# INTRODUCTORY STATISTICS COURSE TERTIARY STUDENTS' UNDERSTANDING OF P-VALUES

ROBYN REABURN
*University of Tasmania*
Robyn.Reaburn@utas.edu.au

## ABSTRACT

*This study aimed to gain knowledge of students' beliefs and difficulties in understanding p-values, and to use this knowledge to develop improved teaching programs. This study took place over four consecutive teaching semesters of a one-semester tertiary statistics unit. The study was cyclical, in that the results of each semester were used to inform the instructional design for the following semester. Over the semesters, the following instructional techniques were introduced: computer simulation, the introduction of hypothetical probabilistic reasoning using a familiar context, and the use of alternative representations. The students were also encouraged to write about their work. As the interventions progressed, a higher proportion of students successfully defined and used p-values in Null Hypothesis Testing procedures.*

**Keywords:** *Statistics education research, p-values, simulation, student understanding*

## 1. INTRODUCTION

This study examined students' problems in understanding *p*-values, and the results of an intervention that aimed to improve this understanding. Null Hypothesis Testing (NHT) is one of the main techniques in inferential statistics, yet previous research has shown that the concept of the *p*-value can be problematic for students (Batanero, 2008; Gliner, Leech & Morgan, 2002; Nickerson, 2000).

*P*-values have come about from the desire to estimate the likelihood that a sample was drawn from a population with a specified value for the population parameter. When a venous blood sample has been taken correctly, the sample will be like the blood in the rest of the venous system. In most sampling situations, however, it is extremely unlikely that a sample will be exactly representative of the population. If another sample were taken, it too is unlikely to be exactly representative of the population, and in addition, unlikely to be exactly like the first sample. Despite this, researchers know that the sample will in some way tend to resemble the population, and that it is still possible to make conclusions about the population, even if it is not possible to be absolutely certain about the accuracy of these conclusions.

One way around the problem of uncertainty is to perform NHT. With this process, a proposition (the *null hypothesis*) is made about a population parameter. A sample is then collected, the relevant sample statistic calculated, and a judgment is made as to how likely the sample statistic (or one even more extreme) would be if the proposition about the parameter were true. In the NHT process, this judgment is made by calculating a conditional probability, the probability of obtaining the sample with the given or more extreme statistic, if the population has the parameter proposed in the null hypothesis. It is this probability that is known as the *p*-value. One way to interpret this *p*-value is to compare it to a pre-set value. If this *p*-value is below the pre-set value, it is concluded that it is unlikely that the sample came from a population with the stated null hypothesis and the null hypothesis is *rejected*. If this *p*-value is above the pre-set value, then it is concluded that the sample could have come from a population with the proposed value and one *fails to reject* the null hypothesis.

Previous research shows that students of statistics can have problems understanding this process, and this lack of understanding can be undetected by their instructors because the students may follow the procedures accurately (Garfield & Ahlgren, 1988). It is only when questions are asked that require students to describe their reasoning that this lack of understanding is detected. The aim of this study was to gain knowledge of students' beliefs and difficulties in understanding *p*-values, and to use this knowledge to develop teaching programs to enhance student understandings of this concept. The research questions were: What are students' understandings of *p*-values? What misconceptions may they hold? And can teaching methods be developed to improve students' understandings?

## 1.1. LITERATURE REVIEW

A null hypothesis test starts with the statement of the null hypothesis containing the proposed value of the population parameter. Previous research shows that students may believe that this null hypothesis refers to both the sample and the population, and are therefore confused about NHT from the very start of the process (Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). It has also been found that students may carry out the procedures for NHT correctly, but then may misinterpret the results through lack of understanding of what rejecting and failing to reject the null hypothesis really indicates. This problem was investigated by Haller and Krauss (2002) who conducted a survey of staff and students, some of whom were statistics instructors, from the psychology departments of six universities. In this survey, an example of an independent samples *t*-test was given where the *p*-value was 0.01. Approximately 26% of the participants (including a small number of statistics methodology instructors) agreed with the statement: "You have found the probability of the null hypothesis being true." Approximately 69% of the participants (including approximately one third of the statistics methodology instructors) agreed with the statement: "You know, if you decide to reject the hypothesis, the probability that you are making the wrong decision." Those who agreed with this statement did not seem to be aware of the conditional nature of the probability the *p*-value represents. That is, the *p*-value is the probability of making the wrong decision if the null hypothesis is true.

The belief that the *p*-value is the probability that the null hypothesis is true appears to be a commonly held misconception. A related misconception is that 1-*P* is the probability that the alternative hypothesis is true. It may also be believed that rejecting a null hypothesis proves the underlying theory that predicted the rejection. It may also be believed that a low value for the *p*-value suggests that the results are replicable (Nickerson, 2000).

## 1.2. WHY USE P-VALUES?

The use of the null-hypothesis test is widespread and *p*-values are reported widely in the literature. The way a *p*-value is used differs and is the subject of debate (Cumming, 2010; Gliner, Leech, & Morgan, 2002; Hubbard & Lindsay, 2008). One way *p*-values can be used, attributed to Neyman and Pearson, is that a pre-existing *level of significance* is chosen, and the null hypothesis is rejected if the *p*-value is less than this level of significance. This form of analysis leads to the possible calculation of Type I and Type II error rates. An alternative (advocated by Fisher) is to look at the level of support a *p*-value gives to a null hypothesis. As the *p*-value decreases, the level of support given for the null hypothesis is also considered to decrease (Wagenmakers, 2007). Recently, however, the question has been asked: should *p*-values be used at all?

One tenet of a scientific experiment is that it should be replicable. Therefore, it would seem not unreasonable to assume that if an experiment should be repeatable then the *p*-value would also be replicable. Cumming (2010) has shown that in fact *p*-values vary much more from sample to sample than many researchers realise. Hubbard and Lindsay (2008) show that *p*-values can vary even with the same data, depending on the method of analysis chosen by the researcher and on whether the researcher has chosen a one- or two-tailed test.

Another problem with *p*-values is that they do not indicate the effect size. A small study with a large effect size can yield the same *p*-value as a large study with a small effect size (Hubbard & Lindsay, 2008; Wagenmakers, 2007). In addition, there is concern about the validity of the way *p*-values are calculated. Assuming the null hypothesis is true, a *p*-value is the probability of the observed data and the probability of more extreme data, yet these more extreme data are not actually observed. It is questionable whether decisions should be made on unobserved data (Hubbard & Lindsay, 2008).

It is for these reasons that it has been suggested that the results of scientific experiments should instead be presented as confidence interval estimates of the parameters. Confidence intervals have the advantage that they are in the same units as the point estimate, and make it easier for the reader to determine if an effect is important, rather than just if it is statistically significant. Of even more consequence is that confidence intervals give an idea of the precision of an estimate via the width of the interval. In addition, the width of the interval gives an idea of what the infinite set of possible results may look like (Cumming, 2010; Wagenmakers, 2007). The contrast between the variation in *p*-values and the variation in confidence intervals is graphically and amusingly illustrated by the "Dance of the *p*-values" (http://www.youtube.com/watch?v=ez4DgdurRPg).

## 1.3. GUIDED DISCOVERY LEARNING AND SIMULATION

Some of the interventions in this study are based on discovery learning. This style of learning is based on the idea that knowledge students find for themselves has more value than knowledge that has been merely explained (Klahr & Nigam, 2004). The aim of discovery learning is to engage students in activities that lead them to "discover" a principle. It has been found, however, that students do not always notice the principle the instructor intended them to discover. This may be because students may not notice regularities in the data. In addition, students may not change previously held beliefs even if the data contradict these beliefs (de Jong & van Jooligen, 1998; Mills, 2002). It has been suggested, therefore, that either guided discovery learning or a combination of guided discovery learning and direct instruction is needed (delMas, Garfield, & Chance, 1999; Lane & Peres, 2006). With guided discovery learning students are given a series of questions or activities that lead them to a predetermined goal (Lane & Peres, 2006).

One way guided discovery learning may be achieved in statistics classes is with the use of computer simulation. Computer simulation allows the drawing of a large number of samples from a hypothetical population or from a data set quickly. In addition, computer simulation can be used to model NHT procedures. Erickson (2006) demonstrates how this can be carried out for a 2-sample *t*-test for the difference in means. Once the difference in means is calculated, the data can be randomly jumbled into two groups of the same size as before. This jumbling can be repeated many times. The distribution of the difference in means can then be compared to the test statistic when the data were divided by the characteristic in question. Using this technique, the students can get a subjective feel for the likelihood of the test statistic under specific conditions compared to the statistic when groups are formed at random.

Computer simulation does not, however, automatically lead to the desired learning outcomes because students may not make the links that the instructor intended. Lipson (2002) found that students failed to make the link between a simulation and the theory, even after these students had appeared to carry out the simulation successfully. Also, delMas, Garfield, and Chance (1999) confirm that students may appear to carry out computer simulations successfully but still demonstrate a lack of understanding afterwards. Lane and Peres (2006) suggest that knowledge acquisition is improved if students are asked to make predictions about what will happen before a simulation, and compare the results to what they expected,  the "query first, answer later" method. This forces students to confront discrepancies between what they expect and what actually occurs and makes it more likely that they will change previously held beliefs (Garfield & Ahlgren, 1998; Hardiman, Pollatsek, & Well, 1986; Posner, Strike, Hewson, & Gertzog, 1982).

## 1.4. RESEARCH CONSTRAINTS ON THE TEACHING AND LEARNING UNIT

The debate over whether or not *p*-values should even be used was not introduced to the students in the teaching and learning unit described in this research. The students were, however, introduced to the connection between *p*-values and confidence intervals, in that a *p*-value of 0.05 or greater is obtained when the confidence interval includes the value proposed in the null hypothesis. The decision not to introduce this debate is due to the nature of the teaching and learning unit in which this research took place, the nature of the students, the time available to teach the content, and the constraints placed on the researcher. The unit is a first-year service unit in applied statistics for students of aquaculture, biomedical science, environmental science and sports science. The purpose of the unit is to introduce students to the principles of inferential statistics, sampling, experimental design, null hypothesis testing and confidence intervals, and to introduce simple statistical techniques such as one- and two-sample *t*-tests, chi-squared tests, analysis of variance, and linear regression. These topics are needed for the many students who will go on to study statistics further in their own areas of study. In addition, although there is a shift away from NHT to estimation, NHT is still in widespread use. The students need to be familiar with the concept of a *p*-value, not only for their own research but also to be able to read journal articles. Also, the unit is one semester (13 weeks) in duration, and therefore the content is limited. Whereas the methods of delivery could be altered by the researcher, the content was fixed and could not be altered. The expectations of the faculties whose students undertook this unit also constrained the choice of computer packages used for the simulations. Whereas there are several useful simulation applets on the internet, it was expected that students would receive extensive experience in the use of *Microsoft Excel*.

There were further constraints placed on the researcher because the researcher was also the instructor. Ethics considerations meant that the researcher did not know who had volunteered to participate in the research until after the students had received their official grades for the unit. This is expanded upon in the discussion.

## 2. METHODOLOGY

### 2.1. PARTICIPANTS

This study was part of a doctorate project (Reaburn, 2011) and was carried out over four consecutive teaching semesters of a one-semester introductory statistics unit at a tertiary institution. The subjects of the study for each of the four semesters were volunteers from this unit. The students were volunteers and there was considerable variation in the proportions of students who volunteered data over the period of the study. In the first semester there were 12 volunteers out of a possible 20, in the second semester there were 23 out of a possible 26, for the third semester there were 6 out of a possible 27, and for the fourth semester there were 12 out of a possible 26.

### 2.2. METHODOLOGY

The study was in the form of action research (Mills, 2007) in which the researcher was the lecturer of the unit. The study was cyclical, in that the results of each semester were used to inform the instructional design for the following semester. The first semester of the study, the pre-intervention semester, was used to gain knowledge of students' understanding of *p*-values before the teaching program was altered. In the teaching program prior to the interventions, the unit had been taught with each week having two lectures, one tutorial session, and one practical session where needed statistical calculations were demonstrated and practiced with *Microsoft Excel*. The lectures used direct instruction, that is, they were used to pass on, and explain the information that the students were required to know. The students were expected to take notes and to ask questions if required. The tutorial sessions were used to give the instructions for their formal assessments, and to answer questions from the formal notes that they were given in place of a textbook.

### 2.3. THE FIRST CYCLE OF THE INTERVENTION

Two new teaching and learning strategies were introduced into the program for the first intervention. The first strategy was to introduce students to simulation and the reasoning behind *p*-values in an informal way. Early in the semester, the students were asked to predict if the ratio of boys to girls would change if the one child policy in China was replaced with a "have children until a boy is born" policy. Using coins, each student determined the number of children born under the new conditions for 10 families and the results were collated. It was then suggested that the ratio of boys to girls would remain unchanged and a series of questions was asked. Would the results from these families be exactly like those for the entire population? Were these results consistent with the population ratio remaining unchanged? How far from the 1:1 ratio would these sample results have to be before they would be convinced that the ratio of boys to girls would change in the population?

The second strategy was to use computer simulation (using *Microsoft Excel*) for guided discovery learning. The topics introduced by simulation were sampling variability, the Central Limit Theorem, the null hypothesis testing procedure in several contexts, the meaning of *p*-values and the effect of measurement error in linear regression. For each simulation, the students were given a scenario, asked to make a prediction of the outcome, and then to test their predictions (delMas, Garfield, & Chance, 1999; Garfield & Ahlgren, 1988; Mills, 2002). Each simulation was carried out before the relevant material was introduced formally in a lecture.

To introduce students to the Central Limit Theorem, the random number generator in *Excel* was used to produce 500 samples from a normally distributed population. The means of these samples were calculated and these means were then plotted in a histogram. The students were then shown a graph of data that were uniformly distributed and asked to predict what the distribution of means would be. After this was completed, the random number generator function was used to produce 500 samples from a uniform distribution. Again the means were calculated and plotted into a histogram. The students then went through a similar process for a binomial distribution with a small sample size and then a larger sample size.

Simulations also introduced students to the two-sample *t*-test using a suggestion by Erickson (2006). The students were given data from a random sample of 40 Grade 12 students obtained from the *Census at School* website at the Australian Bureau of Statistics (www.abs.gov.au). In this sample, the data were grouped by gender and the mean height calculated for each group. The students were then asked whether they would conclude that the mean height for Grade 12 males was higher than the mean height for Grade 12 females, considering they were using a sample. Using *Excel*, the students repeatedly randomly assigned the data to two groups of the same size as before and calculated the difference in means each time. Their results were compiled, results plotted with a histogram, and then compared to the test statistic. The formal theory of the two-sample *t*-test was introduced in the first lecture after the simulation. In this lecture, the calculated *p*-value was compared to the observed proportion of samples where the difference in means equaled or exceeded the test statistic.

Computer simulation was also used to introduce students to a 1-sample *t*-test for proportions and the chi-squared test for independence, each time before the formal lecture on these topics. For these simulations, the students were also required to make a prediction before carrying out the simulation. The number of times the sample statistics equaled or exceeded the test-statistic were compared in each situation, and these values were then compared to the *p*-value obtained by formal methods. In addition, computer simulation was used to demonstrate how random errors in measurement may alter the slope of the line of best fit.

## 2.4. THE SECOND CYCLE OF THE INTERVENTION

In the second cycle of the intervention, the strategies from the previous semester were used with some additions. Because students are generally unfamiliar with the hypothetical probabilistic process used in hypothesis testing (Yilmaz, 1996), an example relating to the weather was used to help create this familiarity. In this example (Shaughnessy & Chance, 2005), the proposition was made that “It was hot outside.” The students were then told that everyone outside was wearing winter clothes. Because it is unlikely that people would be wearing winter clothes on a hot day, the proposition was considered to be incorrect. This problem was presented in a tabular format that was then used for all future hypothesis testing in the unit (Table 1). The formal terminology of hypothesis testing (null hypothesis, *p*-value) was not introduced until later in the semester.

*Table 1. An example of the probabilistic hypothetical process*

| | |
|---|---|
| Hypothesis | It is hot outside today |
| Data | When we look out of the window, everyone we see is wearing winter clothes (wool hats, gloves and coats). |
| What is the probability of seeing people wearing winter clothes if it is hot outside? | Very, very low. |
| Conclusion about hypothesis | It is incorrect. |

There is evidence to suggest that learning and understanding of mathematics is improved if students are encouraged to explain their reasoning (Confrey, 1990; Morgan, 2001; Pugalee, 2001). Therefore the other strategy used in this cycle was to encourage students to write about their understanding of the NHT process, with an emphasis on explaining the meaning of the appropriate *p*-value for each hypothesis test they were introduced to in the semester. If, for example, it was proposed that the null hypothesis was that the population mean is 200g, and the alternative hypothesis was that the population mean is less than 200g, the *p*-value would be the probability of obtaining a sample mean with the value observed or lower if the population mean were really 200g.

## 2.5. THE THIRD CYCLE OF THE INTERVENTION

The third cycle of the intervention included all the teaching and learning strategies from the previous two cycles with two additions. The first addition was the use of diagrammatic representations of the *p*-value. This was in line with the work of Brase (2009), Moreno and Duran (2004), and Ozgun-Koca (1998) who suggested that the use of multiple representations can improve students' learning and understanding. The second addition was the introduction of Popper's (1963) proposition about the nature of science.

An example of a *p*-value in diagrammatic form is shown in Figure 1. In this diagram, the distribution of the *t*-statistic is drawn and the value of the test statistic ($t$ = 3.6) added. From this diagram, the likelihood of the test statistic or one more extreme, if it came from a distribution centred on $t = 0$ can be estimated.  After the visual judgment is made it can be related to the actual *p*-value.

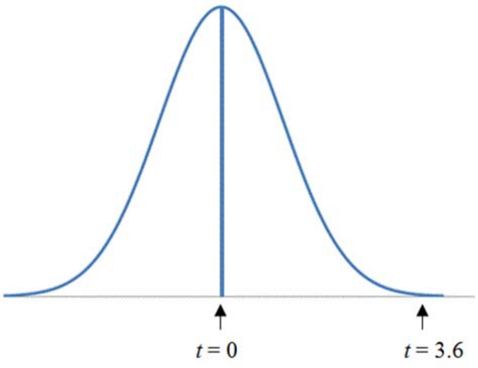

*Figure 1. A diagrammatic representation of the results of a* t*-test.*

During this cycle, there were several questions from the students about the way null hypotheses are written. Popper's work was used to find an explanation of the choice of the null hypothesis that the students could understand. Searching for an answer to what distinguishes science from other fields of knowledge, Popper (1963) proposed the criterion of "falsifiability, or refutability, or testability" (p. 37). By this criterion, Popper stated that "A theory that is not refutable by any conceivable event is non-scientific" (p. 36), so that "statements or systems of statements, in order to be ranked scientific, must be capable of conflicting with possible, or conceivable observations" (p. 39). This reasoning was introduced to the students with the statement that "all swans are white."  It is not possible to prove this statement true, no matter how many white swans are seen, because there is always the possibility that a swan will be found that is not white. In contrast, it only takes one observation of a non-white swan to disprove this statement. It was pointed out to the students that this reasoning is similar to that used in null hypothesis testing. In NHT, it is also acknowledged that it is not possible to prove the truth of the proposed hypothesis unless the entire population is examined, but it is possible to find evidence against a proposed hypothesis. If the sample obtained is highly unexpected from a population with the proposed parameter, then it is considered that the proposed hypothesis may be incorrect.

## 2.6. ASSESSING STUDENTS' UNDERSTANDING

To assess students' understanding, the students were given a test at the end of each teaching semester. Two questions from this test were used to assess the students' understanding of *p*-values:

> *Question 1*: A *p*-value of 0.98 indicates that the null hypothesis is almost certainly true. Is this statement correct? Give reasons for your answer.
>
> *Question 2*: In the test of a null hypothesis that a new drug produces the same expected benefit as the standard drug, versus the alternative hypothesis that the new drug produces a higher expected benefit than does the standard drug, a *p*-value of 0.01 is obtained. Explain what this result means to a patient who has read the result on the web but has no statistical training. Avoid all statistical jargon.

An ideal response to Question 1 would indicate that the *p*-value is the probability of the observed sample statistic or one more extreme if the null hypothesis were true. This *p*-value indicates that the observed sample statistic is very likely to occur if the proposed population parameter in the null hypothesis was correct but does not prove the truth of the null hypothesis. An ideal response to Question 2 would include a definition of the *p*-value in this context: If it were true that the new drug has the same expected benefit as the standard drug, then the probability of the new drug yielding results this good or even better would be only 1%.

The student responses were coded according to the level of reasoning shown in these responses (Table 2). For question 1, responses that were similar to the ideal response described in the previous paragraph received a code of "3." If the response stated that a high *p*-value could be obtained if the true situation was "close" to that of the null hypothesis the response received a code of "2." This code was also given to responses that stated that hypothesis tests only find evidence against the null hypotheses but cannot prove them true. A general statement that nothing is ever proved in inferential statistics received a code of "1."

For question 2, a response similar to the ideal described above received a code of "3." If a standard hypothesis test was carried out without further explanation a code of "2" was given. If it was stated that the new drug "works better" without further explanation a code of "1" was given. It was intended that these scores would be used for a Rasch analysis using the Partial Credit Model (Masters, 1982) but the questions did not fall onto a unidimensional scale that a Rasch analysis requires. These scores were used, however, to compare students' results over the period of the study.

*Table 2. The coding used to score answers to the p-value questions*

| Question | Score | Answer type |
|---|---|---|
| 1 | 3 | False – *p*-value is probability of sample value or a value more extreme if $H_o$ true, so sample likely if $H_o$ true |
| | 2 | False – Can only find evidence against $H_o$/True situation could be a value close to $H_o$ |
| | 1 | False – cannot prove true or untrue in inferential statistics |
| | 0 | True |
| 2 | 3 | If it were true that the new drug has the same expected benefit as the standard drug, then the probability of the results shown by the new drug is only 1%. |
| | 2 | Hypothesis test performed with no further explanation |
| | 1 | New drug will do "better" or similar |
| | 0 | Inappropriate use of *p*-value |

# 3. RESULTS

## 3.1. PRE-INTERVENTION

For Question 1, no student attempted to describe the meaning of the *p*-value. Most of the responses indicated that the students were aware that the process of hypothesis testing does not prove a hypothesis but they gave no further reasoning. Two students indicated the statement was true.

For Question 2, most of the students produced a formal hypothesis test with the null and alternative hypotheses and the *p*-value, even though the instructions stated to avoid statistical jargon. Four students attempted to try to explain the meaning of the *p*-value but were not correct. Their answers indicated that they believed that the *p*-value is the probability of being incorrect or that the *p*-value gives the rate of replication of the results, or that the *p*-value gives the probability of seeing a difference.

## 3.2. CYCLE 1 OF THE INTERVENTION

In the responses to Question 1, five students attempted to explain the *p*-value, but only one was correct. Sixteen students stated that the statement was false because all that happened was that there was insufficient evidence to reject the null hypothesis. For Question 2, seven students attempted to explain the meaning of the *p*-value by suggesting that the *p*-value gives one of the following: the probability of a difference between the treatments, the rate at which the two treatments will give equal benefit, or the probability that the null hypothesis is correct. One student stated "…our test value is very unlikely assuming the same benefit as the standard drug [therefore] we conclude that the new drug gives a greater expected benefit than the standard drug."

Of the other 14 students, nine used the expressions *p*-value, null hypothesis, or alternative hypothesis without further explanation and five students stated that the new drug works "better" with no other explanation.

## 3.3. CYCLE 2 OF THE INTERVENTION

In Question 1, four of the students agreed that the high *p*-value indicates that the null hypothesis is likely to be true. One stated that there was not enough evidence to prove the statement true, and the other answer was incomprehensible. For Question 2, the answers stated that the *p*-value gives the rate at which the new treatment will work better, the *p*-value gives the probability of the observation (partly correct) or a low *p*-value indicates that the alternative hypothesis is true.

For some students, the responses to Question 1 and 2 were inconsistent with each other, suggesting that these students were confused about the nature of the *p*-value. Further evidence of this confusion was shown by internal inconsistencies within some answers. For example, one student stated that the new drug works better 1% of the time, but then stated that the new drug worked better.

## 3.4. CYCLE 3 OF THE INTERVENTION

All twelve of the students in this intervention stated that the statement in Question 1 was false. Of these students eight students correctly explained the meaning of the *p*-value. One student defined the *p*-value as being the probability that the null hypothesis is true. The other students did not explain the meaning of the *p*-value, but stated merely that hypothesis tests do not give proof. In Question 2 three of the students correctly explained the meaning of the *p*-value. Six students stated that the evidence suggested that the new drug worked better than the old drug but did not explain the *p*-value. Two students used the null and alternative testing procedure without further explanation.

Figure 2 shows the distribution of the codes for the Question 1 over the four semesters of the study. In the first three semesters, no students received the highest score of "3," whereas in the last semester no student received the lowest score of "0." An increase in score was confirmed by the results of the Kruskal-Wallis test (Table 5). The difference in scores for the four semesters was significant and the highest mean rank was achieved in the third cycle of the intervention ($p = 0.015$). Table 3 shows the distribution of scores for Question 1 in raw numbers.

*Table 3. Number of students receiving each score for Question 1*

| | Score | | | | |
|---|---|---|---|---|---|
| **Semester** | 0 | 1 | 2 | 3 | Total |
| Pre-intervention (*n*=12) | 5 | 5 | 2 | 0 | 12 |
| Cycle 1 (*n*=23) | 8 | 14 | 2 | 1 | 23 |
| Cycle 2 (*n*=6) | 6 | 0 | 0 | 0 | 6 |
| Cycle 3 (*n*=12) | 0 | 2 | 2 | 8 | 12 |

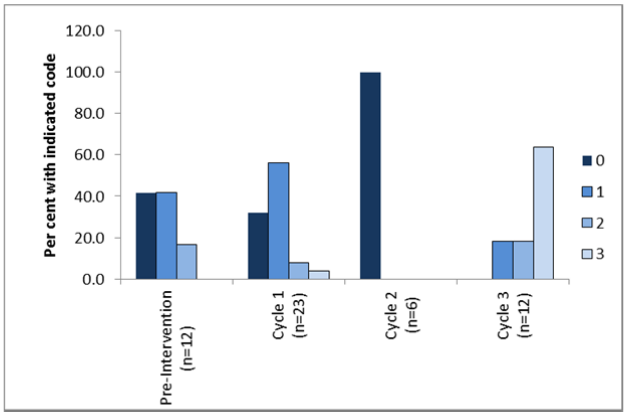

*Figure 2. Percent of students in each semester who received the indicated codes for Question 1.*

*Table 4. Number of students receiving each score for Question 2*

| | Score | | | | |
|---|---|---|---|---|---|
| **Semester** | 0 | 1 | 2 | 3 | Total |
| Pre-intervention (*n*=12) | 1 | 10 | 1 | 0 | 12 |
| Cycle 1 (*n*=23) | 4 | 16 | 5 | 0 | 25 |
| Cycle 2 (*n*=6) | 3 | 3 | 0 | 0 | 6 |
| Cycle 3 (*n*=12) | 0 | 6 | 2 | 3 | 11 |

Table 4 shows the distribution of scores in raw numbers. An increase in code was confirmed by the results of the Kruskal-Wallis test (Table 5). The difference in mean rank scores for the four semesters was significant and the highest mean rank was achieved in the third cycle of the intervention ($p < .001$) (Table 5). Figure 3 shows the distribution by proportion (percent) of the codes given to the answers Question 2 by the students over the study and demonstrates that the proportion of students who correctly answered this question increased over the four semesters of the study.

*Table 5. Mean ranked scores for the p-value questions*

| | Mean Rank | |
|---|---|---|
| Semester | Question 1 | Question 2 |
| Pre-intervention (*n*=12) | 22.14 | 26.00 |
| Cycle 1 (*n*=23) | 26.04 | 26.86 |
| Cycle 2 (*n*=6) | 13.33 | 5.25 |
| Cycle 3 (*n*=12) | 42.54 | 36.33 |

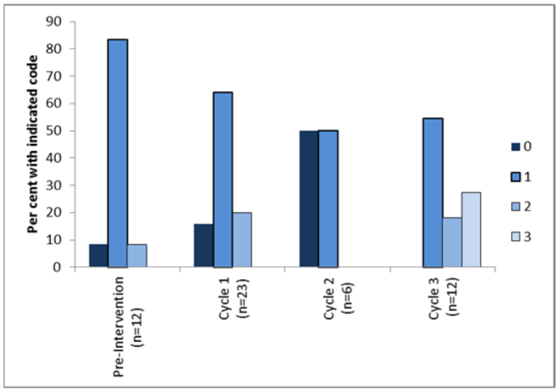

*Figure 3. Percent of students who received the indicated codes for Question 2 of the test.*

## 4. DISCUSSION

One limitation of this research was the variation in the participation rate each semester, as noted in Section 2.1. The researcher was the instructor and assessor for the unit described here, and therefore it was considered essential that students should have no opportunity to suspect that their grade could be influenced by their agreement or lack of agreement to participate. Ethics approval was granted only on the conditions that the researcher had no knowledge of who participated until after the students had received their grades, and that another person handed out and collected any materials that related to the research. As a result, it is not known why there was such variation in the participation rate from one semester to another and it was not possible to ask.

Even with these constraints the researcher believes that this study showed that using a combination of teaching methods other than that of didactic teaching will help in students' understanding of *p*-values. It is believed that the improvement is due to no single strategy, but rather the combination of computer simulation, writing about mathematics, finding a way to make the unfamiliar familiar, and relating the Null Hypothesis Test to Popper's ideas of the scientific method.

Computer simulation allowed students to "discover" some of the statistical principles for themselves. In particular, the "query first, answer later" method introduced an element of surprise into their statistical learning and increased students' interest. Such simulations would have been tedious before the introduction of spreadsheet programs such as *Excel*. It is the ability of such spreadsheets to "resample" and to recalculate statistics such as the mean for a large number of cases instantaneously that makes these spreadsheets so valuable. There are numerous examples of easy to use simulations to develop understanding of statistical concepts on the internet that do not require the setting up of the simulations in *Microsoft Excel*, but this program was used for all the simulations for reasons explained in the introduction. The applet developed at the University of Auckland (http://www.socr.ucla.edu/applets.dir/samplingdistributionapplet.html) is particularly useful as it shows the individual points in the sample, then calculates the mean and places this value on a number line. Two other simulations are also useful because although they do not show the individual sample points, they show the distribution of the means for differing population distributions, and are easily adjusted for different sample sizes. These are the simulations from the Computer-Assisted Statistics textbooks (http://cast.massey.ac.nz/collection_public.html) and from Learning by Simulations (http://www.vias.org/simulations/simusoft_cenlimit.html).

The students appeared to find the connection between Popper's ideas of falsification and the writing of null hypotheses interesting. It was clear that some students had not been previously exposed to questions as to what makes a scientific proof, and therefore had not considered that repeated observations do not prove scientific statements. With Popper's suggestions in mind, students were able to write null hypotheses more intentionally, instead of just by rote.

Some considerable time was spent by the researcher in investigating how to make the probabilistic, hypothetical thinking involved in NHT easier to understand; to make the unfamiliar familiar. The discovery of the "It is hot outside" problem from Shaugnessy and Chance (2005) was considered a breakthrough. The students easily understood this problem and, in general, could use this problem as a template for their later work where *p*-values were involved. It gave them a simple problem to understand that formed a basis for their later work as it became more complex.

The representation of the hypothesis tests in visual form (Figure 1) enabled students to make connections between the probability distribution, what sample statistics would be likely given the population parameter, the likelihood of the test statistic given the parameter, and the numerical result for the *p*-value.

At the end of the study there were still difficulties with students trying to learn a rule and then misapplying it; this has also been noted by Chance, delMas and Garfield (2004). These students stated that if the *p*-value is below 0.05, then the null hypothesis is not rejected. Another misconception was that the *p*-value is the probability of being incorrect (also observed by Gliner, Leech, & Morgan, 2002). In addition, there was one other misconception that was partly correct, that the *p*-value gives the probability of the observation. This is a simpler interpretation than the correct definition of the *p*-value, and appears to make sense to the students.

The most persistent and common misconception was that the *p*-value is the probability that the null hypothesis is true. This misconception has also been observed by Gliner, Leech and Morgan (2002). Once the instructor realised that some students had this misconception, she informed them of their error. This latter strategy, however, appeared to make no difference. This is a simpler interpretation than the correct definition of the *p*-value and appears to make sense to the students with the result it is tenaciously held. This misconception demonstrates that students may be interpreting new material in the light of what they already know, even if it is not accurate. To make conceptual change takes effort, and students will not make this effort unless they see a good reason to do so (Posner, Strike, Hewson, & Gertzog, 1982). Therefore instructors need to be aware of the possible misconceptions students may develop and produce teaching and learning activities to avoid these occurring.

From an instructor's viewpoint, the use of computer simulation appeared to result in students being more engaged in their work and being able to understand some areas of statistics that had been difficult with previous methods of instruction. The element of surprise in the result of the simulation to introduce the Central Limit Theorem resulted in students being much more likely to remember that sample means form a normal distribution. After an exercise in linear regression where the "errors" were simulated the students appeared to cope much more easily with the idea that there can be a non-significant relationship between the variables even when there is a non-zero gradient. Because the students had become familiar with visual representations of *p*-values, the students also had little difficulty in coming to terms with the principles of Type I and Type II errors, and the idea that a distribution of the "true" situation may overlap with the distribution of the proposed situation. The questions on the test that asked students to explain the consequence of a Type I error in a particular context were usually done well.

This study provides a basis for developing instruction to assist in students' understanding of *p*-values. It also suggests possibly fruitful areas for further research, such as why students come to believe that the *p*-value is the probability of the null hypothesis being true.

## ACKNOWLEDGMENT

I thank Emerita Professor Jane Watson for her helpful advice in the writing of this article.

## REFERENCES

Batanero, C. (2004). Statistics education as a field for research and practice. In M. Niss (Ed.), *Proceedings of the 10th International Congress on Mathematical Education*, July 4–11, 2004. Copenhagen: IMFUFA, Roskilde University.

Brase, G. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, *23*(3), 369–381.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dodrecht, The Netherlands: Kluwer Academic Press.

Confrey, J. (1990). What constructivism implies for teaching. In R. Davis, C. Maher, & N. Noddings (Eds.), *Constructivist views on the teaching and learning of mathematics* (pp. 107–210). Reston, VA: National Council of Teachers of Mathematics.

Cumming, G. (2010). Understanding, teaching and using *P* values. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the 8th International Conference on Teaching Statistics, Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistics Institute.
[ Online: http://iase-web.org/documents/papers/icots8/ICOTS8_8J4_CUMMING.pdf ]

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, *7*(3).
[ Online: http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm ]

de Jong, T., & van Jooligen, W. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179–201.

Erickson, T. (2006). Using simulation to learn about inference. In B. Phillips (Ed.), *Developing a statistically literate society. Proceedings of the 7th International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg, The Netherlands: International Statistics Institute.
[ Online: http://iase-web.org/documents/papers/icots7/7G2_ERIC.pdf ]

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, *19*(1), 44–63.

Gliner, J., Leech, N., & Morgan, G. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education*, *71*(1), 83–92.

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

Hardiman, P., Pollatsek, A., & Well, A. (1986). Learning to understand the balance beam. *Cognition and Instruction*, *3*(1), 63–86.

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

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction. *Psychological Science*, *15*(10), 661–667.

Lane, D., & Peres, S. (2006). Interactive simulations in the teaching of statistics: Promise and pitfalls. In B. Phillips (Ed.), *Developing a statistically literate society. Proceedings of the 7th International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg, The Netherlands: International Statistics Institute.
[ Online: http://iase-web.org/documents/papers/icots7/7D1_LANE.pdf ]

Lipson, K. (2002). The role of computer based technology in developing reasoning of the concept of sampling distribution. In B. Phillips (Ed.), *Developing a statistically literate society. Proceedings of the 6th International Conference on Teaching Statistics*, Cape Town, South Africa. Voorburg, The Netherlands: International Statistical Institute.
[ Online: http://iase-web.org/documents/papers/icots6/6c1_lips.pdf ]

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Mills, G. (2007). *Action research: A guide for the teacher researcher*. Upper Saddle River, NJ: Merrill Prentice Hall.

Mills, J. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, *10*(1).
[ Online: http://www.amstat.org/publications/jse/v10n1/mills.html ]

Moreno, R., & Duran, R. (2004). Do multiple representations need explanations? The role of verbal guidance and individual differences in multimedia mathematics learning. *Journal of Educational Psychology*, *96*(3), 492–503.

Morgan, C. (2001). The place of pupil writing in learning, teaching and assessing mathematics. In P. Gates (Ed.), *Issues in mathematics teaching* (pp. 232–244). New York: Routledge Falmer.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Özgün-Koca, S. (1998). *Students' use of representations in mathematics education.* Paper presented at the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Raleigh, NC.

Popper, K. (1963). *Conjectures and refutations*: *The growth of scientific knowledge*. London: Routledge and Kegan Paul.

Posner, G., Strike, K., Hewson, P., & Gertzog, W. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Pugalee, D. (2001). Writing, mathematics, and metacognition: Looking for connections through students' work in mathematical problem solving. *School Science and Mathematics*, *101*(5), 236–245.

Reaburn, R. (2011). *Students' understanding of statistical inference: Implications for teaching* (unpublished doctoral thesis). Tasmania: University of Tasmania.

Shaughnessy, J., & Chance, B. (2005). *Statistical questions from the classroom.* Reston, VA: National Council of Teachers of Mathematics.

Sotos, A., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98–113.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of *P* values. *Psychonomic Bulletin and Review*, *14*(5), 779–804.

Yilmaz, M. (1996). The challenge of teaching statistics to non-specialists. *Journal of Statistics Education*, *4*(1).
[ Online: http://www.amstat.org/publications/jse/v4n1/yilmaz.html ]

ROBYN REABURN
University of Tasmania
Faculty of Education
Locked bag 1307
Launceston
Tasmania
Australia, 7250