# Dynamic Visualizations and the Randomization Test

## 1. INTRODUCTION

For many years, the backbone of statistical inference in most undergraduate introductory statistics courses has comprised the normal distribution, the Central Limit Theorem, and sampling distributions of estimators. However, research evidence suggests that the theoretical and mathematical procedures involved in the inferential process have created a barrier to student understanding. It has long been noted that students have difficulty in following the unfamiliar logic associated with statistical inference – logic that is rooted in normal-based theory. The majority of students who come through traditional introductory courses in statistics fail to gain a true understanding of statistical inference (Jones, Lipson, & Phillips, 1994; delMas, Garfield, & Chance, 1999; Saldanha & Thompson, 2002). In fact, many who teach these courses struggle themselves to understand fully the intricacies involved in the process of statistical inference (Thompson, Liu, & Saldanha, 2007). In today's world, with advances in computing power, we no longer need to rely on normal-based theory to develop students' understanding of statistical inference. With this in mind, many statistics education researchers (e.g., Cobb, 2007) believe that the way forward is to change the way in which the statistical reasoning process is taught.

Motivated by Cobb's (2007) challenge to place the logic of inference at the center of the introductory statistics curriculum, and aligned with recent moves in statistical practice (Hesterberg, Moore, Monaghan, Clipson, & Epstein, 2009), we have taken steps to introduce the randomization and bootstrapping methods as core parts of our curriculum and to develop new visual thinking tools for students (Pfannkuch, et al., 2011). The appeal of both the randomization and the bootstrapping methods is that they are logical, accessible, and conveniently lend themselves to visual processes, which we conjecture may assist students with their understanding of statistical inference. They also do away with the need for distributional assumptions, and can be applied to many different situations. In this paper, we focus on our learning trajectory for the randomization method and we present pilot study findings about student learning outcomes.

## 2. LITERATURE REVIEW

Statistical inference is the term given to procedures that are used to draw conclusions about the world based on data. The three philosophies that underlie statistical inference are Bayesian, Neyman-Pearson and Fisherian. The Bayesian approach has as its core the concept of subjective probabilities and personal decision-making, and will not be dealt with in this paper since it is outside the scope of our curriculum and discussion. Both the Neyman-Pearson and the Fisherian approaches utilise objective measures of probability in order to come to conclusions. In simple terms, the Fisherian framework involves the setting up of a null hypothesis, usually one of no effect, with evidence being gathered against this null model. The Neyman-Pearson decision theoretic framework involves the setting up of competing hypotheses; the null hypothesis, and the research (or alternative) hypothesis. A decision is then made, based on the observed data, as to which hypothesis

to accept (Rossman, 2008). One consequence of this difference is that there is only one type of error defined in the Fisherian approach; rejecting the null hypothesis when it is true (i.e. a Type I error), while there are two errors associated with the Neyman-Pearson approach: a Type I error, and rejecting the alternative hypothesis when it is true (i.e. a Type II error) (Gigerenzer, 1993). For many, having two approaches to hypothesis testing introduces problems in deciding which method to adopt (Lenhard, 2006). One result is that many researchers "most commonly adopt a hybrid approach, combining aspects of both Fisherian inference and Neyman-Pearson decision-making to statistical hypothesis testing" (Quinn & Keough, 2001, p. 39).

Hypothesis testing, regardless of whether it is set within the confines of the Fisherian framework or the Neyman-Pearson decision theoretic framework or a combination of the two, is one of the most difficult topics for students and researchers to understand (Jones, Lipson, & Phillips, 1994). Many researchers (e.g., Carver, 1978; Cohen, 1994; Daniel, 1997; Falk & Greenbaum, 1995; Haller & Krauss, 2002; Hurlbert & Lombardi, 2009; Johnson, 1999; Mulaik, Raju, & Harshman, 1997; Nickerson, 2004; Schmidt, 1996) have documented persistent misconceptions related to hypothesis testing. These misconceptions include, but are not limited to:

- regarding a *p*-value as the probability that the research results are due to chance;
- considering a *p*-value as the probability that the null hypothesis is true (given the data) rather than the probability of the data (assuming that the null hypothesis is true);
- believing that the size of a *p*-value is an indicator of the size of any difference or relationship;
- concluding that ruling out a null hypothesis at a particular level of significance, say $\alpha$, means that the research hypothesis has a probability of 1-$\alpha$ of being true;
- interpreting a statistically significant result as practically important;
- accepting the null hypothesis if the *p*-value is considered to be 'large'.

Acceptance of the null hypothesis on the basis of a large *p*-value is a logical misconception, and arises from a misapplication of deductive syllogistic reasoning (Cohen, 1994). Most people are not attuned to thinking and reasoning probabilistically, and therefore apply a deterministic form of reasoning to such situations (Pfannkuch, et al., 2011). Liu and Thompson (2009, p. 127), through an analysis of some teachers' thought processes, believed that the conceptual obstacles were "rooted in their non-stochastic conceptions of probability and in their lack of understanding of the logic of indirect argument." That is, the teachers did not have a conception of distribution from which they could determine the *unusualness* of an observation, and they had a hidden belief that "rejecting a null hypothesis means to prove it wrong" (p. 142). Other researchers have also noted the difficulties associated with the indirect argument used in hypothesis testing (e.g., Nickerson, 2004; Rossman, 2008). In addition to these misconceptions is the fact that classical presentation and handling of statistical inference in the undergraduate curriculum relies on mathematical theory. Concepts associated with the development of a sampling distribution for any test statistic create obstacles in the path of understanding for all but the most mathematically inclined students (Chance, delMas, & Garfield, 2005; Rubin, Bruce, & Tenney, 1990).

After 20 years of attempting to improve students' hypothesis testing reasoning with little progress (Meletiou-Mavrotheris, Lee, & Fouladi, 2007) there appears to be a consensus that a new paradigm should be considered. The new learning emphasis should be on the reasoning and logic underpinning inference or hypothesis testing and that the mathematical procedures and manipulation of symbols of normal-based hypothesis testing should be replaced with the randomization method, which seems to be more conceptually accessible both visually and verbally (Cobb, 2007; Gould, Davis, Patel, & Esfandiari, 2010; Rossman, Chance, Cobb, & Holcomb, 2008; Tintle, VandenStoep, Holmes, Quisenberry, & Swanson, 2011).

Randomization (or permutation) tests have a long history, largely attributable to works by Fisher (1925) and Pitman (1937). The main premise of the randomization method is that the exact distribution of any test statistic specified under a null hypothesis can be obtained by reshuffling the original data many times. This is in contrast to the majority of other statistical tests which rely on asymptotic approximations which tend towards an exact solution only when sample sizes grow infinitely large. Fisher's exact test, used when analysing categorical data in contingency tables, is an example of a randomization (or exact) test, since the distribution of the test statistic is exactly hypergeometric if the marginal counts of the contingency table are fixed (Agresti, 1992). Karl Pearson's chi-square test provides an approximation to Fisher's exact test, although it will be inadequate if the sample size is small, or if cell counts under the null hypothesis are low. The beauty of Fisher's exact test is that it can be performed quite easily when dealing with a two-by-two table, with computer algorithms capable of dealing with data having more than two categories. Pitman demonstrated that randomization tests could be extended beyond the categorical data example originally described by Fisher, with other statisticians following his lead (Ludbrook & Dudley, 1998).

Recent exploratory research has used the randomization method to teach about both comparative observational and experimental studies. Preliminary results show promise, with Tintle et al. (2011) and Tintle, Topliff, Vanderstoep, Holmes, & Swanson (2012) reporting that students learned and retained significantly more about statistical inference using the randomization method while Gould et al. (2010) cautioned that new reasoning misconceptions could appear such as students believing the re-randomization distribution was the observed data distribution.

The randomization method, moreover, can be mediated through visual representations, which allow concepts to become more accessible to students. According to Clark and Paivio (1991), student understanding can be enhanced by the addition of visual representations and that encouraging students to generate mental images improves their learning. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) include the role of technology as an important tool to develop conceptual understanding (Aliaga, Cuff, Garfield, Gould, Lock, Moore, Rossman, Stephenson, Utts, Velleman & Witmer, 2005). A central contribution of technology to student learning is that it enables students to link multiple representations − visual, symbolic, and numeric − and it facilitates understanding through promoting a visualization approach to learning (Sacristan, Calder, Rojano, Santos-Trigo, Friedlander, & Meissner, 2010). Several researchers have found that the introduction of technology, through computer simulation activities, has enhanced students' understanding, although the improvement was modest (Hodgson, 1996; delMas, Garfield, & Chance, 1999). Dynamic software can allow students to analyze directly the behavior of a phenomenon, to visualize statistical

processes in ways that were not previously possible such as viewing a process as it develops rather than analyzing it from the end result. Exposure to such processes "can help develop the abilities and intuitive thinking that can enhance powerful mental conceptualizations" (Sacristan et al., 2010). Such representational infrastructure allows access to statistical concepts previously considered too advanced for students, as mastery of algebraic representations is not a prerequisite. The computer is able to take on the lower level tasks, such as performing many calculations, whilst the student can attend to the higher level tasks of applying the logic of statistical inference to the problem at hand (Jones, Lipson, & Phillips, 1994). Hence to obtain the full potential for learning inferential reasoning via the randomization method the teaching approach should incorporate dynamic visualizations. The role of hands-on activities as a tool to assist in the development of the inferential argument is also a major consideration. Some researchers argue that technology-driven simulations are more effective when presented in conjunction with appropriate hands-on activities (delMas, Garfield, & Chance, 1999; Lunsford, Rowell, & Goodson-Espy, 2006; Hesterberg, 2006; Pfaff & Weinberg, 2009; Seier, 2010).

# 3. OUR RESEARCH

Reflecting on the research literature and the requirements and constraints of our curricula we have chosen a different learning pathway from other researchers for beginning students. Unlike others (Tintle et al. (2011), (Tintle, Topliff, Vanderstoep, Holmes, & Swanson, (2012)), we decided not to use hypothesis testing structures and language; instead we generated a more natural form of argumentation based on the Fisherian framework. We limited the learning situations to comparative experiments in order to link how the randomization method mimics the data production process, which in turn determines the type of inference that can be drawn, that is, causal (Cobb, 2007). We are not using sample and population ideas for experiments, as is the case in traditional normal-based methods, and a matter which Cobb (2007) described as sleight of hand because in reality experiments do not take random samples from populations. We developed novel free dynamic visualization software for the randomization method to use as both a teaching and analysis tool. Our learning pathway is premised on the fact that for Year 13 (final school year) students it is sufficient for the analyses they are required to perform and that our undergraduate introductory statistics students, because of client department demands, still need to do normal-based inference. In time at the university level we hope to develop a fully-fledged randomization and bootstrapping curriculum. Our research question for the pilot study, the focus of this paper, is: what problems are inherent in the randomization method and our learning trajectories based on an analysis of students' learning and reasoning?

# 4. METHOD

A collaborative research project team of 33 members is involved in the development of these innovative approaches to teaching statistical inference. The team consists of two education researchers, two resource developers, a statistical software conceptual developer, eight university lecturers, fourteen secondary school teachers, five professional development facilitators, and one international advisor. Using design

research principles (Hjalmarson & Lesh, 2008), the development process involves two research cycles with four phases: (1) from an identified problematic situation, understand and define the conceptual foundations of inference; (2) development of new resource materials and dynamic visualizations; (3) implementation with Year 13 and university introductory statistics students; and (4) retrospective analysis followed by modification of teaching materials. The focus of design research is to support and engineer new types of reasoning and thinking in response to problematic situations. As well as being pragmatic through producing an educational product that can be used by teachers, design research can also lead to new educational theories and areas of research (Bakker, 2004).

In the first research cycle a pilot study was conducted with ten students, five from year 13 and five from university. All of the year 13 students had good grades at this level. Of the five university undergraduates, one student had studied a variety of third-year statistics papers and two had taken our traditional introductory statistics course. The remaining two undergraduate pilot study students had not taken any statistics papers while at university. One had good grades at the year 13 school level and one had minimal passes at the year 13 school level. A one-day teaching session was conducted with half of the day devoted to the randomization method, and the other half to the bootstrapping method. Using a mixed-methods approach, data collected for this pilot study were: student pre- and post-tests and interviews, student post-task interviews, video of teaching implementation, and reflections and observations of the project team. A thematic qualitative data analysis using NVivo 9 software (QSR International, 2010) was conducted on the student interviews (Braun & Clarke, 2006), while numbers of students who responded to multi-choice and true/false questions were recorded. In the second research cycle data will be collected from about 3000 students.

# 5. PILOT STUDY DESCRIPTION AND RESULTS

The purpose of the pilot study was to detect problems in the pre- and post-tests, learning trajectories and software. We will describe some of the issues that arose in the pilot study pre-test, during the implementation of the teaching unit, and in the post-test and post-task. For the implementation of the teaching unit we will elaborate on some of the learning experiences and language we used.

## 5.1. Pilot Study Pre-Test

Prior to the teaching session, the students completed the pre-test and were then interviewed in order to probe their reasoning behind their responses. The randomization section of the pre-test described a Fish Oil and Blood Pressure study (Knapp & FitzGerald, 1989) investigating whether a fish oil diet produced greater reductions in blood pressure than a regular oil diet. The Fish Oil and Blood Pressure study participants, 14 male volunteers with high blood pressure, were randomly assigned to the treatment group (fish oil diet) and the control group (regular oil diet). Data were provided numerically and graphically with two questions, one probing the students' understanding of random assignment and another aimed at eliciting the students' initial response in providing two main possible explanations for the observed difference in blood pressure reductions between the two groups (Figure 1).

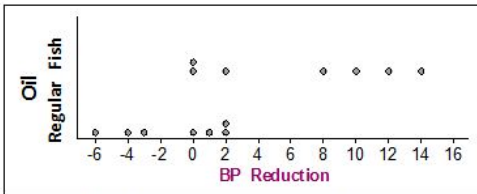



**Figure 1**. Dotplots of reductions in blood pressure

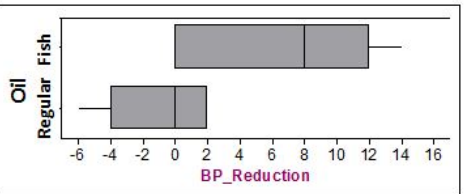



**Figure 2**. Box plots of reductions in blood pressure

8. What was the purpose of the random assignment of the 14 male volunteers to one of the two groups? Which ONE of the following statements gives the best response to this question?

   A. To increase the accuracy of the research results.
   B. To ensure that all male participants with high blood pressure had an equal chance of being selected for the study.
   C. To reduce the amount of sampling error.
   D. To produce treatment groups with similar characteristics.
   E. To prevent skewness in the results.

Figure 1. Part of one pilot study pre-test question, Question 8

Question 8 in Figure 1 was adapted from the Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS) (delMas, Garfield, Ooms, & Chance, 2007). In a large scale pre- and post-test situation, delMas et al. observed poor performance in understanding the purpose of random assignment with 8.5% and 12.3% answering correctly in the pre- and post-tests respectively. In our pilot study pre-test, three students selected the correct response, namely *to produce treatment groups with similar characteristics*. Another three selected the response *to increase the accuracy of the research results.* However, upon questioning in the interview session, all of these students talked about avoiding bias and wanting comparisons to be fair. Even though they selected an incorrect response, they used words and reasoning that indicated an appreciation behind the reason for random allocation to groups with statements such as:

S1: … so you didn't take the seven youngest to do the fish oil or … seven of the same ethnic group

Therefore the problem seems to be their interpretation of the language of the multi-choice options rather than their reasoning. Several students, however, were confused between random assignment and random selection. Three selected the option *to ensure that all male participants with high blood pressure had an equal chance of being selected for the study* with one student stating:

S2: … when they have all the people who volunteered with high blood pressure, they chose a random 14

Such a statement could be true but again it seems that knowing and understanding the language of statisticians is very important.

The next question asked for two main possible explanations for the observed difference in blood pressure reduction between the two groups. All but two of the pilot study students were able to state that one possible explanation for the observed difference in blood pressure reduction was that the fish oil treatment was effective. We were particularly

interested to find out if the students would suggest *chance* as another possible explanation. Attempts at formulating a second explanation ranged from statements about the fish oil study participants knowing they were being treated with fish oil and therefore eating more healthily, to psychological effects or biological properties having an impact on blood pressure, and some partial chance ideas about who just happened to be in which group. For example:

S3: The participants in the fish oil group had a lifestyle better suited to reducing blood pressure than the regular oil group

Results from the pre-test heightened our awareness that we would need to think carefully about the terminology that we used when employing the randomization method. In addition, we would need to be mindful with the development of the chance-alone explanation since none of the students articulated this as a possible explanation for the observed difference.

## 5.2. Pilot Study Teaching Session

Since seven of the students had not previously been exposed to designing comparative experiments, the first activity in the teaching session began with the instructor stating that she believed that some types of words were easier to recall than other types of words. By being careful to state that she was using a convenience sample of volunteers, that is, the pilot study students, she then explained that in order to investigate her hunch she wished to carry out a memory test. Her plan was that each person in the room (which included the students and the researchers) would be allocated to one of two groups, one of which would be given a list of words of one type to memorize, and the other to be given a list of words of another type to memorize. She then suggested that everyone over the age of 18 would be put into the first group, while everyone under the age of 18 would be put into the second group. At the outset nobody disputed this suggestion, which surprised us. Once challenged by the instructor as to whether or not there were any problems with her plan, a few suggestions were made such as "… *the groups are not representative of the population*…", "… *not enough people*…", and "… *the numbers won't be even*…". Having dealt with these suggestions, and with no comments on bias forthcoming, the instructor then continued by asking the students to consider males being given a list of words of one type to memorize, and females being given a list of words of the other type to memorize. Very quickly, someone mentioned that there would be bias, and that a comparison between the number of words recalled between each group would not be fair. Finally one student stated that there would need to be random assignment to each group so that a fair comparison could be made. A link was then made to the original approach of allocating students to each group on the basis of their age, and students appeared to realize that it would not be reasonable. We were surprised that nobody was critical of the original plan given their responses to the pre-test.

The second activity involved an example designed to demonstrate statistical argumentation in action in everyday life (Vickers, 2010). We conjectured that by using an everyday example as the basis for our argumentation, and by encouraging the students to link future examples back to a readily identifiable context, students might find the statistical reasoning process and in particular the nature of the indirect argument more logical. The everyday activity described a situation involving a father and his daughter, and their daily battles with tooth brushing. Alice is told to go and brush her teeth. She

disappears into the bathroom, the tap runs, and she returns several minutes later stating that she has brushed her teeth. The plausibility of Alice's claim is tested by her father retrieving and examining her toothbrush. The following scenarios are then considered: (1) the toothbrush is dry and (2) the toothbrush is wet. The processes of argumentation aligned with each of these scenarios are shown in Figure 2.

| | Scenario One | Scenario Two |
|---|---|---|
| 1. Statement to test. | She has brushed her teeth. | She has brushed her teeth. |
| 2. Collect data (information). | The toothbrush is dry. | The toothbrush is wet. |
| 3. Consider 1. and the data:<br>*If 1. is true, then what are the chances of getting data like that in 2.?* | *The toothbrush-is-dry* would be *highly unlikely* if she had brushed her teeth. | *The toothbrush-is-wet* would *NOT* be surprising if she had brushed her teeth. |
| 4. Review the statement in 1. in light of 3. together with the data in 2. | Therefore, it's a fairly safe bet she has <u>not</u> brushed her teeth.<br>I have evidence that she has not brushed her teeth. | Therefore, she <u>could</u> have brushed her teeth. *Or she could have just run the brush under the tap.*<br>I have <u>no</u> evidence that she has NOT brushed her teeth. |

Figure 2. Development of everyday argumentation

A widely-held misconception, that no evidence against the null model (i.e. the statement being tested, that Alice did brush her teeth) provides evidence in favor of the null model, has been widely documented (Cohen, 1994; Falk & Greenbaum, 1995; Haller & Krauss, 2002). However, when Scenario Two was described to the pilot study students, they were quick to suggest that a wet toothbrush was not definitive evidence of Alice having brushed her teeth. We conjectured that use of this everyday example, the intuitive line of reasoning, and the readiness with which students were able to think of alternative explanations for a wet toothbrush, might allay some common misconceptions.

The argumentation process was then considered in a more probabilistic situation, the third activity, although one to which students could easily relate (Eckert, 1994). With a pack of cards in hand, the instructor asked "What are the chances of drawing a red card from this pack?" Assuming a fair pack of cards, the students were certain that the chances of drawing a red card were ½, that in 10 draws (with replacement) they would expect to get approximately five red cards, but that they were not guaranteed to get exactly five red cards in every draw of ten cards. Their assumption that it was a fair pack of cards, that is, the chances of drawing a red card from the pack was 0.5, was then tested. One volunteer chose ten cards at random, with replacement, from the pack of cards and it was noted that each card drawn was black. When the seventh and eighth black cards were drawn, the students started asking if there were any red cards in the pack at all, a sign that they were skeptical of their initial assumption that it was a fair pack of cards ($p = 0.5$). After all ten cards were drawn, all of which were black, the students were asked if they thought that pack was fair. They emphatically stated no, the pack of cards was not fair. They

acknowledged that one *could* draw ten black cards from a fair pack, but that it was highly improbable. When asked for how many red cards they would want to see in order to accept that the pack was fair, one student answered "… *at least one*…", while another answered "… *three to six*…". Discussion then turned to the distribution of outcomes that one might expect if the process of drawing ten cards from the pack, with replacement, was carried out many times. The students acknowledged that possible outcomes (i.e. the number of red cards in ten draws with replacement) would range from zero to ten, that five would be the most likely outcome, and that outcomes on either side of five would become less likely as they moved away from five. An image of the distribution of outcomes was co-constructed with the students. When the observed outcome of zero red cards was marked on the distribution of outcomes, students could see that although not impossible, it was a highly unlikely result. They also noted that one red card was "…*kind of unusual*…", but appreciated that it was less unusual than no red cards, and that anything between three and seven red cards would not be unusual. We hoped that this intuitive argument with an unusual result displayed within a distribution would assist students when faced with drawing a conclusion from the re-randomization distribution. But, as Rossman (2008) observed, students seem to have no difficulty following the argument with a direct probability situation but do when encountering hypothesis testing.

The fourth activity, the randomization part of the teaching session, then followed, and involved a simplified version of an actual experiment that was designed to investigate whether a program of special exercises for infants for 12 minutes per day could speed up the process of learning to walk (Zelazo, Zelazo, & Kolb, 1972). Experimental data from two of the study groups, Exercise and Control, is displayed in Figure 3 with a red arrow indicating the difference between the measures of center for the Exercise group and the Control group. In this instance, the measure of center is taken as the median. However, it is also possible to use the mean as the measure of center.

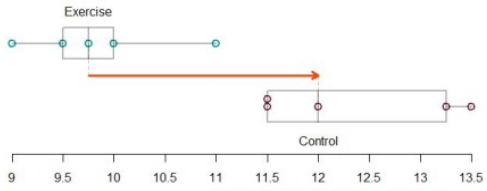



Figure 3. Plot of experimental data. X-axis is walking age in months

Two possible explanations for the observed difference of 2.25 months in the medians of the walking ages were discussed with the students: chance alone factors and both chance factors *and* the treatment factor. The randomization method was presented to the students as a method that allows us to experience the nature and extent of the variability we might expect to see when chance is acting alone, which we can then compare with what we see in our observed data.

A hands-on version of the randomization method was introduced at this stage. By giving students hands-on experience, Holcomb, Chance, Rossman, Tietjen, & Cobb (2010) suggest that students can progress to the next level of the inferential argument by having these hands-on scenarios repeated via software. In pairs, the students were presented with

ten tickets representing the five babies in the Exercise group and the five babies in the Control group and manually plotted the data seen in Figure 3. Under chance alone, we were interested in experiencing what sorts of differences we might expect to see in the measures of center of walking age for the Exercise group and the Control group. This was achieved by randomly reassigning the babies to each group. The tickets were split, effectively breaking the link between group membership and walking age. The group membership tickets and the walking age tickets were then shuffled, and the first observation under chance-alone was obtained by randomly selecting one ticket from each pile. This process, which we call re-randomization, was repeated until all of the tickets were used up. The resulting data were then plotted, and the difference between the measures of center calculated. This was done several times by each pair of students, with students then collating their data as in Figure 4.

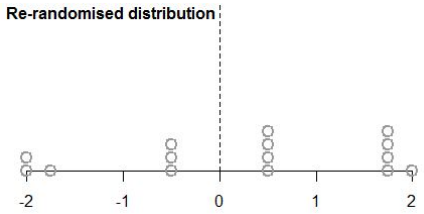



Figure 4. Plot of 15 re-randomizations. X-axis is the difference in median walking ages in months

The students were then presented with a software demonstration, which reiterated the hands-on process, but was able to repeat the re-randomization process many more times. Development of suitable specialized software was a crucial consideration. The software was designed as a teaching tool and an analysis tool, and was required to mimic the hands-on process to allow for an extension to the intuitive logic provided by the hands-on experience. Figure 5 shows an example of an early version of the dynamic visualization tools which were used in the pilot study (http://www.stat.auckland.ac.nz/~wild/VIT/). In the top section of the graphics panel, note the experimental data, with box plots demonstrating walking ages for both the Control group and the Exercise group. This is the same plot as in Figure 3 with the observed difference in medians of 2.25 months indicated by a red arrow. By choosing Run from within the Re-randomization (Idea) portion of the control panel, the data from both groups is combined, and then each walking age is re-allocated at random to a treatment group.

This process of *re-randomization* mimics the hands-on ticket-tearing activity. Note the results from one re-randomization in the data panel and in the middle section of the graphics panel. Each re-randomization difference in the medians can be captured in the bottom section of the graphics panel, allowing a distribution of up to 1000 such differences to be built. The tail proportion, representing the fraction of these 1000 re-randomizations producing a difference in group medians at least as big as the observed difference of 2.25 months, is indicated on this distribution.

We then discussed with the students the conclusion they would make and related it to the Alice story framework Scenario One. Our choice to begin by introducing an example with a small tail proportion was in part due to the suggestion that starting with an example with a large tail proportion might “reinforce students’ natural inclinations to

regard a non-small *p*-value as evidence in support of the null model, rather than a lack of evidence against the null model" (Holcomb, et al, 2010, p. 4).

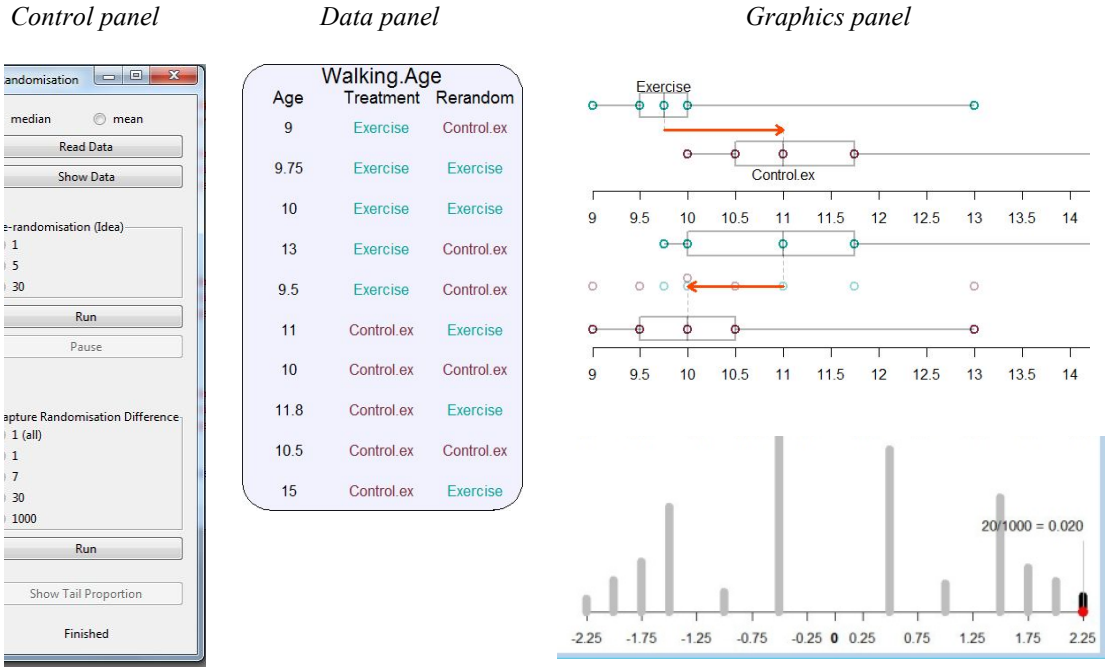



Figure 5. Software panels for performing re-randomization process

The fifth activity involved data from the same walking age study, although this time a comparison was made between the Exercise group and another group. This comparison resulted in an observed difference of 1.4 months, corresponding to a tail proportion of approximately 15%. We discussed with the students how we would interpret this tail proportion including stating that: Chance *could* be acting alone, or the special exercise program *could* be effective. We cannot say, either way. Note a relatively large tail proportion does not lead us to accepting the chance-alone explanation as the only explanation. Our aim was to minimize the possibility of introducing this common misconception. We hoped that the Alice story framework, where tooth brushing was not the only explanation for a wet toothbrush, would provide further reinforcement to the concept that lack of evidence against the tested explanation does not confirm that explanation.

Throughout the entire teaching session, great care was taken with the terminology used. Since the term *randomization* is often used to describe randomness in the data production, and therefore can indicate either random allocation to groups or random sampling from a population, *re-randomization* was the term given to the process of random re-assignment of units to one of the groups. Issues associated with the use of language when applying the randomization method have been published elsewhere (Pfannkuch, et al., 2011).

## 5.3. Pilot Study Post-Test

One week after the teaching session, the pilot study students returned for a written post-test, an interview and a task session. With regard to the randomization method, the same questions that were asked in the pre-test were asked in the post-test. However, there were some new questions in the post-test that we hoped would shed some insight into students' understanding of the randomization process.

There was no increase in the number of correct responses to the question concerning the purpose of randomly assigning the 14 male volunteers to one of the two groups, a finding similar to delMas et al. (2007), albeit with a very small sample. This would appear to confirm our belief that language is very important, and that confusion in the interpretation of random allocation to groups, random sampling from populations and randomization methods exists. However, our questioning of students upheld our original impression that most students who answered incorrectly still appreciated the motivation for random assignment.

With regard to identifying two main possible explanations for the observed difference in blood pressure reduction between the regular oil group and the fish oil group, all students claimed that one explanation would be that the treatment was effective. Furthermore, seven students were able to articulate a chance-alone explanation to some extent. This is evident in both their written and verbal responses, with comments such as:

S2: Chance may or may not be acting on the reduction in blood pressure
S3: That the fish oil diet group was made up of people who were more likely to have their blood pressure reduced for whatever reason
S4: Chance is acting alone. The observed data has resulted in the way it has by chance

Hence a fundamental element of the inference argument seems to be included in these students' reasoning. All but one of the students were able to state that the researchers in the Fish Oil and Blood Pressure study would have been surprised to see a result such as the observed difference in group medians if chance was acting alone. Most of these students made reference to the tail proportion as part of their argument. Compared to Liu and Thompson's (2009) findings, these students seem to have a conception of *unusualness* in a distribution, which we attribute to the software always displaying the tail proportion as part of a distribution, never as a numerical value on its own.

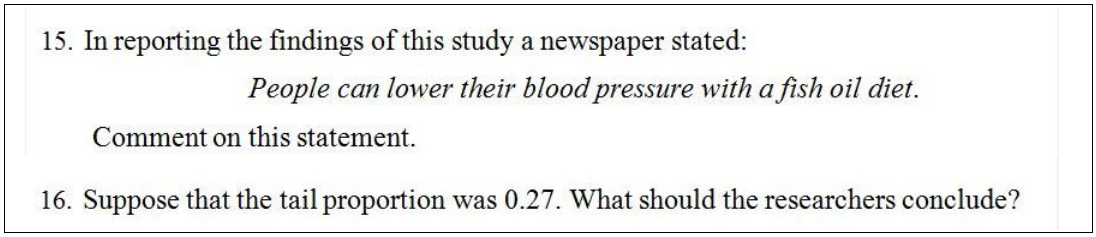
Figure 6 shows two additional questions asked in the post-test. Question 15 was aimed at finding out the scope of the conclusions that the students were prepared to accept. We, perhaps unrealistically, expected the students to break the claim down and restate it so that it was statistically correct. However, some students did not respond in the way we envisaged, and as a result this question has been modified for the main study. Seven students noted that the claim generalized the findings to *people*, commenting that only males, and/or only those with high blood pressure, were studied. For example:

S5: The research was conducted on men and there is a possibility that women may react differently to fish oil consumption
S6: 'People' is too general. As the participants were all male with high blood pressure only inference about this particular group can be made

None considered stating *on average* as a way of dealing with the fact that not all males would lower their blood pressure. Students were generally uneasy accepting what they felt was the definitive aspect of the statement outlined in Question 15.

I: How would you rewrite it?
S1: It is possible…

Even after having taken into account the over-generalizing feature of the statement by acknowledging that inference can only be made about males with high blood pressure, students were still uncomfortable with the word *can*. We speculate that this may be attributed to the tendency to qualify *all* statements with a degree of ambiguity, which Biehler (2011, p. 3) suggested meant nothing in terms of learning about students' statistical reasoning, since "all our knowledge is uncertain."


15. In reporting the findings of this study a newspaper stated:

*People can lower their blood pressure with a fish oil diet.*

Comment on this statement.

16. Suppose that the tail proportion was 0.27. What should the researchers conclude?

Figure 6. Pilot study post-test questions, Questions 15 and 16

We were interested in the responses to Question 16 (Figure 6), particularly in light of the common misconception that a large *p*-value is taken as evidence in favor of the chance-alone explanation. There was a range of responses such as:

S1: Further investigation is required
S7: The result is higher than 10%, showing that it is likely that chance is acting alone [incorrect]
S8: They cannot conclude anything, it means that there's no evidence against chance acting alone and that maybe some other factors could be acting along with chance [correct]

During the teaching session, a guideline had been given for assessing *chance acting alone*. The guideline suggested that if the tail proportion was less than 10%, we had evidence against *chance acting alone*. Perhaps not surprisingly, students had the tendency to grab onto 10% and use it as a rigid cutpoint. One student had the idea that the tail proportion was a measure of how effective fish oil was, with a proportion in the region of 50% indicating that fish oil was definitely not useful, a proportion of around 27% indicating that fish oil may or may not be useful, and that a tail of less than 10% indicating that fish oil was probably useful. Although not probed at the time, we anticipate that his answer to a tail proportion of 1% would be to state that the fish oil was definitely useful. Such a misconception has been consistently documented for hypothesis testing (Falk & Greenbaum, 1995). Their responses, however, were not surprising since the tail proportion idea has not changed through our visualizations, only an appreciation of how the tail area is obtained has changed, that is, it is not a numerical value, rather a part of an understandable distribution. Interpretation of a large tail proportion and the indirect nature of the logic of the argument seem to remain a problem with this method as

it was with normal-based inference (Falk & Greenbaum, 1995; Nickerson, 2004; Liu & Thompson, 2009). We attribute this problem partially to the fact students only had about two hours tuition and we did not successfully link in the Scenario Two everyday argumentation (Figure 2). Even though this type of argumentation is used in everyday life, we think the argumentation will continue to remain difficult as it appears to be an alien way of reasoning (Thompson, Liu, & Saldanha, 2007) particularly when overlaid with chance-alone, the re-randomization distribution of differences in means, and tail proportion ideas.

## 5.4. Pilot Study Post-Task

The purpose of the post-task was to find out how the students handled the software, and whether they understood the dynamic visualization representations, the hands-on components of the activities, and the language used, and whether further issues could be identified in their reasoning. Pairs of students were given a written scenario to consider. The research question was "Does added calcium intake reduce blood pressure?" The scenario described data that had been obtained from a randomized experiment designed to establish whether or not an increase in calcium intake was associated with a reduction in blood pressure. The students were required to enter the data into a file, import it into the software package and to analyze it such that the research question would be answered. They were asked to describe each step of their analysis for the benefit of the interviewer who was to be considered a novice.

All students were able to use the software competently. Of the five student-pairs, three had initial problems entering the data which was not a software problem rather a data structure problem. However, all were able to correct themselves with minimal input from the interviewers. All students correctly interpreted the red arrow (see Figure 3) as the difference between the measures of center for the two groups being compared and understood what the re-randomization distribution represented. The dominance of the visual distribution in students' thinking was noted particularly in the bootstrap method scenario (not reported in this paper), which led us to refine the software to use a fainter color and fade option for the distribution so that visually students' attention was drawn to the observed value, the tail proportion, and its relationship to the distribution. Noteworthy was a comment made by the student who had studied statistics at the third year level – "*now I understand what a p-value is*" –reinforcing our view that presenting the tail proportion as part of a distribution assists learning.

When quizzed on the effectiveness of the hands-on activities that formed part of the randomization teaching sequence, one student noted: "… *having something that you can do and something you can visually see with the graphs, I thought that was really useful*". One student-pair described the connection they made between ripping up the pieces of paper in the Walking Babies example with the automated reassignment of blood pressure reductions in the dynamic visualizations. Another student remarked: "*the randomization experiment (mixing the various pieces of paper) was very straightforward and left no room for confusion.*"

Five students were asked if they considered the line of argumentation associated with the toothbrush story an effective analogy, that is, that the line of reasoning is the same as the intuitive reasoning underlying the toothbrush story. All five agreed, mentioning that it was a readily identifiable situation as everyone brushes their teeth. However one student

noted that she was able to understand the story when the toothbrush was dry, but was confused by the conclusion when the toothbrush was wet. This reinforced our view that the concept of '*no* evidence *against* chance alone' appears to be especially problematic for students. We conjecture that the double-negative in '*no* evidence *against* …' is the source of the problem. This difficulty has been compounded in the conclusion of the Alice story when the toothbrush was wet by rephrasing 'no evidence *against* the statement being tested' as 'no evidence *for* the negation of that statement'. To avoid potential obfuscation of the analogy, we conjecture that 'I have no evidence that she has not brushed her teeth' should be replaced with 'I'm still not sure if she has brushed her teeth'.

Again the language we used manifested itself as problematic in several areas. One area was students' understanding of what *chance is acting alone* means. Several students were unable to articulate their understanding of what it means for chance to be acting alone. It may be that they have grasped the concept, but were unable to express their thoughts. One pair of students described chance as "…*it just so happened*…", and then stated that randomization was used to "… *confidently be able to say there is no chance that chance had any effect on the results*". When questioned on the hands-on activities that formed part of the randomization teaching sequence, another pair of students understood the ticket-tearing procedure as "…*creating chance*." Since the idea that *chance is acting alone* is a central concept in the randomization method, we were concerned about the students' difficulties with this notion. Consequently new software, designed to illustrate *chance is acting alone* unencumbered by experimental data, has been developed.

Another problematic area was the reluctance to state a causal relationship between calcium intake and reduced blood pressure. One pair of students, in finding a tail proportion less than 10%, chose to say that there was "…*likely*..." a difference in blood pressure reduction between the calcium and control groups. It was unclear whether they said *"...likely...*" because they were hesitant to give a strong confirmatory response, or whether they felt that there was still a chance that what they observed in the data was not the truth. Such reasoning is understandable given that the observed difference could be that rare occurrence.

Two further issues that arose in students' reasoning in the post-test were also identified in the post-task. There was a general tendency for students to use the 10% guideline as a strict cut-off, with statements such as "*The decrease in (the) calcium group was lower than the placebo, so ...because it's less than 10% ... it doesn't happen by chance.*" Also interpreting a large tail proportion was difficult for students. Given a tail proportion of 40%, one pair of students stated "…*there's a pretty good chance that ... calcium doesn't necessarily reduce blood pressure…",* while another pair, given a tail proportion of 27% concluded that "*chance is acting*". Thus it would appear that misconceptions associated with interpreting a *no evidence against chance alone* explanation as if it were evidence in favor of a chance explanation is an issue that we need to address.

# 6. CONCLUSION

Experience from the pilot study has been invaluable. The purpose of the pilot study was to detect problems in the pre- and post-tests, learning trajectories and software before trialing the randomization method with over 3000 students. The wording in some pre-

and post-test questions was changed, such as the wording in Question 15 (Figure 6), to better reflect the type of response we were anticipating.

With respect to the dynamic visualizations for inference, students seemed to understand its components and what each was representing. In particular they were able to state what the re-randomization distribution represented with none thinking it was the data distribution (cf. Gould et al., 2010) although we did have a very small sample of students. We speculate that the vertical arrangement of the graphics panel within the dynamic visualization tool, a feature not present in other software such as Fathom (Finzer, 2006), facilitated this understanding. The vertical arrangement allows the students to view the observed data, notice representations of the differences in means after each re-randomization, and subsequently watch these differences dropping down to form the re-randomization distribution. Thus viewing development of the entire process within the same panel may lead to more of an understanding of what the re-randomization distribution represents.

From the students' comments the prior hands-on activities, which were closely aligned to the dynamic visualizations, were a major element in helping them understand them. We believe that the hands-on component is a critical part of the conceptual understanding of the randomization test and that without it students will have difficulty in reconstructing the process when faced with a new scenario. Furthermore, showing the observed difference and consequent tail proportion as part of the re-randomization distribution seemed to help the students conceptualize the probability of obtaining the observed difference or greater under chance-alone (cf. Liu & Thompson, 2009). It is therefore plausible that the combination of technology and hands-on activities form a powerful tool for facilitating understanding of the randomization method. As outlined in the GAISE recommendations technology is a tool that, with appropriate implementation, can remove the distraction of computational minutiae from the path of student understanding and direct focus to the bigger picture, that is, the fundamental principles underlying statistical inference (Aliaga et al., 2005).

The dynamic visualizations, however, need to be accompanied by verbalizations and language that help students to reason and argue from the simulated data. The students' difficulties with drawing a conclusion from a large tail proportion, the same problems identified by other researchers (Nickerson, 2004; Liu & Thompson, 2009), challenges us to find a better way to assist students in this form of argumentation. In fact, after much reflection we have modified the learning trajectory so that two explanations for an observed difference are not seen to compete against one another, but rather evidence is gathered against the *chance acting alone* scenario. We realized we should break with this traditional way of thinking and present the whole argument in a more natural format. Also we intend to make better connections to the Alice story framework in order to reinforce students' understanding of the logic of the indirect argument.

The notion of what *chance acting alone* means is clearly a difficult one for students. Given that this concept lies at the core of the randomization test, it is important that we think carefully about how we might facilitate more of an understanding of what *chance acting alone* might look like. In response to the pilot study finding that emphasized this issue, a new dynamic visualization module has been developed with the aim of facilitating student understanding of *chance acting alone*. The module demonstrates, for example, weights of people being randomly allocated to one of two groups with the

differences in mean weights of the two groups being recorded in the middle panel of the vertical screen and subsequently dropping down to the bottom panel where a re-randomization distribution is built up. Students can then see that differences in mean weights between the two groups can range between -10kg and +10kg simply under chance alone. We anticipate that this module will help to elucidate the *chance acting alone* concept since it stands alone, unencumbered by experimental data or treatment variables.

Causal argumentation remains a problematic area. We conjecture that the reluctance of students to state that "a fish oil results in (causes) blood pressure that tends to be lower in male volunteers with high blood pressure than a regular oil diet" may be in part attributable to the fact that "all our knowledge is uncertain" (cf. Biehler, 2011, p. 3). Another likely contributory factor is the fact that while sample to population inference is familiar, causal inference is new territory for these students. Despite the fact that the fish oil and regular oil treatments were randomly assigned to the study participants, the students were uncomfortable making a causal statement, wanting to use words such as "is likely to" and "may" rather than "results in" or "causes". In order to complete the inferential process, the students are required to reflect on the scope of the inference. A sound understanding of the evidence obtained from the randomization test, which has difficulties due to the probabilistic reasoning required by students, needs to be cognitively integrated with experimental design, causal, and generalization ideas in order to understand how a conclusion is reached about the effectiveness of the experimental treatment. There are several sources of uncertainty at play in the inferential process. In the Fish Oil study, the randomization test produced a very small tail proportion. Thus the observed difference was highly unlikely if chance was acting alone. However, there is still the *possibility*, albeit small, that the observed difference was one of the rare occurrences that *could* be expected if chance was acting alone. Applying a deterministic form of reasoning to this situation may result in a hesitance to make a causal claim owing to the element of uncertainty that is present. Moreover, it is evident from the plots of the data (see Figure 1) that not all of the study participants randomized to the fish oil diet experienced a greater reduction in blood pressure than those randomized to the regular oil diet. Rather it is the tendency of the fish oil subjects as a group to experience a greater blood pressure reduction than the regular oil group that is being considered. Therefore there exists uncertainty about making a causal claim when it cannot be made for all of the study subjects. Finally, the fact that the study was carried out on male volunteers with high blood pressure introduces another source of uncertainty when it comes to generalizing the results to other people such as women, and those with normal blood pressure. The complexity of the language employed in the reasoning process may lead students to use words such as "is likely to" and "may" to soften what they feel is a definitive causal statement in order to deal with these different sources of uncertainty. There is much for the students to come to grips with, and it is perhaps not surprising that they are unable to assimilate all of these new concepts coherently. Thus we need to further explore the issues surrounding students' reasoning in the presence of uncertainty.

The randomization method and our dynamic visualizations are not a panacea for making inferential or hypothesis testing reasoning readily accessible to students. We believe, however, from this pilot study, that more underpinning concepts such as mimicking the data production process, chance is acting alone, the tail proportion, and the re-randomization distribution are more accessible and transparent to students. In particular, the dynamic visualizations allowed students to view the process of re-randomization as it

developed and grew into a distribution, giving students direct access to the behavior of the chance-alone phenomenon (Sacristan et al., 2010). Compared to the mathematical procedures of hypothesis testing, we believe the pilot study students did learn more about statistical inference using the randomization method (cf. Tintle et al., 2011).

# REFERENCES


Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science, 7* (1), 131-153.

Aliaga, M., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P. & Witmer, J. (2005). *Guidelines for assessment and instruction in statistics education (GAISE): College report.* Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/education/gaise/

Bakker, A. (2004). *Design research in statistics education: on symbolizing and computer tools.* Utrecht, The Netherlands: CD-β Press, Center for Science and Mathematics Education.

Biehler, R. (2011). Five questions on curricular issues concerning the stepwise development of reasoning from samples. *Presentation at the International forum on Statistical Reasoning, Thinking and Literacy,* Texel, The Netherlands.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in Psychology , 3* (2), 77-101.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48* (3), 378-399.

Chance, B., delMas, R., & Garfield, J. (2005). Reasoning about sampling distributions. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Clark, J., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review , 3*, 149-210.

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education , 1* (1), 1-15. Retrieved from http://escholarship.org/uc/item/6hb3k0nz

Cohen, J. (1994). The earth is round ($p<0.05$). *American Psychologist , 49* (12), 997-1003.

Daniel, L. G. (1997). Kerlinger's research myths: An overview with implications for educational researchers. *The Journal of Experimental Education , 65* (2), 101-118.

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). Retrieved from http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6* (2), 28-58. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ6%282%29_delMas.pdf

Eckert, S. (1994). Teaching hypothesis testing with playing cards. *Journal of Statistics Education , 2* (1). Retrieved from http://www.amstat.org/publications/jse/v2n1/eckert.html

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

Finzer, W. (2006). Fathom dynamic data software [Computer Software]. Emeryville, CA: Key Curriculum Press.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver & Boyd.

Gigerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. In G. &. Keren (Ed.), *A handbook for data analysis in the behavioral sciences: methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum.

Gould, R., Davis, G., Patel, R., & Esfandiari, M. (2010). Enhancing conceptual understanding with data driven labs. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_C208_GOULD.pdf

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

Hesterberg, T. (2006). Bootstrapping students' understanding of statistical concepts. In G. Burrill, & P. Elliot, *Thinking and reasoning with data and chance: NCTM Yearbook* (pp. 391-416). Reston, VA: National Council of Teachers of Mathematics.

Hesterberg, T., Moore, D., Monaghan, S., Clipson, A., & Epstein, R. (2009). Bootstrap methods and permutation tests. In D. Moore, G. McCabe, B. Craig, D. Moore, G. McCabe, & B. Craig (Eds.), *Introduction to the practice of statistics* (pp. 16-1 - 16-60). New York, NY: Freeman.

Hjalmarson, M., & Lesh, R. (2008). Engineering and design research: Intersections for education research and design. In A. Kelly, R. Lesh, & K. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching* (pp. 96-110). New York, NY: Routledge.

Hodgson, T. (1996). The effects of hands-on activities on students' understanding of selected statistical concepts. In E. Jakbowski, D. Watkins, & H. Biske (Ed.), *Proceedings of the eighteenth annual meeting of the North American chapter of the international group for the psychology of mathematics education* (pp. 241-246). ERIC Clearing House for Science, Mathematics, and Environmental Education.

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia.* Voorberg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_8D1_HOLCOMB.pdf

Hurlbert, S., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici , 46* (5), 311-349.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management , 63* (3), 763-772.

Jones, P., Lipson, K., & Phillips, B. (1994). A role for computer intensive methods in introducing statistical inference. In L. Brunelli, & G. Cicchitelli (Eds.), *Proceedings of the First Scientific Meeting of the International Association for Statistical Education* (pp. 199-211). Perugia, Italy: University of Perugia. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/proc1993/255-263rec.pdf

Knapp, H., & FitzGerald, G. (1989). The antihypertensive effects of fish oil. A controlled study of polyunsaturated fatty acid supplements in essential hypertension. *New England Journal of Medicine , 321* (23), 1610-1611.

Lenhard, J. (2006). Models and Statistical Inference: The controversy between Fisher and Neyman-Pearson. *British Journal of the Philosophy of Science , 57*, 69-91.

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

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician , 52* (2), 127-132.

Lunsford, M., Rowell, G., & Goodson-Espy, T. (2006). Classroom research: Assessment of student understanding of sampling distributions of means and the central limit theorem in post-calculus probability and statistics class. *Journal of Statistics Education , 14* (3). Retrieved from http://www.amstat.org/publications/jse/v14n3/lunsford.html

Meletiou-Mavrotheris, M., Lee, C., & Fouladi, R. (2007). Introductory statistics, college student attitudes and knowledge - a qualitative analysis of the impact of technology-based instruction. *International Journal of Mathematical Education in Science and Technology , 38* (1), 65-83.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. Harlow, S. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-116). Mahwah, NJ. Lawrence Erlbaum Associates.

Nickerson, R. (2004). *Cognition and Chance.* Mahwah, NJ: Lawrence Erlbaum Associates.

NVivo qualitative data analysis software; QSR International Pty Ltd. Version 9, 2010.

Pfaff, T., & Weinberg, A. (2009). Do hands-on activities increase student understanding?: A case study. *Journal of Statistics Education , 17* (3). Retrieved from http://www.amstat.org/publications/jse/v17n3/pfaff.html

Pfannkuch, M., Regan, M., Wild, C., Budgett, S., Forbes, S., Harraway, J., Parsonage, R. (2011). Inference and the introductory statistics course. *International Journal of Mathematical Education in Science and Technology* .

Pitman, E. (1937). Significance tests which may be applied to samples from any population. *Supplement to the Journal of the Royal Statistical Society , 4* (2), 225-232.

Quinn, G., & Keough, M. J. (2001). Hypothesis Testing. In G. Quinn, & M. J. Keough, *Experimental design and data analysis for biologists* (pp. 37-68). Cambridge, UK: Cambridge University Press.

Rossman, A. (2008). Reasoning about informal inference: One statistician's view. *Statistics Education Research Journal , 7* (2), 5-19. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ7%282%29_Rossman.pdf

Rossman, A., Chance, B., Cobb, G., & Holcomb, R. (2008). Concepts of statistical inference: Approach, scope, sequence and format for an elementary permutation-based first course. *Unpublished paper*. Retrieved from http://statweb.calpoly.edu/bchance/csi/CSIcurriculumMay08.doc

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics, Dunedin, New Zealand. 2*, pp. 314-319. Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/18/BOOK1/A9-4.pdf

Sacristan, A., Calder, N., Rojano, T., Santos-Trigo, M., Friedlander, A., & Meissner, H. (2010). The influence and shaping of digital technologies on the learning - and learning trajectories - of mathematical concepts. In C. Hoyles, & J. Lagrange (Eds.), *Mathematics education and technology - Rethinking the terrain: The 17th ICMI Study* (pp. 179-226). New York, NY: Springer.

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

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods , 1* (2), 115-129.

Seier, E. (2010). Hands-on activities to introduce randomization methods and hypothesis testing. In C. Reading (Ed.), *Data and context in statistics education: Proceedings of the Eighth International Conference on Teaching Statistics, Ljubliana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_P31_SEIER.pdf

Thompson, P., Liu, Y., & Saldanha, L. A. (2007). Intricacies of statistical inference and teachers' understandings of them. In M. C. Lovett, & P. Shah (Eds.), *Thinking with Data* (pp. 207-232). New York, NY: Routledge.

Tintle, N., Topliff, K., Vanderstoep, J., Holmes, V., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal , 11* (1), 21-40. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ11%281%29_Tintle.pdf

Tintle, N., VandenStoep, J., Holmes, V., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education , 19* (1). Retrieved from http://www.amstat.org/publications/jse/v19n1/tintle.pdf

Vickers, A. (2010). *What is a p-value anyway?* Boston, MA: Pearson Education Inc.

Zelazo, P. R., Zelazo, N. A., & Kolb, S. (1972). 'Walking' in the Newborn. *Science , 176*, 314-315.