Liu, Y., & Thompson, P. W. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies, 4*(2), 126-138.

MATHEMATICS TEACHERS' UNDERSTANDINGS OF PROTO-HYPOTHESIS TESTING

Yan Liu
Strategic Education Research Partnership Institute

Patrick W. Thompson
Arizona State University

Abstract

Hypothesis testing is one of the key concepts in statistics, yet it is also one of the least understood concepts. The purpose of this study was to investigate teachers' understandings of early forms of hypothesis testing in an effort to generate insights into ways of supporting their design of effective strategies for teaching hypothesis testing. To this end, we conducted a professional development seminar and interviews with 8 high school statistics teachers in 2001 in the southeast US in which we attempted to unpack the difficulties and conceptual obstacles teachers encountered as they thought about methods by which they could test the validity of a claim made about a population that is based on a single sample. We found that teachers' difficulties in such situations were rooted in their non-stochastic conceptions of probability and in their lack of understanding of the logic of indirect argument. We conclude by offering promising pedagogical approaches for developing teachers' understandings of data-based inference.

Hypothesis testing, or quantifying the viability of claims about a population that are based on information gained from a sample from that population, is one of the key concepts in every introductory statistics course. It is also one of the most difficult topics students encounter in their learning of statistics (Albert, 1995; Bady, 1979; Evangelista & Hemenway, 2002; Link, 2002; Moshman & Thompson, 1981). Textbook authors typically present the logic of hypothesis testing as a multi-step procedure that includes stating the null and alternative hypotheses, defining a critical value, calculating a value of the test statistic, finding a *p-value*, deciding about the null hypothesis, and interpret the situation (Yates, Moore, & McCabe, 1998). What often gets lost in discussions of hypothesis testing is the fact that it is a method for testing the viability of data-based claims about characteristics of a population. That is, the larger issue is whether students and teachers see data as a starting point for making quantitative arguments for or against claims someone might make about a population that the data represents.

The ideas of probability and unusualness are central to the logic of hypothesis testing. In hypothesis testing, one rejects a null hypothesis (a claim about a population that is counter to what you suspect is true) when a sample from this population is judged to be sufficiently unusual (improbable, rare) in light of the null hypothesis. This is consistent with a Fisherian approach to hypothesis testing (Fisher, 1956).

This logic demands that we assume the values of the sample statistic of interest are generated by a random sampling process and that the statistic's values have some underlying distribution. Without assuming an underlying distribution for a statistic's values we have no way to gauge any sample's rarity. To declare that a sample suggests we should deny the null hypothesis is like a policy decision: When samples like the one

observed[1] are sufficiently rare according to our assumptions (i.e., should occur less than *x*% of the time in light of assumptions) then we declare that at least one of our assumptions is untenable—either our sampling procedure was not random or values of the sample statistic are not distributed as the null hypothesis states.

A sample is *rare* or *unusual* if, over the long run, we expect the sample and those more extreme to occur a small fraction of the time. Developing this stochastic conception of unusualness is non-trivial (Liu & Thompson, 2007). We observed in high school teaching experiments (Saldanha, 2004; Saldanha & Thompson, 2002) that students had a robust meaning of "unusualness". They meant that an observed sample outcome is surprising, where "surprising" means "differing substantially from what one anticipates". By this meaning, if one has no expectation about what the outcome should be, then no outcome can be unusual. We observed that students rarely made theoretical assumptions about the distribution of outcomes, i.e., about how the outcomes of the process "take a sample from [this population], compute [this statistic]" are distributed, and as such their attempt at applying the logic of hypothesis testing often became a meaningless exercise.

In this article we report an investigation of teachers' understandings of issues surrounding the logic of testing claims about a population that are based on evidence from a sample. Our interests in teachers' understandings stems from our belief that they have a profound influence on teachers' capacity to teach mathematics effectively (Ball & Bass, 2000; Sowder, Philipp, Armstrong, & Schapelle, 1998), and, in turn, on what students end up learning and how well they learn it (Begle, 1972, 1979). We therefore believe that supporting the transformation of teaching practices takes careful analysis of

[1] We use the phrase "like the one observed" to mean that the person making the judgment declares the grounds for similarity. Often, grounds for similarity are "a sample whose statistic's value is at least this extreme."

teachers' understandings of what they teach. Such efforts increase the likelihood that what teachers teach and how they teach have the potential of supporting students to develop coherent and deep mathematical understandings.

Against this background, we conducted a professional development seminar in the summer of 2001 with eight high school statistics teachers. Our overall goal was to unpack the complexities of understanding ideas in probability and statistical inference, especially with regard to the difficulties they experience when they try to understand them more coherently. We believe that the answers to these questions will provide insight into ways of supporting teachers' learning and ways of enhancing teachers' capacity to design instructional activities that will support students' development of coherent understandings of probability and statistical inference. In this article we focus on what teachers understood of questions about how one can think of a sample as providing evidence that supports or disconfirms a claim about a population that the sample represents. We will use the phrase "hypothesis testing" in place of this long description, even though we run the risk of being misunderstood. By "hypothesis testing", we mean more than what a statistics book might include under its umbrella. Instead, we mean ways of thinking about how to quantify the viability of data-based claims.

To elaborate on and address these research questions, we will first explain what we mean by "understanding", and the method we use in developing descriptions of an understanding. Next we will provide an overview of the research design and implementation, followed by the research results in which we highlight the conceptual obstacles teachers experience in knowing hypothesis testing. We will end this article by

offering some suggestions of promising pedagogical approaches that would aid in the development of a deep and coherent understanding of hypothesis testing.

By "understanding" we 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" (Thompson & Saldanha, 2003, p. 99). Building on earlier definitions of understanding based on Piaget's notion of assimilation—for example, Skemp's (1979) definition of understanding as assimilation to an appropriate scheme[2], with special emphasis on appropriate-- Thompson & Saldanha (*ibid*.) extended its meaning to "assimilation to a scheme", which allowed for addressing understandings people do have even though they could be judged as inappropriate or wrong. As a result, a description of understanding requires "addressing two sides of the assimilation—what we see as the thing a person is attempting to understand and the scheme of operations that constitutes the person's actual understanding" (*ibid*., p. 99).

METHOD

In conducting this study, we used a modified constructivist teaching experiment (Cobb & Steffe, 1983; Steffe, 1991; Steffe & Thompson, 2000). The constructivist teaching experiment methodology was adapted from the Soviet-style teaching experiment

[2] In the original texts, Skemp used the word "schema". For a period of time many people used the words "schema" and "scheme" interchangeably. However, now there is a consensus that by "schema" Piaget meant something much smaller than what he meant by "scheme". "Schema" refers to a fairly local organization of action, often meaning stimulus-response, reflex types of organization, but also including irreversible actions. "Scheme" is much broader. Schemes involve mental operations; schemata don't. We interpreted Skemp's meaning as being consistent with the modern use of "scheme".

(Kantowski, 1977) to serve the purpose of developing conceptual models of students' mathematical knowledge in the context of mathematics instruction.

We concur with Steffe (1991) that it is necessary 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, we 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 to build models of subjects' mathematical realities. To create models of subjects' mathematical realities, we must attempt to perturb them so as to reveal both their composition and boundaries.

To construct a description of a person's understanding, we adopted an analytical method that Glasersfeld (1995) called conceptual analysis, 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 "intentionally isomorphic" (Maturana, 1978, p. 29) to that of the person's. In conducting a conceptual analysis, a researcher builds models of a person' understanding by observing the person' actions in natural or designed contexts and asking herself, "What can this person be thinking so that his actions make sense from his perspective?" (Thompson, 1982, pp. 160-161) In other words, "the researcher puts himself into the position of the observed and attempts to examine the operations that he (the researcher) would need or the constraints he would have to operate under in order to (logically) behave as the observed did" (Thompson, 1982, p. 161).

DESIGN AND IMPLEMENTATION

With the purposes of constructing models of teachers' understandings of probability and statistical inference, we designed and conducted a professional development seminar for high school teachers in a metropolitan city in the Southeast part of United States. The seminar was advertised as "an opportunity to learn about issues involved in teaching and learning probability and statistics with understanding and about what constitutes a profound understanding of probability and statistics." From 12 applicants we selected eight who met our criteria—having taken coursework in statistics and probability and currently teaching, having taught, or preparing to teach high school statistics either as a stand alone course or as a unit within another course. Participating teachers received a stipend equivalent to one-half month salary. Table 1 presents the 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.

Table 1. *Demographic information on seminar participants*

Teacher	Years Teaching	Degree	Statistics Background	Teaching experience
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

We prepared for the seminar by meeting weekly for eight months to devise a set of issues that would be addressed in it, selecting video segments and student work from

prior teaching experiments on students' stochastic reasoning to use in seminar discussions, and preparing teacher activities. The seminar lasted two weeks in June 2001, with the last day of each week devoted to individual interviews. Each session began at 9:00a and ended at 3:00p, with 60 minutes for lunch. All sessions were led by a high school statistics teacher, Terry, who had collaborated in the seminar design throughout the planning period. We interviewed each teacher three times: prior to the seminar about his or her understandings of sampling, variability, and the law of large numbers; at the end of the first week on statistical inference; and at the end of the second week on probability and stochastic reasoning. The data for analysis included video recordings of all seminar sessions made with two cameras, videotapes of individual interviews, teachers' written work, and documents made during the planning of the seminar.

The analytical approach we employed in generating descriptions and explanations was 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 were aimed at developing increasingly stable and viable hypotheses and models of teachers' understanding.

RESULTS

Teachers' conceptions of probability and *unusualness*

In the seminar we devoted approximately a week's time to probability. We found that majority of the teachers were not inclined to conceptualize a stochastic process in situations that entail an interpretation on unusualness. For example, on day 3 of the

seminar we engaged the teachers in discussion of the following question, adapted from Konold (1994, pp. 16-18).

> 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?

Ways of thinking about this question that indicates a stochastic conception of unusualness would be:

1. Assuming that people go to the theatre randomly;
2. Thinking of a collection of nights, when random groups of 250 people from the population of 30000 go to the theatre;
3. Recording the number of people known to Ephram attending each night;
4. Plotting a distribution of these numbers, and calculate the density of "at least 2", the chance that at least two people Ephram knows attend the movie he shows;
5. If the proportion is smaller than 5% (a conventional significance level), then concluding that it would be unusual that at least two people known to Ephram attend the movie (assuming that attendance is a random phenomenon).

We wish to highlight from this passage that a stochastic conception of unusualness builds on conceptions of sampling and distribution of values of a sample statistic.

The teachers first gave intuitive answers. All said it would not be unusual for Ephram to see two people he knows. Subsequent discussion focused on the method for investigating the question, and it revealed that only one teacher, Alice, had a conception

of unusualness that was grounded in an understanding of distribution of values of a sample statistic. She proposed, as a method of investigating the question, that "each night [Ephram should] record how many he knew out of the 250 and keep track of it over a long period of time", which suggested that she had conceived of "Ephram sees *x* people he knows" as a random event and would evaluate the likelihood of the outcome "Ephram sees at least two people he knows" against the distribution of a large number of possible outcomes.

Other teachers had various conceptions of unusualness. Three teachers, Sarah, Linda, and Betty stated flatly that something is unusual if it is unexpected, and expectations are made on the basis of personal experience. John's conception of unusualness was also subjective and non-stochastic. He justified his intuitive answer by reasoning that Ephram knows 300 people out of 30,000 people in his town, so 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 representative sample of 30,000 in his town. John employed what we call the *proportionality heuristic*: evaluating the likelihood of a particular value of a sample statistic by comparing it against the population proportion. He did not conceptualize a scheme of repeated sampling that would allow him to quantify unusualness. Henry's conception of unusualness was somewhat stochastic, albeit nonstandard: "Something is unusual if I am doing it less than 50% of the time." Our extensive analysis of teachers' interpretations of probability throughout the entire seminar (Liu & Thompson, 2007) revealed that it is not uncommon for teachers to have non-stochastic conceptions of probability, which pointed to a real challenge in coming to understand inference and hypothesis testing.

Teachers' commitment to null hypothesis

In light of the teachers' difficulties in conceiving probability stochastically, we designed an activity in which we oriented the teachers to think of a distribution of values of a sample statistic when judging the likelihood of a particular sample. Teachers were given this scenario:

> A pollster asked 100 people which they like better, Pepsi or Coca Cola. 55 said "Pepsi". How likely is this result?

and they were then asked, "What is the meaning of 'how likely is this result'?"

The conversation around this question centered on the idea that in order to investigate the meaning of "how likely is it that 55 out of 100 people prefers Pepsi", we must assume some portion of the population actually prefers Pepsi to Coca Cola. If we assume that the population is evenly split between Pepsi and Coke, then 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'?"

Following this discussion, we presented the teachers with results from a computer simulation made to generate 135 randomly generated collections of 100 zeroes and ones (see Appendix). We called "0" a head (Coke) and "1" a tail (Pepsi). We then gave the teachers the following task:

> 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 purpose of this task was to understand different ways the teachers approach investigating the viability of a claim about a population based on one sample.

At the beginning of the discussion, three teachers, Lucy, John, and Henry argued that the statement *there were more people in the sampled population who prefer Pepsi than prefer Coca Cola* was false. They based their claim on the evidence that only 2.96% of the simulated samples had 60% or more favoring Pepsi (i.e. only 4 out of 135 samples). Apparently the teachers had seen the simulation as providing evidence that the population was *not unevenly* split. Their logic seemed to have been that 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.

Terry, the seminar leader, pushed the teachers to explain the tension between 1) we actually got a sample of whom 60% preferred Pepsi, and 2) the sample's occurrence was rare under the assumption that the population was evenly split. Henry suggested that the sample might not have been randomly chosen, which could have explained the tension except that the task stated otherwise, i.e., "the sampling procedure was acceptable".

John suggested the tension should lead one to conclude that the assumption that the population was evenly split was not valid. Linda, however, insisted that the assumption about the population should not be rejected on the basis of one sample. She

argued that no matter how rare a sample was, it *could still* occur. Thus, its occurrence *could not* be used against any assumption. Her opposition to rejecting the assumption about the population revealed that she was concerned about whether the assumption was proven false. She stated that she would reject the assumption only if there was overwhelming evidence against it and therefore proposed to take more samples to see whether the assumption was right or wrong. Her concern for establishing the truth or falsity of a working assumption (the null hypothesis) is inconsistent with the idea of a decision rule. A decision rule does not tell us whether the null hypothesis is right or wrong. Rather, it tells us that if we apply the decision rule consistently, then we can anticipate, over the long run, rejecting the null hypothesis inappropriately only a small percent of the time.

In summary, the discussion around this activity revealed two ways of thinking that the teachers had had that led to their failure to employ the logic of hypothesis testing. The first way of thinking, demonstrated in Lucy, Henry, and John's initial argument, in essence, supported an assumption about a population on the basis of the distribution of values of a sample statistic in light of that assumption, thus tossing away of the observed sample and rejecting the claim about the population suggested by this sample. The second way of thinking was revealed in Linda's reluctance to reject an assumption about the population given that the observed sample was rare in light of this assumption. Both ways of thinking bear a hidden commitment to the working assumption about the population. This is incompatible with the logic of hypothesis testing. In hypothesis testing, we would reject the null hypothesis (working assumption about the population) whenever the observed sample is judged to be unusual in light of this assumption.

Challenges in conceiving hypothesis testing as a tool

In addition to the conceptual challenges discussed above, teachers' failure in employing hypothesis testing was in part a result of their challenges of conceiving hypothesis testing as a tool for making statistical inference, i.e., knowing the kinds of problems that the method of hypothesis testing was created for. In an individual interview conducted at the end of the seminar discussion on hypothesis testing, we asked the teachers this question:

> 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 60% 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 question presents a typical hypothesis-testing scenario: a stated claim about a population parameter; a random sample of 20 graduates from the actual population; and an implied question ("Are samples like or more extreme than 60% sufficiently rare, assuming the administration's claim, to reject that claim?"). When asked how they would test the administration's claim, only one teacher, Henry, proposed to use hypothesis testing. The methods other teachers proposed fall into the following categories:

1. Take many more samples of size 20 from the population of graduates (John, Nicole, Sarah, Alice)
2. Take a larger sample from the population of graduates (Alice)
3. Take one or a few more samples of size 20 from the population of graduates (Lucy, Betty)
4. Survey the entire population (Linda)

All of these methods presumed that the teachers would have access to the population, yet none of them offered well-defined policies that would allow one to make consistent judgments. This led to our conjecture that the teachers had not internalized the functionality of hypothesis testing. In other words, they did not know the types (or models) of questions that hypothesis testing was created for, and how hypothesis testing can be a particularly useful tool for answering these types of questions.

CONCLUSIONS AND IMPLICATIONS

Understanding hypothesis testing builds on a scheme of interrelated concepts including probability (or unusualness), random sampling, distribution of values of a sample statistic, significance level, and the logic of hypothesis testing. In this article we attempt to unpack the meanings of these concepts and their connections, and discuss the difficulties and conceptual obstacles that teachers encountered as they attempted to conduct, or make sense of, hypothesis testing. In doing so we hope to contribute to understandings of ways of supporting students' learning of hypothesis testing as a tool for making statistical inference.

As we have seen, part of teachers' difficulty in understanding and employing statistical inference came from their compartmentalized knowledge of probability and of statistical inference. That is, their conceptions of probability (or unusualness) were not grounded in a conception of distribution, and thus did not support thinking about distributions of sample statistics and the fraction of the time that a statistic's value is in a particular range. The implication of this result is that instructions on probability and on

statistical inference must be designed with the principal purpose that it helps one understand probability statistically and to understand statistics probabilistically. This purpose might be achieved by designing instruction so that teachers develop the capacity and orientation to think in terms of *distributions of sample statistics*, which hopefully would have the salutary effect of supporting a stochastic, distributional conception of probability, and lead to their inclusion of distributions of sample statistics in their understanding of statistical inference.[3] We suspect that teachers who value distributional reasoning in probability and who imagine a statistic as having a distribution of values will be better positioned to help students reason probabilistically about statistical claims.

We also learned that part of the teachers' difficulties in understanding hypothesis testing was a result of their hidden commitment to the null hypothesis, and the belief that rejecting a null hypothesis means to prove it wrong. The implication of this result is that understanding hypothesis testing entails a substantial departure from teachers' prior experience in, and their established beliefs about, inference and reasoning about data. To confront these often hidden yet unhelpful beliefs, we should design tasks and activities that engage teachers in explicit discussions about them, given what we now know from this study. Emphasis should be given to the logic of hypothesis testing, particularly, the clarification on the central ideas:

- The work done by a null hypothesis,
- Sampling as a stochastic process,
- Distributions of sample statistics, and

[3] We disagree with one reviewer that introductory statistics textbooks already emphasize distributions of sample statistics. They emphasize sampling distributions (i.e., the distribution of all values of a sampling statistic), but not the more general idea of distributions of sample statistics. A sampling distribution is just one particular distribution of sample statistics, just as a square is just one particular rectangle.

- Policies that warrant a rejection of a null hypothesis.

We also observed that the teachers did not readily recognize hypothesis testing as a tool for making statistical inferences. This led us to the conjecture that they did not understand the function of hypothesis testing. The implication of this result is that a major component of teaching hypothesis testing entails providing sufficient experience with situations for which hypothesis testing is a tool and for which the logic of hypothesis formation and hypothesis testing is a method. These logics would emerge from learners' dealing repeatedly with situations that lend themselves to investigation through hypothesis formation and testing and explaining to themselves and others why their methods are sensible. These conceptual explanations should be given even more emphasis than the procedures of actually doing hypothesis testing.

Finally, we remind readers that the teachers *already teach statistics in high school!* They have taken courses in statistics from statisticians! We believe it would behoove collegiate statistics programs to examine closely what their students are, in fact, learning about the ideas behind statistical testing.

References:

Albert, J. (1995). Teaching inference about proportions using Bayes and discrete models. *Journal of Statistics Education, 3*(3). Retrieved January 7, 2003, at http://www.amstat.org/publications/jse/v3n3/albert.html

Bady, R. J. (1979). Students' understanding of the logic of hypothesis testing. *Journal of Research in Science Teaching, 16*(1), 61-65.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning* (pp. 83-106). Stamford, CT: Ablex.

Begle, E. G. (1972). *Teacher knowledge and pupil achievement in algebra* (NLSMA Technical Report No. 9). 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.

Cobb, P., & Steffe, L. P. (1983). The constructivist researcher as teacher and model builder. *Journal for Research in Mathematics Education, 14(2)*, 83–94.

Cobb, P., & Whitenack, J. (1996). A method for conducting longitudinal analyses of classroom videorecordings and transcripts. *Educational Studies in Mathematics, 30*(3), 213-228.

Evangelista, F., & Hemenway, C. (2002, July). The use of jigsaw in hypothesis testing. Paper presented at the 2nd International conference on the teaching of mathematics at the undergraduate level, Hersonissos, Crete, Greece.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh, Oliver and Boyd.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New Brunswick, NJ: Aldine.

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

Kantowski, M. (1977). *The teaching experiment and Soviet studies of problem solving*. Unpublished manuscript, Gainesville, FL.

Konold, C. (1994). *Prob Sim User Guide*. Santa Barbara, CA: Intellimation.

Link, C. W. (2002, March). An examination of student mistakes in setting up hypothesis testing problems. Paper presented at the Louisiana-Mississippi Section of the Mathematical Association of America.

Liu, Y., & Thompson, P. W. (2007). Teachers' understandings of probability. *Cognition and Instruction 25(2), 113-160*.

Maturana, H. (1978). Biology of language: The epistemology of reality. In G. A. Miller & E. Lenneberg (Eds.), *Psychology and Biology of Language and Thought* (pp. 27–63). New York: Academic Press.

Moshman, D., & Thompson, P. A. (1981). Hypothesis Testing in Students: Sequences, Stages, and Instructional Strategies. *Journal of Research in Science Teaching, 18*(4), 341-352.

Piaget, J. (1971). *Genetic epistemology*. New York: W. W. Norton.

Saldanha, L. A. (2004). *"Is this sample unusual?": An investigation of students exploring connections between sampling distributions and statistical inference*. Unpublished doctoral dissertation, Vanderbilt University, Nashville, TN.

Saldanha, L. A., & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257-270.

Skemp, R. (1979). Goals of learning and qualities of understanding. *Mathematics Teaching, 88*, 44–49.

Sowder, J. T., Philipp, R. A., Armstrong, B. E., & Schapelle, B. P. (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. In E. v. Glasersfeld (Ed.), *Radical Constructivism in Mathematics education*. Dordrecht, Netherlands: Kluwer.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In R. Lesh & A. E. Kelly (Eds.), *Research Design in Mathematics and Science Education* (pp. 267-307). Mahwah, NJ: Erlbaum.

Thompson, P. W. (1982). Were lions to speak, we wouldn't understand. *Journal of Mathematical Behavior, 3*(2), 147–165.

Thompson, P. W., & Saldanha, L. A. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick, G. Martin & D. Schifter (Eds.), *Research Companion to the Principles and Standards for School Mathematics*.(pp. 95-114). Reston, VA: National Council of Teachers of Mathematics.

Yates, D., Moore, D., & McCabe, G. (1998). *The Practice of statistics: TI-83 graphing calculator enhanced*. New York: W. H. Freeman and Company.

APPENDIX

List of simulated results from repeatedly drawing a random sample of size 100 from a large population that is evenly split between *Heads* and *Tails*.

Sample	Heads	Sample	Heads	Sample	Heads	Sample	Heads
1	52	35	46	69	49	103	47
2	46	36	59	70	59	104	42
3	37	37	42	71	51	105	49
4	54	38	51	72	58	106	40
5	54	39	51	73	49	107	53
6	46	40	45	74	56	108	44
7	49	41	47	75	57	109	47
8	41	42	55	76	46	110	52
9	62	43	57	77	54	111	49
10	60	44	52	78	44	112	46
11	50	45	50	79	45	113	54
12	51	46	44	80	57	114	52
13	52	47	48	81	53	115	60
14	49	48	49	82	44	116	53
15	45	49	49	83	59	117	45
16	55	50	56	84	60	118	48
17	56	51	53	85	45	119	49
18	52	52	49	86	50	120	50
19	42	53	49	87	38	121	52
20	44	54	50	88	46	122	55
21	46	55	52	89	52	123	42
22	38	56	56	90	44	124	45
23	47	57	53	91	48	125	60
24	49	58	53	92	52	126	59
25	50	59	47	93	51	127	50
26	44	60	50	94	57	128	60
27	50	61	45	95	53	129	43
28	58	62	50	96	57	130	57
29	49	63	47	97	57	131	49
30	50	64	47	98	55	132	53
31	54	65	54	99	46	133	53
32	55	66	54	100	56	134	50
33	48	67	46	101	42	135	46
34	45	68	57	102	51		