a null hypothesis, to her, means making a conviction that the null hypothesis was wrong. Because of this belief, she opposed to "reject the null hypothesis on the basis of one sample" because any rare sample could still occur theoretically.

Discussion around Activity 2-3: Rodney King scenario revealed the difficulties the teachers experienced with the idea of decision rule. This difficulty expressed itself in at least three ways: 1) in understanding the question and how a decision rule relates to it; 2) in designing a decision rule before the simulation; and, 3) applying a decision rule to the simulation results.

Discussions in Part I and in Part II (episode 5 segment 4) suggested that the teachers did not have an operational understanding of "chance deployment", and therefore did not know how to answer the question using hypothesis testing. Episode 1, 2, 5 of Part II showed that although the teachers conceived of the Rodney King scenario stochastically and in turn designed the right simulation, they did not know how to use simulation to answer the question: Is the deployment in the Rodney King scenario a "chance deployment"? Episode 3 focused on the difficulty the teachers experienced with coming up with a decision rule. The three active participants in this episode, Nicole, Terry, and Henry, each held different images of the situation and of "decision rules". No one clarified a decision rule that they could apply to the simulation results. Episode 4 summarized the teachers' attempts at interpreting the simulation results of selecting 15 officers at random from the entire 8300 member force 84% of whom was white. Because

of the lack of a decision rule, the teachers could not reach a conclusion about whether the simulation results settled the question of whether an all-white group of 15 officers constituted evidence of bias, and the topic of discussion returned back to the meaning of "chance deployment".

Discussion in Part III showed that a number of teachers including Nicole, John, Linda, Sarah, and Henry, had understood the concept of decision rule. The discussion also revealed that Sarah and Henry, despite their understanding of decision rule, held a non-stochastic interpretation (detective's perspective), and would not make a decision about the Rodney King scenario on the basis of statistical investigation.

CHAPTER VIII

TEACHERS' UNDERSTANDINGS OF VARIABILITY AND MARGIN OF ERROR

This chapter describes four sets of activities and interview questions related to the concepts of variability among sample statistics and margin of error (see Table 50).

Table 50: Overview of the activities and interviews in Chapter 8

Chapter 8 Teachers' understanding of variability and margin of error			
Section	**Activity (A) and Interview (I)**	**Day**	**Duration**
8.1 variability	I1-1 General questions		
	I1-2 Variability of investment		
	I1-4 Accuracy of measurements		
	I1-7 Law of large numbers		
8.2 variability and sample size	A1-7 Fathom investigation	3	104 m.
8.3 variability and population parameter	A1-8 Stan's interpretation Part I	4	53 m.
	I2-4 Purpose of simulation		
8.4 margin of error	A1-8 Stan's interpretation Part II	4	67 m.
	I2-2 Harris poll		

The chapter consists of four sections. The first section focuses on the Pre-Interview in which we probed teachers' understanding of variability. Teachers' responses to four interview questions are summarized and analyzed. The second section describes an activity in which the teachers interacted with the computer program Fathom. The intent of the activity was for the teachers to explore the relationship between variability and sample size. The third and fourth sections center on Activity 1-8, in which a student named Stan offered an interpretation of margin of error that exhibited several subtle inconsistencies. In the first part of Activity 1-8, which I summarize in section 8.3, the theme was the relationship between variability and population parameter. Sections 8.2

and 8.3 together push to the foreground the idea that "variability of distribution of sample statistics has to do sample size, but is largely independent of the population parameter"[15]. This idea is the logical precursor of margin of error: Since we know that variability is independent of the population parameter but dependent on the sample size, it means that in polling situation we can use variability (i.e., margin of error) to reflect how accurate a sampling method (i.e., of a particular sample size) produces estimates of populations. The last section, which is the second part of Activity 1-8, presents a polling situation and queried teachers' understanding of margin of error.

In each section, I begin by elaborating the rationale for the design and implementation of the activity. This is followed by a chronological recap of the discussion unfolded around the activity, highlighting the interactions and teachers' thinking that emerged within them. Then, I highlight teachers' responses to the interview questions that provide additional insights into their understandings of variability and margin of error. Finally a summary of major findings is provided in the end of each section.

Variability

This section focuses on the pre-interviews designed to reveal teachers' understandings of sampling as a stochastic process and of sampling variability. Teachers were asked to read an excerpt from Chapter 4 of Moore's *Basic Practice of Statistics* (1995). In it Moore develops the ideas of parameter estimation, sampling distributions, and the central limit theorem. Table 51 lists summaries of what the teachers thought the chapter was about and

[15] We know that this is not generally true, but in the context of the activity it was essentially true.

what were the important ideas in it. Only John and Henry saw that the excerpt was clearly about sampling distributions, although Henry gave greater importance to the central limit theorem. The other teachers saw less organization than John, focusing more on smaller ideas as if they were a list of topics.

Table 51: Teachers' responses to I1-1: what the chapter was about and important ideas in it.

John	Sampling distributions. Everything else hangs off of it.
Nicole	Law of large numbers, central limit theorem, mean remains the same but standard deviation changes as you take larger samples
Sarah	Statistics vs. parameters; mean and standard deviation; effect of sample size on a sample's distribution
Lucy	Statistic vs. parameter; central limit theorem, law of large numbers
Betty	Population vs. sample; distributing the data shows how the deviation can affect the mean and standard deviation; law of large numbers; central limit theorem
Linda	Population distribution vs. sampling distribution; overall picture of sample and mean; what a mean *is*; problems can be solved with formulas
Henry	Distributions; mean and standard deviation; central limit theorem
Alice	Random sampling; parameter vs. statistic; central limit theorem

Interview 1-2: Variability of investment

Question 2 asked what was varying with regard to the statement, "Buying several securities rather than just one reduces the variability of the return on investments." Moore intended the statement to mean that, for a given period of time, the distribution of average returns on collections of, say, 10 stocks, will cluster more tightly than will the distribution of returns on the population of individual stocks from which they are drawn. However, all teachers initially interpreted the statement as saying that the rate of return on any collection of ten stocks will vary less over time than will the return on any of the individual stocks in it, and they understood "vary" to mean "differ from purchase

price."[16] Only John, after some probing, reconsidered his answer to say that the variability occurred from "investment to investment".

Interview 1-4: Accuracy of measurements

Question 4a repeated a sentence fragment from Moore's text, "The fact that averages of several observations are less variable … " and asked teachers to interpret it. Table 52 shows that only John interpreted the statement distributionally, saying that the averages will cluster more tightly around the population mean than will individual measurements. Linda said that the averages of the samples, speaking of more than one average, would be closer to the true mean than the individual measurements. The remaining teachers all said that when you average the measurements you would get a result that is closer to the "true mean" than the individual measurements that make up the average.

Table 52: Teachers interpretations of Q4a, "average will be less variable."

John	Means of samples (collections of measurements) will cluster more tightly around the population mean than will individual measurements
Nicole	The average will be closer to the mean
Sarah	If you average your data it will be closer to the true average of total population
Lucy	Difference between population mean and sample mean will be less than the difference between individual measurements and the population mean
Betty	Compute running averages as you select a sample and the running averages will be closer to the true mean
Linda	The averages of samples will be closer to the true mean than will individual measures.
Henry	Larger the sample the closer will be the average to the true mean.
Alice	Difference between true mean and calculated average will be less than between true mean and individual measurements.

[16] We realize that another way of examining variability is by computing the variance of a stock's value from its running average rate of return (which is the exponent of an exponential function), but Moore's point still remains that the comparison is between a distribution of average rates of return for collections and a distribution of average rates of return for individual stocks.

Question 4b stated:

The author also says,

> It is common practice to repeat a careful measurement several times and report the average of the results.

Does this mean that if I take one measurement of an object's weight, and you take 4 measurements of its weight and then calculate their average, then your average will be more accurate than my measurement? (Explain.)

Table 53 shows that several teachers were more sensitive to issues of variability in answering Question 4b than in answering 4a, although none of them referred to a distribution of averages. John said that this statement applies only to the long run—that in the long run the average would be closer. Nicole and Sarah said that it should be true theoretically, but the thing you were measuring might change during the measurement process. Lucy, Linda, Henry, and Alice said that it could or should be, but it might not. Only Betty said that the average would definitely be closer to the true measurement.

Table 53: Teachers' responses to Q4b, "accuracy of 1 measurement versus average of 4 measurements."

John	Statement by itself tells us nothing. If we assume this is repeated, then in the long run I will get a good estimate of the actual mean, and you won't.
Nicole	Theoretically, the average of my four measurements should be closer than your one. But also need to measure many times because the thing you are measuring (e.g., air quality) can change over short periods of time.
Sarah	Probably not. Many variables undefined – measuring instrument, time of day, age of person. (Fix them?) Then theoretically, yes, but actually might not.
Lucy	Depends. I pick 4 you pick one. Your one could be closer than any of my four.
Betty	Yes.
Linda	Not necessarily. But you minimize the chance of being wrong by measuring it more times. Less chance of being close when measuring only once (but cannot articulate "less chance").
Henry	Could be. Also, measuring four times gives greater chance to detect measurement error.
Alice	Probably should be, but I don't know whether it would be.

Interview 1-7: Law of larger number

Here is the author's statement of the **Law of Large Numbers**:

"Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the mean $\overline{X}$ of the observed values gets closer and closer to μ."

a. Please explain what this statement says.

b. Assume we are sampling from the females in Nashville, TN and that we calculate a sample's mean height.

- Yan collected a random sample of 50 females and calculated their mean height.

- Luis collected a random sample of 100 females and calculated their mean height.

Whose mean height is closer to the population mean (i.e., the mean height of all girls in the population)?

c. Suppose Luis' sample contains 52 female, would you say the same thing?

In this question, Moore misstated the Law of Large Numbers, saying that $\overline{X}$ necessarily becomes closer to μ as the sample size increases. Table 54 shows that only John noticed this, saying that he disagreed with the statement, that it should say that if they repeated their sampling, Luis would "have the better estimate" (but was unclear about what that would mean). Nicole, Betty, Linda, and Alice interpreted the statement as written. Sarah and Henry initially interpreted it as written, and then qualified their interpretation to say "the likelihood is increased" that the sample mean is closer to μ with increased sample size, although Henry confounded number of samples with sample size. John and Lucy said that the statement said that means of larger samples "should" be closer to μ than means of smaller samples. None of Sarah, Henry, and Lucy thought that their interpretations were in conflict with Moore's statement. It is worth noting that, in this question none of the teachers interpreted the Law of Large Numbers distributionally, in

the sense that means of larger samples will cluster more tightly around the population mean than would means of smaller samples.

Table 54: Teachers' interpretations of Moore's Law of Large Numbers

John	If you go by Moore, then Luis. But I disagree—cannot stop there. Must resample. A sample of size 100 *should* be closer than a sample of size 10.
Nicole	Sounds like a limit.
Sarah	Take a larger sample size and you'll get closer to the mean. (Like a limit?) Like a reality. More times than not it should be closer.
Lucy	Larger sample more likely to be closer than smaller sample. (Likely?) Could be farther but probably closer.
Betty	Take the average of many samples and you'll be closer to the mean than an individual score.
Linda	The more observations the closer the sample mean is to the population mean.
Henry	The more observations and the more samples, the better is the representation of the population. To get the true average you would have to repeat sampling. The larger the sample increases the likelihood that you will be getting the true average.
Alice	As the number of observations increase, calculating a running average, the closer the average is to the population average.

Question 7b asked teachers to compare the accuracies of Yan's sample of size 50 and Luis' sample of size 100. By Moore's s Law of Large Numbers, Luis' sample would be necessarily closer. By the standard Law of Large Numbers, we could say only that Luis' sample is "more likely" to be closer, meaning that a larger proportion of all samples of size 100 would be within a given range of the population mean than all samples of size 50. Table 55 shows that only Nicole stated flatly that Luis' sample mean would be closer to the population mean than Yan's. Sarah, Betty, and Alice conditioned their response on Moore's wording. Each teacher responded consistently with their response to 7b when asked the follow-up question 7c, whether they would say the same thing if Luis' sample was of size 52.

Table 55: Teachers' responses to accuracies of samples of size 50 and 100.

John	Objected to just one sample. Said repeated sampling is necessary (but did not talk about distribution of sample means). "Larger sample is better estimate."
Nicole	Both samples are random? (Yes.) Luis is closer.
Sarah	Based on Moore's statement it should be closer. But most of the time the larger sample should be closer.
Lucy	Luis, most likely. Most of the time the larger sample will have a closer mean, but there can be variability.
Betty	According to this the larger should be closer. But the average of those two would be closer to the true height than either one of your averages.
Linda	Luis. (For sure?) Not for sure ... probably. Probably need more observations to be sure Luis' is closer, but I don't know how many women there are in Nashville to know how many observations you need.
Henry	They both could be just as accurate. You're looking for a breaking point (1/10 the population size) to be sure.
Alice	According to the LLN, the sample of 100 is closer. (Okay with this?) Yes. But the LLN says you should keep going.

Summary

The pre-interviews suggest that the teachers, with the exception of John, were predisposed to think in terms of individual samples and not in terms of collections of samples, and thus distribution of sample statistics was not a construct by which they could form arguments. "Likelihood" of a sample statistic being close to a population parameter was a property of individual samples and not of a distribution of sample statistics. Moreover, when asked to consider what was varying when comparing investments in collections of stocks versus individual stocks, they thought of a single collection of stocks in comparison to individual stocks in it. Only John came to see, after our probing questions, that it was a collection of collections that was less variable than individual stocks. Finally, only John and Linda referred to collections of averages when explaining what "the average will be less variable" meant, and while Linda referred to "averages" in the plural, it was not clear that she had a distribution in mind.

Variability and Sample Size

This section focuses on Activity 1-7: Fathom investigation. This activity was designed for the teachers to explore the relationship between sample size and variability of distribution of sample statistics.

Activity 1-7: Fathom investigation

In this activity, the teachers worked in pairs with Fathom installed in four laptop computers. We prepared a Fathom document that showed a collection of the murder weapons used in 669 murders committed in Chicago from January 1990 through July 1990 (NACJD: http://www.icpsr.umich.edu/NACJD/SDA/chd95d.html). In the first part of this activity, we asked the teachers 1) to first simulate taking a sample of 10 murders from this population and calculate the proportion of murders that were committed with a handgun, and then, 2) take 15 samples of 10 murders from the population and construct a histogram (by hand) that shows the frequency of proportions of handgun related murders in each sample. Our intent was for the teachers to go through the process of constructing a distribution of sample statistics.

In the second part of the activity, of which the purpose was to engage the teachers in exploring the relationship between variability and sample size, we asked the teachers to simulate taking 1) 100 samples of size 10; 2) 100 samples of size 25; 3) 100 samples of size 100, and to create a histogram of sample proportions of handgun-related murders for each collection. The seminar host, Terry, then juxtaposed three histograms and projected onto the front screen. Below are the Fathom windows that show the three histograms.

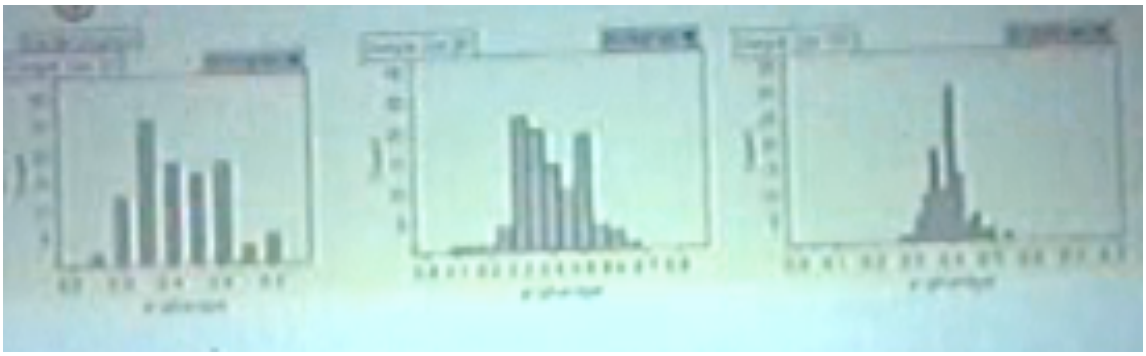


Figure 11: Distributions of proportions of handgun related murders in100 samples of size 10, 25, and 100 from population of 669 murders

We then asked the teachers to discuss the characteristics of these distributions, as well as the similarities and differences among them.

Activity 1-7, Episode 1: Variability and sample size

The teachers noted that the similarity among the distribution lie in that 1) they are relative normal, and 2) they all center around 0.4. Terry then asked the teachers how the distributions were different from each other. Excerpt 85 demonstrated that Sarah and Henry understood the relationship between variability and sample size, i.e., smaller sample size is associated with larger variability.

Excerpt 85: How were the distributions different?

12.	Terry	All right… and then how are those histograms different from each other—how are those distributions different from each other?
13.	Sarah	The lower the sample size, the greater the range.
14.	Terry	Okay. Smaller sample size had greater range than the larger sample sizes? What's another word we can use there?
15.	Henry	Variation.
16.	Terry	Variation, variability. There was less variability among the samples for the larger population, for the sample size 100 than there was for the sample size 10. Big idea. Very important idea that larger sample size means less variability in the distribution of the sample statistics. You have to really think about what you're saying. Okay?

Activity 1-7, Episode 2: Transition to margin of error

In this episode, Terry tried to make the connection between variability, sample size, and sampling error. That is, the variability in a distribution of samples of size *n* can be an indicator of sampling error. Lines 25 and 36 in Excerpt 86 suggested that Lucy and Alice might have understood this idea.

Excerpt 86: Connecting variability to sampling error

17.	Terry	Okay, let's talk real quickly about where the variability issue is important. What is the reason that we go out and take a sample? What's the reason for taking a sample of something?
18.	John	To predict something about the population.
19.	Terry	Because we want to know something about the population. Okay? That's the reason we go get a sample, because we want to be able to make some inference about what's happening in the population. All right, how do you know or how sure can you be that your sample is really representative of the population? How do you know? I mean the samples are not all the same, right? You got different samples each time. How confident or how sure can you be that the one sample that you took, that you're using to describe the population, how sure can you be that your one sample truly represents, in this case, the proportion of handguns that were used in the murders in Chicago?
20.	Terry	If I took a sample of size 10 … when I took a sample of size 10, one sample of size 10. Pick any sample you want=
21.	Nicole	You need to be able to do it over and over again.
22.	Terry	=Just think of all the samples of size ten that we have up there. If I pick one of those samples of size ten, can I be pretty comfortable that my sample is pretty close to the proportion in the population? Lucy you're shaking your head, how come?
23.	Lucy	10 doesn't seem like a very large number to me out of 669.
24.	Terry	Well, relate it back to your histogram. Other than you just feel that it is not big enough.
25.	Lucy	The range and the variability were greater when there were only 10.
…		
32.	Terry	Do you feel a little more confident about any one sample in sample size 100=
33.	John	Yes.
34.	Terry	=being a good estimate? Why do you feel like those are better estimates? Think about where your samples are.
35.	Linda	They look at the area of that shaded portion …
36.	Alice	Less variability.

Line 23, Lucy's first reaction to Terry's question, however, suggested that she was not oriented in thinking about distribution when considering measurement error of a sample (i.e., she does not have a Contractor's perspective).

Summary

The discussion on this activity was relatively straightforward. The ideas that emerged from the teachers' simulation and ensuing discussion included 1) the mean of distribution of sample statistics is close to the mean of the population from which the samples are drawn; 2) smaller samples have larger variability. Terry, towards the end of the discussion, attempted to relate the idea of variability to sampling error, which would build a foundation for later discussions on margin of error.

Variability and Population Parameter

Section 8.2 put forth one of the two central ideas that are foundational to the idea of margin of error, i.e., sampling variability is directly related to sample size, the larger the sample size, the smaller the variability. Another idea, which is the focus of this section, is that sampling variability is largely unaffected by the population parameter.[17] Together, these two ideas lead to the concept of margin of error as an indicator of sampling error.

[17] Note that this is true only if the standard deviation of the population remain the same, but this idea was intentionally ignored in the design of this activity due to the complications it might incur and due to the fact that populations with non-extreme proportions differ relatively little in their variability.

Activity 1-8, Part I: Distributions of Sample Statistics

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%

1) To how many populations does this table refer?
2) The entry in column 5, row 3 is 64.9%. That refers to 64.9% of what (be specific)?
3) The entry in column 1, row 4 is 60%. That refers to 60% of what (be specific)?
4) All the percents in each of columns 4 through 7 are approximately the same. What can we conclude from that?

In the first part of Activity 1-8, we presented the teachers with a table that contains information about collections of samples drawn from four populations having parameters of, respectively, 57%, 60%, 65%, and 32%. Each row of the table corresponds to one population and a collection of samples of size 500 taken from that population. The size of the collection varies as shown in column 3. Columns 4-7 quantify the dispersion of sample statistics relative to the population parameter: The percent of sample statistics that are within one, two, three, and four percentages points of the population percentage. Note that patterns of dispersion hardly vary across changes in population parameters.

This questions following the table queried teachers' understandings of the information displayed in the tables, and their understandings of the patterns in the distributions of sample percents. Table 56 provides a model answer to each of the four questions:

Table 56: Activity 1-8 questions1-4 and model answers

Q1	**To how many populations does this table refer?**
A1	The table refers to four populations. The first column of the table tells the population parameters—65%, 32%, 57%, and 60%.
Q2	**The entry in column 5, row 3 is 64.9%. That refers to 64.9% of what (be specific)?**
A2	64.9% refers to 64.9% of the 6800 sample statistics, each statistic computed for a sample of 500 individuals selected from the population of which 57% of it is "Yes" The value indicates that 64.9% of these sample statistics are within 2 percentage points of the population percent, i.e., in the interval [0.55, 0.59].
Q3	**The entry in column 1, row 4 is 60%. That refers to 60% of what (be specific)?**
A3	60% refers to 60% of people in a population. The value indicates that 60% of people in that population believe "Yes" on some issue.
Q4	**All the percents in each of columns 4 through 7 are approximately the same. What can we conclude from that?**
A4	This pattern suggests that no matter what the underlying population percent, the fraction of sample percents contained within 1 through 4 percentage points of the population 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)

It was hoped that the teachers understood that behind the pattern exhibited in the table was the idea that *for a given sample size, the sampling variability—the patterns of dispersion of sample statistics—is largely unaffected by the population parameter.*[18]

Episode 1: Question 1

To Question 1: To how many populations does this table refer, the teachers offered two types of answers. Linda, Lucy, and Henry believed that the table referred to only one population, while Nicole argued that the table referred to four different populations. The following excerpt revealed the underlying arguments given by both parties.

Excerpt 87

516.	Terry	So question "1". What do you say, Linda?
517.	Linda	One.
518.	Terry	All right, one. Lucy, what do you say?
519.	Lucy	One.

[18] By this I mean that *the data* suggest that variability of sample statistics is independent of those population parameters. It goes without saying that variability of sample statistics is greatest when $p = 0.5$ and decreases as p nears 0.0 or 1.0.

520.Terry One.
521.Nicole I don't know what I'm doing, but I think it is four.
522.Terry Why do you think it is four?
523.Nicole Because the… the percent yes and the population varies so much. Percent of people saying yes.
524.Terry Okay. So, if it were one population … Then—And so the idea here is that they've asked some people in a population a question, a yes or no question, I'm assuming, and they're looking at the percent of people who say yes. If you have a population and you asked them a question one time, and you tally up how many people said yes in the population, is that going to vary?
525.Henry Yes …
526.Terry You're going to go and ask everybody in the whole population?
527.Henry Oh, no, I thought you were talking a sample.
528.Terry No.
529.Henry No.
530.Terry The population parameter is a fixed value at every given time. We're not talking about, you know, tomorrow, but=
531.Linda I was confused about the title of the column. I thought this was the result of … that experiment. The 65% was the result the first time.
532.Lucy That's what I thought too=
533.Terry Okay, so you're interpreting it as being the same population that we asked four different times?
534.Linda Different collection of samples each time, and this is the result we got.
535.Nicole Yeah, see that was confusing because I always think about our not knowing the population parameter=
536.Linda Right, you don't know what it is=
537.Nicole =what we're trying to do by statistical sampling is to estimate.
538.Pat How do you interpret that "percent of yes" in the population?
539.Nicole Well, to me, I thought—I interpreted it as that for some freaky reason we know exactly how many people are saying yes to the population. That's why I found the whole thing confusing, 'cause I didn't think you'd ever know the actual percent of yes in the population.
540.Terry Linda, can you go back to what you were saying about the—lets see, how did you say it—it was the results of four different times, is that what you said=
541.Linda I took this as being—each one of those numbers, 65%, 32..as being the result of our having done different collections of samples.
542.Henry That's what it looks like to me.
543.Linda =So 500 people in a sample didn't do any—2500 times what we got as a number was 65%.
544.Henry And then you did it again=
545.Linda And then you did it again and you get .2=
546.Henry 500 size sample you get 2000 times and you get 32=

547.Linda Mmhm, that's what I took it to mean, but, and again because I don't think we ever know, really, what the actual value's going to be. If we did we wouldn't be doing samples.

This excerpt showed that three teachers, Linda, Henry, and Lucy believed that the table referred to one population. The grey highlights indicated that Linda and Henry did not take the first column "percent of yes in population" as a parameter that would define a population. Rather, they saw the values in the first column of the table as results from collections of samples, although they could not articulate what they meant—of what 65% was a result? Line 75 shed light on why Linda did not take the first column as population parameters. She believed that *the purpose of sampling was to estimate the population parameter, therefore the population parameters were not supposed to be known*. Since the table clearly indicated that samples (collections of samples) were taken, it meant that the population parameter was not known.

Nicole believed that the table referred to four populations. She saw the first column "percent of yes in population" as population parameter, and each value defined a different population. However, her answer did not come free of confusion. She also believed that *the purpose of sampling was to estimate the population parameter, therefore the population parameters were not supposed to be known*. This belief made her confused: why sample if we already know the population parameters?

This discussion revealed that Linda and Nicole had a very fixed and narrow understanding of sampling, i.e., the purpose of sampling as estimating a population parameter. However, sampling (or parameter estimation for that matter) is as much about quantifying variability as about obtaining an estimate for a population parameter. The variability of a particular sampling method can be quantified by the variability of a

distribution of sample statistics generated from repeated sampling using this method. Therefore, in order to understand sampling error, one has to look beyond "take one sample" to study the behavior of distributions of sample statistics. The latter allows one to develop a sophisticated understanding of sampling that includes knowing the relationship between sample size, population parameter, and the characteristics of a distribution of sample statistics. This is essentially the purpose of this activity. Linda and Nicole's conception of sampling during this discussion revealed that they were not at a position to appreciate this purpose. They were not oriented to "study the behavior of distribution of sample statistics".

The discussion ensued and Linda suggested that if the word "parameter" had appeared in the first column, she would have understood it the way it was intended. Terry responded that the word "population" in a phrase such as "percent of population" would suggested that it was a population parameter, and adding the word "parameter" would be redundant. It was then Linda agreed that the first column of the table referred to population parameters.

Episode 2: Question 4

Question 4: All the percents in each of columns 4 through 7 are approximately the same. What can we conclude from that? The following excerpt revealed John's answer to this question.

Excerpt 88

137.John	I was just going to make a comment that the chart's up, if you look yesterday, we changed the number of people in the sample=
138.Terry	Yes.
139.John	=We made it 10, 25, 100, well you see they fix it here. I think what they're trying to show us … If you look at all the other columns, I know they're off by 1 or 2, you know, tenths of a percent, or whatever, but mainly everything is the same. So what they're trying to, I believe,

	I may be wrong, but what they're trying to show us here is, we can sit here and take 500 and do it 2000 times, 500 do it 6800 times, still look everything is just the same. They're just throwing those there just to show you that those, that the population—those can be different.
140.John	I think, because yesterday what we did was we took 10 and did it 100 times, so what would have changed would've been the second column change for us yesterday and we saw that it got closer in, the bound was closer.
141.Terry	So your point about what's varying, do you want to say that again about the 3rd column. You're saying that's varying and what is that illustrating=
142.John	That's just showing you that it really doesn't have the much of an affect on the outcome as it does when you change the size of the sample. Not the repetition of the sample. I mean, you still need to sample a lot of time=
143.Terry	When you say 'it', when you say 'it doesn't have as much an affect' … Can you give me a better word for 'it'?
144.John	Let me say it again because I don't know what I said 'it' for. So, what I'm saying is, the size of the sample that you take is more important—I want to be careful here because I'm afraid I'll make a mistake here—it is more important, I believe, than the repetition of the size of that—taking that. So, for instance, that's what I'm trying to say.

In this excerpt, John reiterated the idea that they had generalized from the previous activity, i.e., *the smaller the sample size, the more variable (or "spread-out") the distribution of sample statistics* (line 140), and he made the connection to the table in question. He observed that in the table the sample sizes were the same and the numbers in column 4-7 were also approximately the same. The fact that John made the connection suggested that he understood the column 4-7 as indicators of the variability of the distribution of sample statistics. John's generalization was that *the population parameter and the number of samples drawn do not have as much an effect as the sample size to the variability of the distribution of sample statistics* (line 139 and 144).

The ensuing discussion (Excerpt 89) revealed that other teachers made only half of this generalization, i.e. the relationship between the number of samples and the variability of distribution of sample statistics (line 156).

Excerpt 89

156.	Nicole	Okay so the answer to "d" is because the number of samples drawn— Since the number of people in a sample is constant, then the number of samples drawn doesn't make that much difference. All right=
157.	Alice	Provided you draw a large number.
158.	Nicole	Yeah.
159.	Alice	And what determines what that number is. I mean if you only do 5 samples versus 1000 samples, somewhere it is got to be—something in there has to determine when those numbers become close like that … True or false?

Alice's comments (line 157) to Nicole's reiteration of John's idea and her follow-up question in line 159 led the discussion to an unexpected direction. After Nicole repeated John's utterance in line 144, i.e. when sample size is constant, the number of samples drawn does not make a different, Alice added that this conclusion was contingent upon that condition that a large number of samples were drawn. I infer that this was a result of her observation that all the numbers in column 3 were relatively large (comparing to the number of repetitions they had done in the Fathom simulation on the previous day). Alice then asked the question: *How large does the number of samples have to be in order for the pattern in column 4-7 to sustain?*

Alice's question, in essence, was about the relationship between variability (of distribution of sample statistics) and sample size, only that the *sample* here is a collection of samples, and the *sample size* is the number of samples in a collection. As I mentioned earlier, the teachers had concluded from the previous day's simulation that *as you change the sample size n, the variability of a distribution of sample statistics from samples of size n would also change*. *Smaller samples are more variable*. To understand the question Alice posed, one has to conceive of "a collection of samples" as a SAMPLE. In other words, it is a SAMPLE of samples, and the "number of samples drawn" would be the SAMPLE size of the SAMPLE of samples, in which case the same relationship between

sample size and variability would apply. The statistics of a SAMPLE would be the percents of samples that within various ranges of the population parameter.

To investigate Alice's question, the teachers decided to simulate taking repeated samples of size 25 from a population and observe how the distribution behaved as the number of samples vary. They first simulated taking 100 samples of size 25, and found that 67% of the sample statistics were within 10% of the population parameter. They then simulated taking 10 samples of size 25. Terry repeated this simulation five times, and for each time respectively, 70%, 100%, 70%, 90%, and 80% were within 10% of the population parameter. Finally they took 1000 samples of size 25 and found that 71% of the sample statistics were within 10% of the population parameter.

Pat took this opportunity to probe whether the teachers saw the connection between their earlier generalization that sampling variability decreases with increased sample size and the current discussion. That is, he was interested in whether they understood that the collection of samples can be seen as a SAMPLE, and smaller SAMPLES are more variable as shown in the simulation results.

Excerpt 90

226.Pat	Can I ask a question? Do you see any connection between what we're doing here now, at this particular moment and what you did yesterday? And a conclusion we drew yesterday?
227.Terry	Yesterday morning or yesterday afternoon?
228.Pat	Afternoon.
229.Sarah	I think in the afternoon we were looking at changing the size of the samples and we were keyed around the .4, is that going to change?
230.Pat	Are you changing the size of, a size of a sample here?
231.Sarah	No. We're changing the number of samples.
232.Pat	Yes, you're changing the sample=
233.Various	[talk over one another]
234.Terry	Oh, Okay. Yeah.
235.Pat	So, what did you conclude yesterday about changing the size of the sample? Smaller samples are more variable, right?
236.Alice	Right.

237.Pat	Didn't you see that here?
238.Terry	Yeah and that's the other layer, is now we've got a sample of samples … So when we saw it—We saw it when we did a bunch of samples of size ten and saw it go … The percent of samples that were within a certain range, I think we had 67,70 we 100%, 90%, 80% … We were actually thinking about the distribution of the collections of samples. And that's going to also vary. And now that we've had more in that collection, we didn't do 1000 again, we could, but we probably won't see … as much. I don't see as much of a change. See?
239.Pat	See now you're taking a sample of size 1000 and=
240.Terry	You're looking at a collection of samples of=
241.Pat	Samples.
242.Terry	Right. A sample of samples.
243.John	Yes. That makes sense now. If you have to think on that second level.
244.Terry	Yeah. But I think you're point that for them to see that the variation is going to stay about the same, you know, regardless of how many samples. There probably is a point at which you're going to see some variation until you get to a certain point, and then you see it sort of settle down. We're actually going to address that in a minute, Okay? That's a good point and that's something students get confused about—the difference between how many samples and sample size, and they think if I take more and more samples I'm going to see something different, not really thinking about the fact that the sample size is probably, I wouldn't say, well we would say, more important than how many samples you do. Because, theoretically, in all the theoretical statistics, you're theoretically taking every possible sample, which we can't ever do in a classroom unless you have a really tiny population.

The highlighted utterances from line 229 to 231 revealed that Sarah did not make this connection. Terry understood that she was simulating taking SAMPLES of samples and changing the SAMPLE size, and thus the relationship between sample size and variability would apply. John's comment in line 243 suggested that he might have understood this as well. What this understanding would imply with respect to Alice's question, as Terry address in line 244, was that the pattern in the column 4-7 did have to do with the number of samples. That is, Alice was right that the pattern was contingent upon the condition that a large number of samples were drawn.

Interview 2-4: Purpose of conducting re-sampling simulations

One important finding emerged from Episode 1 of the above discussion, which we did not anticipate, was about teachers' understanding of the purpose of sampling. As we know, the purpose of taking one sample is to estimate a population parameter. The purpose of a repeated sampling, or re-sampling simulation, however, is to explore the characteristics of a sampling distribution. For example, the purpose of the Fathom activity discussed section 8.2 was to engage the teachers in explorations of distribution of sample statistics, and particularly, relationships between variability and sample size. Discussion in this episode revealed that there seemed to be a predominant view among the teachers that the purpose of sampling was to estimate the population parameter. To further probe teachers' understanding of sampling or re-sampling simulation, in the interview occurring the next day we asked the teacher the following question:

Mrs. Smithey conducted a computer simulation of collecting 100 samples of size 25 from a population having 32% with characteristic X. A student wondered out loud what the point of doing the simulation is when you already know the answer! Please comment: What is the purpose of using a simulation to make collections of sample statistics?

Table 57 summarized teachers' answers to this question.

Table 57: Summary of teachers' answers to Interview 2-4

John	John: A lot of times we don't know what the population percent is, and we need to sample to get a good estimate of it. Since we know what the population percent is, then we know we're doing something correct. Luis: if we know the population percent, why would we take samples? John: To see if it is true, if the population actually is … Luis: but we know John: Yes, we know. (laughing) By going through and doing this, we can know about some other characteristics of the population, e.g. standard deviation, things like that … that are also important.
Nicole	I think she wants to show that there is a variability in terms of what happens when the mean is going to vary … If you were to look at the entire distribution of 100 samples, and you can study all sorts of characteristics apart from variability. You can alter the … you can do all kinds of simulations to see the distribution changes, depending on the size of the sample, or the number of samples.
Sarah	As an instructor, I think the simulation could be a real good thing to show the effectiveness of sampling. You take 100 samples of size 25, hopefully you could come up with the number that exists in the population, and students start to see a connection between this idea of sampling 100 and how it related to what you know is real … this is just a confirmation or affirmation or whatever, the process. You are showing that the process hopefully get close to what the real is. Because if all this statistical process doesn't get close to the real, kids are going to have a hard time knowing why do we do this.
Lucy	To show them that here is p, 95% of the times they are going to fall around that (drew normal curve, margin of error) And if you compute p hat for each sample, and you have p hat + 5%, and p hat –5%. If you do this many times, 95% of these intervals will contain the 32%.
Betty	It depends on what the assignment is. It could be that you're checking how close these samples reflect the parameter. Maybe to show that how many you need to take … what difference might there be in the samples …
Linda	The one reason we do simulation is to understand the behavior of the samples, and what would happen when we do these samples. If you knew the population percentage … [what would you do with that understanding when you gain it?] You could see how the sample fall within that 32%.
Henry	The simulation provides a tangible proof that indeed the population percent is 32%. It will reinforce the idea that the average of averages will indeed approach the 32%.
Alice	Well, the point is we don't always know what the population parameter. So by taking all these samples, more than likely the proportions we got from that samples would be consistent with this population percent, we would be more or less confirming. We could make inference from samples, rather than knowing the true population percent.

Table 57 suggested that the teachers had conceived of a variety of purposes for which simulations would be useful. There purposes can be summarized into the following categories:

1. To confirm or refute that the population parameter is 32%;
2. To find out other characteristics of the population, such as standard deviation;
3. To demonstrate the variability among samples;
4. To learn the behavior of distribution of sample statistics, which includes
   4.1. To study the relationship between variability, sample size, and the number of samples;
   4.2. To show that 95% of the sample statistics fall within a certain range;
   4.3. To show that the mean of the distribution of sample statistics approach to 32%, and therefore make the connection between sampling and population;

Figure 12 presented these categories graphically and in more general terms (p denotes the population parameter).

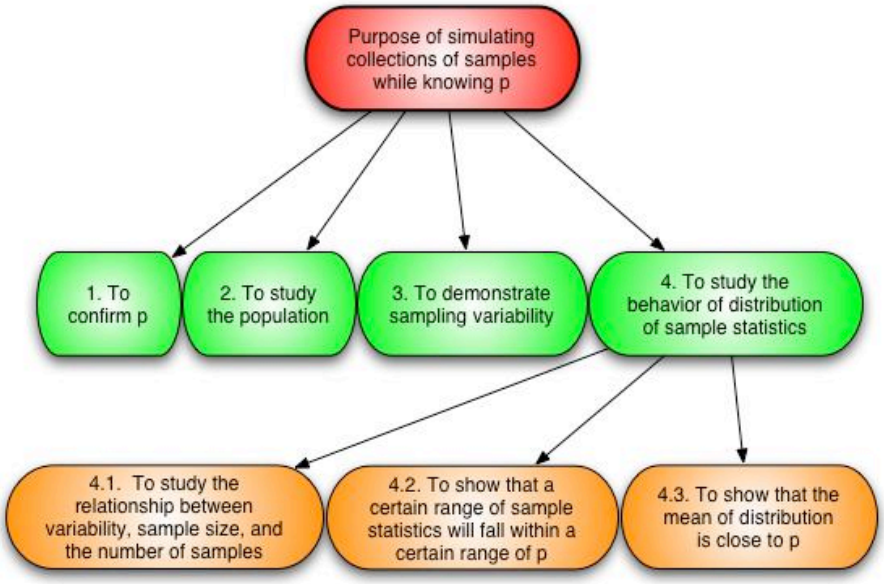


Figure 12: Framework for purposes of simulation

Table 58 coded the teachers' answers using the framework. It showed that three teachers, John, Henry, and Alice, claimed that the purpose of re-sampling simulation was to confirm the population parameter. However, with the exception of John and Alice, all the remaining teachers had identified one purpose of simulation as studying the behavior of distribution of sample statistics.

Table 58: Teachers' understandings of the purpose of simulation

	1	2	3	4		
				4.1	4.2	4.3
John	√	√				
Nicole			√	√		
Sarah						√
Lucy					√	
Betty					√	
Linda					√	
Henry	√					√
Alice	√					
Counts	3	1	1	1	3	2

Summary

The discussion around Activity 1-8 revealed that teachers made the generalization that the variability of the distribution of sample statistics was unaffected by the number of samples drawn. Except John's vague suggestion in line 139 (Excerpt 88), no explicit conclusion was drawn about the relationship between population parameter and variability, i.e. (in this activity) *the variability of a distribution of sample size has to do with sample size, but not population parameter*. Understanding this relationship is one of the necessary conditions for one to understand the rationale behind the idea of margin of error and confidence level as indicators of sampling error in polling situations.

The discussion around Activity 1-8 and Interview 2-4 also revealed the teachers' understandings of the purpose of re-sampling simulation. In Activity 1-8, we found that Nicole and Linda had the belief that the purpose of simulation was to estimate the population parameter, which would suggest that if a population parameter were known, sampling or simulation of sampling would be unnecessary. We designed Interview question 2-4 to further probe teachers' understanding around simulation. We found that while three teachers, John, Henry, and Alice continued to believe that the simulation was

to confirm the population parameter, all teachers except John and Alice had conceived of a purpose of re-sampling simulation to be that of "studying the behavior of distribution of sample statistics".

Margin of Error

Activity 1-8, Part II: Stan's interpretation

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%

5) 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?

Question 5 of Activity 1-8 queried teachers' understanding 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. We coined the scenario so that the information on the table could determine the confidence level associated with the scenario's sampling method and the reported margin of error. A "conventional" interpretation of the reported margin of error is:

> The margin of error ±4% means that if we were to repeatedly sample 500 TN voters, around 91% of the sample statistics will be within ±4% of the true population proportion. We don't know if 36% is within that range.

The same interpretation expressed with the idea of confidence interval is:

> We don't know if the interval 36%±4% will contain the true population proportion, but we do know that if we were to repeatedly sample 500 TN voters, around 91% of the intervals constructed like this will contain the true population proportion.

This question was given as homework on day 3. Teachers were asked to give a written answer to the question prior to the discussion. After a 2-hour discussion on day 4, we asked the teachers to give a second answer to the question. Below I will parse the description of results into two sections. In section I, I will discuss the teachers' written answers using the conceptual framework that I described in Chapter 4. In section II, I will highlight significant episodes of the discussion to explore the underlying thinking behind the teachers' interpretations of margin of error.

Section I: Teachers' written answers

Coding system

The rubrics of the coding scheme (Table 59 and Figure 13) were explained in Chapter 5.

Table 59: Theoretical constructs in the margin of error framework

1	The interval $p\pm r$ contains $x\%$ of s_i;
2	$x\%$ of the intervals $s_i\pm r$ contains p;
3	The interval $p\pm r$ either contains or does not contain s;
4	The interval $s\pm r$ either contains or does not contain p;
5	$x\%$ is the confidence level;
6	The interval $s\pm r$ contains p;
7	There is an $x\%$ probability that the interval $p\pm r$ contains s;
8	The interval $s\pm r$ contains $x\%$ of s_i;
9	The interval $p\pm r$ contains $x\%$ of the intervals $s_i\pm r$.
1or2or8or9	Interpretations that include distribution of sample statistics;
1&3&5	Understanding of margin of error;
2&4&5	Understanding of confidence interval.

Margin of error $\pm r$ ($0<r<1$), confidence level $x\%$, a population parameter p, a sample statistic s from a sample of size n (an estimate of p), a collection of samples of size n: s_i.

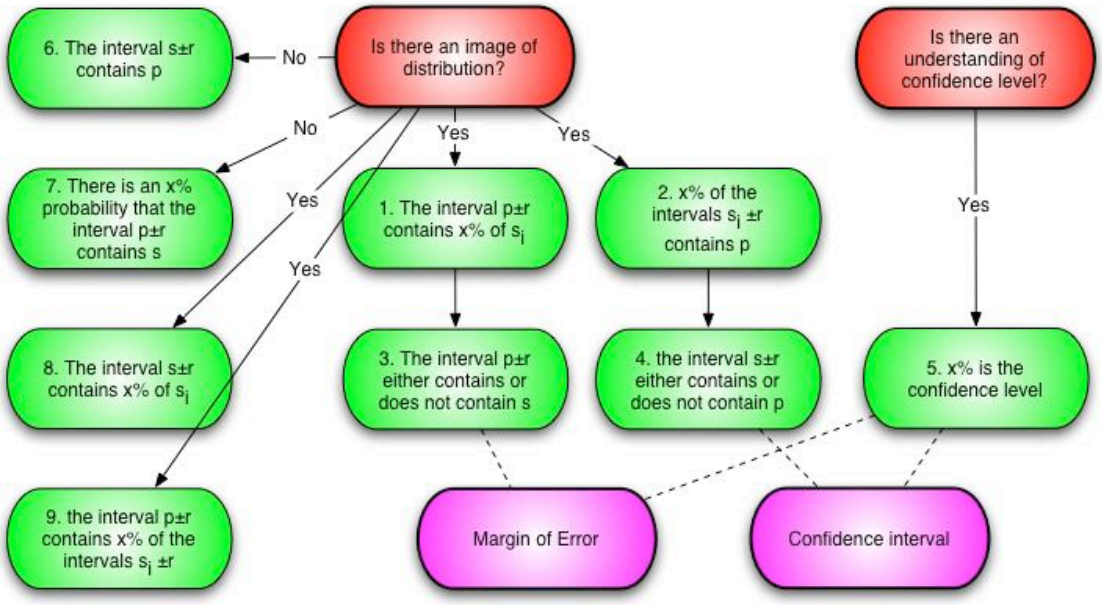


Figure 13: Theoretical framework for understandings of margin of error

Codes 1-4 and 6-8 pertain to teachers' image of distribution of sample statistics. The process of assigning a code to teachers' interpretation is relatively straightforward. Code 5 is less explicit. I assign code 5 to a teacher's interpretation when his/her interpretation exhibits an understanding of the confidence level. In the Stan's Interpretation scenario,

understanding confidence level means that a teacher is able to recognize from the table that the confidence level for the reported margin of error is 91% and understand that 91% means that 91% of the sample statistics fall within the range ±4% of the population parameter.

Teachers' written answers prior to the discussion

Prior to the discussion of Question 5, we asked the teachers to give a written answer to the question. Table 60 summarizes the teachers' initial answers.

Table 60: Teachers' initial answers to Q5 of A1-8 Stan's Interpretation

John	First I assume that the confidence interval is 95%. The answer is Stan's interpretation is not good. If the margin of error is ±.04 then that means that if we sample x amount of people 100 times then 95 of them will fall in an interval that is + or - .04 units away from the proportion of the population. It doesn't mean that the interval is bounded by .32 and .40 because the error is made of assuming that .36 is the actual proportion of population.
Nicole	No, the ±4% margin of error means that if a Gallup poll of 500 TN voters was taken repeatedly, the percentage of people saying they favored a state income tax would be between 32% yes and 40% yes 95% of the time. So 95% of the samples have a mean between 36±4%. (We don't know the actual % of the TN voting pop. who would say yes.)
Sarah	Not necessarily—If I understand what was said yesterday, then: Assuming a 95% confidence level, then there is a 95% probability that 36% of the population is within 4% of the true mean. This would indicate that the mean of the sample is between 32% + 40%.
Lucy	"500 of TN voters think a state income tax is necessary to overcome future budge problems … w/ margin of error of +4%"Stan says this means b/t 32% and 40% of TN voters believe income tax is necessary. I don't believe Stan's interpretation is a good one because he makes the assumption that the statistic found in the sample is the same as the statistic for the population. The sample he is regarding could have one of the 'rare" cases. I would interpret it to say that the survey found that 32% of their samples (500 TN voters) thinks a state income tax is necessary, and Gallup is 95% certain that the result they found for their sample is within ±4% of the actual population's (all TN voters) opinion on state income tax.
Betty	Not really. The interval 32% to 40% is just one sample of 500 people. The ±4% represents an interval that would hold the population proportion. 95% of sample proportions will be within 4% of true population mean.
Linda	Not exactly. Stan's interpretation is not exactly a good one. It shows he has some idea but is wrong on 1 aspect. (Kind of has it backwards). We are 95% sure that the result, 36% is within the interval (m-4, m+4) where m is the actual percentage of TN voters who think an income tax is necessary. (Assuming a confidence level of 95%).
Henry	A margin of error being 4% means that he is 95% sure that the true population mean is

	between 32% and 40%. Because, 95% of all the confidence intervals produce in this matter will contain the true population mean. For this particular interval it either contains the mean or it doesn't. But 95% of all confidence intervals will contain the population mean. Thus, the average of the voters opinion will likely be between 32% and 40% but some voters' individual opinions might be outside this interval.
Alice	No, Stan's interpretation is not a good one because we do not know the center of all the data (if more samples had been taken). We only have information about one sample. However, the margin of error implies that more samples were taken. 95% of the sample proportions should lie between ±4% (4% on either side) of the center of all the sample proportions. Since we don't know the center, we can't be sure if 36% is an accurate representation of the population.

Table 60 showed that none of the teachers agreed with Stan's interpretation. Three teachers, John, Betty, and Alice, interpreted the margin of error ±4% as meaning "95% of sample statistics fall within ±4% of the unknown population parameter". Henry believed that the margin of error ±4% meant "95% of the confidence intervals constructed from this margin of error will contain the unknown population parameter". These two interpretations of margin of error, conveyed by codes 1 and 2, are two coherent interpretations of margin of error, both of which build on an image of a distribution of sample statistics.

Nicole had the misconception that the *interval s±4% contains x% of the sample statistics* (code 8). Three teachers, Linda, Lucy, and Sarah, used the word "probability" to relate the sample statistic and the population parameter (code 7). These interpretations of margin of error did not built on an image of a distribution of sample statistics.

Although none of the teachers agreed with Stan's interpretation, only one teacher, Henry, explicitly stated an idea that countered Stan's interpretation. This idea was: *The interval s±4% does not necessarily contain p* (code 4) or equivalently, *the interval p±4% does not necessarily contain s* (code 3).

All teachers used the number 95% where they hoped to convey their subjective level of confidence. Only three teachers, John, Sarah, and Linda, stated that the 95% was the "confidence level". None of the teachers utilized the table to infer that the confidence level (standard sense: number of sample statistics that are within the interval $p \pm r$) was 91%. Table 61 summarizes and quantifies these results. The heading row of this table corresponds to the theoretical constructs/codes delineated in Table 59.

Table 61: Teachers' initial interpretations of margin of error

	1	2	3	4	5[19]	6	7	8	9	1or2or8or9	1&3&5	2&4&5
John	√				*					√		
Nicole								√		√		
Sarah					*		√					
Lucy							√					
Betty	√									√		
Linda					*		√					
Henry		√		√						√		
Alice	√									√		
Counts	3	1	0	1	0	0	3	1	0	5	0	0

Table 61 also shows that five teachers' interpretations of margin of error were built on an image of a distribution of sample statistics (code 1 or 2 or 8). Codes 1&3&5 or 2&4&5 are used to denote two different ways of understanding margin of error that are both coherent and complete[20]. As we can see from the table, none of the teachers understood margin of error as indicated by either combination.

[19] I assign √ when an answer indicates that the confidence level is 91%. I assign * when a teacher uses the phrase "confidence level" to refer to the percentage of samples that are within the interval $p \pm r$.

[20] By "coherent", I mean understanding margin of error to mean "95% of sample statistics are within the interval [population parameter ± margin of error]". By "complete", I mean understanding of margin of error that also include an understanding of confidence level, and an understanding that "a particular sample statistic might not be one of those 95% of sample statistics that are within the interval [population parameter ± margin of error]".

Teachers' written answers after the discussion

Table 62 summarizes the teachers' answers to Question 5 after the discussion.

Table 62: Teachers' second answers to Q5 of A1-8 Stan's Interpretation

John	No. First of all we need to assume that 95% of all samples taken will fall within ±4% of the actual population proportion. Stan has made the mistake of using the 36% as the population proportion. Stan should have done 1 of 2 possible things. 1. He should have interpreted the 36% as just one out of a large number of possible sample proportions that could be taken. For instance out of 100 samples taken, 95% of them would within 4 percentage points of the true population proportion. OR 2. He could have constructed a 95% confidence interval using the 36%±4% to be (32%, 40%) which means for any sample proportion that is calculated there is an interval of ±4% constructed around that sample proportion. If 100 of these confidence intervals were constructed, then 95% of them would contain the true population proportion.
Nicole	No. The 4% margin of error means that if a Gallup poll of 500 TN voters was taken over and over again, then 95% of the those samples would have the proportion of people saying "I do not agree with State Senator Doug Henry and I favor an income tax" would fall within 4% of the actual (unknown) proportion of voters in TN who supports an income tax. (Diagram) The actual proportion of income tax supporters is likely to be between 32% and 40% of TN voters.
Sarah	Not really, the 36% sample may or may not fall within ±4% of the population proportion and ∴ doesn't limit the range to 32-40%. (However, using a confidence level of 95%, he can have some reasonableness to the observation—scratched) Therefore, based on one sample it is difficult to make such a statement. A better attempt might be: Approx. 36% (±4%) of the population indicated they would support an income tax. This was sampled and is reported with a 95% of confidence. Or, ninety-six percents of those polled indicated that they would support an income tax. Inferring across the entire population, there is a margin of error of ±4%.
Lucy	I still don't believe Stan's interpretation is a good one. I think I have the same opinion as I did on my previous answer to this question. However, I think I may not have worded my answer with 100% correct terminology the first time. So let's try again … If Gallup were to take 99 more polls (or rather samples using the same poll), 95% of those samples would contain the actual population proportion within a confidence interval of their sample proportion. i.e. 95 of 100 times, p would be contained within p hat ±4%. Maybe you could reword Stan's thought + make it correct by saying 95% of the sample proportion (p hat) will fall between (p-4)% and (p+4)% rather than what he said—95% will fall between 32% and 40%.
Betty	No. 36%±4% represents the margin of error for one sample thus giving an interval of (32%, 40%), General assumption (95%)—95% of sample intervals will contain the population proportion (this value we don't know). Or 95% of all sample proportions will fall within 4% of the unknown population proportion. This interval would be one of the intervals that is included in the "95% of all sample intervals," thus containing the population proportion.
Linda	Stan's answer is not correct because as you can see from the graph above, the interval "32 to 40" doesn't tell us anything and it doesn't even fall within the interval determined by the intended meaning of ±4%. He should have said the poll result was

	36% with the knowledge that if we did this sampling 100 times, 95% of our results would fall within ±4% units of the actual population percentage that think an income tax is necessary. Furthermore, he could have said "the interval (32, 40) may contain the true population percentage that think an income tax is necessary. Out of 100 similar polls, 95% of such resulting intervals would actually contain the true population percentage. (to put into perspective whether or not you want to trust that interval.)
Henry	Stan is somewhat confused. 95% of all confidence intervals collected in this manner will contain the true population proportion and 95% of all sample proportion will be within ±4% of the true population proportion. Stan's interval of 32% to 40% will either contain the true population proportion or it won't. But 95% of all such intervals will contain the true population proportion. So Stan's interval of 32% to 40% may indeed contain the true population proportion and it could be said that he is 95% sure that the true population proportion falls within that interval, understanding the process behind the "95% sure' statement.
Alice	No. Stan's interpretation is not good. A 4% margin of error means that 95% of the sample proportions will fall within 4% of the population proportion, since we do not know the population proportion, the numbers 32% and 40% should not even be used. However we can be reasonably sure (95% of the time) that the sample of 36% yes would be within 4% of the population proportion (an unknown measure).

I coded the teachers' answers using the same method as I did with their initial answers (Table 63).

Table 63: Teachers' second interpretations of margin of error

	1	2	3	4	5	6	7	8	9	1or2or8or9	1&3&5	2&4&5
John	√	√			*					√		
Nicole	√					√				√		
Sarah			√		*	√						
Lucy	√	√								√		
Betty	√	√				√				√		
Linda	√	√		√						√		
Henry	√	√		√						√		
Alice	√									√		
Counts	7	5	1	2	0	3	0	0	0	7	0	0

As we can see from Table 63, all the teachers, except Sarah, understood the margin of error ±4% to mean "95% of sample statistics fall within ±4% of the unknown population parameter". Five teachers also understood that the margin of error ±4% could be

interpreted as "95% of the confidence intervals constructed from this margin of error will contain the unknown population parameter". None of the teachers used the word "probability" to relate the sample statistic and the population parameter, or had the misconception that *the interval s±4% contains x% of the sample statistics*. All teachers except Sarah had an image of distribution of sample statistics in their understanding of margin of error. Compared to their written answer prior to the discussion, 3 additional teachers, Nicole, Lucy, and Linda, had a coherent image of the distribution of sample statistics and understanding of how it relates to margin of error.

Three teachers, Sarah, Linda, and Henry, stated explicitly that *the interval s±4% does not necessarily contain p*, or *the interval p±4% does not necessarily contain s,* as opposed to only one teacher (Henry) in prior answers. However, a conflicting result was while no teacher agreed with Stan's interpretation in prior answers, three teachers, Nicole, Sarah, and Betty, held the same interpretation as Stan's interpretation this time around. [Note that Sarah's utterances offered two interpretations, one supporting and the other refuting Stan's interpretation.]

With respect to confidence level, only John and Sarah mentioned the phrase. Like in the prior answers, all teachers used the number 95% where they needed to convey their confidence level. None of them utilized the table to infer that the confidence level was 91%. As a result, once again none of the teachers had a complete understanding of margin of error. Their answers exhibited at most two elements (out of three) of an understanding of margin of error.

To summarize, in this section, I described the teachers' written answers to Question 5: Is Stan's interpretation a good one? The juxtaposition and comparison of

teachers' written answers prior to and after the discussion indicated that while majority of the teachers had developed a coherent understanding of how margin of error relates to a distribution of sample statistics, none of the teachers had an understanding of the idea of confidence level. Teachers' answers also revealed a level of inconsistency with which the teachers responded to Stan's interpretation. Stan's interpretation of margin of error—*the interval s±4% contains p*— exhibited a Carpenter's perspective: An orientation towards the additive difference between a sample statistic and a population parameter. While none of the teachers agreed with Stan's interpretation, few teachers (one in initial answers, and three in second answers) explicitly pointed out the error in this interpretation, and stated that *the interval s±4% does not necessarily contain p*. In the post-discussion answers, we found that three teachers implicitly held the same interpretation as Stan's despite their open disagreement, of which the most conspicuous result being Sarah's interpretations that supported and refuted Stan's interpretation as the same time.

Below I will turn to Section II, in which the teachers engaged in the discussion of the same question. It is hoped that an elaboration of teachers' thinking revealed in this discussion would clarify the inconsistency exhibited in their written answers.

Section II: Discussion

This discussion around Question 5 lasted 67 minutes. I will parse the discussion into the following four episodes:

Table 64: Overview of discussions around Q5 of Activity 1-8 Stan's Interpretation

Episode	**Theme**
1	Nicole and Linda's Contractor's perspective
2	Henry's confusion
3	Sarah's use of probability
4	Henry's question

Episode 1: Nicole and Linda's Contractor's perspective

We started the discussion by sharing with the teachers a list of students' answers (Table 65) to the same question. These answers were collected from a previous teaching experiment (TE 2).

Table 65: Students' answers to Q5 of A1-8 Stan's Interpretation in TE2

1	It could be that we don't necessarily know what "+/-4%" stands for. We don't know because we did not collect poll and neither did Stan.
2	No. Stan is wrong. Margin of error refers to past polls where the company has been within 4% of the actual percent 100% of the time.
3	No. It means when they take collection of samples of 500 TN voters' opinion, every time the percentage of TN voters believe it is necessary…is because…between 32% to 40%. On the survey above, it has 36% voter who believe it is necessary. It doesn't change.
4	This interpretation isn't a good one because the margin of error has nothing to do with the percent from the problem. The ±4% means that the poll has a chance to be up to ±4% off of the 36% of the poll.
5	No. The interpretation is not a good one. The margin of error did not come from that sample. It came from other ones like it with the same sample size number. The margin of error 4% does mean that between 32% and 40% may believe an income tax is necessary.
6	No. This actually means that 36% of TN voters believe an income tax is necessary, and that it could actually be 4% higher than 36% or 4 percent lower than 36% not both.
7	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 actually 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 they know lies between 32% and 40%.
8	No. The ±4% means that 96% when things are sampled with sample size 500 they are accurate.

We engaged the teachers in discussions about which answers demonstrated an understanding, or a lack of understanding, of margin of error.

Excerpt 91

395.Terry	Okay, what essential idea is [student] number one not understanding?
396.John	He doesn't know what the plus or minus 4% stands for.
397.Nicole	No, but he doesn't know anything! He doesn't know that=
398.Terry	But-but why doesn't he know what it means, is what I'm saying? What image or concept does he not have, or she not have?
399.Sarah	He's assuming you have to do your own data collection for your new sample.

400.Terry Well … Okay. We'll come back to that.

401.Nicole I don't think he or she has a clue that you have to do this whole sample repeatedly in order talk about a margin of error, right?

402.Linda I think that they all kind of think that this plus or minus 4% is because we know something special about this poll that we've done. I mean, it is the kind of poll that we do and we know that it always comes out pretty close, so …

…

407.Sarah Are they missing this idea of … this 36%, and this could be what these guys are saying and I could be wrong [trails off inaudibly], but are they missing this idea that this 36% is one sample it is being compared to their overall thing, so it may or may not fall within that 4%, and then nobody makes the statement like we just kind of arbitrarily defined it to 95%... is that there's a 95% probability, and correct me if I'm just wrong here, that this 36% will fall in ±4% of the true mean?

408.Terry I wouldn't use the word "probability", but let's not go into it. The first part of what you said, I think, is important … You said, I think, that they think that … they're somehow getting fixated on that 36%=

Nicole, Linda, and Sarah's comments to student number one's interpretation suggested that they understood the idea that margin of error is not about one specific sample result. Rather, it has to do with repeated sampling (line 401) or sampling method (line 402 "the kind of poll that we do"). In other words, they held a Contractor's perspective about margin of error, i.e., a margin of error measures the variability of a distribution of sample statistics, not particular sample statistics.

Note that in line 407 Sarah stated that the ±4% margin of error meant "There is 95% probability that the sample statistic 36% will fall within ±4% of the population parameter" without saying what she meant by "95% probability". Terry objected to the use of "probability" in line 408. Sarah came back to Terry's comment later and I will highlight the surrounding discussion in Episode 3.

Episode 2: Henry's confusion

In the next excerpt, Henry talked about his understanding of margin of error. His utterances in the highlighted portion suggested that he had an incoherent understanding regarding margin of error and confidence interval.

Excerpt 92

425.Henry	Pretty often I think about the fact that the true population proportion is [microphone problems] and this particular interval and the 95% probability, the 95% sure has to do with the number of times that you're running and getting these confidence intervals and 95% of all those confidence intervals will be within ±4% of the true population parameter, so this particular confidence interval, could be way off the mark. It could be nowhere near the true population parameter and I think they're looking at it and saying that well, within that ±4% the true population parameter is definitely in that, it is hard to see that that entire interval could be off, not just close.
426.Terry	Did you have a comment (to Pat)
427.Pat	Henry you changed your language from talking about margin of error to talking about confidence intervals.
428.Henry	Yes, I did.
429.Pat	Now, is that different or the same?
430.Henry	The margin of error refers to … I have difficulty articulating distinct differences, but I do believe that they are different, they're not the same and that the margin of error refers to the fact that all of the … that 95% of all confidence intervals collected will be within ±4% if the true mean. So that's what that margin … and the margin of error for this … Well, I'm going to shut up on that.

As I said earlier, there are two ways of expressing the same idea:

$$\textit{The interval } p \pm r \textit{ contains } x\% \textit{ of } s_i; \qquad \text{(E1)}$$

and

$$x\% \textit{ of the intervals } s_i \pm r \textit{ contains } p. \qquad \text{(E2)}$$

Henry's interpretation was instead:

$$\textit{The interval } p \pm r \textit{ contains } x\% \textit{ of the intervals } s_i \pm r. \qquad \text{(E3)}$$

This interpretation is incoherent because all intervals are of the same width ($2r$). It doesn't make sense to think that one interval will contain other intervals. However, we do

not know what might have contributed to Henry's understanding, especially considering that in his written answer prior to the discussion, he had a coherent interpretation of confidence interval.

Pat pointed out the incoherence in Henry's thinking, and attempted to make him see E1 and E2 as two ways of expressing the idea of margin of error. Excerpt 93 suggests that Henry remained confused at the end of this episode.

Excerpt 93

431.Pat	If I were to say 95% of the samples that we take of this particular size will be within 4 percentage points of the true population percent … Or if I were to say that a confidence interval centered at the sample proportion of width 8 percentage points, will include the population percent 95% of the time.
432.Henry	No that's … I agree. I disagree with the second one.
433.Terry	We haven't really talked about confidence intervals yet. You may want to hold off=
434.Pat	I'm just saying: Are those stances different or the same?
435.Henry	Different.
…	
445.Terry	I think what Pat's trying to get you to say, realize, or to think about is … .Let me just tell you, Pat and I went through this very same conversation and he had to do this with me about, I don't know how many times, it was about 20 minutes before I finally went 'oh!' That when you talk about a 95% confidence interval, that 95% is related to that ±4% whatever percent and then how is saying 95% of the intervals will contain the population parameter, the same as saying I'm pretty sure I'll be within ±4% of the population parameter. That those are two ways of saying the same thing and they are exactly the same thing, two ways of saying it and to understand that those are two ways of saying the same thing. Okay so let's say we take=
446.Henry	You see, that's what I thought yesterday and you talked me out of it so I have no idea now.

This exchange revealed that Henry did not see E1 and E2 as two ways of expressing the same idea (line 435). Henry's response (line 446) to Terry's explanation suggested that he remained confused about how margin of error relates to the distribution of sample statistics.

Episode 3: Sarah's use of probability

This episode was prompted from Sarah and Terry's exchange in Episode 1 (lines 407 and 408) in which Sarah tried to relate confidence level to the result of a specific sample. For the next 10 minutes, Terry and Pat tried to divert Sarah from fixating on particular sample statistics when *thinking* or *expressing* margin of error. The following excerpt provides a glimpse of the difficulty Terry and Pat encountered when trying to clarify this point with Sarah.

Excerpt 94

497.	Sarah	Okay, any sample. Is it 95% likely that any sample would fall four percent above or below that middle number?
498.	Pat	Are you talking about any particular sample?
499.	Sarah	I'm talking about any sample in general.
500.	Various	(chuckle)
501.	Sarah	Just say yes or no.
502.	Henry	Any particular sample in general.
503.	Terry	Well, you can't talk about any particular sample. You can say that 95% of these samples will fall within 4 … and so …
504.	Pat	No particular sample will fall in there 95% of the time!
505.	Sarah	Okay now, I have a question because kids are going to ask it. Why can I not apply that to that particular sample?
506.	Pat	Because that sample will be either in that range 100% of the time, or out of that range 100% of the time.
507.	Sarah	Well, I understand that. But if just go out here and take a sample, a sample, and well I don't know what the numbers are.
508.	Pat	And that sample will be in there 100% of the time or out of there 100% of the time.
509.	Sarah	But the probability of it being in there …

As we can see, despite Pat's reiteration that a particular sample will either be in or out of the range $p\pm4\%$, in which case, the probability of it falling in $p\pm4\%$ is either 1 or 0, Sarah repeatedly asked the same question: Why couldn't I say "There is a 95% probability that the sample will fall within that range"? The fact that Sarah did not see the apparent incoherence in her question suggested that she embraced a meaning of probability that was incompatible with the one Pat held. For Pat, probability has to do what is going to

occur over the long run. Probability of a single event, in the case of a sample falling within *p±4%,* is either 1 or 0. Saying that there is a 95% probability of one sample falling within *p±4%* is incoherent because "no particular sample will fall in there 95% of the time". A "95% probability" only makes sense when a person has a background image of a collection of sample statistics. The following excerpt revealed that Sarah did not have an image of a collection of sample statistics when talking about probability.

Excerpt 95

516.Terry	What you're saying … and that I understand, this is what the kids deal with … you want to look at that sample and tell me that there's a 95% chance that that sample … You want to talk about … There's not chance related to that sample. That sample either has the population proportion or … it is either within four percent or it is not within four percent. Where the probability come in, is if I went and took 100 samples, 95 out of those 100 would be within 4%, so I'm pretty sure, since 95% of them do, I'm pretty sure that the one that I got is one of those 95%, I don't know that=
517.Sarah	=So how is *that* different from what I'm saying?
518.Terry	Because you're trying to tell me that I'm going to take this one sample, and I'm going to look at it, and there a 95% chance that that one sample is … Here, what does that mean to tell me there's a 95% chance that that one sample's=
519.Sarah	It is more likely to fall in that range then it is outside that range?

In line 516 Terry explained what a "95% probability that *s* falls within *p±4%*" meant for her. There were two parts in what she was saying: 1) *95% of sample statistics fall within p±4%*, and 2) *I'm pretty sure that the one that I got is one of those 95%*. I argue that the word "*that*" in Sarah's question "how is *that* different from what I'm saying" referred to only the second part of what Terry was saying, in other words, Sarah did not have a collection of sample statistics in mind. What she did mean by a "95% probability that *s* falls within *p±4%*"—"It is more likely to fall in that range then it is outside that range"—was clearly an index of her subjective feeling about the particular sample statistic. Terry's view reflected the Contractor Perspective ("I have no specific knowledge of this sample's

accuracy, but the method employed to get it produces results that are within ±4% of the population proportion about 95% of the time"), whereas Sarah exhibited the Carpenter's Perspective ("I want to make a statement about whether *this statistic* is within ±4% of the population percent").

Episode 4: Henry's question

Recall that in Episode 2, Henry held an incoherent image of confidence interval and margin of error, i.e., *the interval p±r contains x% of the intervals* $s_i \pm r$. In Episode 4, Henry had cleared his confusion. In other words, he had understood that a sample statistic of 36% with margin of error ±4% meant that *95% of samples of the same size will be captured by p±4%, but we don't know if 36% is one of those 95%.*

This episode centered on a question that he raised: "Why bother taking a poll if we don't know *for sure* if the poll result will be among those 95% of the poll results that are within a certain interval of the true population parameter?" The conversation around this question revealed 1) Betty, Henry, and John's belief that a sample statistic is right when it equals to the population parameter; 2) Henry's Carpenter's perspective: concern of how far a sample statistic is from the population parameter; and 3) Linda's Contractor's perspective: The idea of margin of error is about collections of samples. Below I substantiate these claims.

Excerpt 96

589. Terry So your interval width is staying the same but the center of your interval is sliding around. Well, 95% of those intervals are going to overlap the population proportion. If I do that over and over again 95% of those intervals are going to contain the … Now there's going to be some intervals that don't contain it. That interval out here doesn't contain the population percent, Okay? So can I say anything about this one interval? Do I know whether this interval is one of these, or this one out here? (shrugs) All I know is that of all the intervals I could've gotten, 95% of them are going to contain the

population proportion. I'm pretty sure then, since 95% of them do, that I got one of those, but I don't know.

590. Betty So it is real possible that any statistic that gives us this, like this 36%, really does not communicate or give us … is not really *right*?!

591. Terry When you say "real possible"=

592. Betty Yeah, not real possible. It is 5% possible.

593. Terry Right, it is five percent possible that=

594. Sarah It was more than four percent from the mean.

595. Terry Don't say about that one … Five percent of the time that you'll get an estimate that not=

596. Betty This could be a 5%.

In this excerpt, Terry reiterated the idea of confidence level and margin of error: *We don't know whether the interval s±r contains p, but we do know that 95% of the intervals* $s_i \pm r$ *contain p*. This idea is built on the background knowledge, as I elaborated in Chapter 4, that the information we obtain from statistical inference carries *less than perfect* certainty. In other words, we could never know if a sample statistic *equals to* the unknown population parameter, and the ideas of confidence level and margin of error provide a way for us to express our confidence in the accuracy of the sample statistic. Betty's comments suggested that she did not share this background knowledge. The highlighted sentence (line 590) revealed that she came from a framework where statistics are supposed to be either right or wrong: It either equals to or doesn't equal to the population parameter.

In the next excerpt, Henry revealed his carpenter's perspective: his concern of how far off 36% is from the population parameter.

Excerpt 97

627. Henry I'm really comfortable now. I have just one question and I don't know if I know the answer to this or not [chuckles] … I'm seeing two things, when we read this question, and this question ended up with 36% as being the average of this particular sample, right? Now, I see two important things if I'm reading the newspaper. I see this 36% and then I'm thinking about the true population parameter, okay? Now, can you make a sentence, a very simple sentence that you'd publish in the

newspaper for the readers, relating that 36% to the true population? Because I think that's the intent here with this person and now everybody is trying to make is, well we got 36%, how does this relate to the *true* population percent?

628. Terry Can I throw that back out and see if somebody here can do it?

…

638. Terry Okay Alice, go ahead Alice.

639. Alice 36% will fall within 4% of the population proportion 95% of the time. Is that correct?

640. Terry Do you all agree with that?

641. Various [mumbles of disagreement]

642. Terry Linda's making a face. Do you want to edit what she said?

643. Linda 36%, there's not a 36% of anything. I think 36% was our answer, right?

644. Alice That was one sample=

645. Linda So, to put our answer into perspective as to whether or not it is the right one, the right parameter, we would say 95% of our sample results could be expected to be within the range—±4 of our true, vary no more than 4 either way, from our true value=

646. Terry So what does that tell me about that 36%?

647. Linda It tells me absolutely nothing about the 36%, but it is just saying that if you want to put it into perspective as to whether or not you believe this, we can expect that if we did this 100 times, 95 of that 100 would give us an answer that would fall 4 units left or right of what we're looking for.

648. Henry This 36 could be one of those. I understand that, but that's *not* what would be printed in the newspaper. What I want to know is, I mean why bother? Why don't we just throw away statistics? I mean if you're going to prove these numbers in the newspaper, I just want to know the math because I've confused myself so much, I just want to try to get unconfused. I just want to hear some simple language about why do I even want to bother getting a number, 36, if I'm just going to write down some arbitrary statement? I just want to know. I could just write down … What I published and what I just said=

649. Terry If you look at that 36%, so if I tell you that I did a poll of, let's say a large number of people, so that you have some sense of it being somewhat accurate. I did a poll of a large number of people in Tennessee and I found that that 36% of them favored, whatever, my margin of error is ±4%, are you saying that that … .Does that tell you anything … .Are you saying that that does not give you any idea about the population?=

650. Henry I understand, I think you're confused. I understand everything now, I think. Okay? What I'm saying is I want to know what is the best way to interpret your sample proportion and relate that to the population proportion in a newspaper format, without using the language, without *saying* 'if I did this 100 times, then 95% of the time I would get one

> that's within 4% of the true population. This time we got 36%, go figure yourself whether it is right or not'. I mean, that's would you would have to print. If *I* were writing it, because I'm not a writer, that's what I would have to write now to get my meaning across. I'm just saying, how is it normally written?

The green highlights suggested that Henry had overcome his confusion exhibited in Episode 2. He had understood was that a sample statistics 36% with margin of error ±4% meant that *95% of samples of the same size will be captured by p±4%, but we don't know if 36% is one of those 95%*. His questions, asked in many different ways in the yellow highlights in lines 627, 648, and 650, can be summarized as

1. How far off is the sample statistic 36% from the unknown population parameter?
2. If we could not have the answer to the above question, then why bother collecting the sample?

These questions revealed his conviction that if it does not matter what sample statistic we obtain we could never know how far off it is from the population, then there is no need for generating any sample. Behind this belief is a search for measurement of certainty in sample results that are expressed from the Carpenter's perspective: How far off is a measurement from the actual measurement? For Henry, the accuracy expressed from the Contractor's perspective—how many measurements are within certain range of the actual measurement— has an inherent arbitrariness in the sense that *for any specific measurement* we do not know where it falls relative to the actual measurement. Overall, this excerpt suggested that although Henry was capable of thinking like a Contractor, he was not able to get over a Carpenter's perspective.

Excerpt 98 revealed three other teachers, Nicole, John, and Linda's thinking on the same issue.

Excerpt 98

651. Henry Yes they do, but my question was, can you write it without talking about the method at all. Can you just relate the number 36 to the true population parameter without the discussion of method, because the readers don't necessarily want to hear about the method process.

659. Pat Okay, now if you, not you as a general reader, but you all as, now, as educated statisticians=

660. Nicole Well, now wait a minute! [chuckles]

661. Pat =How would you feel if a newspaper writer tried doing that to you? How would you react?

662. Terry and Nicole Tried to do what?

663. Pat Just said 'Harris found out that 36% of the voters preferred income tax'.

664. Nicole Actually, I'd be happier if they wrote 'Harris found out that around 36%' I really would be happy if they had said that!

665. Sarah I think that if you understand that Harris is a polling agency and you put some validity in their existing methods, they don't have to be explained to you each and every time. I don't think that I want to read "USA Today" if every time they run a Harris poll result, they explain to me what Harris poll has done, do you understand?

666. Pat Well that's different from explaining what they've done. This is telling you what this ±4% means.

667. Linda I think as a reader, the general reader should be able to say, 'Okay, ±4 as compared to one that's got a ± 8%, I'm going to be more likely to believe the Harris poll, or I'm going pay more attention than I would=

…

676. John Well to me the 36% in there is for the reader. They know that the majority of readers are going to have no idea to really understand the ±4%. What they're putting the ±4% in there for is to cover their butt if things don't work out=

Nicole's comment (line 664) suggested that she was comfortable with the *less than perfect* certainty, and that in fact she did not need a measurement of accuracy from neither perspective. John (line 676) seemed to think the same way as Henry did. What he meant by "if things don't work out" was likely to be "if the sample statistic is not equal to the unknown population parameter", which again revealed a Carpenter's perspective on the accuracy of the sample statistics. Only Linda (line 667) understood the mechanism with which margin of error conveys/quantifies the accuracy of sample statistics: A polling method with margin of error of ±4% will generate more accurate results than one

with margin of error ±8%, accurate in the sense that sample statistics are more closely clustered around a population parameter. Linda's conception of margin of error reflected a Contractor's perspective on accuracy.

At the end of the discussion (Excerpt 99), Pat and Terry reiterated the big ideas: 1) There is no way we can know how far off the sample statistic is from the population parameter, and 2) Margin of error quantifies the accuracy with which *samples of the same size (or the sampling method)* estimates a population parameter.

Excerpt 99

688.Pat	Well, I guess my answer to your question, Henry, would be, well, either you're going to say ' I don't care if I'm fooling them. Let's just write something that satisfies them and give up on trying to gain a sense of accuracy. Or try to write in such a way so that they convey that these percentages are about the method, not about this particular result.
689.	(Group takes a break)
690.Terry	But the other answer to your question is that's why we teach Statistics so that hopefully our students will have some interpretation of that when they go out into the world. It is a revolution.
691.Henry	Off the point of focusing on students, indirectly, I was just wondering if there's a clear succinct way of relating that 36%=
692.Terry	What you said was without typing that method, no, because it is the method=
693.Henry	Or is there, even talking about the method, a clear succinct way, because everything that we've said has involved multiple, long sentences.
694.Terry	But you can't, can't talk about it without saying that if I repeated this process over and over again I would expect 95% of the time to get a result, I just don't know any other way to say it. There's not other way to say it because that's what the 4% comes from.
695.Terry	So … it is a hard, I mean, just thinking about all the stuff we've been talking about, it is hard for the kids. The answer to your question is no. I don't think there's any way you can describe what that statistic means.
696.Henry	So I think there's a systemic error here in this process that needs to be brought out, otherwise it is just going to continue to perpetuate. And we focused a lot of attention on that sample statistic, that 36% stuff … A lot of attention focused on that and there needs to be a nice clear succinct way to explain what that is in relation to what we want. And what everybody wants is the true population.

697.Terry Yeah, you're right. It is, and that's why the kids get fixated on that.

Excerpt 99 revealed Henry's resistance in accepting idea E1, that the interval $p \pm r$ contains $x\%$ of the sample statistics, and his persistent fixation in finding out the distance between the sample statistic and the population parameter (line 696).

Interview 2-2: Harris poll

In the interview occurring the day after discussing Stan's Interpretation, we asked the teacher the following question:

A Harris poll of 535 people, held prior to Timothy McVeigh's execution, reported that 73% of U.S. citizens supported the death penalty. Harris reported that this poll had a margin of error of ±5%.

Please interpret "±5%". How might they have determined this? How could they test their claim of "±5%"?

Teachers' interpretations of the ±5% margin of error are summarized in Table 66.

Table 66: Teachers' answers to I2-2, Q1: What does ±5% mean?

John	Two ways to interpret "±5%"—1) 95% of all possible samples will fall within ±5% of the population percent. 73% could be in or out of that range. 2) Construct a confidence interval 68% to 78%. If you take multiple samples, 95% of the confidence intervals will contain the true population percent.
Nicole	We don't know how many people in the US support death penalty. If we repeatedly take samples of 535 people, the proportion of people who support … will fall in a window of … that mean … the proportion of people who support the death penalty will be … overlap … I need to draw a picture … the proportion of people who support death penalty is within 68% to 78% … 95% of the time.
Sarah	What ±5% means is what the true population mean is, the Harris poll people allows themselves a leeway of error on either side of it … They can go higher than the middle, and lower than the middle. Leeway means you have a range of values. If you have middle, you can 5% above and 5% below. To me that's a 10% leeway. We don't know what that middle is.
Lucy	Assume you do 95% confidence, that would mean, you have a normal curve, a population mean, we don't know what that is. (drew the graph) whatever that *p* is, this would be p+5%, this would be *p*- 5%. They are 95% sure … if they do the samples 100 times, 95 times they would got this within 5% of this true population. So this (73%) is probably in there, but there is 5% that it could be out there (pointed to the outliers) somewhere.
Betty	95% of the percentages will fall between that interval (*p*±5%). This interval 73%±5% will represent our confidence level. This is not right. Let me start again. This 73% could be here, with a confidence interval 73%±5%. It could be there … 95% of these confidence intervals would contain this unknown population proportion.
Linda	There is a 95% probability, or 95% of all the surveys, the results will fall within ±5% of the real percentage. The ±5% is not about the 73%, it is about the actual percent that we don't know. We don't know whether that 73% will fall within that range.
Henry	95% of all polls taken in this manner will report an average, in this case it was 73%, that will be within ±5% of the true population average. We also discussed that you can construct the confidence interval, so the 73% and the ±5% could produce a particular confidence interval, in this case, 68 to 78. That confidence interval is one particular confidence interval. Of all such confidence interval that could be collected in that fashion, 95% of them would contain the true population average. So this particular confidence interval either will or will not contain the true percent of people who support death penalty.
Alice	It implies that they've done more polls, and have discovered that 95% of the data will fall within ±5% of the population proportion.

Table 67 presented the codes of the teachers' answers using the coding scheme described earlier.

Table 67: Teachers' interpretations of margin of error in Interview 2-2

	1	2	3	4	5[21]	6	7	8	9	1or2or8or9	1&3&5	2&4&5
John	√	√								√		
Nicole								√		√		
Sarah			√									
Lucy	√		√		√					√	√	
Betty	√	√								√		
Linda	√		√							√		
Henry	√	√		√						√		
Alice	√									√		
Counts	6	3	3	1	1	0	0	1	0	7	1	0

As we can see from Table 67, all but two teachers, Nicole and Sarah, understood the margin of error ±5% to mean "95% of sample statistics fall within ±5% of the unknown population parameter". Three teachers, John, Betty, and Henry, also understood the margin of error ±5% to mean "95% of the confidence intervals constructed from this margin of error will contain the unknown population parameter". Nicole took up again her understanding that *the interval s±r contains x% of the sample statistics.* All teachers except Sarah built their interpretation of margin of error on an image of a distribution of sample statistics. Four teachers, Sarah, Lucy, Linda, and Henry, stated explicitly that *the interval s±4% does not necessarily contain p*, or *the interval p±4% does not necessarily contain s.*

With respect to confidence level, all teachers used the number 95% where they needed to convey their confidence level (Note that the question did not specify a confidence level). Only Lucy explicitly assumed a confidence level of 95% before using it to refer to the percent of samples what are within the interval $p\pm r$.

[21] In this particular situation, I assign √ only when a teacher explicitly assumes a confidence level when talking about a percentage of samples that are within the interval $p\pm r$.

Table 68 compared the teachers' understandings exhibited in their written answers of Activity 1-8 Question 5 and the Interview 2-2.

Table 68: Comparison of teachers' interpretations of margin of error in A1-8 and I2-2

Counts	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**1or2or8or9**	**1&3&5**	**2&4&5**
A1-8, 1	3	1	0	1	0	0	3	1	0	5	0	0
A1-8, 2	7	5	1	2	0	3	0	0	0	7	0	0
I2-2	6	3	3	1	1	0	0	1	0	7	1	0

Table 68 shows that before and after the discussion of Stan's Interpretation, there was a significant improvement in the number of teachers who could interpret margin of error coherently (captured by codes 1 and 2). There were no significant changes in teachers' understanding of confidence level, and of the idea that *the interval s±4% does not necessarily contain p*.

Teachers' answers to the second question "how might they have determined this?" (Table 69) revealed that out of the seven teachers who were asked this question, none understood how a margin of error was determined.

Table 69: Teachers' responses to I2-2, Q2: How was ±5% determined?

John	I don't know, but I felt it has something to do with this 535. The larger the sample, the smaller the variation, the data more clustered toward center.
Nicole	It is the sample size that has something to do with the ±5%. If they increase the sample size, then the margin of error would be smaller.
Sarah	I think if I understand what we have said in this place, it is arbitrary. They could just pick this number out. Probably they would work with a margin of error that's relatively small.
Lucy	Can't remember the process … I remember in college you have some formulas you can do with this.
Betty	I'm sure there is some kind of formula for figuring it out. I really don't know how they get that. This (±5%) tells us how accurate it is going to be. If it is 3%, then I would be more confident that this (sample result) is closer to an accurate report.
Henry	They based that on the confidence level they were looking for. I have a numerical understanding of this, my conceptual understanding is not strong. If you look for a 95% confidence level, then you do a calculation … that has something to do with the standard deviation.
Alice	I'm not sure.

At best, we saw that John and Nicole understood that a margin of error was associated with sample size, and Henry understood that it had to do with confidence level and standard deviation (of the population).

Summary

This section investigated teachers' understanding of margin of error. Analysis of teachers' written answer to the Question 5 of Activity 1-8 and Interview question 2-2 revealed that there was a significant improvement in the number of teachers whose meaning of margin of error includes an image of distribution of sample statistics. That is, more teachers were able to articulate that a margin of error $\pm r$ means that

$$\textit{The interval } p \pm r \textit{ captures } x\% \textit{ of } s_i. \qquad x \in [0,100] \qquad \text{(E1)}$$

or

$$x\% \textit{ of intervals } s_i \pm r \textit{ contain } p. \qquad \text{(E2)}$$

There was also a slight improvement in the number of teachers who stated that

We don't know if the interval $p \pm r$ captures s. *(E3)*

or

We don't know whether the interval $s \pm r$ contains p
(but we do know that x% of intervals $s_i \pm r$ contain p). (E4)

With regard to the idea of confidence level, in answering both Question 5 and I2-2, the teachers used 95% whenever they referred to the percentage of samples that were within the interval $p \pm r$. Teachers' responses to Question 5 revealed that none of the teachers were able to recognize from the table in question that the confidence level for the reported margin of error was 91%. This suggested their lack of understanding of what confidence level is and its relationship with margin of error and sample size. The second question of I2-2 also showed that the teachers did not understand how margin of error was determined.

The discussion on Question 5 revealed additional insights about teachers' understanding of margin of error. In Episode 1 we saw from Nicole, Linda, and Sarah's comments on students' answers that they understood the idea that the interval $s \pm r$ does not necessarily contain p. Episode 3 analyzed Sarah's use of the phrase "95% probability". It revealed that although Sarah understood that the interval $s \pm r$ does not necessarily contain p, when she spoke about "95% probability that the interval $s \pm r$ contains p", she was not thinking of a distribution of sample statistics of which 95% of the intervals constructed from the sample statistics contains p. Rather, she had a subjective conception of probability, and her image of margin of error was fixated on the particular sample statistic, as opposed to a distribution of sample statistics.

Episode 2 and 4 documented Henry's transition from initially not having a coherent image of margin of error to acquiring it near the end of the discussion. This

coherent image of margin of error is what a person who holds a Contractor's perspective would think, that a sample statistics 36% with margin of error ±4% means that *95% of samples of the same size will be captured by p±4%, but we don't know if 36% is one of those 95%*. However, Henry's questions raised in Episode 4: how far off 36% is from the population parameter, revealed that although he was able to think like a contractor, he continued to hold a carpenter's perspective. Discussions around Henry's questions suggested that 1) Nicole had a lack of orientation to "quantifying sampling error" in parameter estimation, i.e., a lack of concern for the accuracy of estimates; 2) Betty, John, and Henry seemed to have an orientation to "getting the right answer". That is, they live in a world of perfect certainty, in which everything is either right or wrong. 3) Even if 2) is not true (they are not just concerned about right or wrong), they exhibited an orientation towards finding out the (additive) difference between sample statistics and population parameter. This suggested that even when a teacher acquires a Contractor's way of thinking, the Carpenter's perspective could still be very ingrained in his or her thinking.

Chapter Summary

This chapter explores teachers' understandings of the concepts of variability and margin of error. Since variability is a property of a distribution of sample statistics, understanding variability is intrinsically linked to a scheme of distribution of sample statistics. Pre-Interviews suggest that the teachers, with the exception of John, were predisposed to think in terms of individual samples and not in terms of collections of samples, and thus distributions of samples were not a construct by which they could form

arguments. When asked to consider what was varying when comparing investments in collections of stocks versus individual stocks, they thought of a single collection of stocks in comparison to individual stocks in it. Only John came to see, after our probing questions, that it was a collection of collections that were less variable than individual stocks. Only John and Linda referred to collections of averages when explaining what "the average will be less variable" meant.

Activity 1-7 Fathom simulation engaged the teachers in computer simulations of repeated sampling, with the aim that they make connections among ideas of distribution of sample statistics, sampling variability, and sample size. Teachers simulated taking three collections of samples of different sizes, and compared the histograms generated from the simulation. They concluded that 1) the mean of distribution of sample statistics is close to the mean of the population from which the samples are drawn, and 2) as the sample size increases, the variability of the distribution decreases.

Activity 1-8 unfolded in a sequence of two interrelated parts that probed teachers' understanding of 1) the relationship between variability, population parameter, and the number of samples, and 2) margin of error. Discussion of the first part revealed that while the teachers made the generalization that variability of a distribution was independent of the number of samples drawn, they did not explicitly state the relationship between variability of a distribution and the population parameter. An interesting result that emerged from this discussion concerns teachers' understanding of the purpose of simulation. Linda and Nicole, who apparently disagreed on the number of populations that are involved in the table, seemed to share one common belief that the purpose of the simulation was to find out about the population parameter. To further probe this belief,

we designed the interview question 2-2 at the end of the first week. We found that that with the exception of John and Alice, all the remaining teachers had identified one purpose of simulation as studying the behavior of distribution of sample statistics.

The second part of Activity 1-8 investigated teachers' understanding 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. Teachers gave a written answer both prior to and after the discussion. Analysis of the teachers' written answers revealed that there was a significant improvement in the number of teachers whose meaning of margin of error includes an image of distribution of sample statistics. With regard to the idea of confidence level, the teachers used 95% whenever they referred to the percentage of samples that are within the interval $p \pm r$. None of the teachers were able to recognize from the table in question that the confidence level for the reported margin of error was 91%. This suggested their lack of understanding of what confidence level is and its relationship with margin of error and sample size.

The discussion on Question 5 revealed more insights about teachers' understanding of margin of error that complements those obtained from their written answers. The most significant finding concerns with teachers' orientation in sampling error. We found that although the teachers were able to root the interpretation of margin of error in a scheme of distribution of sample statistics, they continued to hold a carpenter's perspective. That is, they were concerned with the additive difference between a population parameter and its sample estimate.

CHAPTER IX

CONCLUSIONS

This concluding chapter begins with a broad summary of the teachers seminar, highlighting central findings on teachers' understanding of probability, hypothesis testing, variability, and margin of error. The chapter then elaborates the study's contributions and limitations. Finally, the chapter concludes with a forward-looking stance, pointing to potentially relevant future research.

Summary

The goal of this dissertation research, as I described in Chapter 1, is to explore and characterize teachers' personal and pedagogical understanding of probability and statistical inference, and to develop a theoretical framework for describing teachers' understanding. To this end, I conducted an extensive review on existing literature on the history of probability and statistical inference, and the psychological and instructional studies of people's understanding of probability and statistical inference (Chapter 2). Against the background of this knowledge, I analyzed the teachers seminar that my research team conducted in 2001 with eight high school teachers. The background theories and methodology of this study were summarized in Chapter 3, and the specifics of this teachers seminar were provided in Chapter 4.

I parsed the results of this study into four chapters. Chapter 5 provided a conceptual analysis of probability and statistical inference. In this chapter, I elaborated

the theoretical frameworks that were developed from reviews of literature and my analysis of the teachers seminar. These theoretical frameworks, consisting of theoretical constructs that I later use to make sense of teachers' understanding, provided an understanding of both what constitute coherent and powerful understandings of probability and statistical inference and how these understandings might develop from relatively less sophisticated conceptions. Using these frameworks, I described teachers' understandings of probability and statistical inference emerged from their engagement in the seminar activities and interviews. I parsed this part of the writing into three chapters, each of which focuses on one, or one set of ideas.

Chapter 6: Teachers' understanding of probability

Chapter 7: Teachers' understanding of hypothesis testing

Chapter 8: Teachers' understanding of variability and margin of error

Below I will summarize the central findings from each of these chapters.

Chapter 6 Teachers' understanding of probability

This chapter explored the teachers' conceptualizations and interpretations of probability situations. I structured the activities and interviews around two important ideas: 1) A stochastic conception of probability is one that supports thinking about statistical inference. Thus, We want to know the extent to which the teachers reasoned stochastically, and what difficulties the teachers might have experienced in reasoning stochastically. 2) We wanted the teachers to understand that a situation is what you conceive it to be. That is, a probability situation can be interpreted either non-stochastically or stochastically, depending on how one conceives of the underlying

process behind the stated situation. Moreover, when a situation is conceived of non-stochastically, there are multiple ways to interpret what probability is in that situation. So is stochastic interpretation. The multiplicity of these ways of thinking and interpretations were presented in the theoretical framework for probabilistic reasoning (Figure 3). We believe that as teachers of probability, they must be aware of, and be able to control, different interpretations of probability situations so that they will be equipped to understand students' various conceptions of probability.

With respect to the first question: to what extent the teacher reasoned stochastically, we found that in the beginning of the seminar, most of the teachers had a non-stochastic conception of probability, specifically, only two teachers interpreted one (out of two) probability situations stochastically. In the beginning of the second week when we began to focus on probability, we saw a turning point during the teachers' discussion on the PowerPoint slide 4: *Rain & Temperature*. This slide presented two probability situations, and while one group of teacher consisting of Henry, Nicole, and John argued that the two situations could be interpreted both non-stochastically and stochastically, another group consisting of Terry and Linda insisted that one had to be non-stochastic and the other stochastic. At the end of this debate, there appeared to have a shared understanding among the teachers that a situation is not stochastic (probabilistic) in and of itself, and that it is how one conceives of it that makes it stochastic.

In the post-interview, teachers' interpretations to the five probability situations revealed that most of the teachers only interpreted the situations in one way. Only two teachers, Henry and Lucy, gave both non-stochastic and stochastic interpretations in a number of occasions (Henry 3 out of 4, Lucy 2 out of 5).

Summative analysis of teachers' interpretations across all situations revealed that three teachers, Nicole, Alice, and Henry, had predominantly stochastic conceptions. Sarah had a non-stochastic conception. The remaining teachers' conceptions of probability were situational: Their interpretations of particular probability situations were contingent upon how these situations were stated.

Activity 2-4 Clown and Cards situation was specifically designed to investigate how the teachers would respond to multiple interpretations of a probability situation. In the long and heated discussion, we observed the teachers experience a large amount of confusion and frustration as they attempted to reconcile the differences between two competing interpretations of the Clown and Cards situation. The teachers exhibited a high commitment to their own interpretations, and a low degree of reflection. As such, it was difficult for them to entertain alternative interpretations. When Henry, came to see the reasonableness of the alternative interpretation, he changed his interpretation, as opposed to acknowledge the reasonableness of both interpretations. The teachers seemed to have an underlying assumption that there should be only one correct interpretation for any situation and they believed that computer simulation could decide which interpretation was correct. Despite Pat's persistent attempt to engage the teachers in conversations about pedagogy (thus entailing reflection on their thinking/interpretations as a collection), the teachers resisted and kept arguing over the correctness of their interpretations. In the post-interview on a similar question, however, we saw signs of a significant change in teachers' responses. All teachers acknowledged the multiple interpretations and stated that they would be conformable to accept them. However, half of the teachers also

expressed a commitment to consensus, and stated that they would avoid such situations in their classroom.

Chapter 7 Teachers' understanding of hypothesis testing

This chapter explored the teachers' understanding of hypothesis testing. The idea of hypothesis testing builds on a scheme of intricately interrelated concepts, including null and alternative hypothesis, probability and *p-value*, distribution of sample statistics, decision rule, and Type I error, etc. In this chapter, I parsed the discussion into three sections. Section 1 focused on the concept of unusualness. What we were trying to probe in this section was whether the teachers had a stochastic conception of *p-value*. Section 2 focused on the teachers' understanding of null and alternative hypotheses, and of the logic of hypothesis testing. Section 3 focused on the teachers' stochastic conception, and understanding of decision rule.

In Section 1, discussion around Activity 1-6 Movie theatre scenario revealed that only one teacher, Alice, had a stochastic conception of unusualness, and two teachers, Alice and Henry, conceived of unusualness as a statistical conception. The remaining teachers held a subjective meaning for unusualness. In this discuss we also found that a teacher (Sarah)'s attempt at developing a stochastic conception of unusualness was hindered by her confounding *a sample percent (relative proportion of some item in a sample s)* with the *relative frequency of samples like s over a large number of times*. The follow up interview, however, showed that six out of eight teachers conceived of the situation stochastically.

Section 2 investigated the teachers' understanding of the logic of hypothesis testing. Discussion around Activity 1-3 Pepsi scenario revealed that the teachers did not conceive of the situation as entailing hypothesis testing. They exhibited a high commitment to the assumption about the population parameter. In other words, they did not understand that the assumption was made intentionally so that an alternative hypothesis about the population could be confirmed. As depicted by the framework (Figure 1), the teachers made non-conventional decisions when there was a small *p-value*, i.e. a collected sample was unusual in light of the initial assumption. These decisions include: 1) John, Lucy, and Henry rejected the alternative hypothesis on the basis of a small *p-value*; 2) Henry suggested that the sample was not random, and thus no decision about the initial assumption/null hypothesis would be made based on the *p-value*; 3) Linda refused to reject the initial assumption, which was likely caused by her concern over wrongly reject a true assumption/null hypothesis, and her belief that in order to reject the null hypothesis she would need overwhelming evidence against it and a small *p-value* calculated on the basis of one sample does not constitute overwhelming evidence. This collection of decisions and their underlying reasoning revealed the difficulties the teachers with understanding the logic of hypothesis testing.

Activity 2-3 Rodney King scenario in Section 3 investigated teachers' stochastic conception and understanding of decision rule. Successful completion of the task requires that the teachers conceive of an underlying stochastic process, and come up with a decision rule to determine whether the event in question (the police deployment in Rodney King's scenario) was a random or biased occurrence from this stochastic process. Discussion around Rodney King's scenario revealed the difficulties the teachers

experienced with the idea of decision rule. The teachers did not have an operational understanding of "chance deployment". As such, they did not know how a statistical method would answer the question: Whether the police deployment was a chance deployment. Although they conceived of the deployment stochastically and designed the right simulation, they did not know how to interpret the simulation results. Later discussion on decision rule revealed that Henry and Sarah held a non-stochastic conception of the situation, and that they refused to make a decision about the situation using the simulation results.

Overall, Chapter 7 revealed that the teachers experienced difficulties in understanding almost every concept that is entailed in understanding and employing hypothesis testing. Beyond the complexity of hypothesis testing as a concept, I conjecture that part of teachers' difficulties was due to their lack of understanding of hypothesis testing as a tool, and of the characteristics of the types of questions for which this tool is designed. This conjecture was supported by the evidence revealed in Interview 2-1 where we presented a situation that entails hypothesis testing, and only one teacher, Henry, proposed hypothesis testing as the method of investigation.

Chapter 8 Teachers' understanding of variability and margin of error

In this chapter, we investigated teachers' understanding of variability and margin of error, particularly, the idea of variability as a property of distribution of sample statistics; the relationship among variability, sample size, population parameter, and number of samples; and the ideas of margin of error, confidence interval, and confidence level.

Pre-Interviews suggested that only one teacher, John, had a distributional understanding of variability. The remaining teachers were predisposed to think in terms of individual samples, as opposed to collections of samples. Activity 1-7 (Fathom simulation) engaged the teachers in computer simulations of repeated sampling and in discussions about how the resulting histograms vary as the sample size varies. There seemed to be a shared understanding about the relationship between variability and sample size: As the sample size increases, the variability of the distribution decreases.

Activity 1-8 unfolded in a sequence of two interrelated parts that probed teachers' understanding of 1) the relationship between variability, population parameter, and the number of samples, and 2) margin of error. Discussion of the first part revealed that while the teachers made the generalization that variability of a distribution was independent of the number of samples drawn, they did not explicitly state the relationship between variability of a distribution and the population parameter. Discussion of the second part revealed teachers' various interpretations of margin of error. Comparison of teachers' interpretations of margin of error prior to and after the discussion found that there was a significant improvement in the number of teachers whose meaning of margin of error included an image of distribution of sample statistics. Teachers' interpretations, both in this activity and the follow-up interview, also exhibited their lack of understanding of confidence level and of its relationship with margin of error and sample size. The most significant finding in this section concerns with teachers' orientation in sampling error. We found that although the teachers were able to root the interpretation of margin of error in a scheme of distribution of sample statistics, some of them (Henry, John, and

Betty) continued to hold a carpenter's perspective. That is, they were concerned with the additive difference between a population parameter and a sample's estimate of it.

Overall conclusion

Looking across the chapters, we observe a complicated mix of understandings, both within individual teacher's thinking and among the group of teachers, that are often situationally triggered, which are often incoherent when the teachers try to reflect on them, and which do not support their attempts to develop coherent pedagogical strategies regarding probability and statistical inference. It seems that the teachers were untroubled by the understandings they had developed through doing mathematics because in doing mathematics they could compartmentalize their understandings around patterns of activity in response to different types of performance-requests, e.g., find the probability of x, show the sample space of x, perform this test for x, etc. This study revealed a principle source of disequilibrium for the teachers in this seminar: They were being asked to develop understandings of probability, sample, population, distribution, and statistical inference, that cut across their existing compartments.

Contributions and Implications

The most salient findings of this study were the theoretical frameworks that emerged from the analyses of teachers' understandings of probability and statistical inference. These theoretical frameworks, in comparison to prior relevant research studies, open up the "black box" of probabilistic and statistical reasoning. They advanced our understanding of probabilistic and statistical reasoning in that they explicated what

constitutes a coherent and powerful understanding of probability and statistical inference, and non-conventional ways of understandings that people might have, and how these various levels of understandings relate to, or develop into, a coherent understanding. These frameworks provide a tool for understanding and supporting people's learning of probability and statistical inference by allowing us to model their understandings of these concepts and develop insights about possible instructional interventions to support their learning.

This study also provides a rich description of the kinds of difficulties the teachers experienced in developing coherent and powerful understandings of probability and statistical inference. It also develops, whenever possible, conjectures about what it is that might have hindered the teachers' attempts in doing so. The set of descriptions and conjectures provide an insight into the complexity in understanding probability and statistical inference, and what we should reasonably expect of the understandings of the content knowledge of high school teachers who teach, or are going to teach, statistics. In the general population, we should expect a complicated mix of understandings of probability and statistical inference that are often incoherent and highly compartmentalized, which do not support teachers' attempts to develop coherent pedagogical strategies regarding probability and statistical inference.

These theoretical frameworks and our knowledge of the teachers' understandings, together, provide many insights as to how instructions of probability and statistical inference should be designed in future professional development in order to support teachers' learning of probability and statistical inference. For example, we learned that teachers' understandings of probability and statistical inference were highly

compartmentalized: Their conceptions of probability were not grounded in the conception of distribution, and thus did not support thinking about statistical inference. The implication of this result is that instructions of probability and statistical inference must be designed with the principal purpose as that of helping the teachers develop understanding of probability and statistical inference that cut across their existing compartments. In Chapter 6, we learned that a powerful conception of probability that supports reasoning in statistical inference built heavily on the conception of distribution of outcomes. To develop a stochastic conception, one has to develop a series of ways of thinking that include 1) conceiving of an underlying repeatable process, 2) understanding the conditions and implementations of this process in such a way that it produces a collection of variable outcomes, and 3) imaging a distribution of outcomes that are developed from repeating this process. We have seen that some teachers failed in 1) and ended up with different kinds of incoherent interpretations of probability, or some teachers succeeded in 1) but failed in 2) and also ended up with incoherent interpretations. In Chapter 7 and 8, we also learned the foundation of distribution of sample statistics in understanding the concept of hypothesis testing, margin of error, and confidence interval. This suggests a strategy for instructional design for professional development for probability and statistical inference: Start by engaging teachers in activities that support their building an image of *distribution of outcomes* from a random experiment, and ask probability questions about these distributions. The purpose of these activities is to broach the concept of *probability* and *distribution of outcomes*, and to help teachers developing a [stochastic] conception of probability as long run expectations and probability "regions" as regions of a distribution. Then, engage teachers in actual process

or simulation of repeated sampling, and in discussions of features of the resulting distributions of sample statistics. In doing so, we can help teachers developing a stochastic conception of probability in the context of repeated sampling, and building connections among concepts of *probability*, *distribution of sample statistics*, and *p-value,* the ideas essential to statistical inference. Finally, we move on to topics in statistical inference. This strategy, by exerting a great amount of coerced effort in helping teachers develop the capacity and orientation in thinking of a *distribution of sample statistics*, allows them to develop a stochastic/distributional conception of probability, and incorporating the image of distribution of sample statistics in their thinking of statistical inference.

In Chapter 7 and 8, we learned that a great amount of the teachers' difficulties in understanding hypothesis testing and margin of error were results of the teachers' tacit beliefs or assumptions about statistical inference, e.g., the belief that rejecting a null hypothesis means to prove it wrong, the assumption that the measurement error in parameter estimation could be cast in terms of the additive difference between sample statistic and population parameter, and etc. The implication of these results is that understanding statistical inference and teaching effectively entails a substantial departure from teachers' prior experience and their established beliefs. Below I will illustrate an example of how I would use the theoretical frameworks developed in this study to engage the teachers in confronting and reflecting on their established yet tacit beliefs with respect to hypothesis testing.

In Chapter 7, we learned that the teachers made non-conventional choices when a small *p-value* was found in hypothesis testing scenarios. This was, in part, because they

did not understand the logic of indirect argument, which suggests that the logic of indirect argument (or proof by contradiction) should become an explicit topic of discussion in the instructions of hypothesis testing. When working with teachers on hypothesis testing, we could engage the teachers in two parallel conversations: one on the logic of indirect argument, and another on the logic of hypothesis testing. In proof by contradiction, the logic is, "If p is true, then q is true", we assume the truth of p and $\sim q$, and deduce that $\sim p$ is also true or we derive $\sim r$, where r is a statement already taken to be true, e.g. an axiom or theorem. This contradiction leads us to the tension between two choices: either (I) we insist that $\sim r$ is true, or (II) we conclude that our assumption of $\sim q$ is false. The implication of choice (I) could be non-sensible or catastrophic for the system in which r is true, and it is this implication that would eventually lead one to make choice (II) and conclude that the original statement q must be true when p is true. By the same token, in hypothesis testing, we could design an instructional activity in which the following different choices were given when a small *p-value* was found: 1) rejecting alternative hypothesis, 2) concluding the sample was not random, 3) not rejecting null hypothesis and needing more evidence against null hypothesis, and 4) rejecting null hypothesis (from theoretical framework for the logic of hypothesis testing). By engaging teachers in discussions about the implications of making each choice, we could have the teachers reflect on the tacit beliefs that might lead them to non-conventional choices, and come to appreciate the logic of hypothesis testing.

This study also contributed to our understanding of teacher communication and teacher reflection. As we have observed from the discussion around the Clown and Cards scenario, simply exposing the teachers to alternative interpretations did not elevate the

discussion into a reflection conversation. The teachers were so deeply ingrained in their own ways of thinking that it was very difficult for them to entertain alternative interpretations. In the mean time, they also experienced such a great amount of frustration that they believed this kind of discussion should be avoided in their classrooms. This points to a serious challenge facing teacher educators: In what ways can we facilitate reflective conversation without making the teachers feel overly uncomfortable, and how can we make them see the importance of engaging themselves in reflective abstraction of their own understandings?

We note that the source of teachers' frustration was the incompatibility between our request (of the teachers engaging in conceptual analysis) and the teachers' conceptions of learning and teaching. In this study, we found that the teachers had a conception of learning as "knowing how to solve problems" and teaching as "displaying that expertise of problem solving". We have observed that these conceptions of learning and teaching prevented the teachers from engaging in conceptual and reflective discussion in probabilistic and statistical reasoning.

The implication of this result is that if conceptions of learning and teaching are unaddressed, then in future professional development we would encounter the same difficulty in engaging the teachers in conceptual analysis as we did in this study. In other words, teachers' conception of learning and teaching should become part of an explicit agenda in the design of future professional development. The strategy that I propose is to engage teachers in reflective abstraction *after* they have come to see, through designed instruction, the power of understanding probability, distribution of sample statistics, and statistical inference as a scheme of interconnected ideas.

It will also be extremely important that teachers see that we did not help them develop powerful understandings of probability and statistical inference by "displaying" to them correct ways of solving problems. That is, it is important that teaches create a didactical transposition in which they move from identifying themselves as learners of what is taught to designers of what is taught. They need to understand that we had an elaborated design that takes into account what it was that we wanted the teachers to understand, what they might have understood prior to the instruction, and informed conjectures of what we might do to reach our instructional agenda. I would hope that this reflective conversation will not only change the teachers' conceptions of learning and teaching, but also help them see the principle with which we designed our instruction and be able to utilize this principle in their own instructional design.

Limitations

In hindsight, there are many limitations of this study, and many of these are unavoidable due to the very nature of the study. Due to the deficiency of any systemic attempt at unpacking teachers' probabilistic and statistical understanding in the field of statistics education, this study is highly exploratory. This means two things.

First, we must work with a small group of teachers so that we give a fair amount of opportunity for each teacher to reveal their understandings in the seminar. As a result, this small sample size does not support making claims about the prevalence of this study's central findings to a broader population of teachers.

Second, when we designed and conducted the seminar, we did not have as much understanding, as we do now as depicted in the theoretical frameworks, about what it

means to understand probability and statistical inference coherently and how such understandings develop. As a result, on occasions the activities and interviews were not carried out in its optimal sequence or manner. This is suggested by the non-chronological order with which I organized the description of activities and interviews within and across the chapters. Some data seemed to be weak or inadequate in hindsight. For example, Post-Interview 3-1 provided an opportunity for us to explore the teachers' stochastic conception of probability. At the time when the interview was conducted, the only distinguishing element that the interviewers believed that separated a stochastic conception from a non-stochastic conception was whether one conceived of a repeatable process for a probability situation. Therefore the questioning stopped once that information was collected. As such, the data turned out to be insufficient in probing how well the teachers understood these repeatable processes and whether they had an image of distributions of outcomes generated from the processes, which we learned at the end of the analysis are important benchmarks for stochastic conception.

Next steps

As I mentioned in Chapter 1, this study is an early step of a larger research program, which aims to understand ways of supporting teachers learning and their transformations of teaching practices into ones that are propitious for students' learning in the context of probability and statistics. As a precursor, this study has developed initial frameworks for understanding teachers' (in general, statistics learners') understandings of probability and statistical inference. In the mean time, it also opened many doors to many different directions of future research. The most immediate follow-up work would be to refine

these frameworks, i.e. to test their viability through working with a broader audience and revising them accordingly. One of the results from this work would be to generate insights about the prevalence of people's particular conceptions and understandings. These results could then inform instructional design of probability and statistics.

Although I documented the chronological change of teachers' thinking, in most cases I do not know how these changes occurred. This was partially due to the fact that this study was not designed as a traditional design experiment that takes teacher change as its primary focus. Naturally, one of the follow up studies would be to design a teaching experiment that explicitly focuses on ways of supporting teachers' development of coherent probabilistic and statistical understanding, given what we learned in this study about what it means to for them to have such understanding and what difficulties they might encounter in develop this understanding.

REFERENCES

Albert, J. (1995). "Teaching Inference about Proportions Using Bayes and Discrete Models." Journal of Statistics Education, 3(3).

American Statistical Association (1998). What is margin of error? What is a survey? S. o. S. R. Methods. Alexandria, VA.

Ayer, A. J. (1972). Probability and evidence. New York, Columbia University Press.

Bady, R.-J. (1979). "Students' Understanding of the Logic of Hypothesis Testing." Journal-of-Research-in-Science-Teaching 16(1): 61-65.

Ball, D. and G. W. McDiarmid (1990). The subject matter preparation of teachers. Handbook of research on teacher education. W. R. Houston. New York, Macmillam: 437-449.

Ball, D. L. (1990). "The mathematical understandings that prospective teachers bring to teacher education." Elementary School Journal 90: 449–466.

Ball, D. L. and H. Bass (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. Multiple perspectives on mathematics teaching and learning. J. Boaler. Stamford, CT, Ablex: 83-106.

Bauersfeld, H. (1980). "Hidden dimensions in the so-called reality of a mathematics classroom." Educational Studies in Mathematics 11(1): 23–42.

Bauersfeld, H. (1988). Interaction, construction, and knowledge: Alternative perspectives for mathematics education. Effective mathematics teaching. T. J. Cooney and D. Grouws. Reston, VA, National Council of Teachers of Mathematics.

Bauersfeld, H., G. Krummheuer, et al. (1988). Interactional theory of learning and teaching mathematics and related microethnographical studies. Foundations and methodology of the discipline of mathematics. H. G. Steiner and A. Vermandel.

Bayes, T. (1763). "An Essay towards solving a Problem in the Doctrine of Chances." Philosophical Transactions of the Royal Society of London(53).

Begle, E. G. (1972). Teacher knowledge and pupil achievement in algebra. Palo Alto, CA, Stanford University, School Mathematics Study Group.

Begle, E. G. (1979). Critical variables in mathematics education: Findings from a survey of the empirical literature. Reston, VA, National Council of Teachers of Mathematics.

Behr, M., G. Harel, et al. (1992). Rational number, ratio, and proportion. Handbook for research on mathematics teaching and learning. D. Grouws. New York, Macmillan: 296–333.

Best, J. (2001). Telling the truth about damned lies and statistics. Chronicle of higher education.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning, and the search for causes - DO incoming goods need A probabilistic revolution after incoming goods have taught DATA analysis? ICOTS 4, Marrakech 1994 Minneapolis: University OF Minnesota.

Biggs, J. B. and K. F. Collis (1982). Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome). New York, Academic Press.

Blumer, H. (1969). Symbolic interactionism: Perspectives and method. Englewood Cliffs, NJ, Prentice-Hall.

Borko, H., M. Eisenhart, et al. (1992). "Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily?" Journal for Research in Mathematics Education 23(3): 194–222.

Cobb, P. (2000). Conducting teaching experiments in collaboration with teachers. Research design in mathematics and science education. R. Lesh and A. E. Kelly. Dordrecht, The Netherlands, Kluwer.

Cobb, P., A. Boufi, et al. (1997). "Reflective discourse and collective reflection." Journal for Research in Mathematics Education 28(3): 258-277.

Cobb, P. and L. P. Steffe (1983). "The constructivist researcher as teacher and model builder." Journal for Research in Mathematics Education 14: 83–94.

Cobb, P. and J. Whitenack (1996). "A method for conducting longitudinal analyses of classroom videorecordings and transcripts." Educational Studies in Mathematics 30(3): 213-228.

Collins English Dictionary, t. (2000). The Collins English Dictionary, HarperCollins Publishers.

Cortina, J. L., L. A. Saldanha, et al. (1999). Multiplicative conceptions of arithmetic mean. Twenty First Annual Meeting of the International Group for the Psychology of Mathematics Education - North American Chapter. F. Hitt and M. Santos. Cuernavaca, Mexico, ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, OH. 2: 466-472.

D'Agostini, G. (2003). Bayesian reasoning in data analysis: A critical introduction, World scientific publishing.

David, F. N. (1962). Games, Gods, and Gambling: The Origin and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era. New York, Hafner Publishing Company.

Davis, P. J. and R. Hersh (1981). The mathematical experience. Boston, Houghton Mifflin.

de Finetti, B. (1937). La Prevision: ses Lois logiques, ses sources subjectives. Studies in subjective probability. H. E. Kyburg and H. E. Smokler. New York, Robert Krieger: 53-118.

de Finetti, B. (1974). Theory of Probability: A Critical Introductory Treatment. London, New York, Sydney, Toronto, John Wiley & Sons.

Dewey, J. (1910). How we think. Boston, D. C. Heath & Co.

Dewey, J. (1981). Experience and nature. John Dewey: The Later Works, 1925 - 1953, Vol. 1. J. A. Boydston. Carbondale, Southern Illinois University Press.

Doob, J. L. (1996). "The development of rigor in mathematical probability." American Mathematical Monthly 103(7): 586-595.

Dooren, W. V., L. Verschaffel, et al. (2002). "The impact of preservice teachers' content knowledge on their evaluation of students' strategies for solving arithmetic and algebra word problems." Journal for Research in mathematics education 33(5): 319.

Eisenhart, M., H. Borko, et al. (1993). "Conceptual knowledge falls through the cracks: Complexities of learning to teach mathematics." Journal for Research in Mathematics Education 24(1): 8–40.

Evangelista, F. and C. Hemenway (2002). The use of jigsaw in hypothesis testing. 2nd International conference on the teaching of mathematics at the undergraduate level, Hersonissos, Crete, Greece.

Falk, R. and C. Konold (1999). The Psychology of Learning Probability. Statistics for the twenty-first century. F. S. G. Gordon, S. P., Mathematical Association of America: 151-164.

Fidler, F. and S. Finch (2000). Students' understanding of confidence intervals: descriptive or inferential statistics? OZCOTS-3.

Fine, T. L. (1973). Theories of Probability: An Examination of Foundations. New York, London, Academic Press.

Fischbein, E. (1975). The intuitive sources of probabilistic thinking in children. Dordrecht, The Netherlands, Reidel.

Fischbein, E. and A. Gazit (1984). "Does the teaching of probability improve probabilistic intuitons?" Educational Studies in Mathematics 22: 523-549.

Fischbein, E. and D. Schnarch (1997). "The evolution with age of probabilistic, intuitively based misconceptions." Journal for Research in Mathematics Education 28(1): 96-105.

Fisher, R. A. (1956). Statistical methods and scientific inference. Edinburgh, Oliver and Boyd.

Garfield, J. and A. Ahlgren (1988). "Difficulties in learning basic concepts in probability and statistics: Implications for research." Journal for Research in Mathematics Education 19: 44-63.

Garfield, J. and D. Ben-Zvi (2003). Research on statistical literacy, reasoning, and thinking: issues, challenges, and implications. The challenges of developing statistical literacy, reasoning, and thinking. D. Ben-Zvi and J. Garfield. Netherlands, Kluwer.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is relevant for psychology (and vice versa). Subjective probability. G.Wright and P. Ayton. New York, Wiley.

Gigerenzer, G. (1996). "On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996)." Psychological review 103: 592-596.

Gigerenzer, G. (1998). Ecological Intelligence: An adaptation for frequencies. The evolution of mind. D. D. C. C. Allen. Oxford, Oxford University Press.

Gigerenzer, G., Z. Swijtink, et al. (1989). The empire of chance: How probability changed science and everyday life. Cambridge, Cambridge University Press.

Gillies, D. (2000). Philosophical Theories of Probability. London and New York, Routledge.

Glaser, B. G. and A. L. Strauss (1967). The discovery of grounded theory: Strategies for qualitative research. Chicago, Aldine.

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

Good, I. J. (1965). The Estimation of Probabilities: An Essay on Modern Bayesian Methods. Cambridge, Massachusetts, The M.I.T. Press.

Gould, S. J. (1991). Bully for brontosaurus: reflections in natural history. New York, Norton.

Green, D. (1979). "The Chance and probability concepts project." Teaching Statistics 1(3): 66-71.

Green, D. (1983). "School pupils' probability concepts." Teaching Statistics 5(2): 34-42.

Green, D. (1987). "Probability concepts: putting research into practice." Teaching Statistics 9(1): 8-14.

Green, D. R. (1989). School pupils' understanding of randomness. Studies in mathematics education. R. Morris. Paris, UNESCO. 7: 27-39.

Grossman, P. L., S. M. Wilson, et al. (1989). Teachers of substance: Subject matter knowledge for teaching. Knowledge base for beginning teachers. M. C. Clinton. New York, Pergamon Press: 23–36.

Hacking, I. (1965). The logic of statistical inference. Cambridge, UK, Cambridge University Press.

Hacking, I. (1975). The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference. London; New York, Cambridge University Press.

Hacking, I. (2001). An introduction to probability and inductive logic, Cambridge University press.

Harel, G. and J. Confrey, Eds. (1994). The development of multiplicative reasoning in the learning of mathematics. Albany, NY, SUNY Press.

Hawkins, A. and R. Kapadia (1984). "Children's conceptions of probability: A psychological and pedagogical review." Educational Studies in Mathematics(15): 349-377.

Hendricks, V. F., S. A. Pedersen, et al., Eds. (2001). Probability Theory: Philosophy, Recent History and Relations to Science. Studies in Epistemology. Dordrecht, Boston, London, Kluwer Academic Publisher.

Hertwig, R. and G. Gigerenzer (1999). "The "conjunction fallacy" revisted: How intelligent inferences look like reasoning errors." Journal of Behavioral Decision Making 12: 275-305.

Hong, E. and H. F. J. O'Neil (1992). "Instructional Strategies to Help Learners Build Relevant Mental Models in Inferential Statistics." Journal-of-Educational-Psychology 84(2): 150-59.

Horvath, J. K. and R. Lehrer (1998). A model-based perspective on the development of children's understanding of chance and uncertainty. Reflections on statistics: learning, teaching, and assessment in grades K-12. S. P. Lajoie. Mahwah, New Jersey, London, Lawrence Erlbaum Associates, Publishers.

Howson, C. and P. Urbach (1991). "Bayesian reasoning in science." Nature 350: 371.

Hunting, R. P. (1983). "Emerging methodologies for understanding internal processes governing children's mathematical behaviour." The Australian Journal of Education 27(1): 45–61.

Johnson, D. H. (1999). "The insignificance of statistical significance testing." Journal of Wildlife Management 63(3): 763-772.

Jones, G. A., C. W. Langrall, et al. (1997). "A framework for assessing and nurturing young children's thinking in probability." Educational Studies in Mathematics(32): 101-125.

Kahneman, D., P. Slovic, et al. (1982). Judgment under uncertainty: heuristics and biases. Cambridge; New York, Cambridge University Press.

Kahneman, D. and A. Tversky (1972). "Subjective probabaility: A judgment of representativeness." Cognitive Psychology 3: 430–454.

Kahneman, D. and A. Tversky (1973). "On the psychology of prediction." Psychological Review 80: 237–251.

Kahneman, D. and A. Tversky (1982). Variants of uncertainty. Judgment under uncertainty: Heuristics and biases. D. Kahneman, Slovic, P., & Tversky, A. Cambridge, London, New York, New Rochelle, Melbourne, Sydney, Cambridge University Press: 508-520.

Kantowski, M. (1977). The teaching experiment and Soviet studies of problem solving. Gainesville, FL, University of Florida.

Kieren, T. E. (1992). Rational and fractional numbers as mathematical and personal knowledge: Implications for curriculum and instruction. Analysis of arithmetic for mathematics teaching. G. Leinhardt, R. Putnam and R. A. Hattrup. Hillsdale, NJ, Erlbaum: 323–372.

Kieren, T. E. (1993). The learning of fractions: Maturing in a fraction world. Conference on Learning and Teaching Fractions. Athens, GA.

Kolmogorov, A. N. (1956). Foundations of the Theory of Probability. Bronx, New York, Chelsea Publishing Company.

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

Konold, C. (1991). Understanding Students' Beliefs about Probability. Radical Constructivism in Mathematics Education. E. von Glasersfeld. Dordrecht, Kluwer Academic Publishers: 139-156.

Konold, C. (1991). Understanding students' beliefs about probability. Radical constructivism in education. E. von Glasersfeld. Dordrecht, Kluwer: 139–156.

Konold, C. (1994b). "Teaching probability through modeling real problems." Mathematics Teacher(87): 232-235.

Konold, C. (1995). "Issues in assessing conceptual understanding in probability and statistics." Journal of Statistics Education, 3(1).

Konold, C., A. Pollatsek, et al. (1993). "Inconsistencies in students' reasoning about probability." Journal for Research in Mathematics Education 24(5): 393-414.

Konold, C., A. Pollatsek, et al. (1993a). "Inconsistencies in students' reasoning about probability." Journal for Research in Mathematics Education 24(5): 393-414.

Krüger, L. (1987). The Probabilistic revolution. Cambridge, Mass., MIT Press.

Lesh, R. and A. Kelly (1996). Multi-tiered teaching experiments. American Educational Research Association. New York City.

Lesh, R. and E. Kelly (2000). Multitiered teaching experiments. Handbook of research design in mathematics and science research. A. E. Kelly and R. A. Lesh. Mahwah, NJ, Lawerence Erlbaum Associates: 197-230.

Link, C. W. (2002). An examination of student mistakes in setting up hypothesis testing problems. Louisiana-Mississippi Section of the Mathematical Association of America.

Mack, N. K. (1990). "Learning fractions with understanding: Building on informal knowledge." Journal for Research in Mathematics Education 21(1): 16–32.

MacKay, D. M. (1955). The place of 'meaning' in the theory of information. Information theory. E. C. Cherry. London, Buttersworth: 215–225.

MacKay, D. M. (1964). Linguistic and non-linguistic "understanding" of linguistic tokens. Conference on Computers and Comprehension, Santa Monica, CA, The Rand Corporation.

MacKay, D. M. (1965). Cerebral organization and the conscious control of action. Brain and conscious experience. J. C. Eccles. New York, Springer: 422–445.

Maturana, H. (1978). Biology of language: The epistemology of reality. Psychology and Biology of Language and Thought. G. A. Miller and E. Lenneberg. New York, Academic Press: 27–63.

McDiarmid, G. W., D. Ball, et al. (1989). Why staying ahead one chapter just won't work: subject-specific pedagogy. Knowledge base for the beginning teacher. M. C. Reynolds. New York, Pergamon Press: 193-205.

Mead, G. H. (1910). "Social consciousness and the consciousness of meaning." Psychological Bulletin 7: 397-405.

Metz, K. (1998). Emergent ideas of chance and probability in primary-grade children. Reflections on statistics: Learning, teaching, and assessment in grades K-12. S. P. Lojoie. Mahwah, NJ, Erlbaum: 149-174.

Moore, D. (1990). Uncertainty. On the shoulders of giants: new approaches to numeracy. L. A. Steen. Washington, DC, National Academy Press: 95-137.

Moore, D. S. (1995). The Basic Practice of Statistics. New York, NY, Freeman.

Moshman, D. and P.-A. Thompson (1981). "Hypothesis Testing in Students: Sequences, Stages, and Instructional Strategies." Journal-of-Research-in-Science-Teaching 18(4): 341-52.

Newport, F., L. Saad, et al. (1997). How polls are conducted. Your frequently asked questions answered, Gallup Corporation.

Nilsson, P. (2003). Expeimentation as a tool for discovering mathematical concepts of probability. Proceedings of the Third Conference of the European Society for Research in Mathematics Education, Bellaria, Italy.

Nisbett, R. E., D. H. Krantz, et al. (1983). "The use of statistical heuristics in everyday inductive reasoning." Psychological Review 90: 339–363.

Piaget, J. (1971). Genetic epistemology. New York, W. W. Norton.

Piaget, J. (1977). Psychology and epistemology: Towards a theory of knowledge. New York, Penguin.

Piaget, J. (1977/2000). Recherches sur l'abstraction reflechissante. Paris, Presses Universitaires de France.

Piaget, J. and B. Inhelder (1975). The origin of the idea of chance in children. New York, W. W. Norton.

Piatelli-Palmarini, M. (1994). Inevitable illusions: How mistakes of reason rule our minds. New York, Wiley & Sons, Inc.

Piccinato, L. (1986). de Finetti's logic of uncertainty and its impact on statistical thinking and practice. Bayesian Inference and decision techniques: essays in honor of Bruno de Finetti. P. K. Goel and A. Zellner. North-Holland-Amsterdam, New York, Oxford, Elsevier science publishers B.V.

Pitkethly, A. and R. Hunting (1996). "A review of recent research in the area of initial fraction concepts." Educational Studies in Mathematics 30(1): 5-38.

Popper, K. R. (1959). The logic of scientific discovery. New York, Basic Books.

Public Agenda (2003). Best estimate: A guide to sample size and margin of error.

Ramsey, F. P. and R. B. Braithwaite (1931). The foundations of mathematics and other logical essays. London, New York, K. Paul Trench Trubner & co. ltd.; Harcourt Brace and company.

Reid, C. (1970). Hilbert. New York, Springer-Verlag.

Richards, J. (1991). Mathematical discussions. Radical constructivism in mathematics education. E. von Glasersfeld. The Netherlands, Kluwer: 13–51.

Ritson, I. L. (1998). The development of primary school children's understanding of probability, The Queen's University of Belfast.

Saldanha, L. A. (2003). "Is this sample unusual?" An investigation of students exploring connections between sampling distributions and statistical inference. Teaching and Learning. Nashville, TN, Vanderbilt University.

Saldanha, L. A. and P. W. Thompson (2001). Students' reasoning about sampling distributions and statistical inference. Twenty Third Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education. R. Speiser and C. Maher. Snowbird, UT, ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, OH.

Saldanha, L. A. and P. W. Thompson (2002). "Conceptions of sample and their relationship to statistical inference." Educational Studies in Mathematics 51(3): 257-270.

Savage, L. J. (1954). The foundations of statistics. New York, Wiley.

Schwartz, D. L. and S. R. Goldman (1996). "Why people are not like marbles in an urn: An effect of context on statistical reasoning." Applied Cognitive Psychology 10: S99-S112.

Schwartz, D. L., S. R. Goldman, et al. (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. Reflections on statistics: Learning, teaching, and assessment in grades K-12. S. P. Lojoie. Mahwah, NJ, Erlbaum: 233-274.

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

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. Handbook for research on mathematics teaching and learning. D. Grouws. New York, Macmillan: 465–494.

Shaughnessy, J. M. (1993). "Probability and statistics." Mathematics Teacher 86(3): 244-248.

Shulman, L. (1986). "Those who understand: Knowledge growth in teaching." Educational Researcher 15: 4–14.

Simon, J. L. (1998). The philosophy and practice of resampling statistics, http://www.resample.com/content/teaching/philosophy/index.shtml.

Simon, M. A. (1994). "Learning mathematics and learning to teach: Learning cycles in mathematics teacher education." Educational Studies in Mathematics 26(1): 71–94.

Simon, M. A. (1995). "Reconstructing mathematics pedagogy from a constructivist perspective." Journal for Research in Mathematics Education 26(2): 114–145.

Simon, M. A. (2000). Research on the development of mathematics teachers: The teacher development experiment. Handbook of research design in mathematics and science education. A. E. Kelly and R. A. Lesh. Mahwah, NJ, Lawrence Erlbaum Associates: 335-359.

Skemp, R. (1979). "Goals of learning and qualities of understanding." Mathematics Teaching 88: 44–49.

Sowder, J. T., R. A. Philipp, et al. (1998). Middle-Grade Teachers ' Mathematical Knowledge and Its Relationship to Instruction: A Research Monograph. Albany, State University of New York Press.

Steffe, L. P. (1991). The constructivist teaching experiment: Illustrations and implications. Radical constructivism in mathematics education. E. von Glasersfeld. The Netherlands, Kluwer.

Steffe, L. P. (1991). "Operations that generate quantity." Journal of Learning and Individual Differences 3(1): 61–82.

Steffe, L. P. (1993). Learning an iterative fraction scheme. Conference on Learning and Teaching Fractions. Athens, GA.

Steffe, L. P. and J. Richards (1980). The teaching experiment methodology in a constructivist research program. Fourth International Congress of Mathematics Education, Boston, Birkhauser.

Steffe, L. P. and P. W. Thompson, Eds. (2000). Radical constructivism in action: Building on the pioneering work of Ernst von Glasersfeld. London, Falmer Press.

Steffe, L. P. and P. W. Thompson (2000). Teaching experiment methodology: Underlying principles and essential elements. Handbook of research design in mathematics and science education. R. Lesh and A. Kelly. Hillsdale, NJ, Erlbaum: 267-307.

Streitz, N. A. (1988). Mental models and metaphors: implications for the design of adaptive user-system interfaces. Learning Issues for Intelligent Tutoring Systems. New York, Springer-Verlag New York, Inc.: 164-186.

Thompson, A. G. (1984). "The relationship of teachers' conceptions of mathematics teaching to instructional practice." Educational Studies in Mathematics 15: 105–127.

Thompson, A. G., R. A. Philipp, et al. (1994). Calculational and conceptual orientations in teaching mathematics. 1994 Yearbook of the National Council of Teachers of Mathematics. Reston, VA, National Council of Teachers of Mathematics.

Thompson, A. G. and P. W. Thompson (1996). "Talking about rates conceptually, Part II: Mathematical knowledge for teaching." Journal for Research in Mathematics Education 27(1): 2-24.

Thompson, P. W. (1979). The constructivist teaching experiment in mathematics education research. NCTM Research Presession. Boston, MA.

Thompson, P. W. (1982). A theoretical framework for understanding young children's concepts of whole-number numeration. Department of Mathematics Education, University of Georgia.

Thompson, P. W. (1982). "Were lions to speak, we wouldn't understand." Journal of Mathematical Behavior 3(2): 147–165.

Thompson, P. W. (1994). The development of the concept of speed and its relationship to concepts of rate. The development of multiplicative reasoning in the learning of mathematics. G. Harel and J. Confrey. Albany, NY, SUNY Press: 179–234.

Thompson, P. W. (2000). Radical constructivism: Reflections and directions. Radical constructivism in action: Building on the pioneering work of Ernst von Glasersfeld. L. P. Steffe and P. W. Thompson. London, Falmer Press: 412-448.

Thompson, P. W. (2002). Didactic objects and didactic models in radical constructivism. Symbolizing, modeling, and tool use in mathematics education. K. Gravemeijer, R. Lehrer, B. V. Oers and L. Verschaffel: 191-212.

Thompson, P. W. and Y. Liu (2002). What is the probability that my car is red: tensions in the development of probabilistic reasoning. Annual conference of American Educational Research Association, New Orleans.

Thompson, P. W. and L. A. Saldanha (2000). Conceptual issues in understanding sampling distributions. Twenty Second Annual Meeting of the International Group for the Psychology of Mathematics Education - North American Chapter. M. L. Fernandez. Tucson, AZ, ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, OH. 1: 332-332.

Thompson, P. W. and L. A. Saldanha (2000). Epistemological analyses of mathematical ideas: A research methodology. Twenty Second Annual Meeting of the International Group for the Psychology of Mathematics Education - North American Chapter, Tucson, AZ, ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, OH.

Thompson, P. W. and L. A. Saldanha (2002). Fractions and Multiplicative Reasoning. Research companion to the principles and standards for school mathematics. J. Kilpatrick, G. Martin and D. Schifter.

Thompson, P. W. and A. G. Thompson (1992). Images of rate. Annual Meeting of the American Educational Research Association. San Francisco, CA.

Thompson, P. W. and A. G. Thompson (1994). "Talking about rates conceptually, Part I: A teacher's struggle." Journal for Research in Mathematics Education 25(3): 279–303.

Tiles, M. (1991). Mathematics and the image of reason. London; New York, Routledge.

Todhunter, I. (1949). A History of Mathematical Thoery of Probability: From the Time of Pascal to that of Laplace. New York, Chelsea Publishing Company.

Truran, J. M. (2001). The teaching and learning of probability, with special reference to south Australian schools from 1959-1994. Graduate school of Education, Department of Pure Mathematics, University of Adelaide.

Tversky, A. and D. Kahneman (1973). "Availability: A heuristic for judging frequency and probability." Cognitive psychology 5: 207-232.

Tversky, A. and D. Kahneman (1983). "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement." Psychological review 90(October): 293-315.

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

Von Mises, R. (1951). Positivism; a study in human understanding. Cambridge, Harvard University Press.

Von Mises, R. (1957). Probability, statistics and truth. London, New York, Allen and Unwin; Macmillan.

Von Plato, J. (1994). Creating modern probability: its mathematics, physics, and philosophy in historical perspective. Cambridge [England]; New York, Cambridge University Press.

von Glasersfeld, E. (1992). "An Exposition of constructivism: why some like it radical."

Yates, D., D. Moore, et al. (1998). The Practice of statistics: TI-83 graphing calculator enhanced. New York, W. H. Freeman and Company.