Liu, Y., & Thompson, P. W. (2005). Teachers' understanding of hypothesis testing. In S. Wilson (Ed.), *Proceedings of the Twenty-seventh Annual Meeting of the International Group for the Psychology of Mathematics Education*, Roanoke, VA. Vicksburg, VA: Virginia Tech.

TEACHERS' UNDERSTANDINGS OF HYPOTHESIS TESTING

Yan Liu
Vanderbilt University
Yan.liu@vanderbilt.edu

Pat Thompson
Arizona State University
Pat.Thompson@asu.edu

The goal of this study was to explore eight high school mathematics teachers' understandings of hypothesis testing as they engaged in a two-week professional development seminar. To this end, we analyzed data collected from the videotaped seminar discussions and the follow-up interviews. We found that teachers' difficulties with hypothesis testing could be explained by a conflation of two sources: their non-stochastic conceptions of probability and their unconventional logics of hypothesis testing. Follow-up interviews suggested that the teachers did not understand the work that hypothesis testing is meant to do. Following these results, we proposed a number of strategies for future professional development.

Research Topic

Past research has found that people have profound difficulties with both understanding and employing the method of hypothesis testing. Albert (1995) and Link (2002) found that students have difficulty recognize the population parameter to be tested in inference scenarios. Albert (1995) also found that the idea of sampling distribution, fundamental to hypothesis testing, was too hard for students to learn. Bady (1979) found that people have a strong tendency to test hypothesis by seeking information that would verify the hypothesis instead of falsifying it.

Examination of popular statistics textbooks suggested that hypothesis testing is typically taught as a multi-step procedure (eg. Yates *et al.*, 1998). Doing hypothesis testing, as shown from the following excerpt, seemed like executing a sequence of actions which does not require any reasoning on the part of students:

> "The first part in this procedure is the statement of the null and alternative hypothesis. Students look for key words and phrases such as "less than", "decreased", "reduces", "greater than", "increased", "improved", and "is different from", as guides in stating the null and alternative hypothesis. In the second part, the critical value of the test statistic necessary to reject the null hypothesis is asked for, which requires that the student recognize the appropriate test statistic, locate the correct tabled value based on the stated level of significance, and supply the correct sign…Finally, the p-value is either read from a table, or is displayed on a graphing calculator screen." (Link, 2002)

We argue that this way of conceptualizing and teaching hypothesis testing might have contributed to students' confusion about hypothesis testing. In this paper, we will explore the difficulties people have as they try to understand hypothesis testing conceptually. To this end, we examined data collected from a professional development seminar, conducted with a group of eight high school teachers, that was designed to investigate their understanding of probability and statistical inference (Liu & Thompson, 2004).

Theoretical Framework & Methodology

Our study was guided by a radical constructivist perspective on human knowledge and human learning. Radical constructivism entails the stance that any cognizing organism builds its own reality out of the items that register against its experiential interface (Glasersfeld, 1995). As such,

in our study that aimed to understand others' mathematical understanding, it is necessary to attribute mathematical realities to subjects that are independent of the researchers' mathematical realities. This is what Steffe meant when he described the researcher' activity in a constructivist teaching experiment, as that of performing the act of de-centering by trying to understand the *mathematics of the [other]* (Steffe, 1991).

To construct models of others'/teachers' understanding, we 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 intentionally isomorphic to that of the person. In conducting conceptual analysis, a researcher builds models of a person's 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?" 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).

Research Design & Data analysis

We designed a two-week summer seminar for high school teachers. 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." Of 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.

The research team 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 to use in seminar discussions, and preparing teacher activities.

Table 1 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.

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 seminar lasted two weeks in June 2001. Each session began at 9:00a and ended at 3:00p, with 60 minutes for lunch. All seminar sessions were led by a high school AP 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. This paper will focus on week 1, in which issues of inference were prominent.

Results

For the purposes of this paper we will focus on two episodes of teachers' discussions during the first week of the seminar, and their responses to an interview question at the end of that week.

The first discussion focused on the idea of *unusualness*. We focused on *unusualness* for several reasons. First, 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 logic demands that we assume the sample statistic of interest has some underlying distribution, for without assuming a distribution we have no way to gauge any sample's rarity. This assumption is made *independently* of the sample. It is like a policy decision: "If, according to our assumptions, we judge that samples like the one observed occur less than *x*% of the time (i.e., are sufficiently unusual), then either our sampling procedure was not random or values of the sample statistic are not distributed as we presumed." Second, we observed in high school teaching experiments that students had a powerful sense of "unusual" as meaning simply that the observed result is surprising, where "surprising" meant differing substantially from what they anticipated. By this meaning, if one has no prior expectation about what a result should be, no result is unusual. Since students infrequently made theoretical commitments regarding distributions of outcomes, their attempts to apply the logic of hypothesis testing often became a meaningless exercise.

To understand the teachers' conceptions of *unusualness*, we adapted the following question from Konold (1994).

> 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 teachers first gave their intuitive answers. All said it was not 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 a scheme of distribution of sample statistics. She proposed, as the method of investigating the question, "Each night 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 outcomes "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-quantitative. He justified his intuitive answer: 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. John employed what we call the *proportionality heuristic*: evaluating the likelihood of a sample statistic by comparing it against the population proportion or a statistic of a larger sample. He did not conceptualize a scheme of repeated sampling that would allow him to quantify unusualness. Henry's conception of unusualness was quantitative: He defined unusualness as "something's unusual if I'm doing it less that 50% of the time" This discussion

revealed that the teachers, with exception of Alice, had a mostly subjective conception of unusualness, and this conception did not support their thinking in hypothesis testing.

The second discussion focused on the logic of hypothesis testing. The logic of hypothesis testing is similar to the logic of proof by contradiction. In proof by contradiction, we reveal the truth of a statement in question by assuming its logical negation and then bringing this assumption into question by deriving a result that is contrary to the assumption or contrary to an accepted fact. In hypothesis testing, we test the plausibility of h_1 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. A small chance of the factual data with h_0 being true casts doubt on the plausibility of h_0 and in turn suggests the viability of h_1.

To understand the teachers' understanding of the logic of hypothesis testing, we engaged them in discussion of the following question:

> 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.

This question was accompanied by a list of 135 simulated samples of size 100 taken from a population split 50-50 in preference. Four of the 135 sample statistics exceeded 60%.

Three teachers, Lucy, John, and Henry, initially 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. 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.

Terry, the seminar leader, pushed the teachers to explain the tension between 1) a sample occurred, and 2) the likelihood of the sample's occurrence is rare under a given assumption. Henry suggested one explanation: The sample was not randomly chosen. John offered another: The assumption (that the population was evenly split) was not valid. Under intense questioning, both Henry and John eventually concurred that the data suggested that samples of 60% or more were sufficiently rare that something must be wrong about the assumption.

One teacher, Linda, insisted that the assumption should not be rejected on the basis of one sample. Her argument was that no matter how rare a sample is, it *can* occur, thus it cannot be used against any assumption. A mixture of beliefs and orientations helped explain why she was opposed to rejecting the null hypothesis, including: 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 "rejecting the null hypothesis on the basis of one sample" because any rare sample could still occur theoretically. Linda's concern for the truth of null hypothesis is inconsistent with the idea of 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 this decision rule consistently, over the long run we can keep the error rate at a reasonably low level.

In sum, the discussion revealed the spectrum of choices the teachers made when facing the question: Do we reject a null hypothesis when a sample is unusual in light of it?

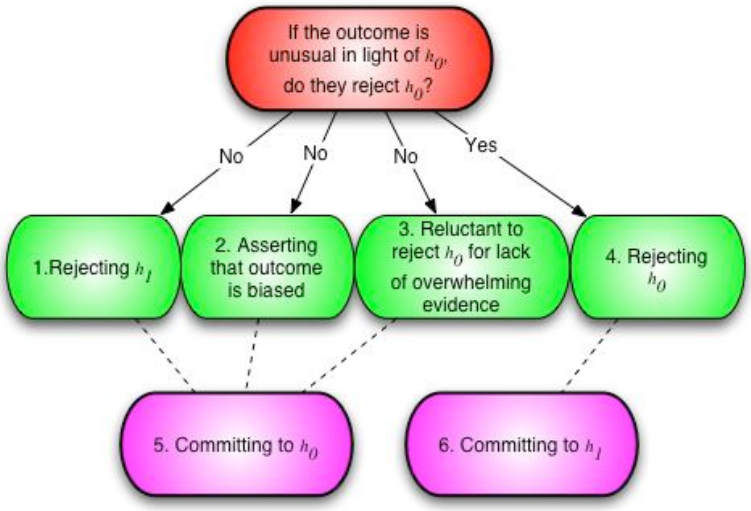

Figure 1: Theoretical framework for the logic of hypothesis testing

This framework captures the varieties of choices the teachers made when a small *p-value* was 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). The results of the discussion suggested that most of 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.

During the interview at the end of the first week, we gave 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 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 60% of them said they were satisfied. The implied question was, "Will the samples like or more extreme than 60% be sufficiently rare for one to reject the administrations' claim that 80% of all graduates are very satisfied with the education they received?"

Almost all the teachers noticed the large difference between 60% 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. 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.

When asked how they would test the administration's claim, only 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)

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 at the end of the seminar, they did not understand the functionality of it. In other words, they did not know the types (or models) of questions that hypothesis testing was created for, and how hypothesis testing became a particularly useful tool for answering these types of questions.

Overall, the results revealed that the majority of the teachers embraced conceptions of probability and logic of hypothesis testing that are incompatible with meanings that will support using it in ways that its inventors intended. Only one teacher conceptualized unusualness within a scheme of repeated sampling, and thus the others did not incorporate the idea of a distribution of sample statistics in their thinking of statistical inference. Most of the teachers did not understand the logic of hypothesis testing. This was revealed in the non-conventional decisions they made when a collected sample fell into the category of "unusual" in light of the initial assumption. These decisions revealed their commitment to the null hypothesis in question. Beyond the complexity of hypothesis testing as a concept, we 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 the interview data where only one teacher proposed hypothesis testing as the method of investigation.

Conclusions and Implications

The results revealed that teachers' understandings of probability and statistical inference were highly compartmentalized: Their conceptions of probability (or unusualness) 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. This purpose could be achieved by exerting a great amount of coerced effort in helping teachers develop the capacity and orientation in thinking of a *distribution of sample statistics*, which allows them to develop a stochastic/distributional conception of probability, and incorporate the image of distribution of sample statistics in their thinking of statistical inference.

We also learned that part of the teachers' difficulties in understanding hypothesis testing was a result of their tacit beliefs or assumptions about statistical inference, e.g., 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 and their established beliefs. To confront these hidden beliefs, we could, for example, design activities that incorporate the theoretical framework (Figure 1) and engage the teachers in discussions of the implications of each potential choice they might make. In having the teachers

reflect on the tacit beliefs that might lead them to non-conventional choices, we could help them come to appreciate the logic of hypothesis testing.

Reference:

Albert, J. (1995). Teaching inference about proportions using bayes and discrete models. *Journal of Statistics Education, 3*(3).

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

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

Link, C. W. (2002). *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. (2004). *Teachers' personal and pedagogical understanding of probability and statistical inference. Proceedings of the 26th PME-NA, Toronto, Canada.*

Konold, C. (1994b). "Teaching probability through modeling real problems." Mathematics Teacher(87): 232-235.

Steffe, L. P. (1991). The constructivist teaching experiment: Illustrations and implications. In E. von Glasersfeld (Ed.), *Radical constructivism in mathematics education*. The Netherlands: Kluwer.

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

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