STUDENTS' REASONING ABOUT *p*-VALUES

BIRGIT C. AQUILONIUS
Stockholm University
birgit.aquilonius@mnd.su.se

MARY E. BRENNER
University of California, Santa Barbara
betsy@education.ucsb.edu

ABSTRACT

Results from a study of 16 community college students are presented. The research question concerned how students reasoned about p-values. Students' approach to p-values in hypothesis testing was procedural. Students viewed p-values as something that one compares to alpha values in order to arrive at an answer and did not attach much meaning to p-values as an independent concept. Therefore it is not surprising that students often were puzzled over how to translate their statistical answer to an answer of the question asked in the problem. Some reflections on how instruction in statistical hypothesis testing can be improved are given.

Keywords: *Statistics education research; College students; Introductory statistics*

1. INTRODUCTION

Many college students are required to take statistics courses for their majors. Hypothesis testing is often taught as the last part of such an introductory course and in one sense becomes the goal of the course. The first author has, during her twenty years as a statistics instructor, received mixed messages about her students' understanding of hypothesis testing. The students in her classes sometimes said or did things which made her believe that her students had a good understanding of the topic. At other times, the same students made mistakes on tests and homework that made her doubt their understanding. For example, surprisingly often, students worked a problem correctly all the way through a process given by their instructor only to give an incorrect answer in the last step.

Falk and Greenbaum (1995) reported that hypothesis tests are commonly taught in introductory statistics classes and also used in psychological research. The authors showed how textbooks often present hypothesis tests in potentially misleading ways. In addition, the authors found that sometimes even psychological researchers misinterpret the results from the tests. Considering the authors' results, it is not surprising that, in line with the experience of the first author of this paper, students exhibit several misconceptions about the *p*-value (Castro Sotos, Vanhoof, Van den Noortgatestudy, & Onghena, 2009; delMas, Garfield, Ooms, & Chance, 2007; Haller & Krauss, 2002).

The student participants in the aforementioned studies were presented with alternative interpretations of the *p*-values and asked to decide which interpretations were correct and which ones were incorrect. This kind of methodology allows the researchers to get

responses from many participants. Therefore those researchers can argue convincingly about their conclusions. However, when students later in their career need to apply hypothesis testing to research questions, they will not be given a set of *p*-value interpretations from which to choose. Therefore, as a complement to the earlier quantitatively-oriented studies, this article reports on a study in which students reason about *p*-values in the context of problem solving. Students were videotaped while solving problems and being interviewed by the first author. As noted by Gordon, Reid, and Petocz (2010), articles published in *SERJ* during the 10 years preceding their publication show the important role qualitative methods can play in investigating statistics education issues.

2. LITERATURE REVIEW

2.1. SOME PROFESSIONAL VIEWS ON THE USE OF *p*-VALUES

Cumming (2010) describes two approaches to hypothesis testing. One is called N-P because of its proponents Neyman and Pearson. The other approach is attributed to its proponent, Fisher. The N-P method calls for specifying a value of α, stating an alternative hypothesis, and then comparing p with α. If $p < \alpha$, one rejects the null hypothesis – otherwise, one fails to reject the null hypothesis. Fisher saw the *p*-value as an indicator of strength of evidence against the null hypothesis. Cumming as well as Haller and Krauss (2002) suggest that the way those two approaches often are presented as equivalent in textbooks might be confusing to students.

In both approaches to hypothesis testing described above, *p*-values play a crucial role. Hubbard and Lindsay (2008), however, have serious objections regarding using *p*-values to measure evidence in statistical significance testing. Some of the authors' arguments are mathematical and rather technical in nature. The authors do, however, make a logical argument against using *p*-values because "a valid measure of strength of evidence cannot be dependent on the probabilities of unobserved outcomes" (p. 79).

2.2. MISCONCEPTIONS IN HYPOTHESIS TESTING

The most pervasive misconception about *p*-values seems to concern the *p*-value as a conditional probability. Falk (1986) described how students often confuse the conditional probabilities $P(H_o|R)$ and $P(R|H_o)$, where H_o stands for the event that the null hypothesis is true and R for the event of rejecting the null hypothesis. Even researchers sometimes fall prey to this misconception. $P(H_o|R)$ is the quantity that would be very helpful to know, because $P(H_o|R)$ would give the probability that the null hypothesis was true though it had been rejected. To actually find $P(H_o|R)$, however, one would need to do some Bayesian computations requiring quantities that are likely to be unknown. The statistical decision rule is instead based on $P(R|H_o)$, which is the *p*-value of the test.

Falk and Greenbaum (1995) provided an example, originally from Pauker & Pauker (1979) using real data from Down's syndrome diagnosis, of how different $P(H_o|R)$ and $P(R|H_o)$ can be.

> Thus, if we substitute 'The fetus is normal' for H_0, and 'The test result is positive (i.e. indicating Down's syndrome)' for D, we have $P(D|H_0) = .005$, which means that D is a significant result, while $P(H_0|D) = .82$ (p. 78)

Falk's (1986) main point was that researchers often misinterpret hypothesis testing and that the method does not answer the questions that the researchers really want to answer.

In addition, she maintained from her own teaching experience that hypothesis testing is confusing for students. Still, as Falk (1986) predicted, researchers continue to use hypothesis testing, which means that this method will continue to be taught in introductory statistics classes. Her discussion of the conditional probabilities involved in hypothesis testing theory sheds some light on why the hypothesis-testing concept is so hard for students. The ideas behind hypothesis testing might not be as straightforward as statistics instructors might like to believe. In particular, a closer look at how students reason about *p*-values is needed, because the *p*-values are the basis for decisions in hypothesis testing.

As noted in the introduction, several authors have documented lists of student misconceptions concerning hypothesis tests. Based on such lists, Castro Sotos et al. (2009) designed a questionnaire for university students to answer following their introductory statistics course. One of the items in the questionnaire addressed the concept of *p*-value. The 144 participants in the study were presented with seven statements from which they had to select the one that correctly defined the *p*-value in hypothesis testing. A relevant result to our article was, for example, that only 52% of the students in the study by Castro Sotos et al. selected the correct response among the seven to complete "Given a *p*-value of 0.01 …." , i.e. "the probability of the same or more extreme data assuming the null hypothesis is true". Forty-six percent of the students selected the correct response *alone* though the students were instructed to select *the* correct option. Seven percent of the students simply ignored the conditioning event. Twelve percent selected that "Given a *p*-value of 0.01" meant that the probability of the null hypothesis is 0.01 and 16% that it meant that the probability of making an error when rejecting the null hypothesis is 0.01.

The CAOS test (delMas et al. 2007) was developed to measure students' conceptual ideas in statistics. The participating students had to respond to 40 multiple choice questions. Students were correct on just a little more than half of the items on average at the end of their introductory course.

Relevant to the present study, the delMas et al. (2007) study had results that indicate that the introductory statistics course had not substantially improved students' understanding of hypothesis testing. On the pretest, 63% of the students answered on an item in a way that showed that they understood that no statistical significance does not guarantee that there is no effect. However, that percentage rose only to 64% on the corresponding item on the posttest. On a different test item of the post test, delMas et al. found that a little over half of the students recognized a correct interpretation of a *p*-value. However, "the majority of those students also responded that an incorrect interpretation was valid, indicating that many students hold both types of interpretation without recognizing the contradiction" (p. 49). That students show so little improvement after an introductory statistics course motivates studies like the one reported in this article. The statistics education community needs to know more about how students reason about topics like *p*-values after these topics have been taught in introductory statistics classes.

2.3. A STRATEGY TO TEACH CONDITIONAL PROBABILITY

Falk (1986) showed that the students often confused the conditional probabilities involved in hypothesis testing. Pfannkuch, Seber, and Wild (2002) write that at their university, as often is the case, they have to cater to a wide variety of backgrounds in their elementary statistics courses. The authors wrote: "Probability theory was the crunch point in the class, and our help facilities became overloaded." (p. 24) The authors then started to teach their probability based on two-way tables rather than starting with formulas. Their strategy turned out to be very successful and gave students tools to solve problems involving conditional probabilities, problems that otherwise would have

required Bayes' Theorem. The authors suggested that at one extreme, probability theory could be taught without any formulas, and at the other extreme, all the usual probability formulas could be developed in parallel and formulas formally used to justify the operations that are performed on the tables. In the context of hypothesis testing, the authors of the present article believe that working with tables in the way suggested by Pfannkuch et al. (2002) very well could help demystify the notations of the conditional probability and help students distinguish between $P(H_o|R)$ and $P(R|H_o)$.

2.4. TWO USEFUL STRATEGIES FOR TEACHING HYPOTHESIS TESTING

The Aquilonius (2005) study, from which the data for this paper were collected, also included interviews with college instructors about their teaching of hypothesis testing. One of the instructors said in his interview that, the *p*-value "is basically the probability that if the null hypothesis is true that you would get the sample result you got, or greater than it." This statement is quite close to the statement assigned to be the correct one in the Castro Sotos et al. (2009) study reported on in section 2.2.

The instructor interviews also made it clear that the instructors relied heavily on thinking about graphical representations of sampling distributions to make sense of the whole hypothesis testing process. The instructors explained to the interviewer that they would frequently use those representations in their teaching to talk about *p*-values. That this teaching strategy would be considered successful is consistent with the findings of Hong and O'Neil (1992).

Hong and O'Neil included two groups of students in their study. The "Intermediates" consisted of 18 doctoral students, of which four had completed most of the statistics courses offered by the School of Education. The other 14 were taking the educational statistics course at the time of the study and had studied the hypothesis-testing chapter. In the study, the Intermediates were asked to solve hypothesis problems. Through protocol analysis, the authors found that 14 Intermediates (of 18) used graphical (called "diagrammatic" by the authors) representations because it helped the students understand and solve the problems.

The study also contained another result that is relevant to the present study. Of the 18 Intermediates, 14 had had difficulty interpreting the result or making an inference about the population even though some of them followed the proper procedures and decision rules for hypothesis testing. This latter result is consistent with the first author's teaching experience and one reason why the present authors wanted to look more closely at how students reasoned about hypothesis testing.

Hong and O'Neil called their second group "Novices". This group consisted of 27 graduate and 29 undergraduate students who were taking introductory statistics courses at a California research university. The experiment started before the topic of hypothesis testing was covered in the participating students' courses. The students were divided into four subgroups of approximately the same size. Each subgroup received a different kind of computerized instruction about hypothesis testing. After the instruction had been completed, the students in all four groups took a paper-and-pencil posttest. When taking the posttest, the students were asked to think aloud during their solution process and an investigator reminded them to talk if they lapsed into silence.

The Hong and O'Neil (1992) study showed that two instructional strategies were beneficial in teaching statistical hypothesis testing. One of those strategies tested by the researchers consisted of presenting the ideas behind hypothesis testing before teaching the procedure. Another strategy was to teach students to graph the sampling distribution as part of their solution process. Both those strategies produced better student performance on statistical hypothesis testing exercises than if the strategies were not used.

The present authors' experience is that the strategy of presenting the idea of hypothesis testing before teaching the hypothesis testing procedure is already common among statistics instructors. However, a more hands-on approach than is normally used may be more effective. Lawton (2009) describes a simulation exercise called "Wheel of Destiny" that he found helpful to illustrate the logic of hypothesis testing. The exercise originates from a story about a woman who runs a small-time gambling operation on campus. She has purchased a new spinner with the numbers 1 through 5 on it and sells each number for $1. Before she begins using the spinner, she wants to make sure the spinner is fair. The statistics students log into a spinner site to simulate the spinner and record their results. Those results are then used to have discussions about probabilities and the chi-square hypothesis test. "When the students begin to show signs of confusion as we present other types of hypothesis tests, we have found it helpful to refer them to the Wheel of Destiny example and to think about how and why we acted as we did" (p. 8).

2.5. A FRAMEWORK FOR THE RESEARCH QUESTIONS

Tall and Vinner (1981) present a lens through which the students' work with hypothesis questions can be viewed. The authors wrote about *concept definitions* and *concept images*. Concept definitions are the words by which a mathematical term is defined. In our context we are interested in the concept definitions of α and the *p*-value. The probability α is defined as the probability of a type I error, i.e., of rejecting the null hypothesis when it is true. The *p*-value has several equivalent definitions. The definition given in one of the participating students' textbooks is: "For the distribution described by the null hypothesis, the P *value* is the smallest level of significance for which the observed sample statistic tells us to reject H_o." (Brase & Brase, 2003, p. 479).

According to Tall and Vinner (1981), the *concept image* is "the total cognitive structure that is associated with the concept, which includes all the mental pictures and associated properties and processes" (p. 152). Tall and Vinner point out that mathematical definitions are subjects of great precision, on which concepts can be defined accurately to provide a firm foundation for mathematical theories. Human brains, however, do not work that logically. We build our concept images from all the different experiences we have with a given concept. For example, when students come to a first course in probability and statistics, they have already met probability and statistics in less formal contexts and have formed their own concept images within those subjects. Those students' concept images might be at odds with those of their instructors, but more likely simply incomplete.

It is clear from Tall and Vinner's (1981) exposition that, in elementary mathematics, informal definitions and reasoning might function well. As mathematics becomes more abstract, however, students' concept images will have to be consistent with formal definitions – and those formal definitions will have to be included in the students' concept images. The authors take subtraction as an example. Young children might think about subtraction as "take away", but at some time the definition of subtraction as the inverse operation of addition has to be incorporated in students' concept images.

Tall and Vinner (1981) also wrote that a student's concept image might not be coherent at all times and call the portion of the concept image activated at a particular time the *evoked concept image*. For a researcher, different evoked concept images by a student might seem contradictory, but the student will experience a cognitive conflict only if conflicting aspects are evoked simultaneously. An example of such conflict was cited from delMas et al. (2007) in section 2.3.

As noted above, both α and the *p-value* have formal mathematical definitions. Tall and Vinner would call those definitions concept definitions. To be competent hypothesis test

problem solvers, students need to know those definitions. To answer hypothesis questions correctly, however, students need some additional cognitive structures – that is, a richer conceptual image than memorization of the definitions can provide.

2.6. RESEARCH QUESTIONS

Our study focused on the following questions:
(1) How did the students use *p*-values in the context of their graphical representations?
(2) What was the students' understanding of *p*-values?
(3) What kinds of problems regarding *p*-values surfaced in the students' reasoning?

3. METHOD

3.1. RESEARCH DESIGN

The study was designed to uncover students' statistics reasoning a layer below what a statistics teacher sees in the classroom. The teacher will see the students' statistical problem solving, but will rarely know *why* the students do what they do. Are students mimicking the teachers' procedures? How much meaning do the students attach to their work? To explore such issues, a qualitative research approach was used. We used a method developed by Alan Schoenfeld's research group at the UC Berkeley School of Education as an adaptation for mathematics education research of methods used by artificial intelligence researchers (Ericsson & Simon, 1984) and other researchers in the cognitive science tradition. The method, called a Two-Person Problem Solving Protocol, consists of videotaping students solving mathematical problems together, transcribing their talk during problem solving, and analyzing the resulting protocols (Schoenfeld, 1982, 1985a, 1985b, 2007). Schoenfeld (1985a) advocated use of student pairs in cognition research rather than the individual talk-aloud protocols common in artificial intelligence research. He wrote, "In single-person 'speak-aloud' protocols, what appears is often the trace of a solution: One sees the results of decisions but gets little insight into how the decisions were made, what options were considered and rejected, etc. When students work together, discussions between them regarding what they should do next often bring those decisions and the reasons for them out in the open" (p. 178). The method with Two-Person Problem Solving Protocols is now commonly used in mathematical cognition research.

The students in this study were encouraged to talk about their work and results, but were also allowed to work silently when they so chose. When students work together, they naturally switch back and forth between working silently, and talking about their work with the partner(s). There is much for a researcher to learn from those spontaneously occurring conversations, especially when the information from student conversations are complemented, as was done in this study, by students' written work and follow-up interviews with the students.

3.2. SETTING AND PARTICIPANTS

The research project was conducted at a community college in Silicon Valley, California, where the first author was a mathematics instructor at the time of the study. Students from five parallel sections of an elementary statistics course participated. The students were business and social science majors who needed the course in order to transfer to a four-year college or a university. The prerequisite for the course was high school

algebra. The students showed varying degrees of statistical competence as measured by their final grades. They also exhibited different demographic characteristics. Their textbooks differed between sections.

At the community college where the study was conducted, the introductory statistics course was called *Elementary Statistics*. Table 1 is an overview of the students in the study with some information about them. The five pairs listed first were still enrolled in their statistics class, while the three later ones had completed their classes.

Table 1. Student participants

Name (pseudonym)	Textbook Used by Pair	Instructor	Approximate Student Age	Course Grade
Alex	*Understandable*	A	20s	B
Ben	*Statistics*	A	20s	A
Cindy	*Understandable*	C	20s	A
Dana	*Statistics*	D	20s	A
Ellie	*Understandable*	A	40s	C
Fran	*Statistics*	D	30s	A
Gus	*Understandable*	A	20s	B
Hal	*Statistics*	A	20s	C
Mary	*Workshop Statistics*	B	20s	B
Nan		B	20s	B
Rose	*Workshop Statistics*	I	30s	A
Sylvia		I	40s	B
Tracy	*Workshop Statistics*	I	40s	A
Ursula		I	40s	B
Vera	*Workshop Statistics*	B	30s	B
Zoe		B	20s	C

The letters A, B, C, D, and I in the instructor column indicate each student's instructor. The letter I in the column means that the first author was the student's instructor. All of the instructors whose students participated in the study were experienced teachers with more than twenty years of teaching experience. The instructors were aware of recent reform efforts in statistics education and incorporated reform ideas, such as simulations, in their teaching.

Half of the students in this study were in classes using *Understandable Statistics* (Brase & Brase, 2003) as their textbook. The other half of the students were in classes using *Workshop Statistics: Discovery with Data and the Graphing Calculator* (Rossman, Chance & von Oehsen, 2001). Both textbooks used by the students in the study tried, in different ways, to stress meaning above formalism.

All research sessions involved pairs of students for the purpose of soliciting verbal and written data. Six students volunteered with partners to form three pairs. The other ten students volunteered as individuals and pairs were formed from those students. Several statistics instructors helped recruit students from their statistics classes for the research project. Most of the videotaping sessions and interviews took place in empty classrooms.

The remaining sessions were conducted in a mathematics study center not used by other students at the time.

3.3. PROCEDURES AND DATA COLLECTION

The main sources of data for the project were videotapes of students solving problems and being interviewed. The tape content was transferred to digital files that were watched repeatedly on a computer. The students' answer sheets were also collected and analyzed.

The students came to two research sessions, both of which lasted two hours. In the first hour of both sessions, the students were given hypothesis test problems to solve. Some of those problems were from the students' textbooks and others were from tests that the first author had given earlier in her classes. First, some standard problems were given to put the students at ease so they would start talking about their work. Those problems also gave information about which procedures the students used for hypothesis testing. One such problem was the Home Value problem:

> *Home Value problem.* A city council member said that 18% of all homes in the city had been undervalued by the assessor's office. The local newspaper conducted a random sample of 98 homes and found that 26 had been undervalued. At $\alpha = 0.05$, test the claim that the proportion of undervalued homes in the city is different from 18%.

Later some problems were given that were not as straightforward. They were chosen to give information about how students think about some of the issues that confuse introductory statistics students. One such problem was the Tranquilizer problem.

> *Tranquilizer problem.* In an experiment with a new tranquilizer, the pulse rates of 25 patients were taken before they were given the tranquilizer, then again five minutes after they were given the tranquilizer. Their pulse rates were found to be reduced on the average by 6.8 heart beats per minute with a standard deviation of 1.9. Using the 0.05 level of significance, what could we conclude about the claim that this tranquilizer will reduce the pulse rate on the average by 7.5 beats per minute?

Another challenging problem was the *Coin problem*, which is shown in the Results section. All students were given the same problems to solve.

Students were told that they would not be given any feedback during the first hour of the research sessions, because their problem solving was part of a research project. Also, students were told that if during the first hour they arrived at a point where they could not continue with problem solving they would simply be given another problem.

During the second hour of each session the students were interviewed about their work and asked more theoretical questions. All students were asked the following two questions:

#1: Why do you reject the null hypothesis when $p < \alpha$?

#2 Why is it not good to say "Accept the null hypothesis" when $p \geq \alpha$?

The purpose of those questions was to solicit students' thoughts about *p*-values in context of their problem solving. When those questions did not give enough information to satisfy the interviewer, she would ask the students for definitions and/or explanations of *p*-values.

3.4. ANALYSIS

This article describes an exploratory study of students' statistical thinking in the qualitative research tradition. The analysis was mainly data driven (e.g., Marshall & Rossman, 1999). The analysis consisted of pursuing emergent themes in ways that are

important parts in qualitative traditions such as grounded theory (Charmaz, 2002). To find those themes, detailed transcripts of the students' conversations and the interviews with students, were made into tables with a place for notes relevant to the research questions. In the Results section, those themes are presented with relevant examples. Tall and Vinner's (1981) framework, as presented in section 2.5, was then used for discussing the results of the study in the Discussion section.

4. RESULTS

4.1. CURRICULUM LEADING UP TO HYPOTHESIS TESTING

Before reporting on the student data related to the *p*-value concept, it is prudent to make a few comments about the material in the elementary statistics curriculum that led up to hypothesis testing. In both of this study's textbooks a treatment of probability precedes inferential statistics. First, the textbooks gave a general exposition of probability. However, the two textbooks used by the students in the article had very short treatments of conditional probability, with one of the textbooks, *Understandable Statistics* (Brase & Brase, 2003), using only one page for its treatment. Thus the students had received very little instruction in conditional probabilities and a lack of understanding of those probabilities could impair their reasoning regarding hypothesis testing. After the general probability sections, a treatment of the binomial and normal distributions followed. The normal distribution was always described in geometric terms, where probability was expressed as the area under the curve. This view of the normal distribution was enforced in the study's participants' classes by frequent use of the TI-83 graphing calculator, which provided an option to graph a normal curve with the desired probability shaded.

After the probability curriculum, the textbooks covered the Central Limit Theorem, which also was described in geometric terms showing how the spread of the sampling distribution shrinks as sample size increases, e.g. figure 7-2, p. 343 in *Understandable Statistics* (7th ed.). *Workshop Statistics,* had several simulation activities, such as 17-1 and 17-2 (pp. 367–375, 1st ed.), showing how as sample size increases, any sampling distribution approaches the normal distribution. The ideas of the Central Limit Theorem were described in geometric terms, with probability represented as "the area under the curve."

Inferential statistics then began with a treatment of confidence intervals. In developing the theory of confidence intervals the textbooks continued to depend on graphs for their arguments (e.g., figure 8-3, p. 377 in *Understandable Statistics,* 7th ed.). In addition to the normal distribution graphs, *Workshop Statistics* also referred to box plots and dot plots in an attempt to build students' intuitions.

Hypothesis testing followed the confidence interval topics. After some introductory material, *Understandable Statistics* summarized its information by giving a four-step procedure for hypothesis testing on page 463. All the students in the study had been given versions of this procedure by their instructors. As an example of a procedure given to the study's participants, below is a list the steps the first author gave her students:

(1) List the information given in the problem
(2) Set up the hypotheses
(3) Decide on a level of significance
(4) Sketch the graph showing the critical region
(5) Use the calculator to compute the *p*-value
(6) Decide to reject or fail to reject the null hypothesis
(7) Answer the question asked in the problem.

As they moved into the hypothesis testing curriculum, books and instructors continued to describe probability in terms of area. Since the mathematical derivation of the probability measures was beyond an elementary statistics course, an appeal to graphs seemed a reasonable approach.

4.2. OVERVIEW OF THE STUDY'S RESULTS

None of the N=16 students could cite a formal definition of *p*-values without consulting their notes. Using the framework from Tall and Vinner (1981), concept definitions of *p*-values were not part of the students' concept images. Still, all students in this study used *p*-values to make statistical decisions for their problems. The only exceptions were two solutions by Mary and Nan. This pair made decisions on two problems by drawing the sampling distribution and rejecting the null hypothesis based on the sample proportion being more than three standard errors away from the hypothesized proportion. In all other cases the students (including Mary and Nan) based their statistical decisions on a *p*-value given to them by their calculators.

Students made the correct statistical decisions based on their calculators' *p*-values. A few times a student became temporarily confused and did not know if he or she should reject the null hypothesis or not. Before the problem was completed, however, that student was always set straight either by his or her partner or by looking in his or her notes. This kind of temporary confusion is not surprising if one considers that *p*-values did not carry much meaning for many of the students. On the other hand, the students knew their rules sufficiently well that those temporary confusions did not prevent the students from correctly rejecting or failing to reject their null hypotheses.

In spite of being able to make correct decisions based on *p*-values, the students could not explain to the interviewer what *p*-values were. Several students thought about the meaning of a *low p*-value as a *small* likelihood for the null hypothesis to be true. This finding can be compared to the delMas et al. (2007) study in which a high percentage of the students understood that low *p*-values were desirable in research studies. Several students in this article's study also said that a *high p*-value meant, "There was not enough evidence to reject the null hypothesis."

For most of the students, the *p*-value did not have any meaning if no significance level was given. For those students, the *p*-value was something to be compared with alpha to get an answer. For two pairs of students, drawing graphs with sampling distributions helped in this process and might also have added some meaning of the *p*-value as a probability represented by area. One student drew her *p*-value incorrectly on her graphs and could have had difficulties if her algebraically inclined partner had not dominated the decision process. Considering that most weak students would not volunteer for a study like this one, there are most likely a higher percentage of weak students with the same confusion in introductory statistics classes than in this study. The first author has also seen the larger-means-to-the-right concept incorrectly being applied to *p*-values in her statistics classes. This difficulty in moving from one abstraction, the number line, to another, probability as area, is an example of how the abstract thinking required in inferential statistics can be a hindrance for students to gain competence.

This overview shows that the students have parts of the needed concept images of *p*-values to solve hypothesis problems, but that crucial parts are also missing. The students did not know the formal definitions of *p*-values and also did not have enough meaning attached to *p*-values for the hypothesis process to make sense to them.

4.3. STUDENTS' GRAPHICAL REPRESENTATIONS

As was mentioned earlier, Hong and O'Neil (1992) showed that using a graphing approach to teach hypothesis testing was productive. Therefore, the analysis of the student data will start with an overview of how students used graphs when talking about *p*-values.

Table 2 below summarizes the frequency with which students in the study used graphs of the normal curve and geometric arguments in making statistical decisions.
For the two first rows, two small check marks in one cell means that *both* students in the corresponding column had made graphs of the sampling distribution on their answer sheet. A small check mark on the *upper* part of a cell means that the student listed *first* in the column heading had made graphs of the sampling distribution on her or his answer sheet, while a small check mark on the *lower* part of a cell means that the student listed *second* in the column heading had done so. For the third and fourth rows, a large check mark means that the *pair* listed in the corresponding column exhibited the behavior listed in the corresponding left heading. As can be seen in the table, all but one pair drew graphs on their answer sheet. However, not all of those graphs had *p*-values marked on them. The second row in the table shows which students had *p*-values marked on their graphs, because only those students could be expected to use their graphs to make their statistical decisions.

Table 2. Students' graphical representations

	Alex Ben	Cindy Dana	Ellie Fran	Gus Hal	Mary Nan	Rose Sylvia	Tracy Ursula	Vera Zoe
Graphs but no *p*-values	√		√	√		√	√	
	√	√	√	√		√	√	
Some graphs with *p*-values	√		√	√	√			
	√	√		√	√			
Geometric arguments for several answers	√			√				
Geometric arguments in explaining decision rules	√		√	√	√		√	

The first two rows of the table contain information collected from the students' answer sheets. Therefore, it was possible to give information for individual students. For the last two rows, the information comes from the videotapes. In most of those instances, the decision to use graphs seemed to be more of a joint decision by the pair than by individuals. Therefore, one large check mark was used to denote that the pair used graphs or geometric arguments.

As can be seen from Table 2, a large majority (81%) of the students drew graphs when they solved hypothesis test problems. The graphs mostly consisted of a normal distribution sketch with shading of the tail(s).The only pair that did not draw graphs at all was Vera and Zoe. Most graphs, however, had neither *p*-values marked on them, nor areas shaded corresponding to *p*-values. Only half of the students used detailed graphs that included some representation of *p*-values. Students with good final grades (Ben and Dana) as well as students (Ellie and Hal) with final grades of C used detailed graphs.

Student conversations revealed diverse attitudes towards the usefulness of graphs in hypothesis testing. At one end of the spectrum was Alex who said, "The visual thing helps a lot" when he was solving his first problem during the research session. At the other end of the spectrum were Rose and Sylvia who said that the graphs "did nothing for them" as

far as understanding hypothesis testing. Even within pairs, opinions about the sampling distribution graphs differed. For example, Zoe suggested to Vera that they should do a diagram during their work on one of the problems. Vera responded: "I never liked the diagrams. I always thought they were tedious. They annoyed me." To which Zoe responded: "They are nice."

Alex and Ben used geometrical reasoning more than any other pair did. The pair called their graphs "the visuals" and used expressions such as "landing in the reject region" and "landing in the accept region". Those expressions were used both during the problem solving session and when the pair was interviewed. When asked about the decision rules, they also used the expression "passing the point where you are willing to accept". The graphical approach served the pair well in their problem solving. From a strictly procedural view, the pair's problem-solving work was flawless. Their two incorrect answers were due to other reasons: one due to a lack of care in reading the text of one problem and the other to an incorrect alternative hypothesis.

The only other pair that used geometric reasoning for their statistical decisions was Gus and Hal. For example, when the pair was asked why you reject the null hypothesis when *p* is less than alpha, Gus answered, "Because if it is in the alpha region, from alpha all the way to infinity, that's the reject region." As with Alex and Ben, the graphical approach served Gus and Hal well. As sometimes happened to other students in the study, the pair was temporarily confused about the statistical decision rules during one of their problem solving attempts. While other students either referred to their notes or checked with their partners, Gus and Hal took advantage of their geometrical approach as can be seen in the excerpt below. In this excerpt, as well as some later ones, the word “it” is italicized to make the reader notice a pattern in the students´ speaking of using “it” as a placeholder for different statistical terms.

Hal: Is it less than when you reject *it*, or is it when *it* is greater than. It is when *it*'s greater than, that you reject *it*, right?
Gus: *It* is right-tailed, right?
Hal: Yes. So you reject *it*, when *it* is greater than or less than?
Gus: You reject *it* if *it* is greater than.
[There is a pause, during which Gus keeps writing.]
Hal: So reject *it*. Even though *it* is on the right side?
Gus: But *it* will be in the reject region.
Hal: *It* will be way out here [points to the right tail of Gus' graph].
Gus: Yes.
Hal: So reject *it*.

After this exchange, Gus and Hal silently wrote up their answers. It seems like Gus and Hal's geometric approach helped them to make a correct statistical decision when their memory failed them regarding the algebraic version of the decision rule. Hal says, "it is all the way out here" and points far to the right of the center on Gus' sampling distribution graph and draws the correct statistical conclusion from this idea. Still, the students used *it* as a placeholder for both the null hypothesis and the *p*-value, as shown in the transcript. This way of speaking distinguishes itself from their instructors’ consistent use of statistical terms. The students’ way of speaking suggests they think locally and do not have the complete hypothesis model in mind.

In the student interviews, most students said that they had made more detailed graphs when they were in the beginning of the hypothesis test curriculum than at the time of the study. However, this study's data do not allow for any conclusions whether the students

actually relied more on graphical representations earlier in their courses. When analyzing the students' conversations, the authors realized that most students depended more on algebraic rules than graphical representations when making their statistical decisions. This reliance on algebraic rules seems unfortunate in light of Hong and O'Neil's (1992) findings.

Creating graphs as part of problem solutions seemed to help students to make correct statistical conclusions based on *p*-values. However, there was almost no evidence in the student conversations that those graphs helped build meaning for the whole hypothesis testing process. As with most of the students' work, shading the tail on the graph to represent the *p*-value was just another step leading to the answer of the problem.

4.4. STUDENTS' UNDERSTANDING OF *p*-VALUES

The students were usually asked what *p*-values were in the context of some problem. None of the students were able to give a definition of *p*-value without consulting notes. Such reference to notes suggests that the students did not attach much meaning to the *p*-value concept. The concept definitions given to the students in class did not seem to help them in this respect.

Another indication of this lack of meaning could be seen in the occasional, but usually temporary, confusion that most students showed regarding the statistical decision rules for hypothesis testing based on *p*-values. One such example is shown below in an exchange, in which Rose and Sylvia discuss the answer to the Home Value problem. The exchange between Rose and Sylvia is also typical because of its reference to the significance level alpha.

Sylvia: So it is .008. So this time it is less [than alpha]
Rose: So we *cannot reject.*
Sylvia: Oh, wait a minute!
Rose: Right?
Sylvia: Doesn't that mean that we have to?
[Rose quietly reflects on Sylvia's question and then responds.]
Rose: Yes, yes, you are right. I got it backwards.

Before the interviewer (I) raised the *p*-value issue, she had discussed the concept of the significance level or alpha with Cindy and Dana. The pair seemed to have a fairly good understanding of the level of significance, a fact that was stated by I in initiating the following exchange.

I: So you pretty much knew what alpha was. It seemed that you had a pretty good idea. But you don't have the same intuition for what the *p*-value is. [Both students shake their heads in agreement.]
Cindy: Yes, I honestly don't know what it is. I just know that if it's greater than or less than, it's just a rule
Dana: It's just what you compare to the alpha
I: Yeah, it takes you through the problem fine. One way to think about it is graphically …
Dana: It's the shaded part under the curve
Cindy: Oh, yeah.

When Cindy said, "I honestly don't know what it is", referring to the *p*-value, she represented the majority response of the participants in this study. The interviewer also asked Cindy and Dana, like she did with all the student pairs: "Why do you reject the null hypothesis when *p* is less than alpha?" Then Cindy answered, "Because our teacher told us to", and both students laughed. In the pair Vera and Zoe, Zoe said, "That's because it's what the teacher told me to do." Similarly, when Mary and Nan were asked the same question, Nan answered, "Oh my gosh, because that's the way we were taught. …It was a rule [the instructor] told us in the very beginning, but I don't think [the instructor] ever talked about why." Thus, it was common for the students in this study to consider *p*-values as a tool to get to the answer without actually understanding *why* the comparison with alpha would lead to an answer.

When students were pressured to produce explanations, those explanations often sounded confusing. The excerpt below shows such an example, in which Nan's partner Mary tried responding to the request for more explanation.

Mary: Maybe because the value that we are testing is 0.05 [she holds up her hands in the air, as to show the two "cut off lines" that alpha creates on the normal curve, when you do a two-tailed test]. I don't know …
Nan: No, no, I think you are on the right track, keep going with that.
Mary: It's our little comfort zone to see if it fits in there [while making a hand gesture suggesting a confidence interval].
Nan: The sampling variability.

For quite some time, Nan and Mary tried to find some sense in the statistical decision rule without much success.

There were exceptions to the mechanical applications of *p*-values and unclear rationales for those applications. For example, Dana was able to show that for her a *small p*-value meant something more than a step towards the answer. When asked for an explanation of why one rejects the null when *p* is less than alpha she said, "Because there is such a small chance that *it* is true". Prompted, she filled in that *it* referred to the null hypothesis.

Other students were able to express similar ideas as Dana about why a small *p*-value implied rejection of the null hypothesis. For example, Ursula answered that one rejected the null hypothesis when *p* was less than alpha because "when *p* is very small *it* cannot have happened by chance." Gus started to answer the question with a geometric argument, "It is in the rejection region". Then he added, "It means that the probability is slim to none for *it* to even happening, so you reject *it*." When asked about what *it* is, he and Hal answered, "The μ that you are testing." Hal also added, "The null hypothesis."

The examples above show that in the context of making statistical decisions, some of the students attached meaning to a *small p*-value, even though they did not know the *p*-value definition. A few of the students also expressed something similar to Dana when she said that a *high p*-value meant "there is not enough evidence to reject the null hypothesis". However, for a majority of the students in the study, a *p*-value only made sense as a quantity to compare with alpha when making statistical decisions. Also, the fact that the students required alpha values from the interviewer to do the hypothesis test problems when alpha values were not given, indicates that most of the students did not see *p*-values as a self-contained concept.

4.5. STUDENT DIFFICULTIES WITH *p*-VALUES

As stated in the overview of the results, students knew how to use *p*-values for making their statistical decisions. It was not in application of *p*-values to statistical decisions where the students had difficulties. Rather, it was the lack of understanding for *p*-values as an independent concept. Because a *p*-value did not mean much unless one compared it to an alpha value, the students often had problems using the *p*-value to answer the question asked in the given problem. The students' answer sheets showed several instances of incorrect final answers although the students had computed the *p*-values correctly. Those mistakes were all done by students having *Understandable Statistics* as their textbook.

Even when students arrived at a correct final answer, they often spent a long time arriving at that answer. The students did not seem to consider the concept of sampling variability that is intimately connected with the concept of *p*-value. For example, all student pairs had lengthy discussions before they arrived at the final answer to the Coin problem.

> *Coin Problem.* You suspect that a certain coin, when tossed, favors heads. You toss it 50 times and find 31 heads. At the 0.05 significance level, does it favor heads or is it a fair coin?

Alex and Ben worked the Coin problem as a right-tailed test and their calculators gave them a *p*-value of 0.04 and they knew that with a given alpha of 0.05 they needed to "Reject". Probably following a general instruction from their teacher, Ben went back to the formulation of the problem to find what they were rejecting. He then picked up the claim in the problem that the coin favored heads and wanted erroneously to reject this claim. This mistake took the pair on a long detour before they arrived at the correct final answer. Still, Ben was not happy with their answer, reasoning as follows:

> Ben: Because if we are rejecting H_o, H_o is .5 in here [refers to calculator] then we are rejecting that it is a fair coin, agreeing that it favors heads. Which is weird because no coin favors heads, it is always fifty-fifty [shakes his head in disbelief].

Rose and Sylvia worked the Coin Problem as a two-tailed test and their calculator gave them a *p*-value of 0.089. The pair concluded that they cannot reject H_0 that $p_0 = 0.5$. Contrary to Alex and Ben, Rose and Sylvia paid considerable attention to the sample proportion $\hat{p} = 0.62$ and did not like their statistical decision, with Sylvia saying: "But it is not a fair coin. Because they got it 31 times." Rose agrees: "Yes" and a lengthy discussion follows in which the students try to make sense of their counterintuitive answer. Below is an excerpt from this discussion.

> Sylvia: That's what's throwing me off. It is 62.
> Rose: That must be 62 here, and not .5
> Sylvia: But we already know it's 62. So why would that be a big deal? Either it's 62 or not, but 62 isn't fair.
> Rose: Right
> Sylvia: It has to be 50.

Of course, if Rose and Sylvia had worked the coin problem as a right-tailed test, they may have arrived at an answer that would have been more pleasing to them. As in the case with Alex and Ben, however, the interesting result here is the lack of inclusion of sampling

variability in the students' reasoning. If the students had been taking the sampling variability in account, they might not have been so puzzled over their answers.

Some students found difficulties with notations used for *p*-values. Zoe's and Vera's instructor used a particular notation in his class to stress that the *p*-values were conditional probabilities. For example, to denote the probability of getting a sample proportion of .62 or more if the hypothesized proportion was .5 he had the students write $P_{H_0}(\hat{p} > .62)$ = 0.048. Vera and Zoe dutifully used this notation for all their problems. Below is an exchange between the interviewer and Zoe about the *p*-value notation used in her class.

I: You had some nice notation in your work, where you had P and then it said a parenthesis and then it had *x*-bar or *p*-hat greater than something. So what does that stand for?

Zoe: Yes, I know what you are talking about. And I have no idea what it stands for. What I know is that when [the teacher] showed us the process that is how we wrote it out and that is how he expected it to be written out on the test or we would get docked points. I don't know what it means. It just gives us the answer and we had to write it out. The p-h-oh-thing, something with the H-O.[H_0]. Oh, there goes bye-bye [She makes a gesture with her hands showing how the abstractions elude her.]

Zoe's and her partner Vera's work on the answer sheets supports Zoe's claim that the conditional probability notation did not carry much meaning for the pair. For example, in the work on the *Tranquilizer problem*, both students wrote $P_{H_0}(\bar{x} \neq 6.8) = 0.0779$. Since the students did not understand the ideas behind the notation, they mimicked what they had seen for one-tailed tests when solving a two-tailed test problem. The resulting statement was faulty and suggests that the pair had a poor understanding of *p*-values. A correct statement would have been that the *p*-value was $2\ P_{H_0}(\bar{x} < 6.8) = 0.0779$ (cf. Rossman et al., 2002, p. 449). To consider the probability of having a sample mean that is not 6.8 as the students' notation suggested, is useless, and probably not even what the students had in mind.

One student, Ellie, consistently drew all her graphs with the *p*-values in the wrong position, i.e., the *p*-value was written on the wrong side of alpha. All the graphs illustrated were either two-tailed tests or right-tailed tests. Thus in the cases where the *p*-value was less than alpha, the *p*-value should have been listed to the right of alpha. Ellie, however, listed the *p*-value to the left of alpha. In the interview with Ellie it became clear that she had marked the alpha and *p*-values on the horizontal axis as if those values had been values on a real number line. Ellie did not think through how those values being represented by areas ordered them differently than, for example, *z*-values on graphical representations. When graphing on the number line, smaller numbers are graphed to the left of larger numbers. Alpha values and *p*-values are graphed as *areas* in the tail of the sampling distributions. Therefore, a *p*-value that is smaller than an alpha value should be listed to the right of that alpha value. It has not been uncommon for the first author to see students in her classes reverse the order of *p*-value and alpha value notations on sampling distribution graphs when they are just learning hypothesis testing. When Ellie made such a mistake two weeks after the relevant instruction was completed, however, it indicated a lack of understanding.

Ellie would not have been able to derive correct answers from her incorrect graphs if she had tried. Her partner Fran, however, preferred making the statistical decisions using algebraic arguments, such as "you reject H_0 if *p* is less than alpha". Ellie followed her partner's lead and the pair solved half of the problems correctly and did some good work on the other half.

On the subsequent problems Ellie made some comments about large *p*-values being "way over there", indicating that marking those *p*-values so far from the center of the graph did not make complete sense to her. Since her partner Fran used the algebraic rule to make her statistical decisions she only politely acknowledged Ellie's comments without reflecting much over them. By letting Fran taking the lead and the pair finishing the problems using the algebraic decision, the pair was able to solve 3 of the 6 problems correctly, but Ellie was intermittently bothered about not being able to match the answers with the graphs.

Students' procedural approach to *p*-values caused other problems. In particular, the students ran into difficulties when they tried to memorize steps of hypothesis testing without understanding the rationale for those steps. One example of such a step was instructor instructions to divide the probabilities given by the calculator by two, when working two-tailed tests. Mary and Nan divided all their *p*-values by two, independent of whether they were working a two-tailed test or not.

In the pair Alex and Ben, Ben seemed to have better understanding of the "dividing by two" issue. In the excerpt below, Alex and Ben discuss the issue of dividing by two.

Alex: When you don't use the visuals, you don't have to divide by two?
Ben: Yeah, you don't. Yes, you never really do. Only for the visual thing.
Alex: Only for the visual?
Ben: Yes. Because if you are going to divide alpha by half, then you have to divide *p* by half. But if *p* is already bigger than … then you already know it is going to be bigger. I don't think it's going to be bigger, though. So we don't have to divide by anything.

As Alex and Ben continued to work the problems, Alex seemed to increase his understanding of the hypothesis process by watching Ben draw the sampling distribution and divide the *p*-value in two when it was appropriate.

5. DISCUSSION

The Discussion section starts with reviewing the definitions of *p*-values that were given to the students in their classes, followed by a commentary on how the students related to those definitions. The discussion continues with a report on the role concept images played in the students' problem solving. A short comparison of the study's results to the common misconception pointed out by Falk (1986) follows. Most of the discussion then consists of suggestions on how to improve the teaching of hypothesis tests.

Half the students were given the following *concept definition* of *p*-value in their textbook: "For the distribution described by the null hypothesis, the P *value* is the smallest level of significance for which the observed sample statistic tells us to reject H_o." (Brase & Brase, 2003, p. 479) The other half of the students was given the following concept definition: "The *p*-value is the probability, assuming the null hypothesis to be true, of obtaining a test statistic at least as extreme as the one observed" (Rossman, Chance, & von Oehsen, 2001, p. 449). None of the students, however, could give a concept definition of the *p*-value without consulting their notes.

Instructor B, who provided his students with a concept definition close to the one used in the delMas et al. (2007) study, also tried to enforce the idea of *p*-value as a conditional probability by having students use notation showing this characteristic of the *p*-value. However, when the student using the notation was asked about it she said that she did not understand the meaning of it. No other students used it in their work. Thus, it does not seem

that any concept definition of *p*-values played a role in how the participating students reasoned about *p*-values.

With respect to *concept images*, graphing the sampling distribution and marking the *p*-value as area seemed to be of help to the students in this study. This result is consistent with Hong and O'Neil (1992), who showed that students who were instructed to graph the sampling distribution as part of their solution process did better than the ones who did not. The question remains though, why several students did not draw more detailed graphs, which might have helped them not only to make the correct statistical decision, but also would have added meaning to the solution process. As it was, comparing *p*-values with α-values did help students to make correct statistical decisions. This algebraic approach, however, might have been partly the reason why some of students transformed their correct statistical answer into an incorrect final answer to some of this study's hypothesis test problems.

It is difficult to decide to what degree the students in this study were subject to the misconception described in Falk's (1986) study. Falk wrote that students often, and researchers sometimes, confuse the conditional probabilities $P(H_o|R)$ and $P(R|H_o)$. Here, H_o stands for the event that the null hypothesis is true and R for the event of rejecting the null hypothesis and the statistical decision rule is based on $P(R|H_o)$.

One reason why the students' reasoning is hard to analyze in this respect has to do with students' frequent use of the word *it* as a placeholder in their answer to the interviewer's questions. (Recall that in the results section of this article, the word *it* was written with italics for emphasis in the transcripts.) An exchange between the partners Gus and Hal and the interviewer is typical. Gus stated first that the probability is slim to none for *it* to even happen so you reject *it*. When asked by the interviewer what "it" is, he and Hal answered, "The μ that you are testing." Hal also added, "The null hypothesis." This kind of reasoning was very typical in that there seems to be no trace of *conditional* probability thinking in the students' reasoning. Thus we cannot say that they exhibited Falk's misconception. On the other hand, they did not express the view that Falk considers correct either. That students do not explicitly consider the conditional part of the *p*-value definition might explain why students in the delMas et al. (2007) study did not see the contradiction in selecting both a correct and incorrect response to what a *p*-value is.

Falk (1986) suggested some alternative methods to hypothesis testing and discussed pros and cons for them. Mostly she suggested using confidence intervals instead of hypothesis tests. Kalinkowski et al. (2010) studied statistical cognition and found confidence interval better understood than hypothesis tests. There is also support for focusing on confidence interval rather than hypothesis testing in introductory statistics classes in delMas et al. (2007) study. In that study, the percentage of students being able to correctly interpret a confidence interval increased from 47.1% to 74.3% between the pretest and the posttest. This result compares quite favorably to the percentage of students being able to recognize an incorrect interpretation of *p*-value, which rose from 42.3% to only 52.7% between pretest and posttest.

A counterargument to omit teaching hypothesis testing in introductory statistics classes can be summarized in the title of the Falk and Greenbaum (1995) article, "Significance tests die hard. The amazing persistence of a probabilistic misconception." Hypothesis tests are ubiquitous in the social sciences and – as Kalinkowski et al. (2010) pointed out – even when confidence intervals are used, researchers often interpret their results in the language of hypothesis testing.

If we continue to teach hypothesis testing in introductory statistics classes, we need students to come away with a concept definition of *p*-value that makes sense to them. The confusion about the two different ways of using *p*-values to make the statistical decisions

highlighted by Cummings (2010) as well as by Haller and Krauss (2002) is important for statistics instructors to consider. The present authors find it hard to make a pedagogical recommendation of how to handle this problem. On one hand, articles such as Hubbard and Lindsay (2008) point to considerable problems in using *p*-values as a measure of evidence in hypothesis testing. Also, students in the present study clearly favored using comparisons of *p*-values with a preset α for their statistical decisions, despite being taught both methods. On the other hand, the practice of using *p*-values as a measure of evidence is used widely by researchers, which means that students will need to understand what those researchers are claiming. What is clear, though, is that if both methods are taught, students need to become aware that comparing p with α builds on deciding on an acceptable type I error before the experiment, while decisions based only on the size of *p*-values are data-driven.

Our study showed that the students' concept images of *p*-values lacked the critical part of conditional probability. As stated in the literature review, Tall and Vinner (1981) stress the role of earlier experiences when students create and use concept images. In our results, we saw students being able to assign meaning to small and large *p*-values, because the students can associate those to everyday expressions such as "I will probably be home tomorrow" or "the probability to win the Grand Prize is extremely small."

The students surely must have had earlier experiences of conditional probabilities as well. Unfortunately, those experiences were probably not been framed in a way that made it possible for students to connect them with the rather abstract formulations of *p*-values. Therefore, those experiences did not help the students incorporate conditional probability into their concept image of *p*-values.

How could statistics instructors make conditional probability part of the way students reason about *p*-values? Working with simulations of sampling distributions, as Lawton (2009) and other proponents of reform statistics recommend, may be a good start. The *Workshop Statistics* textbook (Rossman, Chance, & von Oehsen, 2001) was built more around simulation exercises than the *Understandable Statistics* textbook (Brase & Brase, 2003). It is therefore worth noticing that, in some cases, only students using the latter textbook completed the hypothesis procedure correctly but then answered the question in the problem incorrectly. Still, the fact that all the instructors in the study used simulation exercises did not prevent their students from having difficulties with hypothesis testing.

We suggest that instructors follow the recommendation of Pfannkuch, Seber, and Wild (2002) and teach probability with help of two-way tables. When computing probabilities from such tables, conditional probability becomes transparent. Then by introducing the notation of conditional probability in such a context, the notation would have some meaning.

As stated earlier, graphing the sampling distribution helped students to solve the hypothesis problems. Then, if the students have a better understanding of conditional probability, the instructor should stress the conditional probability role in the hypothesis testing process. More concretely, as in the case of testing a mean, the instructor might insist on students graphing the hypothesized distribution and marking the hypothesized mean on the horizontal axis to show the condition on which the test is done.

To have a conceptual understanding of inferential statistics is important for students, whether they are going to do research or just read research reports. It is also becoming more important for us if we want to be informed citizens. Without such understanding, we cannot properly judge many of the claims politicians make.

REFERENCES

Aquilonius, B. C. (2005). How do college students reason about hypothesis testing in introductory statistics courses? *Dissertation Abstracts International, 66*(02). UMI No. 31631059.

Brase, C. H., & Brase, C. P. (2003*). Understandable statistics: Concepts and methods* (7th ed.). Boston, MA: Houghton Mifflin.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education*, *17*(2).
[Online: http://www.amstat.org/publications/jse/v17n2/castrosotos.pdf]

Charmaz, K. (2002). Qualitative interviewing and grounded theory analysis. In J. F. Gubrium & J. S. Holstein (Eds.), *Handbook of interview research: Context and method.* Thousand Oaks, CA: Sage.

Cumming, G. (2010). Understanding, teaching and using p values. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8).* Voorburg, The Netherlands: International Statistical Institute.
[Online: http://iase-web.org/documents/papers/icots8/ICOTS8_8J4_CUMMING.pdf]

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistical Education Research Journal*, *6*(2), 28–58.
[Online: http://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf]

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: The MIT Press.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*, 83–96.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*, *5*(1), 75–98.

Gordon, S., Reid, A., & Petocz, P. (2010). Qualitative approaches in statistics education research. *Statistical Education Research Journal*, *9*(2), 2–6.
[Online: http://iase-web.org/documents/SERJ/SERJ9(2)_Editorial.pdf]

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*(1), 1–20.

Hong, E., & O'Neil Jr., H. F. (1992). Instructional strategies to help learners build relevant mental models in inferential statistics. *Journal of Educational Psychology*, *84*(2), 150–159.

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*(1), 69–88.

Kalinkowski, P., Lai, J., Fiddler, F., & Cumming, G. (2010). Qualitative research: An essential part of statistical cognition research. *Statistical Education Research Journal*, *9*(2), 22–34.
[Online: http://iase-web.org/documents/SERJ/SERJ9(2)_Kalinowski.pdf]

Lawton, L. (2009). An exercise for illustrating the logic of hypothesis testing. *Journal of Statistics Education*, *17*(2).
[Online: http://www.amstat.org/publications/jse/v17n2/lawton.pdf]

Marshall, C., & Rossman, G. B. (1999). *Designing qualitative research*. Thousand Oaks, CA: Sage Publications.

Pfannkuch, M., Seber, G.A.F., & Wild, C.J. (2002). Probability with less pain. *Teaching Statistics*, *24*(1), 24–30.

Rossman, A. J., Chance, B. L., & von Oehsen, J. B. (2002). *Workshop statistics: Discovery with data and the graphing calculator.* Emeryville, CA: Key College Publishing.

Schoenfeld, A. H. (1982, March). On the analysis of two-person problem-solving protocols. In J. M. Shaughnessy (Chair), *Investigations of children's thinking as they go about solving mathematical word problems.* Symposium presented at the annual meeting of the American Educational Research Association, New York.

Schoenfeld, A. H. (1985a). Making sense of "out loud" problem-solving protocols. *Journal of Mathematical Behavior*, *4*, 171–191.

Schoenfeld, A. H. (1985b). *Mathematical problem solving*. New York: Academic Press.

Schoenfeld, A. H. (2007). Method. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning: Vol. 1* (pp. 69–107). Charlotte, NC: NCTM & Information Age Publishing.

Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, *12*(2), 151–169.

BIRGIT AQUILONIUS
Barks väg 10 lgh 1203
17073 Solna
Sweden