`

LEARNING STATISTICAL INFERENCE THROUGH COMPUTER-SUPPORTED SIMULATION AND DATA ANALYSIS

by

Javier Alejandro Corredor

B.A. Universidad de los Andes, 2000

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Javier Alejandro Corredor

It was defended on

March 18th, 2008

and approved by

Dr. Alan Lesgold, Dean, School of Education

Dr. Christian Schunn, Associate Professor, Department of Psychology, School of Education (Adjunct)

Dr. Jorge Larreamendy-Joerns, Profesor Asociado, Psicología, Universidad de los Andes

Dissertation Advisor: Dr. Gaea Leinhardt, Professor, Department of Instruction and Learning. School of Education

LEARNING STATISTICAL INFERENCE THROUGH COMPUTER-SUPPORTED SIMULATION AND DATA ANALYSIS

Javier Alejandro Corredor, PhD

University of Pittsburgh, 2008

This dissertation explored the effects of two different interventions on the learning of statistics. Each intervention corresponded to a different conception of statistical learning and used a particular type of computer-tool. One intervention used data analysis tools and focused on authentic situations of statistical activity. The other intervention used simulations and focused on formal aspects of probability. Data Analysis (data) and Probability (chance) are the constituent parts of statistical inference and the two lens from which is possible to present this topic. In this study, both perspectives were compared in their effectiveness to teach ANOVA, a central topic in inferential statistics. The results of this study showed that the intervention that used simulations improved students' knowledge about probability, sampling and sample size effects. Protocol analysis of students' answers indicated that the gains in probability knowledge did not alter the way students explained group differences. The intervention that used data analysis tools showed no significant effects on students' data analysis knowledge. Studying the evolution of a sub sample of students suggested that data analysis knowledge develops over periods of time longer than those of this study. Additionally, protocol analysis of students solving statistical questions showed that students use simple decision rules to evaluate sampling and data analysis problems. These rules allowed students coordinating simple descriptions of the problem's elements with conclusions about significance and sampling effects.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 PERSPECTIVES IN STATISTICAL EDUCATION

Studies have not found significant differences in learning between online-supported statistics education and traditional face-to-face courses (Field & Collins, 2005; Wisenbaker, 2003). Another strand of research reports differences in quality, user satisfaction, and learning among different online courses (Alldredge & Som, 2002, Larreamendy et al, 2005). Taken together these findings suggest that the outcome of instruction is determined by specific course characteristics and not by the media of instruction. Therefore, it makes sense to begin to look seriously for the characteristics of computer-supported education that influence learning in specific content domains.

This study explores the relationship between the conception of statistic activity entailed in computer-based instruction and the resulting learning of statistical inference. In particular this study focuses on two different interventions aimed at teaching statistical inference in the context of group mean differences. One intervention requires students to use simulations that permit the learner to conduct limitless trials while controlling sampling and population parameters. The other intervention asks the students to engage in data analysis that uses authentic data as the source of information. While the first type of intervention allows students to conduct many simulations at a very low computational and procedural cost, it restricts the possibilities of choice about the raw data source, organization, and representation. The second type of intervention allows students to control the representation of the data and pushes them to define the structure of the task, but it has higher computational costs and so restricts the number of exercises that can be conducted. The first type of intervention is more efficient but yet not authentic, while the second type is authentic but constrained by time and computational costs. Both types of instruction are found in online statistics courses.

This study rests within a broader debate. There is a shift in scientific education that questions the extent to which instruction that does not resembles authentic tasks prepares students for scientific activity (Gravemeijer, 2002; Petrosino et al, 2003; Snee, 1993). Tasks devoid of scientific complexity preclude students from experiencing fundamental aspects of science such as the definition of measurement and representation frameworks, and the argumentative processes around evidence (Ford & Forman, 2006; Petrosino et al, 2003).

This larger debate takes a specific form within the realm of statistics. In some way what is at stake in this comparison is portrayal of statistical activity that must be used to teach statistical inference. The first intervention conceives statistical inference as a structured task where representational ambiguity is not relevant, and where procedures can be understood and performed in a syntactic way. On the other hand, the second intervention depicts statistical activity as part of a broader range of activities suggested by social scientific practice. Each portrayal of statistical activity privileges certain aspects of the discipline. Students taught in the first approach learn the assumptions and basis of formal procedures; they have had contact with the proofs that support the mathematical structure of inferential statistics. Students taught in the empirical tradition conceive statistics a mechanism to produce theory using statistical tools to organize the complexity of the world. From this second point of view, limiting inferential statistical reasoning to *p*-values' interpretation might underestimate the complexities of statistical inference. There is a call to teach students how to carry out multiple steps of data representation

(Wild & Pfannkuch, 1999), produce global views of data (Ben-Zvi & Arcavi, 2001), and coordinate theory and evidence (Lehrer & Schauble, 2004).

The case of statistical inference is especially interesting because there is a tradeoff between the authenticity of the tasks and the number of simulations or examinations that can be conducted. The effects of this tradeoff need to be explored carefully because statistical inference is at the crossroad of data analysis and probability theory. Therefore, understanding it grows from experiencing both data analysis in authentic contexts and repeated simulation in random environments. Students need to be able to conduct representational permutations, identify global tendencies in data, isolate data patterns and interpret variability; but they need also to recognize the sampling process as a random situation and identify the consequences of its random nature on the information they receive from the sample (Garfield & Ben-Zvi, 2002). Ignoring the probabilistic nature of the sampling process can lead students to use flawed reasoning mechanisms when producing conclusions from data. Nisbett et al (1993) showed that when students do not assign random attributes to a problem, they use other heuristics to make sense of the situation, such as, using causal mechanisms that do not allow variability. Recognizing the sampling process as random is fundamental in order to activate adequate probabilistic reasoning. If the instructional situation favors data analysis in comparison to probabilistic training, students can underestimate the effects of probability on their conclusions. If the instructional situation focuses on fostering probabilistic understanding but ignores the data analysis aspect of inference then students lose perspective on the applied underpinnings of the discipline.

1.1 THE CHALLENGE OF TEACHING INFERENTIAL STATISTICS

Since the empirical evidence regarding reasoning about statistical inference is limited, it is necessary to build a conceptual framework from which understanding the effects of computer-based education. The difficulty of learning inference comes from the way statistics produce inductive arguments by comparing patterns in data with expectations in chance. At a psychological level, learners have to coordinate the space of data and the space of chance to solve problems and give meaning to the methods and tools provided by statistical inference. In statistical inference students have to coordinate schema properties that are superficially similar but essentially different. At the educational level, statistical inference confront teachers with a tradeoff between authenticity and repeated simulation, where the former is necessary to understand applied tasks, and the later is fundamental for the development of probabilistic heuristics.

Formal training in mathematics is not enough to understand the use of probability in inferential statistics. Formal probability describes the properties of random behavior models assuming that the models are completely specified, that is, that all parameter values are known explicitly. On the other hand, statistics uses data to estimate unknown parameters (Tappin, 2000). In this sense, the reasoning involved in traditional probability is different from the reasoning involved in applications of probability to statistical science: both involve an underlying probabilistic model but statistical inference uses this model to answer empirical questions; based on the sample data, inferences can be drawn about the nature reality using the underlying probabilistic mechanism as a mean (Tappin 2000). Part of the problem with learning inferential statistics is that it requires to reason from uncertain empirical data (Moore 1992). What happens is that while the space of data is inductive, the space of chance is deductive, and coordinating both for statistical inference is difficult (Moore 1992). Nisbett et al (1993) found that people may be able to apply statistical rules in one setting (for example, to random generating devices) but rarely or never to similar problems that involve social content. People might not use statistical test even when they have been trained to do so, because they do not appreciate its role in the description of data (Williams, 1998).

Statistical inference can be challenging at many levels. Learners might find it strange that data behave according to the predictions of probability theory. It is not clear why some attributes (e.g. height) have a normal distribution instead of different type of distribution. It might be hard for learners to differentiate the properties of the distributions in the space of data and in the space of chance. Student can assume that sampling distributions' and data distribution represent the same type of object when actually they represent different categories of objects. A data sample graph represents the distribution of a given set of data; a sampling distribution graph (e.g sampling distribution of the mean) represents a set of possibilities organized around probability values. Understanding that the distribution of data in a study corresponds to the distribution of just one sample within the sampling distribution, and not to the whole sampling distribution, can be challenging for learners.

The way in which inferential statistics connects probability and data analysis is ambiguous and it can create problems for learners. The relationship between data and chance in statistical inference can be seen in two ways: One is to use chance as the background against which you compare the actual data (*p*-value); that is, finding the probability of the data results in a random distribution; the other way is start from the real data and see how chance generates ranges of error (Confidence Intervals) around the patterns in data. Both ways represent the same idea, the effects of sampling on the descriptions of data, but they do it from different perspectives. The coordination of both perspectives can be challenging for students and difficult the learning of statistical inference.

1.2 WHAT A SUFFICIENT STATISTICAL UNDERSTANDING SHOULD LOOK LIKE

A sufficient understanding of statistical inference should permit students to coordinate three processes: the comparison of distribution graphs, the interpretation of statistical test results (e.g. ANOVA table), and generation of conclusions in context. Comparing distributions is the basic level of inferential analysis (Ben-Zvi, 2004; Lehrer & Shauble, 2007; Watson, 2002; Watson & Moritz, 1999). It requires students to connect representations that contain both central tendency indicators and variability, in order to draw conclusions about the strength of differences in situations where the information provided by the data is ambiguous. In group comparison situations, the groups have different means but they overlap due to variability, thus, the larger the standard deviation of the groups, the lower the certainty of the differences. The skill of adequately comparing distributions is just one part of statistical inference. Additionally, learners need to be able to understand the effects of sampling on the process of comparing distribution and on the results of inferential tests. Students with inadequate conceptions of sampling misunderstand the way samples resemble the population (Saldanha and Thompson, 2003). Students must also generate conclusions from data representations and statistical tests results. For this, students should be able to make sense of representations and results in the context of research. Having command of sampling and data representation would be useless without the competency to produce thoughtful conclusions from data. Ideally a student knowledgeable of inferential statistics should be able to go back and forth through these three outcomes. This coordination requires understanding that the differences and the variability seen in the distributions together with the sample conditions determine the statistical tests outcomes. This coordination requires also being able to translate the statistical results in meaningful conclusions, knowing that the relationships between mean differences and variability are equivalent in the data representations, in the mean differences *p*-values and in the conclusion in context.

At the same time students move through these outcomes, they need to be able to coordinate knowledge on data analysis and probability (Garfield and Chance 2002). Data analysis knowledge produces descriptions of the patterns and the variability in the data sets. Probability knowledge produces descriptions of how samples should behave randomly. Coordination between data and chance means being able to either locate the actual data patterns in the random space, or to add the random variability of sampling to the pattern and variability found in the data. This coordination is a challenge for most learners. They need to recognize that even if data and chance are represented in similar ways through distribution graphs, they have different properties.

The concept that permits such coordination is variability. At a specific level, students that understand variability are able to see that variability is constant across the three types of outcomes. For instance, graphs that present a small within group variability and a large between group variability come from samples that produce significant differences –once sampling variability is controlled- and they appear in theoretical situations where the model explains largely the behavior of the data. At a general level, understanding variability gives students the main tools to operate in statistics (Garfield 2005). Variability affects the informative value of measures of center; the presence of outliers or unusual distributions of data (e.g. high skewness) warns researchers about the precision of their results. Variability is also useful to make comparison among groups, or to establish the precision of the model; residuals represent how imprecise the model is (Garfield 1995). In some way, statistical inference intends to explain variation by seeking the systematic effects behind random variability of individual and measurements (Moore 1990).

1.3 INTERVENTION TYPES AND STATISTICAL INFERENCE

This study compares two instructional conditions that can affect learning outcomes given the challenge of coordinating data and chance and given the tradeoff this relationship creates for teaching inferential statistics. One intervention relies on the use of simulations and dynamic visualizations to develop understanding of ANOVA. Tools for this type of intervention are very common in the statistical learning literature (Blejec, 2002; Cramer & Neslehova, 2003; Darius et al, 2002; DelMas, Garfield & Chance, 1999; Drier, 2000; Harner & Hengi Xue, 2003; Nicholson et al, 2000; Sanchez, 2002; Shaughnessy & Ciancetta, 2002; Wilensky, 1999; Wood, 2005). The other type of intervention relies on data analysis tasks to teach statistical inference. Examples of this type of intervention are found in different settings. They normally require participants to use statistical packages to analyze data sets coming from authentic or simulated data (Connor 2002; Conti & Lombardo, 2002; Hooper, 2002; McClain, 2002; Wilensky & Stroup, 1999).

The first type of intervention should provide students with understanding of statistical inference as a random process because knowledge and recognition of random situations does not appear spontaneously (Chance, DelMas, & Garfield, 2004; Konold, 1995; Konold, Well, Pollatsek, & Lohmeier, 1993), but grows from the contact with random mechanism (Nisbett et al, 1993). Simulations can show students how probability represents tendencies in events aggregated over several trials (Cramer & Neslehova; DelMas, Garfield & Chance, 1999; Snir, Smith & Grosslight), and provide students with intuitive proofs of the way the statistical tests work. They show in a graphical manner how, for example, ANOVA compares within and between variability. Change in the parameters of the simulation produce changes in the relationships among graphical representations and in numerical indicators (Darius et al, 2002; West and Ogden, 1998; Godino et al, 2003; Mittag). This type of intervention however has costs in terms of the student skills to deal with data analysis situations because there is not transfer from training in probability theory to applied activities in statistics (Cobb & Moore 1997; Lovett & Greenhouse, 2000; Snee,1993; Tappin, 2000). Additionally, the lack of authenticity in this type of situations can create disbelief about the accuracy of the simulation as a representation of actual situations (Velleman and Moore, 1996).

The other type of intervention uses data analysis to provide learners first hand experience with authentic scientific situations. This experience will allow learners to develop the ability to organize data complexity in patterns that isolate signal from noise (Biehler 1995) Authenticity requires students to learn how to build measure and representation methods, to make sense of the actions and results during the statistical process (Lehrer & Schauble, 2004; Burgess, 2002) and to deal with variability in data (Kazak & Confrey, 2004; Petrosino et al, 2003). Experience with data analysis situations should push learners and to develop the skill to conduct representational permutations (Wild and Pafnnkuch), to build global views of data (Ben-Zvi & Arcavi, 2001) and to develop a deeper understanding of the statistical situation (Ben-Zvi, 2002).

Computers provide visual representations that can be used as analytical tools (Garfield, 1995) without the huge computational costs (Finzer & Erickson, 2005). The low computational costs of computer-supported statistics modify the goal of the task and allow the students to focus more on making representational decisions and less on the calculation process (Ben-Zvi, 2000; Rubin, 2002). The cost of this type of intervention is that students are not exposed to the random process that by its own nature relies not on the analysis of one data set, but on the repeated generation of samples. Students in this condition should know more about analyzing data and less about the effects of sampling on the results.

1.4 PURPOSE OF THE STUDY

This study aims to compare the effects of two computer-based interventions on the learning of inferential statistics knowledge. One intervention focuses on the probabilistic aspects of inferential statistics through the use of simulations; the other intervention focuses on the data analysis aspects of inferential statistics. According to the educational literature, these interventions should have different effects on students' knowledge and image of statistical activity. While using simulations privileges understanding of sampling and probability, data analysis fosters the understanding of statistics as a tool to organize information in authentic contexts. Specifically, this study seeks to evaluate the effects of these two types of intervention on the ability to coordinate three types of data outcomes: graphic representation, statistics test results, and conclusion in context. This study aims to validate a set of measures that evaluate this coordination and show that this coordination participates in the process of reasoning in statistical inference. Finally, this study examines effects of new technologies on the distribution of knowledge at a global scale by studying the effects of online resources on the learning of statistics.

1.5 RESEARCH QUESTIONS

This study aims to answer the following questions:

1. Can the ideas of data analysis and sampling be captured in a reliable and valid measurement system?

a) What are students thinking when they respond to the measurement system? What domains are they accessing?

b) Does this system capture change in performance in data analysis, sampling and inference knowledge?

2. Is it possible to design online instruction that reflects the advantages of each perspective in statistical education, and also manages to teach the target ideas of data analysis, and sampling equally effectively?

a) Do data analysis, sampling or inference develop organically over time through sustained interaction with instructional resources and with other people?

b) Does the evolution of statistical knowledge over time constraint the effectiveness of online education when teaching statistical content in short periods of time?

3. Can either perspective be used to equal effect to teach either content (data analysis or sampling)?

a) Is there a sacrifice in the understanding of data analysis if chance is used as an instructional medium to teach statistical inference?

b) Is there a sacrifice in the understanding of probability if data analysis is used as an instructional medium to teach statistical inference?

c) What are the tradeoffs between the authenticity of data analysis instruction and the sustained activity of computer-based sampling?

Question 1 explores whether or not a measurement system can assess ideas and skills typical to the different perspectives in statistical education considered in this study: data analysis and chance. Each perspective highlights certain aspects of statistical inference while overshadows others. Particularly, data analysis is more authentic and situated; while chance is more mathematical and abstract. The challenge for a measurement system in this case is to be able to include ideas from these two sources and to capture the interaction of both knowledge bases when applied to statistical inference. Two related questions need to be answered also. The first question is how participants solve items within this measurement system; that is how different domains are accessed while solving questions about data analysis, sampling and inference. The second question is whether or not this measurement system can capture change in students performance in any of those statistical spaces.

Question 2 evaluates the real potential of online education to teach statistical inference. It is possible that important target ideas and skills of statistics develop over time in the interaction with instructional resources and with other people. If data analysis and sampling knowledge bases develop over time, and not through isolated experiences of computer use, change in statistical knowledge will be observed only in long time frames. For this reason, it is necessary to track the evolution of a sub-group of students from the beginning of instruction to the intervention point. The effects of the computer-based interventions need to be understood in relationship with broader instructional contexts and with the evolution of students within those contexts.

Question 3 evaluates the possible tradeoffs between the two perspectives of statistical teaching considered in this study. It is possible that the target ideas of each statistical space may be sensitive to the type of instruction used to teach them, particularly, to the statistical perspective used in the intervention. For example, teaching sampling ideas might rest on the repeated observation of sampling distributions, and the data analysis perspective could not be able to provide this type of experience. It is possible also that there is a tradeoff between the number of exercises that sampling simulations can provide and the depth of data analysis exercises. In other words, it might not be possible to have equal amounts of practice in both perspectives because the average time of an authentic data analysis activity is higher than the average time of a simulation-based exercise.

1.6 CONTRIBUTION TO THE FIELD

This study expects to contribute to the field of e-learning by showing that effects of online education can be understood by attending to specific instructional characteristics, instead of to the general difference between online and face-to-face education. This study seeks to contribute to the field of statistical education by showing that statistical inference teaching requires both data analysis and simulation tasks. To demonstrate that, the expected results must show that the sampling condition produces larger effects in people's skill at identifying the effects of sampling on the observed data results, but at a cost in their skill at conducting exploratory data analysis (e.g. as seen in the exploratory task of the pretest). The results must show that the opposite effect happens when data analysis is used in the instructional condition. Learners in this condition must be able to conduct exploration of data but they must have problems to understand sampling. Additionally, this study intends to show that computer tools can help learners to understand inferential statistical concepts. In particular, the study aims to show that the use of computer-based simulations boosts the understanding of probabilistic concepts and sampling, and that the use of data analysis packages permits students to analyze large sets of authentic data and to develop data sense. Finally, this study seeks to shed light on the way that new technologies can change the distribution of knowledge at a global scale, by investigating the effects of computer-based resources on a group of Colombian students.

2.0 REVIEW OF THE LITERATURE

In this chapter data and chance, the two basic spaces of statistics, are introduced. After that, a historical review of the origins and development of statistics is presented; this review shows how data and chance evolved separately until they collided to produce statistical inference. Then, the relationship between data and chance in statistical education is reviewed. Finally, the current theories on reasoning, learning and teaching in both data and chance are examined, and the role of computers in each area of statistical content is analyzed.

2.1 DATA AND CHANCE: THE SPACES OF INFERENTIAL STATISTICS

Statistics is divided by its attention to data and to chance (Garfield, 2002). This distinction is not arbitrary, but reflects the implicit historical structure of statistics. Each branch represents a different way of defining statistical problems, as well as, conceiving the object of the discipline itself. One statistical tradition focuses on probability and on formal representations of events that contain uncertainty. Going back to Gauss, Laplace and De Moivre, this tradition emphasizes the assumption that, when aggregated, events that contain uncertainty tend to uncover tendencies in the form of typical distributions. That idea is sustained by two basic findings: First, that the sum of any number of given random variables tends to be distributed according to a typical distribution. Second, that, given a sample big enough, the characteristics of the sample tend to resemble those of the actual population. When a sample gets larger, the probability of an event in that sample approaches the actual probability of the event in the population (Nickerson, 2004).

The other tradition, closest to applied fields such as biology and social sciences, focuses on finding patterns in data, modeling phenomena in ways that maximize the explanatory power of theories. From a historical point of view, this tradition has its roots in the need to solve applied problems in conditions where multiple factors and unknowns determine the observable attributes of a situation, situations in which it is necessary to get a typical pattern that represents in some way the average behavior of phenomena that contains variability (Nickerson, 2004; Leinhardt & Larreamendy-Joerns, 2007). A typical pattern reduces a world that is full of variability to a manageable representation creating an informational gain. The interest in such typical representations grew initially from the need to deal with several observations of the same phenomena in physics, and later to account for large social phenomena during the industrialization process and the political consolidation of European states in the nineteenth century (Nickerson, 2004; Stigler, 1986).

The distinction between data and chance is not trivial and it encompasses two very different ways of understanding statistics. The view focused on chance is more Platonic because it privileges the formal structure of distributions over the need of finding patterns in actual phenomena. The second tradition, data, is more Aristotelian in that it focuses on modeling, assuming that the statistical task rests fundamentally in the use of multiple representational means to make the underlying nature of reality visible. Of course, good understanding of statistics implies coordinating both spaces, especially when it comes to inferential statistics. Any teaching restrained exclusively to one of those two spaces leaves students with fundamental gaps in their knowledge of the disciplinary content.

2.2 HISTORY OF STATISTICS: HOW DATA AND CHANCE EVOLVED AS SEPARATED DISCIPLINARY AREAS

Initially problems of data representation and modeling were considered independent of questions related to the mathematical properties of chance. On one hand, games of chance were known and studied intensively in Europe since the 17th century (Nickerson, 2004). On the other hand, data modeling started to be relevant with the flourishing of astronomy, because finding and using the right set of observations was critical to the attempts of modeling the astronomical phenomena (Stigler, 1986). Both areas of statistical thinking followed independent paths and developed separately of each other until they collided in some specific applied problems, in particular, finding a way to aggregate empirical observations (in astronomy) in the 18th century (Stigler, 1986). However, the integration of both areas, data and chance, to produce systems of statistical inference took until the invention of ANOVA and Regression in the first quarter of the twentieth century (Nickerson, 2004).

The first developments in probability came from the intent of understanding abstract games of chance. By the end of the seventeenth century and beginning of the eighteenth century, there was extensive work conducted by famous mathematicians such as Fermat, Pascal, Leibniz and Bernoulli, that described the abstract properties of games of chance. Those theoretical exercises inspired the first notable findings in probability. Among them, the work of Bayes in inverse probability, and the law of large numbers proposed by Bernoulli in 1713. These pioneer insights created an environment of intense intellectual exchange, that, in the end, led to the isolation of the properties of a subset of random distributions (binomial distributions). From the knowledge of those properties, deMoivre developed a proof of a special case of the central limit theorem that was extended by Laplace around 1800 (deMoivre, 1796; Nickerson, 2004; Shafer, 1993; Stigler; 1986). These mathematicians were working on solving specific abstract problems. The deMoivre books were devoted to solving specific probability problems. Both “De mensura sortis” and “the doctrine of chances” were collections of problems and solutions, but they did not include extensions of these solutions to principles, or connections with applied problems (deMoivre, 1796; Nickerson, 2004).

It was the need of having reliable astronomical observations that brought mathematicians initially to deal with applied problems, and, in this way, created the connection between the space of data and the space of chance. The flourishing of astronomy in the eighteenth century created a symbiotic relationship between theorists and observers. Basically, theorists needed to obtain reliable observations to carry out any type of theoretical modeling. The collaboration between Kepler and Tycho Brahe is one of the more known of these associations. However, until 1745, the accuracy of the observations depended on the credibility of the observer and his telescope. Methods of aggregation were non-existent.

The first attempt at developing a reliable way of aggregating observations was done by Legendre who was working on several projects for the French government. He participated in large-scale measurement projects (e.g., the distance from Dunkirk to Barcelona) that required combining several astronomical observations. It was because of these projects that he came up with the initial formulation of the method of least squares, published in 1805. The method of least squares was a way of finding a typical value, in the center of diverse observations that were supposed to contain errors caused by multiple imponderable factors. The method was quite successful and quickly extended through Europe (Stigler, 1986).

Legendre’s method provided an important tool for aggregating observations, but it was not probabilistic in nature. It was Gauss and Laplace that developed a probabilistic structure for the observation errors, before data and chance could come in contact for the first time. In 1809, Gauss published *Motus Corporum Coelestium in Sectionibus Conicis Solum Ambientium* that explored how the motion of planetoids was affected by large planets. In an appendix of this book Gauss proposed that the observation errors distributed normally around the mean. However, that was more an assumption than a formal proof. In fact, as Laplace noted, the argument provided by Gauss was circular. Laplace had the tool to make it non-circular. Laplace used and perfected the central limit theorem, and when he read Gauss' book, he realized that there was a relationship between both his central limit theorem and Gauss' idea (Stigler, 1986). This connection permitted Laplace to connect the central limit theorem to linear estimation, and in this way to create the connection between probability and data examination. He published this finding in the *Théorie Analytique des Probabilités* in 1812, and started to use the method in applied problems (e.g., studying the tides of the atmosphere) around 1823. There was still a distance before probability theory and data examination completed their intertwining. Two processes were fundamental. First, the contact of different disciplines with statistics, contact that leaded to the widespread practice of using statistical indexes to describe natural and social phenomena. Second, the creation of systems of inference that would transform the knowledge of random distributions into reliable inferential methods.

The first process was helped by the growing interest among ninetieth century governments in the measurement of social variables. Demographic data (e.g., births and deaths) had been gathered in England from the 17th century, and around 1860, there was an increasing number of statistical periodicals that produced tabulations of different types of data. The availability of this information produced an explosion of early attempts to make sense of them. Emergent disciplines known by different names (political arithmetic, social mathematics, moral statistics) intended to produce valid interpretations of this information. However, people interested in these new disciplines were more focused on the political implications of the data than on any mathematical understanding of them. This trend continued even after physics and biology had started to use complex statistical tools to model data in the first quarter of the twentieth century. The tendency to look for patterns in data, and to ignore the mathematical aspects of modeling, outlived the birth of mathematical statistics. Data was more popular than chance in statistics, even when the connections between data and chance (observations and probability) were understood. Oberschall (1987) notes that in the second half of the nineteenth century, "the same impressionistic and arbitrary eyeballing techniques were used to argue" for or against relationships among variables (Oberschall, 1987, p.107).

There were some mathematicians involved in serious efforts to apply mathematical tools to data analysis in natural and social fields. In France, Quetelet worked, at the beginning of the nineteenth century, analyzing the distributions of conscripts' heights, convictions, and suicide rates in search of what he called "the average man." Around the same time, Laplace discovered that the number of dead letters in Paris' postal system was constant from year to year. However, these efforts did not permeate the mainstream of "political arithmetic," nor did they modify the way theories were built until almost 100 years later when the tools necessary for statistical inference were developed (Nickerson, 2004).

Probability influences statistics and data analysis in at least two ways: First, it is fundamental to the treatment of errors of observation. Second, it is the base for inference. The first influence was developed and used before 1900; the second one comes from the developments made in the first half of the twentieth century by Galton, Fisher, and Pearson (Nickerson, 2004; Stanton, 2001; Stigler, 1987). These systems of inference were built to test specific models in different content areas, such as biology, and they connect the space of data and the space of chance in a totally different way: They compare actual patterns of data with what would be expected by randomness. This procedure of connecting expectation and observation is the common denominator of all the hypothesis-testing methods and it represents the final contact point between data and chance.

The structure of statistics as a discipline reflects its historical development. Both data and chance are studied separately in most courses, and they collide when inference and hypothesis testing are introduced. In the next section, we will explore from the perspective of statistical education how the structure of statistics as a domain reflects its history.

2.3 DATA AND CHANCE IN STATISTICS EDUCATION

The distinction between data and chance is not only a historical one. It delineates the way we think about statistics and shapes the way we teach it. This distinction permits statistics to produce a type of argument that combines data modeling and probabilistic theory to produce inferences about reality. Statistics cannot be restricted to probability because there is in the statistical argument the need to deal with reality and discover regularities. It cannot be restrained to find patterns in data because statistics implies the theoretical exercise of looking for generalizability. Statistics' goal extends beyond both. It is concerned deeply with the use of datasets to evaluate hypotheses in the context of theoretical debates. Statistics resolves theoretical disputes that otherwise would go on forever between different scientific perspectives (Abelson, 1995).

The statistical argument makes an inductive case for science feasible. It does so by using chance as a contrast medium for data. Statistical inference assesses the extent to which the pattern observed in data could be attributed to random effects. Another way of seeing this is that statistical inference procedures evaluate how much variation in the estimators that describe the data can be expected by randomness (Confidence Intervals), and to what extent we can be sure our conclusions represent the underlying nature of the data. By understanding the relationship between sample and population, and between statistics and parameters, we can produce theories without having observed all instances of a phenomenon. By addressing and measuring variation through chance theory, statistical inference permits us to talk about typical behavior in the presence of data that does not behave totally in accordance with the predictions of our conjectures.

These two ideas are in some way counter-intuitive and represent a challenge for learning and teaching. Dealing with variation is complicated because variation nullifies or at least challenges generalization. When statistics produces results that average several scores and minimize error, it is violating assumptions that we share and use in non-scientific discussions. The second idea, the process of extracting conclusions from a sample and attributing them to general abstract populations produces a cognitive challenge for most learners. The coordination between the information obtained from the sample and the conclusion about the population is governed by rules that are unfamiliar to most people.

To understand statistics, learners have to comprehend the nature of sampling and the omnipresence of variation, as well as the relationship between sample and population (Garfield & Ben-Zvi, 2002). Students need to understand probability as a measure of uncertainty. They need to know also how to develop and use models to simulate random phenomena and how to produce data to estimate probabilities (Garfield, 1999). More important, they need to coordinate and apply the probabilistic knowledge to the data they are analyzing, to the preliminary conclusions they are working on, and to the theoretical spaces in which they are involved.

Good statistics instruction includes both data and chance, whether instruction is in texts or online. Sampling, hypothesis testing, the measurement of uncertainty, and the quantification of variability rest on probabilistic ideas (Cobb & Moore, 1997; DelMas, Garfield & Chance, 1999; Garfield, 2002; Tappin, 2000). Data, chance, and inference are parts of most statistical courses. "The big picture" of statistics includes understanding the process of producing data, conducting exploratory data analysis (EDA), and making inferences from the sample to the population. More specifically, good statistics instruction includes EDA, the concept of sampling variability, the different approaches to producing data, and the logic of inference (Lovett & Greenhouse, 2000; Meyer & Lovett, 2002). Understanding these topics requires a basic understanding of probability. This idea seems to be a shared one for authors of online courses. Corredor & Leinhardt (2006) in a content review of six online courses showed that there were independent sections devoted to data, chance, and inference in all of them (See Appendix A).

For more than 70 years, statistics was conceived of and taught as a branch of mathematics that rested on probability. Just as with other branches of mathematics (e.g., algebra), however, statistical instruction now focuses on the way real phenomena can be quantified and then examined. Cobb and Moore (1997) suggest that the best way to teach statistics is through EDA first, and then to describe the mathematical complexities of statistical inference. Focusing too much on the abstraction of complex mathematical arguments and probability theory is currently thought to be the wrong path for learning (Tappin, 2000). Lovett and Greenhouse (2000) point out that they focus their curriculum on "the use and interpretation of data analysis techniques without teaching all the probabilistic or mathematical underpinnings." In the same sense, Snee

(1993) proposes to teach statistics in a way that emphasizes data collection, understanding, and modeling, instead of focusing on the probabilistic aspects of inference.

2.4 CHANCE

There are two main positions regarding the nature of statistical thinking in the space of chance. On one hand, one position considers that statistical thinking is an unnatural act. People think about probabilistic concepts in ways that contradict the predictions of probability theory. From this point of view, understanding probability does not arise spontaneously and training has a modest effect in correcting misconceptions (Konold, 1995; Kahneman & Tversky,1972). On the other hand, the other position asserts that people develop intuitive notions of probabilistic concepts and possess heuristics that permit them to deal to some extent with basic statistical activities. These intuitive notions arise from the contact with situations that present random behavior (e.g., a random number generator) (Nisbett et al, 1993) and they are used in several reasoning processes (e.g., everyday induction). From this point of view, the results of Kahneman and Tversky (1972) can be explained by the structure of questions, and the experimental contexts in which those results were found. People possess heuristics to solve statistical problems, but they do not apply that knowledge to Kahneman and Tversky's tasks. From the perspective of Nisbett et al (1993), people do not recognize those tasks as random and they use other explanative mechanisms to understand them (e.g., deterministic models). People's use of statistical heuristics depends on factors such as the clarity of the sample space or cultural prescriptions on which situations are random.

Kahneman and Tversky (1972) showed that people used non-probabilistic heuristics to calculate gains and loses, and that these heuristics did not follow what would be a standard probabilistic algorithm (Kahneman, Slovic, & Tversky, 1982). These studies suggest that probabilistic thinking is unnatural and that people possess misconceptions that are resistant to change even after training (Chance, DelMas, & Garfield, 2004; Konold, 1995; Konold, Well, Pollatsek, & Lohmeier, 1993).

2.4.1 Misconceptions about Probability

2.4.1.1 The Equiprobability Bias

There is evidence that people tend to believe that all events have the same probability in spite of the population characteristics. In other words, people tend to think that everything is equally likely to happen (Konold, 1995). For example, when extracting a random student from a class with more women than men, people tend to think that the probability of selecting a man is equal to the probability of selecting a woman, or that the different outcomes of rolling dice are equally probable (Lecoutre, 1992). The equiprobability misconception might be addressed by providing resources that simulate random phenomena in a rapid and easy form.

2.4.1.2 Outcome Orientation

Students have an intuitive model of chance in which probability values represent single event outcomes instead of tendencies in series of events (Garfield, 2002; Konold, 1989). When told, for instance, that there is an 80% chance of having a sunny day, people assume that there is going to be a sunny day (Konold, 1995). The outcome orientation is resistant to change under instruction (Garfield, 2002). There are several reasons for this phenomenon: Interpreting probability values in terms of single event outcomes can be produced by a global misunderstanding of probability theories. Probability is only observable in aggregated events; it is never visible in single events. But everyday life is made up of single events and people are accustomed to thinking in those terms. Events, to say it in other words, have a p=1. As we will see in the following section, the outcome orientation presents an opportunity for computer remediation in the learning of chance.

2.4.1.3 Sample Homogeneity and Law of Small Numbers

Students tend to think that all samples resemble the population from which they were obtained regardless of the sample size. This idea contradicts probability theory which states that the smaller the samples, the higher the variability among them and the higher the probability of finding extreme sample means (Kahneman, et al, 1982). The idea that after some point increasing the sample sizes does not improve the power of the analysis seem to be hard to understand for most people. The power of a well-drawn small sample is not easily understood (Kanhneman et al, 1982). Students often consider samples simply as arbitrary subsets of the populations that do not include variability or sampling effects. This erroneous conception of sample leads students to reason in flawed ways about statistical inference (Saldanha and Thompson, 2003).

2.4.2 Role of Computers in Chance

Misconceptions that surround probability are present even in students who undergo training (Chance, DelMas, & Garfield, 2004; Konold, 1995). This result is confirmed across cultures and settings (Jun & Pereira-Mendoza, 2003). It is almost as though there is something deeply unnatural in probabilistic thinking, or that, not having a conception of chance people reason from everyday experience (Fischbein, 1975). Addressing these misconceptions is something that we argue can be done especially well by certain affordances available in computer-supported environments. Computers can help people to understand probabilistic concepts by providing learners with simulations of random situations. Even if the idea of Nisbett et al (1993) is more correct than Kahneman et al's one (1982), and people do possess adequate statistical heuristics but fail to recognize many situations in probability instruction as random, then, computers can help to build the bridge between people's heuristics and instructional tasks.

The misconceptions that people have can be classified into two groups. The first group of misconceptions involves problems with understanding the process and characteristics of sampling. This group of misconceptions includes the sample homogeneity, and the large-sample misconceptions, and the law of small numbers. Computers can provide instantiations of sampling processes and make it easier for people to experience the sampling process.

The second group of misconceptions relates to flaws in understanding the meaning of probability values. These misconceptions include outcome orientation and the equiprobability bias. Probability values are hard to understand because probability is only visible when events are aggregated over several trials. Thinking in probabilities is an unnatural act because it means thinking about something that could or could not happen. A statement like the following: "next year, 15% of Americans will be sick with flu," points to many different configurations of events . Probabilities represent stable properties of chance setups, but they do not represent steady facts. What does exactly it mean to have a probability of 1/2? It means that, when aggregated, you will have the event in half of the trials. But if you only have one trial, then what? Actual events have probability 1, and we learn to use heuristics that do not involve uncertainty when solving problems (e.g. explaining events in terms of causal mechanisms). Most instruction leads us to look for informational gains that produce exact answers. Computers can help people to understand probability values because they can reproduce random behavior in basically time-free frames. In some sense, a large part of traditional instruction led us to think most phenomenon as not being random; for example, using a deterministic causal model to describe natural situations.

The origin of both these groups of misconceptions relates to problems in the comprehension of mathematical ideas that require a knowledge of mathematical concepts (e.g., proofs) in order to be understood. Why should one believe that, for example, the sample size influences the likelihood of obtaining extremes values? Of course you can use a mathematical structure to demonstrate why you should expect random sampling to behave that way. But teaching through mathematical proofs has two problems. First, it requires sophisticated prior knowledge and mathematical skills from students, knowledge and skills that many students lack or that they do not consider an important part of their statistical training. Second, focusing on formal proofs takes instructional time away from more grounded authentic activities. Formal mathematical structures are important for students who have interests in mathematical proofs and logic; but they are not as relevant for students with more applied interests. In fact there is a tendency in statistics education to emphasize mathematical formalisms less because there does not appear to be transfer from formal probability theory to the applied activities in statistics (Cobb & Moore 1997; Lovett & Greenhouse, 2000; Snee,1993; Tappin, 2000).

Computers may in fact fulfill a role once reserved for the mathematical proof. It is possible to create dynamic representations of mathematical objects in which students can actively interact with the mathematical properties of those objects. More important, there is now the capacity to represent randomness in a concrete way. Randomness is only visible when you aggregate events over several trials. In the past, that fact was often ignored, represented through the mathematical formalism, or in the best cases explained by physical simulations of random process (e.g., tossing coins) (see, for example Schwarz & Sutherland, 1997). But there was no way, until multimedia computational technology made it possible, to show how events aggregated over several occasions create tendencies (Cramer & Neslehova, 2003). Now, with computer-supported tools, it is possible to simulate random processes, condense events over time, and see how tendencies change as a function of different parameters. Tedious physical demonstrations can be replaced by quick, focused, computer-based simulations. The scope of computer simulations goes beyond replacing physical simulations and facilitating the presentation of randomness in the instructional processes. By its characteristics computer simulations permit students to connect the random processes with different graphical representations and numerical indexes. In this sense, simulations serve as devices to observe phenomena that cannot be observed under normal conditions (DelMas, Garfield, & Chance,1999; Snir, Smith, & Grosslight, 1995).

For the misconceptions described in the literature, computers seem to provide a possible solution to the instructional problem. For the first type of misconception, computers permit one to conduct simulations of sampling processes and to demonstrate many of the sampling distribution characteristics, characteristics that would be otherwise stated abstractly. Blejec (2002) has proposed that graphically supported simulations can be used as proofs of statistical concepts for students what do not have sophisticated mathematical knowledge.

For the second type of misconception, the type that involves misunderstanding the meaning of probability values, computers permit aggregation of events over several trials and in that way to clarify the meaning of probabilistic indexes. Running random simulations permits students to see how probability values represent not steady 1-0 distributions, nor chaotic distributions of data, but something in the middle: distributions that over time tend to fall in tendencies defined by the probability values. In the following paragraphs, we present recent experiences in the use of computer simulations to teach probability.

2.4.2.1 Random Simulations as Tools to Teach Chance

In the literature, two types of activities are referred to as "use of simulations." First, there is the use of ready-to-use simulations that contain all the elements necessary to be used by the students (programmed as Java Applets, or as objects in other programming environments) (e.g., Blejec, 2002). Second, there are built-up simulations, that is, simulations that the same students, or the professor, build using spreadsheets, statistical packages with random number generators, or programming languages[1]. Both types of simulations usually depict probabilistic concepts by selecting several samples of a given size from different distributions to create graphical and numerical displays of the repeated sampling process. Each type of simulation has advantages and disadvantages. An advantage of available graphical representations in the first type of simulation is that they permit one to map abstract statements into the representations directly. A disadvantage of the first type of simulation is that students skip over steps in the representational process, steps that can be critical for understanding and performing authentic statistical activities. Skipping these representational steps, when for example sampling by random selection from an abstract normal distribution, can reduce the confidence of learners in the representation

[1] This version simplifies the one proposed by Mills (2002) by condensing in the first category the three categories that require student programming, and in the second the types of simulations that are ready-to-use in Mill's taxonomy.

as an authentic representation of the world, creating the sort of disbelief in the simulation discussed by Velleman and Moore (1996).

Table 1. Examples of Random Simulations

Experience	Type	Pedagogical Experience	Assessment
VESTAC (Darius et al, 2002)	Ready-to-use	NA	NA
Nicholson et al, 2000	Ready-to-use	NA	NA
EMILeA-stat (Cramer & Neslehova, 2003)	Ready-to-use	NA	NA
MyJavaStat (Harner & Hengi Xue, 2003)	Ready-to-use with parameters	NA	NA
Wood (2005)	Ready-to-use with microworld	NA	NA
The Probability Explorer (Drier , 2000)	Ready-to-use with microworld	NA	NA
Shaughnessy & Ciancetta, 2002	Ready-to-use with microwrold	Students run several simulations with virtual "fair spinners"	Measured by NAEP Items
SAMPLER (Wilensky, 1999)	Ready-to-use with microworld and students' emergent input	NA	NA
DelMas, Garfield and Chance (1999)	Ready-to-use	Problem solving	Improving when evaluation items similar to task
Sanchez, (2002)	Built-up	Students solve questions about random distributions	Students considered the simulation useful
Blejec (2002)	Built-up	Students use simulated data to solve problems	NA

Among the topics covered by simulations are the central limit theorem, *t*-distribution, confidence intervals, binomial distribution, regression analysis, sampling distributions, hypothesis testing, and more recently, ANOVA , regression, *t*-test, and chi-square. In all cases, stochastic processes are used to give students inductive experiences with the concepts involved in the simulation (Kersten, 1983). A problem with this type of simulation is that it lacks authenticity and concentrates on the abstract characteristics of probability. This is not a problem when one is dealing with abstract probabilistic concepts (e.g., sampling distribution) but it becomes more serious when the objects of the simulation are close to real-world activities.

Authenticity requires teaching statistical concepts linked to data analysis activities. Simulations of the first type tend to ignore this requirement when dealing with topics such as hypothesis testing, ANOVA and *t*-tests. Table 1 presents a summary of the available programs with random simulation described in the literature. A more detailed description of these projects is available in Apendix B.

2.5 DATA

Data refers to activities whose goal is to explore data and discover the underlying patterns and tendencies in it. The data space also deals with the articulation of the process of data representation with disciplinary questions and theory. In other words, the space of data has to do with manipulating information to see patterns, relationships, and explanations in the context of disciplinary practices (e.g., theoretical background, measurement systems, representational tools, etc.) (Tukey, 1977).

Interest in the space of data as a legitimate area of statistical learning research has been growing since the 1970s (Leinhardt & Leinhardt, 1980; Tukey, 1977). The so-called "data analysis revolution" (Biehler, 2003; Shaughnessy, Garfield & Greer, 1996) has advocated that EDA (Exploratory Data Analysis), not probability, should be the focus of statistical education. Several factors have coincided to favor this change of focus. The first factor is the realization that transfer does not happen spontaneously and that it depends to a great extent on the similarity between the source and target task (Singley & Anderson, 1989). If this is so, formal training in statistics does not transfer necessarily into adequate use of statistical knowledge in applied contexts. The second factor is discovering that authentic tasks, tasks that contain relevant disciplinary traits and that resemble real contexts, predict better long-term learning and performance in non-school settings. A third reason supporting the interest in data is the availability of user-friendly computer software that permits learners and practitioners to move away from procedural tasks, and concentrate on the conceptual aspects of statistics (Velleman & Moore, 1996). New statistical packages and other computational tools change the goals of instruction because many tasks that used to be carried out by people, from finding sums of squares to looking up critical values in statistical tables, now can be conducted by computers.

A primary idea of EDA advocates is that the nucleus of the statistical practice and teaching is the ability to organize the complexity of real data into clear recognizable patterns (Leinhardt & Leinhardt, 1983). This idea emerged as a reaction to the extreme formalism of the probabilistic approach that often built models that did not account for any practical situation and were only valid when unrealistic assumptions were accomplished. The assumptions behind the preference for real data are that real data possesses some type of structure and that the task of researchers is isolating that structure from the surrounding noise (Biehler, 1995). The data structure exists in the overlap of data configurations and theoretical perspectives, (e.g. the categories that you are using to analyze the data). The EDA perspective assumes that core elements of the practice and activity of data analysis cannot be captured by the mathematical formalism or taught by explicit direct instruction to inexperienced learners. Working with simplified data sets can be beneficial to learning of chance, because, for example, it facilitates the calculations and permits observation of the underlying mathematical mechanisms of hypothesis testing; but working in artificial settings does not help learners to develop essential skills for data analysis and interpretation (Snee, 1993). Working with real data, on the other hand, forces students to deal with messier sources of information and, in this way, it forces them to develop critical statistical skills.

Dealing with real data is only one of the aspects that the EDA perspective proposes. More generally, EDA tasks ask students to be involved in authentic statistical activities. Authentic activities resemble actual research pursuits and therefore they highlight critical aspects of statistical inquiry, such as defining the measurement system, choosing and modifying the scales, conducting the necessary transformations in the data, and selecting the representational means. Authentic activities imply more than that, they require learners to go beyond the statistical context and face many traits of actual scientific tasks. Authentic tasks embedded in disciplinary contexts prompt students to deal with aspects of learning that remain obscure when students are just presented with artificial simplified versions of statistical problems, with activities centered in mechanic calculations, and with the mathematical formalism. They push students to participate in the definition of the problem space (e.g., picking the attributes to be measured), the selection of representational tools, and the process of coordinating empirical evidence and theory. Instead of employing ready-to-use (ready-made) measurement tools and calculation methods, educational interventions based on authentic activities require learners to make sense of the actions in terms of the theory being evaluated and the questions being solved (Lehrer & Schauble, 2004). It is assumed by EDA supporters that dealing with authentic activities and participating in problem solving situations develops students' data sense (Burgess, 2002).

Data sense refers to the ability to interpret in a logical sense, the statistical results and the graphical representations of data. It refers also to the ability to generate information "on which graphs and statistics are constructed" (Burgess, 2002; Friel, Bright, Frierson, & Kader, 1997). In other words, data sense refers to an intuitive sense that permits students to put data in context, give meaning to it, and create sensible interpretations of it, as well as to understand what the data means in the context of a theory.

An argument in favor of introducing authentic activities in the curriculum is that the skills and knowledge needed to interpret and conduct meaningful use of data do not arise spontaneously. It is well known that when people are faced with data analysis tasks, they tend to underuse the data. In a strict sense, they do not make mistakes (e.g., they do not make errors in the calculations) but they often organize the activity in the wrong direction, in ways that ignore the nature of statistical inquiry. Several of these problems with data interpretation have been identified in the literature. Ben-Zvi and Arcavi (2001) have elaborated the distinction between local and global views of data. Local views of data happen because learners focus on individual values or small subsets of data, instead of building an interpretation on the whole available information. Global views refers to interpretations that identify patterns and tendencies that enclose the complete data set. These authors found that statistics experts combine local and global views of data when building interpretations whereas novices focus exclusively on local views. This assertion is consistent with the idea of Konold et al (1997) who asserts that what makes a task statistical is that it encompasses building descriptions and conducting comparisons of non-homogeneous set of data. Global descriptions of non-homogeneous data sets require accounting for the variability inherent to those data sets. If you have to compare, for example, the weight of two men measured once, there is not much room for confusion, debate or interpretation. If you have to compare the weight of two groups of men or the weight of two men measured several times, then, you have to make use of to more sophisticated tools and reasoning skills. Variability in the data makes the comparison hard. Learners tend to focus on particular scores and formulate interpretations based on restricted data spaces. EDA proposes that learners need to be exposed to authentic activities of data interpretation in order to develop the skill to think in terms of group tendencies and look at data globally; skills that are necessary for understanding variability, building generalizations, finding relationships among variables and using hypothesis testing procedures (Burgess, 2002; Ben-Zvi, 2002).

Many problems that students experience when dealing with sets of data arise from their tendency to think about data in terms of properties of individual cases or homogeneous subsets of data (Konold et al, 1997). This tendency can be an extension of everyday and school practices and strategies, where dealing with individual cases and homogenous representations is the rule (Lehrer, 2002; Petrosino et al, 1997), to the realm of statistical activities, where that kind of reasoning is not longer valid. Statistics experts think in terms of propensities, that is, in terms of properties of non-homogenous "data aggregates" (Konold et al, 1997). EDA supports the skill of thinking in terms of "propensities" through participating in complex statistical tasks in the early stages of instruction.

Another possible extension of limited reasoning acquired through non-statistical everyday practices is people's tendency to prefer concrete, punctual representations of data when carrying out statistical analysis. Konold (1995) reports that students have a strong preference for two-way tables and absolute frequencies with precise values over other types of representations as histograms and boxplots. It is possible that learners stick with concrete representations due to problems with considering permutations in the data representations. These permutations, called transnumerations by Wild and Pfannkuch (1999), are considered an important part of statistical thinking. The ability to work with multiple representations is central to the understanding of mathematics in general (Dreyfus & Eisenberg, 1996; Leinhardt et al, 1990) and statistics in particular (Burgess, 2002; Cai & Gorowara, 2002). In the case of statistics, representational permutations permit learners to generate different interconnected data representations in order to attain a deeper understanding of the statistical situation (Ben-Zvi, 2002; Burgess, 2002).

Concrete representations mask the existence of variability in data sets. Frequency counts and central tendency indicators without any corollary or context are blind to the variability that surrounds them. Students focus on this type of concrete representation because everyday and school practices to which they are exposed neglect this facet of data and reality. School-based instruction ignores learners' need to be exposed to and taught to deal with variability in data (Kazak & Confrey, 2004; Petrosino et al, 2003; Wild & Pfannkuch, 1999). Statistical reasoning requires acknowledging variability as a central feature of data and reality (Garfield & Ben-Zvi, 2005). Variability is what creates uncertainty in the conclusions and, therefore, it is what gives statistical character to descriptions and comparison with multi-case datasets (Konold, 1997). Konold and Pollatsek (2002) point out that dealing with variability is what differentiates generalization about variable relationships in the context of variable manipulation (e.g., Masnick et al, in press) from generalization in the context of statistical research. In this sense, statistical instruction must expose students to activities that foster understanding of variability.

Understanding variability has several interconnected facets. It implies understanding that any reasonable account of data includes both the description of typical patterns and references to the surrounding variability. It implies acknowledging that variability nuances the interpretation of data but that variability falls into recognizable patterns when adequately represented. In other words, understanding variability implies believing that variability is not totally unsystematic, and that, as with chance, its form only becomes visible when you look at data globally. Understanding this is basic to master the process of iterative data representation, a process that produces the data distributions that describe the predictable aspects of variability (Petrosino et al, 2003).

Understanding variability means understanding that there are two basic sources of variability in statistical models: the variability generated by measurement error (e.g., the error of astronomical observations) and the variability that comes from the object being studied (e.g., phenotypical variation) (Leinhardt & Larreamendy-Joerns, 2007; Petrosino et al, 2003). Both sources of variation distribute normally but they represent different aspects of scientific research. We explore sources of variation to gain information and to reduce uncertainty, on natural and social phenomena, for example, by aggregating predictors in a model that explores through covariation the relationship between two variables (Garfield & Ben-Zvi, 2005; Wild & Pfunnkuch, 1999).

Variability is not just some part of the nature of scientific research, it is the organizing concept of statistics. Garfield and Ben-Zvi (2005) consider that understanding variability gives students the main tools to operate in statistics. Graphical representations of data are to a large degree designed to represent variability and they help to separate signal from noise. Students learn to handle the data in ways that maximize the informative power of graphical and numerical summaries, controlling unwanted forms of variability (e.g., outliers, observational biases) and conditioning the use of certain tools to the presence of certain characteristics in the data (e.g., parametric tests and normal distributions). Describing variability is also fundamental to comparing data sets, to exploring the accuracy of a model (for example, when classifying explained and unexplained variability), and to qualifying the informative power of statistical indexes (Garfield & Ben-Zvi, 2005).

Ignoring variability as well as focusing exclusively on concrete values when conducting data analysis are problems that have a common root: the lack of conceptual understanding of statistics' aim and tools caused, in part, by traditional teaching that is focused mainly on algorithms (Cai & Gorowara, 2002). Students accustomed to mechanical calculations are unable to give conceptual meaning to different types of statistics (mean, standard deviation) even when they command the algorithms necessary to come up with numerical values for these statistics (Batanero et al, 1994). It has been shown also that use of statistics (e.g., regression, ANOVA) does not happen spontaneously even when the students have the procedural knowledge necessary to conduct the calculations (Ben-Zvi, 2002; Gal, Rothschild, & Wagner, 1990). Among the causes for this situation are an incomplete or non existent understanding of the need for global views of data (Ben-Zvi & Arcavi, 2001), limited experience with the conditions of use of the statistics (e.g., mean vs median in the presence of outliers), and the inability to find meaningful representations of what the statistic means both in graphical and numerical form (Watson & Moritz, 2000).

2.5.1 Computers and Data

Given this situation, the requirements of instruction for data analysis and the limitations of existent teaching practices to develop this knowledge in students, how can computers enhance statistical education in the space of data? There are several possibilities. One is that computers can provide dynamic visualizations that enhance the understanding of statistical concepts. In the same way that random simulations can work as proof of concepts in chance, dynamic visualizations can work as proofs, or at least as enhanced demonstrations of statistical concepts in the data space. A second possibility is that the capacity of computers to generate representations of data can help learners to operate beyond concrete numerical values (e.g., frequency counts), and to create several representations of statistical concepts or situations. In this sense, computers provide visual representations that can be used as analytical tools (Garfield, 1995). Third, statistical packages, the Internet and the large memory capacity of computers permit one to conduct authentic research in classroom settings. Computers permit students to access large data sets collected from real situations and to explore those sets without the huge computational costs that existed before (Finzer & Erickson, 2005). There are two basic uses of computers for teaching in the space of data: exemplification of statistical concepts using dynamic visualizations, and data analysis activities powered by statistical packages and the available large databases.

2.5.1.1 Dynamic Visualizations

During instruction, it is often necessary to exemplify statistical concepts and their properties. In traditional instruction, concepts are normally defined by their conceptual and algebraic formulas. The conceptual formulas permit one to derive the characteristics of the concept, by deduction, for those students who have the knowledge and skills to deal with the mathematical definitions. The computational formulas permit one to exemplify the concepts and their characteristics inductively from a case. However, the computational costs of exploring the concepts using the computational formulas, when done by hand, are large and the task seems meaningless for most students. Computers offer an alternative to traditional methods: dynamic visualizations.

Dynamic visualizations are displays that work graphically and numerically on sets of data. In a dynamic visualization, learners operate on objects or icons to produce changes in the relationships among objects in the visualization and in numerical indicators associated with the concept being represented. For example, a representation of the arithmetic mean can show the different cases in the data set as points in a Cartesian plane, and the mean as a line in the middle of them. Learners can move the line up and down and can see how the sum of the distances from the points to the line changes. More examples of dynamic visualizations are presented in Appendix C.

A problem with dynamic visualizations is that they are artificial. The data sets are often created by the programmers; even when the data come from real situations, the context of the tasks is fixed and the decisions to be made by learners are limited. It is important to remember, as Hawkins et al (1992) proposed, that adding context to a data set is not beneficial for learning or assessment unless there is a meaningful purpose in the tasks.

2.5.1.2 Exploring Data through Computers

A more sophisticated use of computers to teach statistics is the use of databases and statistical packages to explore data in the context of authentic activities. There is a growing number of data sources on the internet and the offer of statistical packages for general and pedagogical purposes is huge (SPSS, Minitab, Tabletop, Fathom, Dynamic Statistics, Tinkerplots). It is important not to confound this type of activity with the use of dynamic visualizations. The difference between both is analogous to the difference between ready-to-use and built-up simulations. While dynamic visualizations present fixed situations, data analysis situations through statistical packages are flexible and permit learners to define several aspects of the task. The cost, of course, is that students need more background knowledge to use statistical packages. In spite of their differences, use of dynamic visualizations and data exploration using statistical packages help learners to move beyond concrete representations of data. Both types of tools permit learners to oscillate among different representations of statistical situations and in this way these tools help to develop statistical reasoning skills. Visualizations and data exploration through computers help students to see statistical problems from multiple perspectives, as well as, to learn how to make informed decisions among different representations and numerical summaries (Ben-Zvi, 2000; Biehler, 1993; Garfield, 1995).

Data exploration through computers is the computer-based equivalent to design experiments that require student to participate in authentic research activities (see, for example, Gravemeijer, 2002; Petrosino et al, 2003). In both cases, the activity requires students to analyze sets of data to come up with structures of distributions, to account for both underlying patterns and variability, and to connect the results with the questions being explored. A difference must be noted: In the case of computer-based activities, the measurement process is skipped and the representational process limited by the options provided by the statistical package. When computers work as graphical devices, it is easy to switch from one representation to a different one and to enhance the use of multiple representations without having a large workload to produce them (DelMas, Garfield & Chance, 1999; Snir, Smith & Grosslight, 1995). The cost is that you have to "believe" in the machine (Bakker, 2002). In this sense computers not only provide computational and representational power but they also change the structure of the instructional task (Ben-Zvi, 2000): The task is no longer to calculate a statistic or construct a graph; the task is to decide what to do. In the next pages, we will see some examples of uses of computers to teach data analysis and descriptive concepts in statistics.

Mori, Yamamoto, and Yadohisa (2003) focus on the pedagogical situation that the tools allow. They introduce DoLStat, a group of courses that use examples from online databases and authentic cover stories. In the perspective of these authors, using real data as well as credible authentic stories creates a new pedagogical situation and challenges students with the complexities of real statistical activity.

Other authors hold a similar belief. A practice that is getting increasingly common is using census or demographic data in class activities. This type of activity has the advantage of presenting an authentic situation that is also socially and personally relevant for students. CensusAtSchool (Connor 2002; Conti & Lombardo, 2002; Hooper, 2002) is a project that collected demographic and other type of data from students between the ages of 7 and 16 to create a national UK database (Connor, 2002). The project provided two questionnaires (ages 7-11 and 11-16) and the virtual infrastructure necessary to combine the data collected in schools across the country. The project's website received information from the schools and gathered it in a unified database. The website also distributed the data to schools and allowed students to draw random samples of 200 cases. Students participated in the collection of information in schools and in the organization of the information in Excel spreadsheets that were mailed to CensusAtSchool right after the adult census was conducted in the UK. Curricular materials for making use of the information were available in the CensusAtSchool website. Hooper (2002) reports that a similar version of this project was conducted in New Zealand using the same virtual infrastructure. Among the interesting aspects of this second project was that students could compare results from the UK and New Zealand. Other uses of census data are reported by Conti and Lombardo (2002), Frey (http://www.ssdan.net/tarek_test/), and Finzer and Erickson (2002). Finzer and Erickson (2002) describe the use of U.S. national census records in a Fathom-based curriculum. They developed a curriculum that exploited the representational and computational properties of Fathom to analyze U.S demographic data (obtained from the Minnesota Population Center). Using this program, students represented and visualized the characteristics of diverse statistical concepts to build interpretations of the Census data.

Most studies mentioned to this point present either computer tools or computer-based curricula but they do not present in-depth descriptions or assessments of pedagogical experiences. However, studies that provide evidence regarding the effects of computer tools on statistics learning are also available. Bryc (1999) presents a very interesting curricular experience in which students have to decode messages in scrambled texts (texts where the characters have been substituted by other characters using word tools). To solve the mysteries, students must use statistical tools. They have to use counters to identify the most frequent characters and then to compare them with the Standard English frequencies. Although only anecdotic evidence is presented, this study shows how an engaging activity can be created using basic computer tools.

McClain (2002) used TinkerPlots (a statistical software program for 4th- to 8th-grade students) to explore and visualize authentic sets of data. She found that participants developed strategies to manage the statistical complexity in the context of group comparisons. Students came up with methods to reduce variability by grouping data and methods to observe the characteristics of the new distributions produced by the blocking process. Participants started to use proportions (e.g., proportional reasoning) to interpret the relationship between bin size and group size in the representation of group distributions.

Rubin (2002) describes an episode of statistics teaching for teachers that asked them to compare two types of batteries using a small data set. The importance of Rubin's study rests on the fact that it compares the responses of participants under two conditions: doing the calculations by hand or using TinkerPlots. The findings show that when computational support is not available, participants tend to focus on the algorithmic process of calculation. When using TinkerPlots, they tend to focus on the representational and conceptual aspects of the task, such as, for example, evaluating the presence of outliers (which are invisible in hand calculations).

From this perspective, computers not only facilitate the activity but they also modify the requirements, goals, and constraints of the activity.

A more complete version of this idea has been presented by Ben-Zvi (2000; 2002). For him, there are several metaphors to understand the role of computers in statistical education. One is the amplifier metaphor, which sees computers as tools that carry out people's tasks but faster and more precisely. The second metaphor is the reorganization metaphor, which sees computers not as tools that facilitate the tasks, but as elements that modify the structure of the whole activity. From this perspective, computers have a similar role to that of memory devices and other material elements in distributed cognition (Hutchins, 1995). Ben-Zvi (2000) describes the "CompuMath" project, in which students have to use computers to analyze authentic data (e.g., 100 meter race times analysis) by graphing in different ways the available information (e.g., graphing with and without outliers) in order to support statements. Ben-Zvi shows that the goal of the activity changes: in CompuMath, the activity focuses on transforming representations of data patterns and variability. This task requires creating multiple representations and dealing with representational ambiguity. In this way, CompuMath encourage the process of interpretation by creating cognitive conflict between different representations. In the process of solving the conflict between different representations and building coherent interpretations, students come to global views of data (Ben-Zvi, 2002).

2.6 CHALLENGES TO SIMULATIONS, VISUALIZATIONS AND OTHER COMPUTER-BASED TOOLS

Ready-to-use simulations have become increasingly popular; they are easy to access and use, and they require little prior knowledge on the part of learners. However, they face to challenges. The first one is that the origin of the representation, the process and algorithm, and the connections between users' actions and numerical changes are not evident in the situation. An advantage of simulations and visualizations is that they permit the direct mapping of theoretical statements onto the representation. A disadvantage is that students skip over critical steps in the representational processes. If mapping is not performed adequately, the interaction between parameters and graphics is not clear. That is, learners might modify parameters without knowing how or why they affect the representation and the results in the screen. A more important concern is that, when the visualization or simulation is presenting counterintuitive results, learners dismiss it as not realistic (Velleman & Moore, 1996). A solution to this problem is, in the case of simulations, to highlight the relationships between the virtual and physical simulations (Drier, 2000; Shaughnessy,1992). In the case of visualizations, the solution requires students to explain the steps of representation and calculation and to provide some knowledge of the basic principles upon which the simulation is built, even when total knowledge of the algorithm is not necessary (Nicholson et al, 2002). One can, for instance, explain the algorithm with some a small number of cases and use computers to explore large sets of authentic data.

The second danger is that visualization and simulations can be used without performance standards by changing parameters without knowing what the task goal is. To avoid this danger, exploration of data through statistical packages can be more useful than use of free-exploration applets (Mills, 2002), because it pushes students to understand the inner logic of the used procedures. For this reason, statistical packages that permit and, in some cases, require students to write down formulas to conduct certain procedures, create a better learning situation for some topics (Finzer & Erickson, 2005). Building and programming a model to test a hypothesis can be far more interesting than using an applet whose internal logic is unknown for learners. Computer-based statistical teaching should create interventions similar to the work that Burill (2002) conducted with physical simulations. She asked a group of students to analyze the age distribution of the employees laid off by a company in order to see whether or not age had a significant effect. Students conducted simulations (by physical means) and after that compared the distribution of the data in the company's records to the distribution produced by random simulation. This example is interesting for two reasons: First, it shows how students can be introduced to modeling situations that push them to deal with ill-defined tasks. Second, it shows how data and chance collide in a concrete instructional situation: At the end, this task is an inferential one. Students had to deal both with data and with chance. They compared data distributions produced by random simulations and by data analysis, and in this process they explored the inner logic of statistical inference. This last example shows how rich pedagogical environments could be built around simulations. The use of computer tools in the teaching of statistics should follow a similar path.

This review of the literature suggests that understanding statistical inference requires understanding data analysis and probability, that simulations can enhance the understanding of probabilistic theory, and that exploration of data with authentic data sets and computer tools can develop students' data analysis skills. The next chapter describes the way in which this study examines the questions elaborated in Chapter One according to the literature reviewed in Chapter

Two. That is, the effects of simulations and data analysis tasks on the learning of inferential statistics.

3.0 METHOD

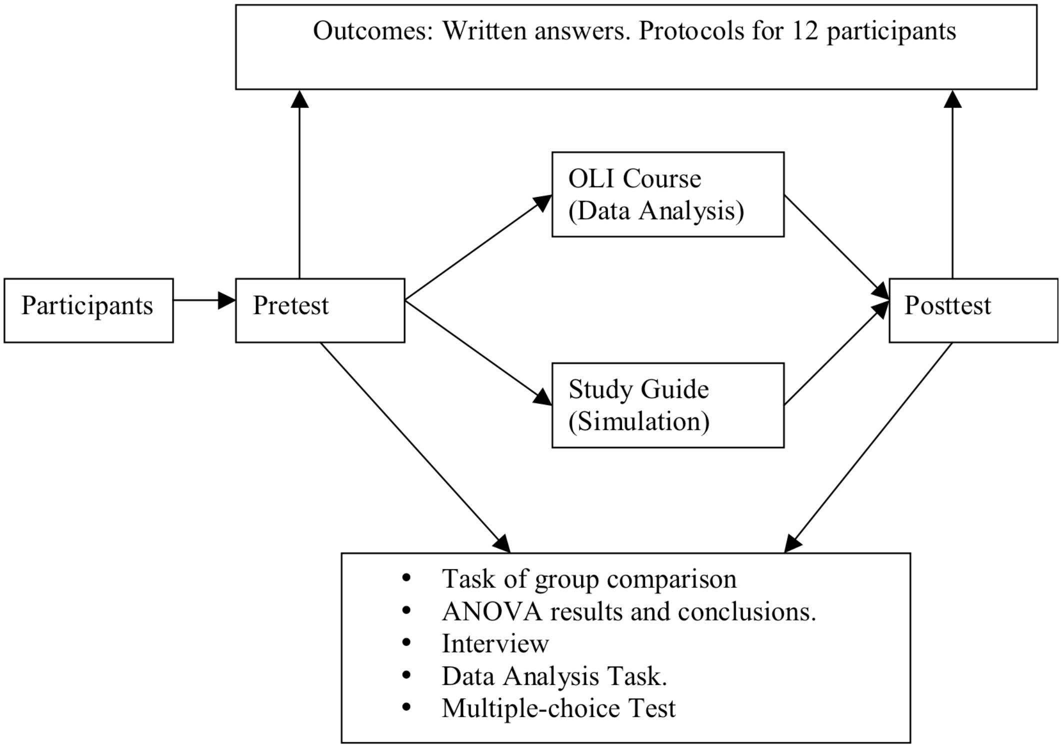
This chapter describes the methods used to address the questions posted in Chapter One. Based on the stated purpose of this study, the design and methods aimed at informing our understanding of the relationship between the conception of statistical activity implied in two computer-based interventions and the learning of statistical inference. One computer-based intervention focused on the understanding of sampling by repeated simulation; the other computer-based intervention focused on data analysis by using authentic data sets and statistical packages. Additionally, this design tracked the evolution of data analysis and sampling knowledge from the beginning of instruction to the intervention point in a subgroup of participants and collected protocols of students answering the pre, posttest items. This chapter describes the population, measures, interventions, and data analysis used to explore these issues.

Before doing so, it is necessary to clarify the terms that are going to be used in the rest of the text. Until now, the term *chance* had been used to refer different aspects of probability that are necessary for the learning of inference. In the following sections, *sampling* and *sample size* effects will be mentioned to refer particular applications of probability to the problems of this study. All these terms are operational uses of *chance*. In the case of *data analysis*, the terms *comparison of distributions, center and central tendency measures* will refer operational applications of *data analysis*. *ANOVA*, *ANOVA tables interpretation*, and the *elaboration of conclusions in context* will be equivalent to *inference*.

3.1 GENERAL DESCRIPTION

The study was a randomized design with two conditions. It was conducted in Latin America with 84 bilingual college students taking inferential statistics courses. Students engaged in a pretest task, after which they were assigned randomly to one of two interventions aimed at teaching ANOVA. After they completed the intervention, they were evaluated again with a set of activities equivalent to those used before the intervention. A subset of the students (*n*=12) was asked to participate in the pre and posttest tasks while thinking aloud (see Figure 1). A different subgroup of participants (*n=14*) was studied in depth to capture the evolution of data analysis and sampling knowledge throughout a statistics course that used computer-based activities.

Figure 1. General Structure of the Study



These measures were used to evaluate both the interventions' effects and the constraints of the instructional situation presented in this study. In order to clarify the relationship between these measures and the research questions of this study, a correspondence table was built. This table presents the questions, the data, the analysis and the conclusions provided by this study (Table 2). Question 1 asked whether or not statistical knowledge could be captured in a measurement system, and what students thought while answering the systems' questions. Reliability and validity analysis of the main questionnaire are presented, side by side with a protocol analysis of students' reasoning process while solving the tasks in the measurement system.

Table 2. Relationship between Research Questions, Measures, and Analysis

Question	Data Source	Analysis
Can the ideas of data analysis and sampling be captured in a reliable and valid measurement system? What are students thinking when they respond to the measurement system?	Main Questionnaire Protocols	Reliability and Validity Analysis. Protocol Analysis
Is it possible to design online instruction that reflects the advantages of each perspective in statistical education, and also manages to teach the target ideas of data analysis, and sampling equally effectively?	Main Questionnaire Class' Quizzes	Mixed ANOVA
Can either perspective be used to equal effect to teach either content (data analysis or sampling)?	Main Questionnaire, Interventions' Coding	Mixed ANOVA

Question 2 asked about the actual potential of online instruction to teach data analysis and sampling knowledge. To solve this question, students' gains from pretest to posttest are analyzed and effects in the different spaces of statistical instruction are described. Additionally, the evolution of a group of students is described using the political sciences course's quizzes. Question 3 asked whether or not there was a tradeoff between both perspectives in statistical education. Gains from pretest to posttest in the main questionnaire are examined, and compared in the different spaces of statistical thinking.

3.1.1 ANOVA and Inferential Statistics

Analysis of Variance (ANOVA) was selected for this study for several reasons. First, ANOVA is a statistics topic of central relevance for students in psychology and education at both the graduate and undergraduate levels (Curtis & Harwell, 1998). Indeed, ANOVA is the predominant statistical method used in educational publications (Elmore & Woehlke, 1996) and in statistics education for psychology students, with 88% of the doctoral programs in psychology offering at least one course on this topic (Aiken et al, 1990). Second, ANOVA is a simple linear model that allows students to experience the complexity of statistical inference in a task that resembles basic group comparison activities. In this sense, ANOVA connects with mechanisms of inferential reasoning: the intuitive notions of central tendency and variability, and the representation of data distributions. Additionally, understanding ANOVA requires the comprehension of sampling variability, and in this sense, using an ANOVA task reveals the extent to which students understand probabilistic concepts in inferential statistics. Finally, ANOVA allows a clean mapping of the three types of data outcomes (data representations, inference test results, and conclusions in context). The connection between variability and central tendency in data graphs, in statistical test results, and in conclusions in contexts is transparent in ANOVA. Explained and unexplained variance connects easily with the sums of squares in the ANOVA tables and with the different parts of the distribution graphs.

3.1.2 Pre and Posttest task

In the first section of the pre, posttest, students solved a task that evaluated their ability to coordinate three processes: the comparison of distribution graphs, the interpretation of statistical test results (ANOVA tables), and the generation of conclusions in context (Appendix D). These three processes are basic for the understanding of statistics (Ben-Zvi, 2004; Lehrer & Shauble, 2007; Saldanha & Thompson, 2003; Watson, 2002; Watson & Moritz, 1999). After the pre posttest, students responded to a questionnaire devoted to aspects of group comparison and sampling that affect the significance of mean differences (e.g., variance, sample size), and they engaged in an exploratory data analysis task. Finally, students were asked to solve a selection of items from the AP exam and the CAOS test in order to evaluate the use of these ideas in a more traditional statistical setting. The total time of the pre, posttest task was about 2 hours; 1 hour at the beginning of the study and 1 hour at the end.

3.1.3 Interventions: Data Analysis and Sampling Simulations

Students were exposed to one of two interventions, either data analysis or sampling. In the *data analysis* condition students were asked to go through the Sampling Distribution and the ANOVA sections of the statistics course of the Open Learning Initiative (OLI). The OLI course was developed by Carnegie Mellon University (CMU) and funded by the Hewlett Foundation. This course explains the concept of ANOVA by placing it within the broader range of activities that test hypotheses for relationships. In particular the OLI course explains that the mechanism by which ANOVA compares differences is the contrast between explained and unexplained variances. The course provides several occasions for Exploratory Data Analysis (EDA) in which students are given data sets and asked to conduct statistical analysis using either Minitab or Excel (Appendix E). The exploratory data analysis situations are modeled throughout the course using authentic examples.

In the *sampling* condition, students were asked to go through a study guide that follows similar steps to those in the OLI course, but instead of using exploratory data analysis, students were asked to pull random samples using several simulation Applets (Appendix F). To illustrate why it is necessary to test mean differences, students were asked to extract several samples from a population with a fixed mean (e.g., extract a sample of 20 cases from this distribution). The sample means varied due to the sampling process. The *sampling* intervention asked students to establish whether or not there was any observable difference between samples that belonged to the sample population and samples that came from different populations. Students were encouraged to test whether or not samples from a single population varied within certain limits and whether or not they fell according to the sampling distribution of the mean. After that, students were asked to play with the sample size and the standard deviation of the population and see how that affected the distribution of the sample means. In a final section, students explored what factors affected the confidence of the observed difference (e.g. the smaller the sample size, then, the larger the likelihood of finding a difference by chance). After that students were asked to use several Applets to visualize the relationship between explained and unexplained variance in ANOVA.

3.2 PARTICIPANTS

This study was conducted with 84 students from an upper-middle class university in Latin America. All participants were native Spanish speakers with high English and computer proficiencies. Students were part of three middle-level statistics courses for social sciences majors. Specifically, participants came from three groups: one group was participating in a middle-level statistics course for psychology majors during the spring term of 2007; the second group was enrolled in an equivalent course during the fall of 2007; the third group was participating in a middle-level statistics course for political science majors taught during the fall of 2007. The first psychology course had about 35 students; the second psychology course had about 60 students, and the political science course had about 15 students. About 60 percent of students were women. All courses covered topics from basic descriptive statistics to basic inferential methods including ANOVA and Regression. All three courses had an EDA-based approach and students were accustomed to work with computer packages for data analysis. The activities necessary for this study were presented during the ANOVA sessions of the courses between the 12th and the 16th week of instruction. Protocols of students solving the pre, posttest tasks were recorded for a sub-sample of 12 students (6 in each condition). Additionally, the evolution of 14 students participating in the course for political science majors was registered from the beginning of instruction to the intervention point.

3.3 INSTRUMENTS

3.3.1 Pre- and Posttest Questionnaire

In the pre and posttest, participants were asked to respond a questionnaire that had four sections (see Appendix D). The first section asked students to connect three levels of data use -- distribution graphs, inferential test results, and conclusions. The second section presented open-ended questions on sampling and data analysis in which students explained several statistical concepts. The third section was an exploratory data analysis activity in which students were asked to interpret a full set of ANOVA results that included graphs, descriptive statistics and ANOVA tables. The fourth section was a multiple-choice questionnaire that combined items from the AP exams and from the CAOS test on sampling, data analysis and inference. The pre- and posttest questionnaires had the same structure and they varied only in the cover stories and in the absolute numerical values of the parameters but the relationship among variables and the presence of significant differences was the same for both versions of the test. Students solved the pretest all at once in the first hour of the intervention, and the posttest in the last hour of the intervention after they had finished the computer-based activities.

3.3.1.1 Section 1: Coordinating Distributions, ANOVA Results, and Conclusions

The first section of the questionnaire evaluated students' skill in coordinating distribution graph pairs, test results, and conclusions in context. Specifically, this section assessed students' skill both in identifying distribution graph pairs that had significant differences, and in connecting these distribution pairs with ANOVA tables that displayed significant results and with statements that presented valid conclusions in context. When combined with the students' protocols, these measures provided information on the students' strategic use of central tendency and variability in the process of coordinating different data outcomes.

Specifically, this section was divided into two tasks. The first task required participants to compare six pairs of distribution graphs. These distributions were produced using a random number generator and they varied in their size (*n*), mean, and standard deviation. The distribution pairs were presented graphically, and the *n*, the mean, and the standard deviation values were displayed at the right of each distribution graph. In the first item of the first task, students compared two distribution pairs that varied in the mean difference but had the same standard deviation. For example, in one pair both distributions had a standard deviation of 10 and a mean difference of 15; in the other pair both distributions had a standard deviation of 10 and a mean difference of 30. In the second item of the first task of this section, students compared distribution pairs that had the same mean differences but different standard deviations. In the third item of this task, they compared two distribution pairs with equal mean differences and standard deviations, but with different sample sizes. This task specifically asked participants to identify in each item the distribution pair that presented the most significant difference, and to explain the reasons for selecting some pairs over others.

The second task of this section required students to interpret conclusions in context. In this task, participants read a research case that presented two findings. One of them was a significant mean difference; the other one was a non-significant mean difference. For both findings, the mean difference was the same. After reading the case, participants were required to solve three items. The first item required students to pair the findings in the case with two ANOVA tables. The second item required students to pair the findings with two graphical representations similar to those in the first task of this section; and the third item required students to explain the difference between the two findings by picking a statement about the possible sample size of the results reported in the case.

3.3.1.2 Section 2: Open-ended Questionnaire

Section 1's tasks provided important information on learners' statistical reasoning. However, an open-ended questionnaire on sampling and data analysis was included to have a more direct measure of statistical knowledge. The first part of the open-ended questionnaire was devoted to sampling variability and data distributions. The open-ended questionnaire explored initially how participants understood the concepts of standard deviation, explained and unexplained variance, within-groups and between groups variability, and sample variability. The second part of the open-ended questionnaire was a task that required students to interpret a pseudo-authentic case using two distribution graphs, both of which had a mean of 50. The distribution graphs differed in that one distribution had a normal distribution (mean = 50), and the other distribution combined data from two smaller distributions (means = 80 and 20) and had two peaks. This part of the questionnaire required participants to explain why an ANOVA did not identify significant differences between those two distributions. To do so, participants read a paragraph that presented a research case that compared a co-ed school and a single-sex school in mathematical learning. The two-peaked distribution corresponded to results from the co-ed school and the single-peaked distribution corresponded to results from the single-sex school. The paragraph explained that researchers sampled only a part of the students in each school (e.g., 100) and did not find significant differences in mathematical learning between the two schools. Participants were asked to assess the results of the study, and to propose options for redesigning it in order to improve the quality of the conclusions. Possible answers to this question were conducting a study with a larger sample or dividing the co-ed school data into two subsets – one for men and one for women. Finally, students were asked whether or not the same results would be found if the study were conducted again.

3.3.1.3 Section 3: Data Analysis Task

A data analysis task was included to evaluate students' actual ability to deal with data analysis in an ill-defined task. Including this measure was consistent with the idea that contact with data analysis activities during learning develops data handling skills and knowledge that would not be developed otherwise. In this section, students had to analyze and produce conclusions from a set of data analysis results. The data set had been produced by random simulation but the results were presented as part of a real research case. The students were asked to explore the data analysis results, draw conclusions about the case, and explain where the conclusions came from.

3.3.1.4 Section 4: Multiple-choice Test

In order to evaluate students' performance in a more traditional environment of statistics learning, the last section of the pre- and posttest tasks required students to answer a multiple choice test. This test was constructed with items from the CAOS and AP tests. These items were selected because they assessed distribution description and comparison, the effects of random sampling, the relationship between *p*-values and hypothesis testing, or the use and interpretation of ANOVA results. The CAOS test is a new tool for the assessment of statistical thinking, reasoning, and literacy that has been evaluated both for validity and reliability (DelMas et al, 2006). The AP exam items are part of a large item database available at http://apcentral.collegeboard.com/apc/public /courses /teachers_ corner /2151.html. The AP statistics exam evaluates content whose level is at that of a non-calculus-based statistics college course. Both tests are well-known assessments of statistics learning and are used in several contexts.

3.3.2 Rationale for the Measures in Sections 1 through 4.

The two first sections of the pre- and posttest questionnaires required students to describe and compare distributions. Comparing distributions presents several advantages for the study of statistical learning and reasoning. First, there is a growing body of research in statistics education that uses this task because it has the fundamental elements of statistical inference (e.g., contrasting typical indicators and variability) but it requires little specialized knowledge (e.g., Ben-Zvi, 2004; Lehrer & Shauble, in press; Watson, 2002; Watson & Moritz, 1999). Second, comparing distributions provides a bridge from basic exploratory tasks, as for example group comparisons, to advanced uses of statistical procedures in scientific contexts. Third, this type of task requires students to build global views of data, to understand diverse equivalent representations, and to produce conclusions about differences in distributions.

In both tasks, four elements were manipulated: mean difference, standard deviation, sample size, and cover story. These elements participate in the assessment of *mean differences'* significance. In the process of assessing the significance of a difference, researchers challenge the existence of a true difference against other possible sources of variation. Variation can be attributed to sampling, to the natural variation of the elements involved in the experiment, or to measurement process. Ideally, students should do the same.

Natural variation was introduced in this design by having different *standard deviations*. The higher the standard deviation, the lower the confidence in the difference. The variation produced by the sampling process was introduced by varying the *sample size*; the lower the n, the lower the confidence in the results. In both cases, high standard deviation and low *n*, the variation that can be attributed to factors different from those included in the model is high, thus the students' confidence in the results should be low. However, evidence shows that this is not the case. Understanding of variation as a factor affecting the size of the mean difference appears only through instructional situations such as, for example, guided discovery (Lehrer & Schauble, in press) or cognitive conflict (Watson, 2002). In the same way, understanding that the sampling process generates variation, and that the lower the *n,* the larger the confidence interval, emerges only through precisely designed instructional situations (Saldanha & Thompson, 2003).

Inferential tests are simply a more sophisticated version of this kind of reasoning. In different ways, they test mean differences while accounting for the variability that can be attributed to within-group variance (that is, the natural variation of the objects in a category and the measurement error) and to the sampling process. They do so by adding variance in the case of confidence intervals, or by contrasting the statistic value observed in the sample against the sampling distribution of that statistic in the case of *p*-values. Cover story was introduced for two reasons: one to account for the fact that the interpretation of a statistical result can be affected by the context in which data is obtained (Nisbett et al, 1993). Second, to see how providing an interpretative context affects the understanding of statistical test results.

The last three measures in the pre- and posttest questionnaires were selected for different reasons. The open-ended questions were included because it provided direct information on learners' knowledge of probability and data analysis. The data analysis task was included because it reviewed the effects of the intervention in open statistics tasks that required organizing information and producing conclusions in research cases. According to the literature, active exploration and open representation of data in statistics education provide students with opportunities for authentic use of data that formal training does not. The data analysis task should make these effects visible. The multiple-choice tests items evaluated the effects of the interventions in an assessment situation that was not clearly related to one of the two types of instructional procedures in this study (e.g., sampling simulation and data analysis), and in this way, these items evaluated reasoning out of the instructional setting.

3.3.3 Main Questionnaire

From all these measures, a main questionnaire was elaborated. This main questionnaire is the source of all the quantitative findings reported in this study. From now on, the items in this questionnaire will be referred as item, and they will numbered between 1 and 16. Questions not belonging to this questionnaire will be reported as open-ended questions. Items 1 to 3 are part of the first task of the first section of the measures (3.3.1.1.); Items 3 to 6 are part of the second task of the first section of the measures (3.3.1.2). Item 7 is the answer to the data analysis task (3.3.1.3). Items 8 to 16 are multiple choice items included in the measures (3.3.1.4).

In terms of origin, items 1, 2, and 3 were designed by the researcher, and required students to identify the more significant of two distributions given the means' difference, the group variance and the sample size (Table 3). Items 4, 5 and 6 were designed by the researcher and required students to connect a case with different data outcomes. In particular, these items required students identifying the data distribution graphs, ANOVA test results and sample sizes that connected with significant results in a hypothetical case. Item 7 was written by the researcher and required students to interpret the difference between two data distributions in two hypothetical situations. Items 8 to 16 were taken from the CAOS test. In content levels, the items 1, 7, 8, 9, 10 required students to compare several data distributions presented in a graphic form.

Item 2 and 4 required students to compare data distributions accounting for group variance. Item 3, 5, 13, 15, required students to understand the role of sample size in different statistical situations. To solve items 6, 11, 12, it was necessary to coordinate significance *p*-values either with graphical representation of data or with conclusion in cases. Item 14 asked students to figure out the relationship between sample and population characteristics. Item 16 involved understanding the difference between sampling and data distributions. Items 10, 11, 12, 13, and 14 included sample size in at least one element of the problem solution.

Table 3. Item Characteristics

Name	Item	Goal	Accounting for	Space
Data1	1	Identify the more significant difference	Different central values	D
Data2	2	Identify the more significant difference	different spreads	D
Sam1	3	Identify the more significant difference	different sample size	S
Data3	4	Produce a conclusion	different spreads	D
Sam2	5	Produce a conclusion	Different sample size	S
Inf1	6	Produce a conclusion	Different p-values	I
Data4	7	Produce a conclusion in Context	Different central values	D
Data5	8	Evaluate a conclusion	Different central values	D
Data6	9	Evaluate a conclusion	Different central values	D
Data7	10	Evaluate a conclusion	Different central values	D
Inf2	11	Interpret a significant result	Sample size	I
Inf3	12	Interpret a significant result	Different central values	I
Sam3	13	Connect population and sample characteristics.	Sample size	S
Sam4	14	Connect population and sample characteristics	Different spreads	S
Sam5	15	Connect population and sample characteristics	Sample size	S
Sam6	16	Different sampling distribution and data distribution	Different spreads	S

3.4 INTERVENTIONS

For the intervention phase of the study, students were divided in two groups. In one group, participants were asked to go through a section of a statistics course that provides opportunities for data analysis. In the other intervention, students learned ANOVA with a study guide that had the same course structure but that supported the learning of ANOVA with ready-to-use simulations instead of data analysis. Both interventions explained that ANOVA compared within and between group variance to determine when there were systematic mean differences, that were not due to random sampling. The literature suggests that these two types of intervention have different effects on the learning of statistical inference. Ready-to-use simulations allow students to interact with probabilistic concepts at a low procedural cost; they allow for the drawing of multiple samples and for visualizing how samples distribute according to the population parameters. However, ready-to-use simulations do not allow students to build the representational frameworks in which to interpret the data, and the lack of authenticity in these simulations can lead students to feel they are not dealing with real data. On the other hand, activities based on data exploration, as in the data analysis condition, require students to handle raw data and to control the representational frameworks in which data is organized. However, this type of activity (data exploration) makes connecting statistical concepts and representations difficult. Several steps separate data from representation, and representation from concept. Data exploration activity also constrains students experience to just one set of data, neglecting the experience of multiple sampling that is necessary to understand probabilistic concepts.

3.4.1 Elaboration of the Interventions

Both interventions followed the structure of the ANOVA section of the Open Learning Initiative (OLI) statistics course. The instructional time for both interventions was about two hours including the reading of the text and the completion of the exercises. The *data analysis* intervention was a simplified version of the ANOVA section of the OLI course and had all the same features except the feedback system. To produce this intervention, the text and exercises of the OLI course were exported from the OLI website to a ".html" file that students could access from a CD. To produce the *Sampling* intervention, the main ideas of the ANOVA section of the OLI course were isolated and the possible locations for simulations in this instructional sequence were identified (see Table 4). Then, several Applets that could serve as simulations for this intervention were found on the Internet, and three of them were selected. With this information in place, a study guide was prepared. Finally, the study guide and the links to the simulation Applets were exported in the same .html format used for the *data analysis* condition. Additionally, a fragment of the sampling distribution section of the OLI course was added to the *data analysis* intervention, in order to provide students in this condition with information on random sampling that students in the other condition would obtain from using the simulations.

3.4.2 General Structure of the Interventions

Both interventions were elaborated around a group of eight ideas obtained from the ANOVA and Sampling Distribution's sections of the OLI course (Table 4). Although both interventions presented the same ideas, the specific presentation of the content varied from one intervention to the other. The data analysis condition tended to use more worked-out examples, while the sampling condition tended to use more simulation exercises.

Table 4. Basic Ideas of the Interventions from the OLI Course

Idea
1. ANOVA evaluates the relationship between a categorical and a continuous variable.
2. ANOVA is necessary to evaluate the relationship between categorical and continuous variables.
3. ANOVA compares within and between group variances to elaborate conclusions on the sample means
4. The variation among group means is considered negligible when within and between variances are similar.
5. The degrees of freedom affect ANOVA's interpretation.
6. The elements of ANOVA tables
7. Interpretation of p-values
8. Interpretation of the results of ANOVA

Although this situation compromises the comparability of both interventions, it mirrors the differences between the perspectives in statistical education compared in this study. A probability-based perspective in statistical education privileges the teaching of probability, and naturally it concentrates activity around this topic; a data analysis perspective focuses on the teaching of data analysis, and for this reason it gives more room for data analysis activities than for the use of probability simulations. To make explicit these differences, this text will provide in the next pages a detailed account of the similarities and differences between both interventions attending particularly to the amount and type of text and activity devoted to each idea.

3.4.2.1 Interventions' Characteristics

Interventions were coded in terms of the type of pedagogical resource they contained (e.g, texts, exercises, examples), and the statistical space they aimed at (e.g., data analysis, sampling, inference) (Table 5). This information makes it possible to compare the interventions according to the amount and type of activity devoted to each space of statistical content, and more generally it permits a more complete understanding of the intervention effects on participants. In the next section, the comparative results of both interventions are presented.

Table 5. Idea units and Questions in each Intervention

	Type of Activity					
	General Text		Examples		Exercises	
Statistical Space	Data Intervention	Sampling Intervention	Data Intervention	Sampling Intervention	Data Intervention	Sampling Intervention
Data Analysis	21	10	43	0	7	9
Sampling	34	19	42	5	14	35
Inference	38	16	43	1	21	25
Total	93	45	128	6	42	69

The results for general text show that the data analysis intervention presented more text than the sampling intervention. The number of idea units of general text in the data analysis intervention was 93, compared to the 45 idea units presented in the sampling intervention. This is true for all the statistical spaces considered in this study; Figure 2 shows that while the data analysis intervention presented 21 idea units of general text in the data analysis space, the sampling interventions had 10 idea units in the same dimension. The same configuration appeared for the inference space where the data analysis intervention had 38 idea units, and the sampling intervention had 16 idea units; and for the sampling dimension where the data analysis intervention had 34 idea units and the sampling intervention had 19 idea units.

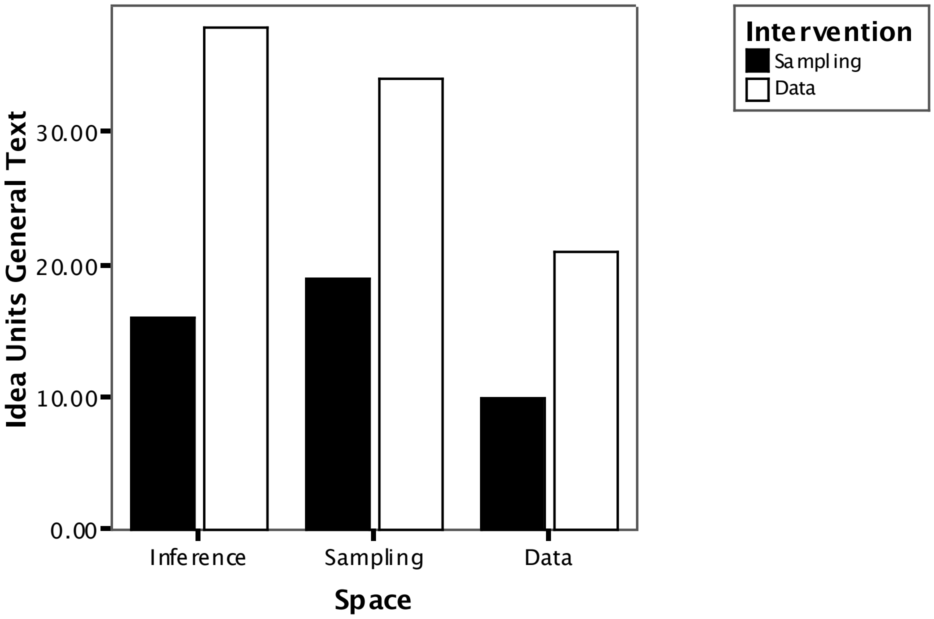


Figure 2. Idea Units in General Text per Statistical Space

A similar pattern appeared for the *example* dimension: the data analysis intervention had 128 idea units in total and the sampling intervention had 6 idea units in the same type of pedagogical resource. This difference was transversal to all dimensions. 43 to 0 in the data analysis space, 43 to 1 in the inference space and, 42 to 5 in the sampling space (Figure 3).

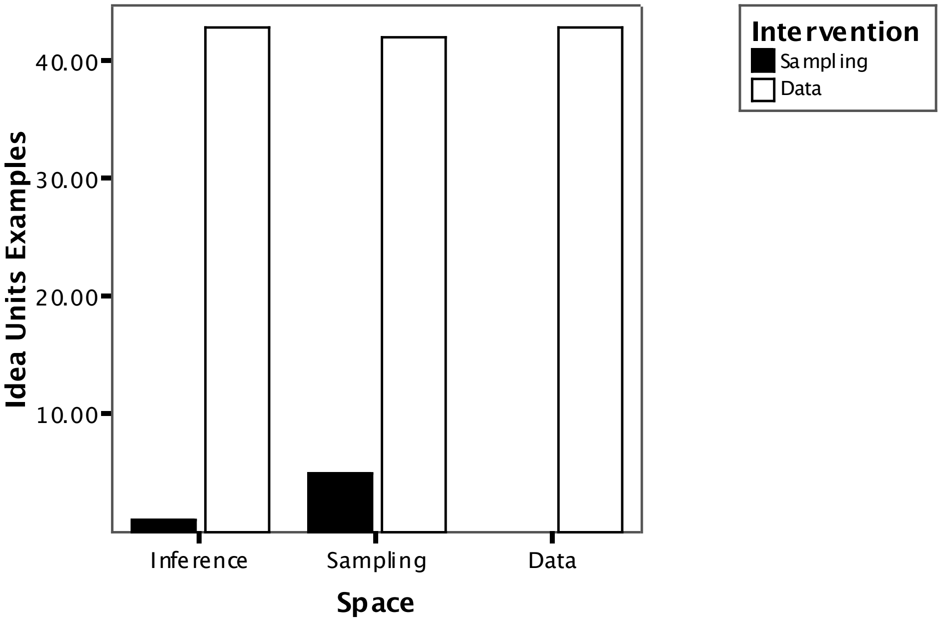


Figure 3. Idea Units in Examples per Statistical Space

The opposite situation happened with the exercises. The sampling intervention provided more exercise's questions (69) than the data analysis intervention (45). This was true for the three statistical spaces (figure 4): For data analysis, the sampling intervention provided 9 questions and the data analysis provided 7. For the inference category, the sampling intervention provided 25 and the data analysis provided 21 questions. Finally, for the sampling space, the sampling intervention provided 35 questions in the exercises and the data analysis intervention provided just 21 questions.

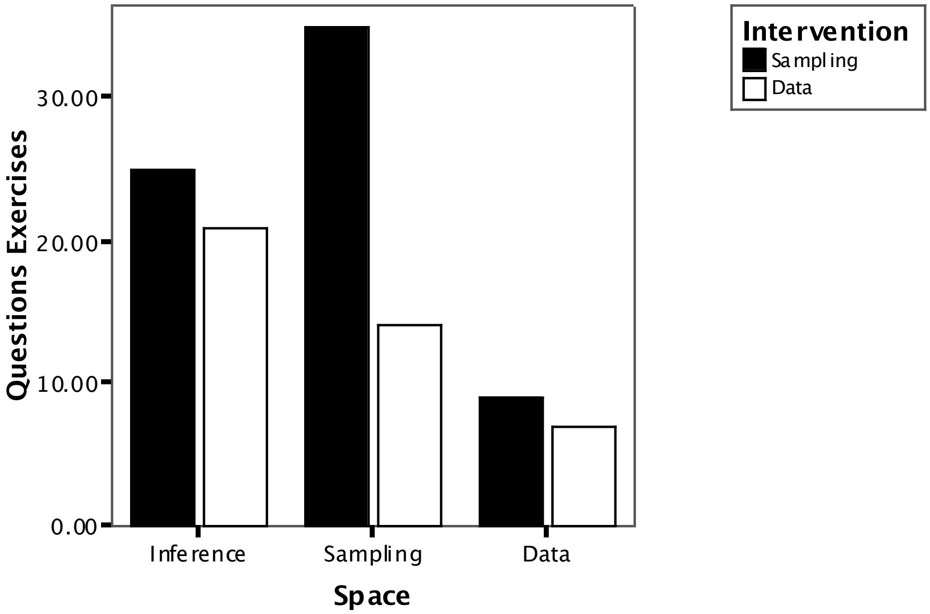


Figure 4. Questions in Exercises per Statistical Space

Differences in the amount of work per space and type of resource between the two conditions were expressed as ratios between the number of idea units and questions in the sampling condition and in the data analysis condition. To test the significance of these differences, a Chi square was calculated for the ratio between the number of idea units or questions in the sampling intervention and the total number of idea units or questions in both interventions. It was assumed that equal work levels in both interventions will produce a .50 ratio, that higher work levels in the sampling condition will produce ratios over .50, and that lower work levels in the sampling condition will produce ratios under .50.

These results show that all the differences were significant except the differences between the number of exercises' questions devoted to data analysis and inference (Table 6). In other words, the data analysis intervention had significantly more text and examples in any space than the sampling intervention, and the sampling intervention had significantly more exercises devoted to sampling than the data analysis intervention. A detailed account of the coding process and a description of the interventions arguments are presented in Appendix G.

Table 6. Ratios for the Number of Idea Units and Questions in the Sampling and the Data Analysis Intervention.

	General Text		**Examples**		**Exercises**	
	Sam/Dat	Sam/Tot	Sam/Dat	Sam/Tot	Sam/Dat	Sam/tot
Data Analysis	10/21=.47	.32*	0/43=0	.0**	9/7=1.28	.56
Inference	16/38=.42	.29**	1/43=02	.02**	25/21=1.19	.54
Sampling	19/34=.55	.35*	5/42=.11	.10**	35/14=2.5	.71**
Total	45/93=.48	.32**	6/128=.04	.04**	69/42=1.64	.62*

3.5 COURSE DESCRIPTIONS

The data for this study were collected in three different courses. This section describes both the general and specific characteristics of these courses. The three courses were EDA based courses that focused on basic and intermediate level statistics contents. The first two courses were intermediate level statistics courses for psychology majors taught in the first and second semester of 2007 by different instructors. The third course was an intermediate level statistics course for political science majors taught by a third instructor. The interventions were conducted in all courses between the 12th and the 16th week of classes. Prior to the intervention, students underwent training on data representation, variable types and relationships, center and spread measures, and correlation. The complete list of the courses topics and the order in which they were presented can be seen in table 7.

Interviews with the courses' instructors indicated that students had extensive practice with data representation and interpretation, including the construction of histograms, and the comparison of different data distributions. During the courses, students were often required to characterize distributions in terms of central tendency and spread measures, as well as, in terms of other distribution characteristics (e.g., skewness).

Table 7. Comparison of Statistical Courses

Topic	Psychology I	Psychology II	Political Science
Item Construction	2	2	
Sampling Techniques	3	3	
Data Bases	4	4	2
Histograms	5	5	7
Descriptive Center Measures	6	6	4
Measures of Spread	7	7	5
Box Plot			6
Variables Continuous Categorical Independent etc	9	9	3
Contingency Tables	10		9
Normal Distribution	11	10	8
Design Types	12	11	2
t-test	13	12	11
Correlation	14	13	10
Intervention moment	16	14	12

The interviews indicated also that formal training in mathematical proof and mechanical calculation was avoided in all three courses; SPSS competency was an important part of all courses. The courses privileged conceptual explanation of statistical techniques and measures (e.g., spread measures) over the mathematical and procedural aspects of them. Typical activities in the classes were considering a problem and a solution in the light of a data set and its representations. Questions like identifying the distribution with the largest spread among several data representations, or describing the elements of a graph were recurrent. Class activities that required students to use SPSS were typical also. No probability training was given prior to the intervention. The sampling topic was restricted to sampling techniques (e.g. stratified, cluster) in each three courses.

3.6 PROCEDURE

Students in the think aloud sub-sample for the pre and posttests were interviewed individually. Participants in this group met with the researcher in a private office; the consent forms were given to participants, and, if they consented to participate in the study, the general instructions for the study were given. These instructions basically asked participants to conduct the tasks in the study while talking aloud. An example and a pilot task for the talking aloud was provided and participants were asked to engage in a short warm-up task. The warm-up task consists of three two-digit multiplication problems that participants had to solve while talking aloud. After this, participants in this sample solved six items of the pretest while being audio recorded, and all the other items in written form. No identification information was associated either with the audiotapes, or with the written documents. Following completion of the pretest, participants in the think aloud group were randomly assigned to one of the two conditions: a *data analysis* condition, where they analyzed authentic data in the context of an statistics course (Appendix E); or a *Sampling* condition that required the students to go through a study guide and to use several simulation Applets (Appendix F). After the interventions, participants solved six items of the posttest while thinking aloud, and all the other items in written form. Students in the general sample followed the same procedure except that they were not interviewed individually and they were not required to think aloud. Only written answers to the pre- and posttest were collected for this larger sample.

In addition to the large study focused on the intervention effects, a sub study aimed at exploring in depth the change in students knowledge prior to the intervention point was conducted with the political science course. The decision to conduct this sub-study was taken because during the initial data collection in the spring of 2007, the intervention effects were restricted to sampling knowledge in the sampling condition, and the data analysis intervention did not produce change on data analysis knowledge. A plausible explanation for this result was that gains in data analysis skills and knowledge required a more intense amount of practice than the amount of practice provided by the data analysis intervention. Moreover, it was hypothesized the practice necessary for developing data analysis skill was provided prior to the intervention point through the courses' representational and interpretative exercises. In fact, the initial levels of data analysis knowledge were higher than the initial levels of sampling knowledge. To track the effects of sustained activity on students knowledge, the following procedure was designed. The political science class evolution was registered during a period of 11 weeks from the beginning of the class to the intervention point. Instructor actions and students' gains were documented. This section describes in detail the instructor actions and the evaluation instruments used during the initial 10 weeks; the results and the changes on students' knowledge are described in the results section.

Class actions are divided in three types: explanation, SPSS use, and interpretation. Explanation was the introduction to the class topics and the presentation of the main elements of the concept by the instructor. SPSS use was the moment in which students practiced by analyzing data bases with SPSS routines. Interpretation was the class moment where students put concepts in context, and read the SPSS results in order to solve questions. Each moment had a different emphasis: explanation focused on the basic ideas of each statistical concept (e.g. SD is related to the distribution spread); SPSS use aimed at developing competency in statistical package use and output reading; and interpretation worked on students skill to solve theoretical questions using statistical ideas. Class actions were registered by collecting the instructor SPSS Power Point presentations.

Table 8. Studying the Evolution of a Class

Topic	Week	Class Action	Weekly Quiz
Design Types	2		
Data Bases	2		
Types of Variables	3		
Descriptive Measures of Center	4	Explanation (Measures of Center) SPSS (Measures of Center) Interpretation (Measures of Center)	
Measures of Spread	5	Explanation (Measures of Spread) SPSS (Measures of Spread) Interpretation (Measures of Spread)	Measures Spread and Center
Box Plots	6	Explanation (Box Plots) SPSS (Box Plots) Interpretation (Box Plots)	Boxplots with context
Histograms	7	Interpretation (Box Plots) Explanation (Histograms) SPSS (Histograms)	Boxplots and Histograms with context
Normal Distribution	8	Interpretation (Histograms) Explanation (Normal Distribution)	Boxplots and Histograms with context
		EXAM	
Contingency Tables	9		
Correlation	10		
T test	11		Boxplots and Histograms

The results of this sub-study come from 5 quizzes taken between week 5 and week 8 and from a follow-up quiz taken during week 11 (Table 8). Quizzes had one or two questions and no reliability measures were obtained for them. The first quiz evaluated the understanding of center and spread measures without attending to any graphical representation. The second quiz evaluated the interpretation of boxplots, particularly, the ability to compare two distribution represented as boxplots. The third quiz was similar to the second one but it included also histograms; students were asked to evaluate the similarities and differences between two variables; one represented as a histogram, and the other represented as a Box plot. The forth quiz tested both students skill to interpret histograms and boxplots in context (comparing several distributions and concluding about them). The follow up quiz was presented to students in week

11 right before the intervention and it was very similar to quiz 4. Additionally, students performance was registered by asking students to email the results of class exercises; class exercises in most cases required students conducting transformations on data using SPSS, and to produce short results interpretation.

3.7 PROTOCOL CODING

3.7.1 Coding of Protocols

3.7.1.1 Coding of Data Analysis Items

Students' verbal protocols on data analysis items were parsed into idea units and then coded according to cognitive actions. An idea unit was defined as a non-redundant proposition with complete meaning. The cognitive actions considered here are *describing, comparing, explaining, inferring, answering* and *metacognition*.

Describing is defined as an idea unit where the participants produced an statement about some directly observable feature of the distribution. Operationally, the describing code required a sentence whose subject (direct or implied* Spanish allows sentence without subject that can be implied through the conjugation of the verb) was an element of the distribution, or the distribution itself, and that contained the verbs "to be" or "to had" (it is, it seems to be, it appears to be, it appears to have).

Comparing is defined as an idea unit where features from different distributions are put side by side and contrasted along some given dimension. Operationally, this code required an idea unit containing references to two or more distributions and the presence of a contrasting verb (“compared to” “different than”) or a comparative adverb (“more than” “less than” “equal”). For the *describing* and *comparing* codes, a distinction was made between participants producing right answers and participants producing wrong answers. Right answers were marked with an asterisk. Additionally, a sub code was created to identify the element being described or compared. For the *describin*g code the values of this argument could be center, spread, distribution or problem feature. For the *comparing* code the values could be center, spread or distribution.

Explaining was a code created to account for cognitive actions in which people presented further justification to some statement. *Explaining* was coded for idea units starting with “because” or an equivalent word or phrase (e.g., this is due to). *Inferring* was the cognitive action of concluding from prior information. This action was coded when a statement was placed after a connector like “thus” “therefore” or an equivalent word.

Answering was used for idea units in which the participant presented direct answer to one of the items used in this part of the study. When an answering code was preceded by a connector like thus, therefore or equivalent word, or followed by a connector like “because” or equivalent word, both the answer and the accompanying statement were considered as part of a single answering code. This decision was taken to avoid confusion between inferring and explaining codes, and the answering code. This system permitted also to treat answers and explanations as single units in order to identify the decision rules used by participants.

Three additional codes were created to facilitate the coding system. A *metacognition* code was created for idea units that showed knowledge self-assessments. Operationally, this code required that the object or subject of the idea unit were either an statement or the participant’s knowledge. When the participant was assessing a final answer or description with only a verb (e.g., I believe that), the idea unit was coded as answering, not as meta cognition.

A *positioning* code was created for idea units that inform about the part of the problem the student was solving. For coding purposes, this code required the idea unit to include phrases like “solving this”, “solving that”, or textual paraphrasing from items. A last code was created for affirmations not included in this system; such idea units were coded as “*other*”.

Additionally, protocols of students were classified into two types depending on the length of the protocol. It was assumed that a short protocol implied a superficial examination of the item (strategy 1), and that a long protocol entailed a deeper exploration of the problem elements (strategy 2). Strategy 1 was defined as an answer that was given before 20 idea units; strategy 2 was defined as an answer that was given after 20 idea units. These definitions provided a simple way to discern between both strategies on the assumption that the difference between students using both strategies lied in elements difficult to identify, such as the presence or absence of certain decision rules, the level of trust on those rules, or the persistence during the problem solving process. The decision to take 20 idea units as the cut point was made based on preliminary analysis of the protocols that indicated that protocols under 20 idea units tended to have a simple structure in which no development or discovery of new decision rules was possible.

3.7.1.2 Coding of Sampling Items

Sampling items were codified with same categories that were used for data analysis items. Some changes were made to adapt the system to the type of answers produced by participants. In the data analysis items, describing and comparing codes had arguments that pointed out what aspect of the distribution was being described; for the sampling items, “sample size" was included among the possible arguments for describing and comparing codes. For the coding system, the propositions coming directly from the items text were annotated and used in the analysis of the decision rules. The assumption here was that the information coming from the text was still active in the working memory and it was being used to response the items.

3.7.1.3 Coding Reliability

In data analysis items, inter-rater reliability was obtained on the coding of 50.3% of student's cognitive actions. The percentage of agreement was 82.0% (Kappa 0.76) for the coding of cognitive actions. The reliability in the identification of the object of the describing codes (e.g., center, spread) was calculated on a sub sample of 52.6% of the describing codes, and the agreement rate was 88.7% (Kappa=.79). The coding reliability in the classification of answering codes into decision rules was calculated on a sub sample of 66.2% of the answering codes, and the agreement rate was 86.7% (Kappa=0.82). For sampling items, the reliability of the classification of answering codes into misconceptions was 88.9% (Kappa=0.85) based on the total sample of answering codes.

4.0 RESULTS

This section presents the main findings of this study. After reviewing the characteristics of the pre, posttest items, this text explores the effects of computer-based interventions on students knowledge. In the second part, the text describes the evolution of students' knowledge prior to the intervention point, as well as, some hypotheses about the factors that determine students performance on data analysis tasks. These hypotheses were based on information about class activities and students' responses registered in one of the courses observed in this study. The third part of this section presents the conclusions of several protocol analysis of students solving the assessment tasks. From this protocol analysis, a plausible explanation of the intervention effects on students knowledge is proposed. In the final part of this section, the answers' explanations provided by the students in the pre, posttest are analyzed and connected with the protocol analysis results.

4.1 A MEASUREMENT SYSTEM FOR STATISTICAL KNOWLEDGE

Students answers were analyzed in order to evaluate the quality of the measures used in this study. Classic measures of reliability were obtained, and, to test validity, a factor analysis on the pre, posttest measures was conducted to test whether or not the items behaved according to the underlying constructs they were supposed to measure; a complete description of the process of item classification is provided in the methods section. To obtain a measure of external validity, the correlation between measures designed by the researcher and items obtained from standard statistics test (e.g., CAOS test) was calculated.

4.1.1 General Descriptive Measures

Table 9 presents item means and standard deviations. In the pretest, no item was answered correctly by more than 77% of participants. Some items, however, were answered correctly by a small percentage of participants. Items Inf1, Inf2, Sam3, Sam4, Sam5 and Sam6 were answered correctly by less than 25% of participants. All these items focused on inference and sampling that were statistical areas in which students had little prior training. A more careful look at table 9 shows that 3 items were answered correctly by less than 20% of participants; 4 items were answered correctly by between 20% and 40% of participants; 5 items were answered correctly by between 40% and 60% of participants; and 4 items were answered correctly by more than 60% of participants.

In the posttest, no item was answered correctly by more than 72% and by less than 36% percent of students. The percentage of right answers for inference and sampling items increased, and the floor effects observed in the pretest disappeared. Items that had had very low response rates in the pretest obtained higher scores in the posttest. Only 2 items were answered correctly by between 20% and 40% of students; 7 items were answered between 40% and 60 % of participants, and 7 items were answered by more than 60% of participants.

Table 9. Items' Descriptive Measures

Name	Item	Space	N	Pre		Post	
				Mean	S. D.	Mean	S. D.
Data1	1	Data	84	.64	.48	.67	.46
Data2	2	Data	84	.40	.49	.46	.50
Data3	4	Data	84	.46	.50	.54	.50
Data4	7	Data	84	.77	.42	.66	.47
Data5	8	Data	84	.55	.49	.60	.49
Data6	9	Data	84	.61	.48	.72	.44
Data7	10	Data	84	.42	.49	.65	.47
Inf1	6	Inference	84	.65	.50	.63	.48
Inf2	11	Inference	84	.23	.42	.57	.49
Inf3	12	Inference	84	.16	.37	.36	.48
Sam1	3	Sampling	84	.33	.47	.54	.50
Sam2	5	Sampling	84	.45	.50	.55	.49
Sam3	13	Sampling	84	.15	.36	.63	.48
Sam4	14	Sampling	84	.22	.42	.48	.50
Sam5	15	Sampling	84	.08	.27	.38	.48
Sam6	16	Sampling	84	.27	.44	.46	.50

4.1.2 Reliability and Validity

To check reliability, a Cronbach's alpha was calculated for the test and subtests involved in this study. For the posttest, the Cronbach's alpha was .76. This value is acceptable given that the Cronbach's alpha reported for the whole CAOS test is .77 (DelMas et al, 2006). Reliability was calculated also for the parts of the test devoted to specific areas of content. For the posttest, the Cronbach's alpha was .74 for items devoted to data analysis, .60 for items devoted to sampling, and .44 for items devoted to inference. To test the effect of the number of items on the Cronbach's alpha and to show that the values found for the sampling and the inference part of posttest were low due to the small number of items included in those subscales, a Cronbrach's alpha was calculated for the sampling and inference items altogether, and the result was .66.

4.1.3 External Validity

To evaluate validity, the correlation between items from standardized tests and items developed specifically for this study was calculated for each sub area of content. Correlations between the CAOS and non-standardized items were significant for the whole test and for all areas of contents (Table 10). For the whole test, this correlation was .59**; for the data analysis part, it was .52**; for the sampling part, it was .25*; and for the inference part, it was .26*. The improvement on students knowledge explain the observed increase in the correlations for the sampling and inference parts of the test.

Table 10. Posttest Correlations between Standardized Items and Items designed for this Study

	Posttest	Researcher Items			
		Total	Sampling	Data Analysis	Inference
CAOS ITEMS	Total	.59**	.42**	.46**	.28**
	Sampling	.25*	.25*	.16	.11
	Data Analysis	.55*	.25*	.52**	.26*
	Inference	.52**	.43**	.36**	.26*

4.1.4 Item Discrimination

Item discrimination measures were obtained for the posttest items. The discrimination index was calculated using the procedure proposed by Kelley (Engelhart, 1965). This formula takes participants with scores below the 27 and above the 73 percentile, subtracts the number of right answers for any given item in the high-scoring group from the number of right answers in the low-scoring group and then divide this number by the size of the groups. This process was conducted for the global score in the posttest, for the sampling, data analysis and inference scores. The results of this process are presented in table 11. This table presents the items' numbers and discrimination indices for the whole posttest, and for the sampling, data analysis and inference parts of the test in the first five columns. In the following three columns, the table presents the difference between the discrimination index for the whole test and the discrimination index for each subarea. In the last column to the right, the table presents the average of differences between the discrimination index for specific parts of the test and the discrimination index for the whole test (C7, C8. C9). In the same column in the parenthesis, it is presented the average of the difference between the posttest and the sampling discrimination index (C7), and the difference between the posttest and the data analysis index (C8). Low averages in this column indicate that the item discriminates between good and poor learners but it does not discriminate between different levels of learning in specific content areas.

Table 11. Item Discrimination

Name	Items	Space	Post	Sam	Data	Inf	Tot-Sam	Tot-Dat	Tot-Inf	Averages
	C1	C2	C3	C4	C5	C6	C7=3-4	C8=3-5	C9=3-6	[7+8+9]/3 ([7+8]/2)
Data1	1	Data	0.59	0.23	0.77	0.36	0.36	0.18	0.23	0.26 (0.27)
Data2	2	Data	0.77	0.36	0.86	0.59	0.41	-0.09	0.18	0.23 (0.25)
Sam1	3	Sam	0.55	0.82	0.27	0.41	-0.27	0.27	0.14	0.23 (0.27)
Data3	4	Data	0.59	0.32	0.77	0.64	0.27	-0.18	-0.05	0.17 (0.23)
Sam2	5	Sam	0.50	0.91	0.36	0.36	-0.41	0.14	0.14	0.23 (0.27)
Inf1	6	Inf	0.50	0.32	0.41	0.82	0.18	0.09	-0.32	0.20 (0.14)
Data4	7	Data	0.59	0.32	0.86	0.45	0.27	-0.27	0.14	0.23 (0.27)
Data5	8	Data	0.50	0.50	0.55	0.32	0.00	-0.05	0.18	0.08 (0.02)
Data6	9	Data	0.45	0.27	0.64	0.55	0.18	-0.18	-0.09	0.15 (0.18)
Data7	10	Data	0.64	0.50	0.82	0.55	0.14	-0.18	0.09	0.14 (0.16)
Inf2	11	Inf	0.73	0.82	0.45	0.91	-0.09	0.27	-0.18	0.18 (0.18)
Inf3	12	Inf	0.55	0.23	0.36	0.73	0.32	0.18	-0.18	0.23 (0.25)
Sam3	13	Sam	0.73	0.82	0.27	0.36	-0.09	0.45	0.36	0.30 (0.27)
Sam4	14	Sam	0.55	0.32	0.18	0.36	0.23	0.36	0.18	0.26 (0.30)
Sam5	15	Sam	0.18	0.50	0.00	0.05	-0.32	0.18	0.14	0.21 (0.25)
Sam6	16	Sam	0.55	0.41	0.32	0.36	0.14	0.23	0.18	0.18 (0.18)

Discrimination indexes below .2 are consider to be low and above .4 to be high (Ebel, 1954). According to this criteria, only Sam5 would be considered as having a low discrimination index for the entire posttest score. Additionally, no item would be considered as having a low discrimination index in its own specific statistical sub area (in gray). When the differences between discrimination indexes for the whole posttest and specific areas are compared, only Data 5 seems to have a low average of differences in column 10 (.08(.02)). This low average indicates that this item has the same discrimination level for the whole test and for specific content sub areas; in other words, this item might be considered a general knowledge item.

4.1.5 Structure of the Measures

To explore the structure of the measures, a factor analysis was conducted on the posttest items. This analysis was restricted to the posttest because the initial analyses showed that low levels of student knowledge affected the structure and consistency of the measures in the pretest. The factor analysis used a orthogonal varimax rotation and the number of factors was constrained to two.

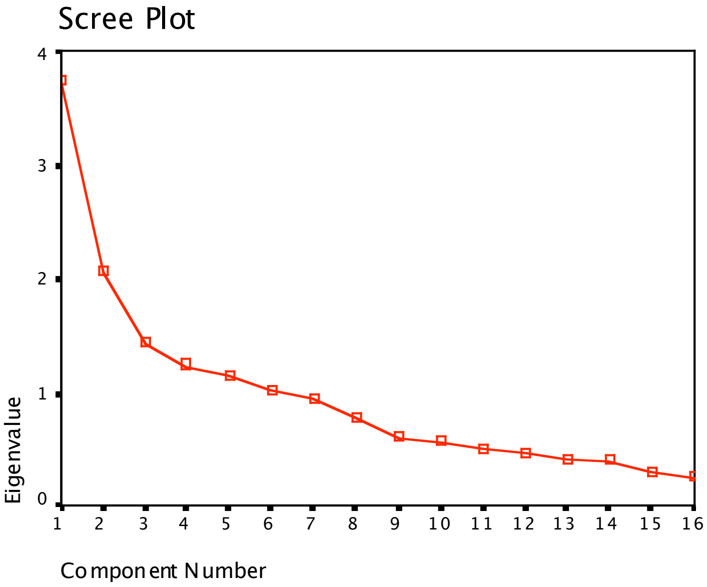


Figure 5. Factor Analysis Screeplot

The results of this analysis showed that items evaluating data analysis grouped in the same factor with exception of Data5. It showed also that the items evaluating sampling grouped in the same factor. Both factors had eigenvalues over 1 (Figure 5), and they explained altogether the 36% percent of the data variance. The first factor is the data analysis factor. As can be seen in table 12, it groups items Data1, Data2, Data3, Inf1, Data4, Data6 and Data7. All of them except Inf1 are data analysis items according to the classification explained in the methods section. Inf1 is an inference item; it required students to interpret an ANOVA table and produce conclusions in the context of a case. It is hard to establish why this item falls in the data analysis category. A possibility is that Inf1 required interpreting the context of a case to produce a meaningful conclusion, and, in this sense, this item shared several characteristics with other items in the data analysis category (e.g., items Data3, Data4, Data6). It is possible also that, as proposed in the theoretical framework of this work, statistical inference requires combining data analysis and sampling knowledge and, for this reason, inference items could fall in either category.

Table 12. Factor Analysis Rotated Component Matrix

Name	Item	Space	Component	
			1	2
Data1	1	Data	.71	-.06
Data2	2	Data	.62	.16
Sam1	3	Sampling	.03	.64
Data3	4	Data	.55	.17
Sam2	5	Sampling	.00	.68
Inf1	6	Inference	.49	.13
Data4	7	Data	.80	-.04
Data5	8	Data	.16	.45
Data6	9	Data	.63	.03
Data7	10	Data	.65	.13
Inf2	11	Inference	.34	.53
Inf3	12	Inference	.24	.36
Sam3	13	Sampling	.05	.69
Sam4	14	Sampling	.24	.32
Sam5	15	Sampling	-.14	.34
Sam6	16	Sampling	.08	.53

The factor 2 groups items that evaluate knowledge about sampling and sample size effects. Items included in this factor are items Sam1, Sam2, Data5, Inf2, Inf3, Sam3, Sam4, Sam5, Sam6. In this list, items Sam1, Sam2, Sam3, Sam4, Sam5 and Sam6 were initially classified as sampling items (Table 3). Items Inf2 and Inf3 were classified as inference items but they involve understanding sample size effects. Data5 was clearly classified as a data analysis item as it did not required students to understand sampling or sample size effects. A possible reason for this result is that this item has a high correlation with both factors. Table 13 shows that Data5 has a higher factor loading with factor 1 than with factor 2 before the rotation is conducted, and table 12 shows that the factor loadings for this item are positive with both factors after the rotation is conducted. It is possible that this item discriminates between high and low performance students despite of the knowledge on specific content areas. This statement is consistent with the findings described in table 11 that show that Data5 has similar discrimination scores for the whole test, and for the data analysis and sampling parts of the test.

Table 13. Factor Analysis Unrotated Component Matrix

Name	Item	Space	Component	
			1	2
Data1	1	Data	.55	-.45
Data2	2	Data	.61	-.21
Sam1	3	Sampling	.39	.51
Data3	4	Data	.55	-.16
Sam2	5	Sampling	.38	.56
Inf1	6	Inference	.48	-.16
Data4	7	Data	.63	-.48
Data5	8	Data	.39	.28
Data6	9	Data	.54	-.33
Data7	10	Data	.61	-.25
Inf2	11	Inference	.58	.24
Inf3	12	Inference	.40	.15
Sam3	13	Sampling	.43	.54
Sam4	14	Sampling	.38	.13
Sam5	15	Sampling	.06	.36
Sam6	16	Sampling	.36	.39

The analysis conducted on the items shows that the measurement system used in this study has reasonable levels of reliability, internal consistency and external validity. The structure of the measures corresponds in general terms to the structure of the content domain. The distinction between data analysis and sampling that will be used in the subsequent analyses of this text is supported by the review of the measurement system. The item discrimination indexes have values within the limits of a decent measurement system. For future uses of this instrument focusing on the distinction between data analysis and sampling, it is recommended to erase Data5 because it relates to both spaces. It is then time to review the changes in students' knowledge detected by this measurement system.

4.2 MEASURING THE TRADEOFF BETWEEN DATA ANALYSIS AND SAMPLING SIMULATIONS IN STATISTICAL EDUCATION

This section describes the analysis conducted on students answers in order to compare the effects of both interventions on students performance, and the potential of the measurement system to capture the students change during the study. The first part of this section describes analyses conducted on the pretest to assure that there were not prior differences between the treatment groups. The second part of this section describes the changes produced by the interventions on students knowledge. From now on, treatment, intervention or group will refer to the different conditions involved in this study; treatment or intervention effect will refer to differences among the groups; pre, posttest change, pre-post variable, occasion or time will refer to the change between the pretest and the posttest. Setting will refer to the different courses in which the data for this study was collected; for example, differences among settings will mean that there were differences among the participants coming from different courses. Finally, amount activity and completion will refer to the percentage of exercises solve by students during the study.

4.2.1 Differences in the Pretest

The global score and the sampling, data analysis, and inference sub scores did not differ significantly between the participants in the sampling and data analysis conditions (Table 14). There were no significant differences between these groups of students in any item. The only item that had a *p*-value below .10 was Sam6; in this case, the participants assigned to the data analysis condition had a slightly higher mean.

Table 14. Treatment Differences in the Pretest

ITEM/SCORE	Intervention	N	Mean	Std. Deviation	Sig.
SAMPLING	Sampling	42	1.33	1.37	.19
	Data	42	1.71	1.27	
INFERENCE	Sampling	42	1.07	.71	.89
	Data	42	1.04	.93	
DATA	Sampling	42	3.80	1.78	.66
	Data	42	3.97	1.75	
PRETEST	Sampling	42	6.21	2.94	.40
	Data	42	6.73	2.83	

4.2.2 Intervention Effects: Effectiveness of Computer-based Statistical Education

The intervention effects were analyzed using a Mixed ANOVA model in which the treatment was assigned as the between subjects factor, and the pre, posttest change was assigned as the within subject factor. These analyses indicated that both the pre-post variable and the interaction between the pre-post variable and the intervention had significant effects on the global scores. In other words, the results showed that there was a significant change in the global scores from the pretest to the posttest (F(1,82)= 33.13, p=.00), and that the trajectories of change were different for participants in different treatment conditions (F(1,82)=9.84, p=.00) (Figure 6).

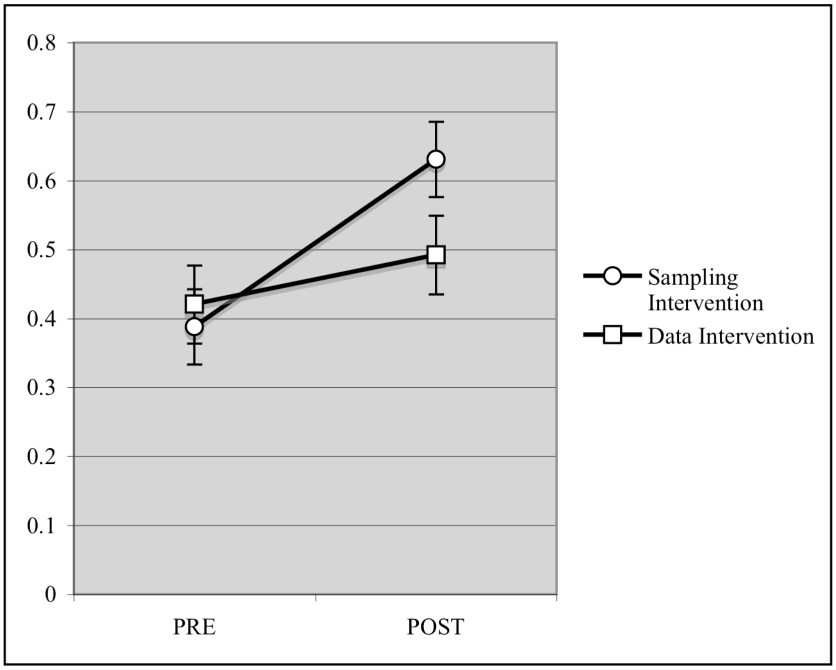


Figure 6. Pre, Posttest Change by Intervention (Global Score)

However, when pre and posttest scores were disaggregated into the sampling, data analysis, and inference parts of the test, it was clear that the change from pretest to posttest was produced only by gains in the sampling knowledge. When a Mixed ANOVA was calculated for the sampling scores, the pre-post change continued being significant (F(1,82)=57.93, p=.00), as well as the interaction between occasion and intervention (F(1.82)=11.53, p=.00). The situation was different for the data analysis scores. The Mixed ANOVA results indicated that neither the change in time (F(1,82)=2.66, p=.10), nor the interaction between time and intervention (F(1,82)=2.39, p=.13) were significant (Figure 7).

With respect to the inference scores, the Mixed ANOVA results indicated that there is a strong change from pretest to posttest (F(1,82)=17.39, p=.00), and a moderate interaction between intervention and occasion (F(1,82)=4.15, p=.05). The change is stronger for participants in the sampling intervention than for the participants in the data analysis condition. A possible explanation for this result is that change in the inference scores was related to gains in sampling knowledge produced by the interventions, but it was not related to gains in the inference knowledge itself.

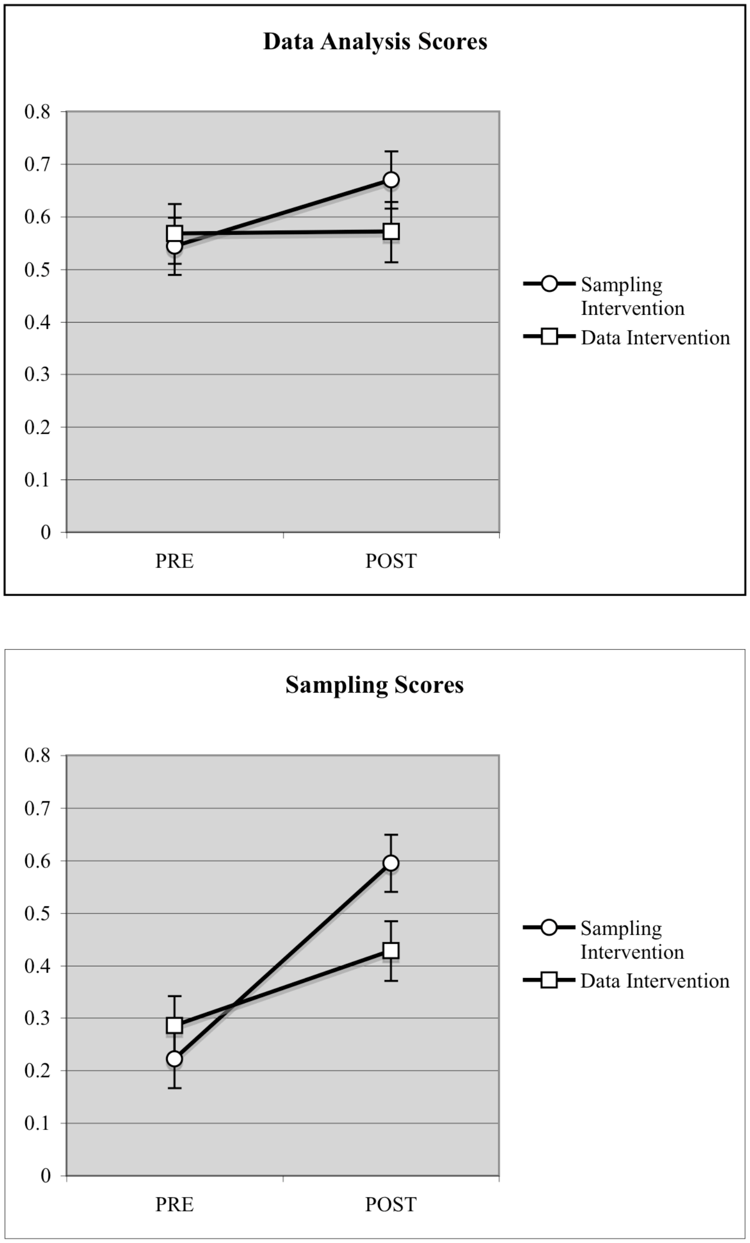


Figure 7. Pre, Posttest Change by Intervention (Data and Sampling Scores)

All the analysis presented in this section were double-checked using adjusted ANCOVAs on the posttest scores, controlling for pretest scores and using intervention as between-subjects factor. This analysis showed the same picture regarding the changes of participants during the study: strong effects for sampling knowledge, moderate effects for inference knowledge and no effect for data analysis knowledge. The effect sizes for this analysis as reported by the eta squared values in the adjusted ANCOVAs are .10 for the global scores, .11 for the sampling scores, .03 for data analysis scores, and .07 for the inference scores.

4.2.3 Testing for Differences among Settings

The data for this study was collected in three different settings: two psychology courses and one political science course. To ensure there were not differences among these settings that influenced the results or interactions between the settings and the treatment conditions, Mixed ANOVAS were calculated using the pre-posttest change as the within subjects variables, and the treatment conditions (Intervention) and settings as between subjects factors. For the global score, the introduction of the setting in the ANOVA analysis did not alter the original results: the pre-posttest change was significant, and the intervention groups had different trajectories. The Mixed ANOVAs on the sampling and data analysis scores show the same pattern: only occasion and the interaction between occasion and intervention had significant effects (Table 15).

For the inference scores, the pre-post change was significant, as well as the interaction between the pre-post change and setting. The interaction between pre-posttest change and intervention was not significant. Overall, it is possible to assert that the differences between the trajectories of both groups are weak and that means that the introduction of a new variable alters the significance of the F-statistics. The same happens with the trajectories of the students belonging to different courses: the effect of the interaction is weak and therefore the significance changes the intervention is introduced as a new between subjects factor. This issue will be explored later in the holistic models section.

Table 15. Intervention, Setting and Pre, Posttest Change

	Global Score	Sampling	Data	Inference
Source	F	F	F	F
Pre/Posttest Change	25.78**	40.52**	2.2	18.03**
Pre/Posttest Change*Intervention	7.71**	8.45**	2.0	3.71
Pre/Posttest Change*Setting	.01	.28	.07	3.13*
Pre/Posttest Change*Setting*Intervention	1.02	.09	1.4	2.05

4.3 COMPLETION

In this section, the effects of completion are explored. It is possible that students' motivational levels during the study moderate the effects of the interventions. Highly motivated students can respond better to the interventions' exercises than students with low motivational levels. It is possible that students having higher levels of completion experience larger gains during the interventions. These possibilities explored in the following pages. The amount of activity during the study was determined by examining the answers that students gave to the activity exercises on the electronic forms. A variable was then constructed by calculating the percentage of completion in the intervention activities for each student. The pretest measures did not correlate with the completion scores. The differences between participants in the sampling condition and in the data analysis condition were evaluated using a Welch statistics and were not significant (Levene=6.16*, Welch(1,66.72)= 1.8, *p*= .174). The same result was found for participants belonging to different courses (Levene=5.8**, Welch(1,28.3)=1.69, p=.202).

There was no correlation between completion and the pre, posttest change, either in the total score, or in the sampling, data or inference sub scores. The same result was found with a repeated measures design in which completion was introduced as a covariate: there was no significant interaction between completion and the change from pretest to posttest in any of the measures (total, sampling, inference, or data). A possible explanation is that above certain number of exercises, the interventions had the same effect: when the correlations are calculated for students who completed less than 60% of exercises, the correlations between the completion scores and the change in the global score (.42*) and in the sampling score (.34*) are marginally significant.

4.4 HOLISTIC MODELS

In the search of a more integrative interpretation of the results, two holistic models were constructed. The first model explains the change from pretest to posttest in terms of the complete set of variables included in this study. The second model reviews the relationship among sampling, data analysis and inference knowledge under the premise that sampling and data analysis knowledge determine the change in inference scores as suggested by the analysis of the literature.

As pointed out in the theoretical review, statistical inference is, from a cognitive point of view, the product of combining data analysis and sampling knowledge bases. Therefore, it is expected that there is a strong relationship between changes in data analysis and sampling knowledge with the gains in the ability to solve inference problems. This hypothesis was tested by building a mixed linear model with an autoregressive covariance structure that specified pre-posttest change in inference scores as the dependent variable, change in sampling and data analysis scores as covariates, and the intervention as between subjects factor.

Table 16. Holistic Model: Data Analysis and Sampling Change predict Inference Change

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	139.03	229.52	.00
Intervention	1	139.03	4.07	.04
Data Analysis Change	1	590.44	22.74	.00
Sampling Change	1	544.18	12.21	.00
Data Analysis*Intervention	1	590.44	1.39	.23
Sampling*Intervention	1	544.18	.33	.56

-2log likelihood= -503.361
AIC= -499.362

The results of this analysis, displayed in Table 16, show clearly that change in data analysis and sampling scores, relates to the change in students' inference scores. The results of this analysis indicate also that intervention plays a role in this process. The broader picture depicted by the holistic models presented in this section shows that changes produced by the intervention were significant for sampling knowledge. The inference scores, although affected moderately by the interventions, were related to small changes in students data analysis knowledge and to strong changes in students sampling knowledge.

According to the results of this study, hypotheses considered in the theoretical review regarding the relationship between sampling, data analysis and inference knowledge in statistical activity were confirmed; hypotheses regarding the effects of simulation-based instruction on the learning of students in the sampling condition seemed to find strong support also. However, the results in the data analysis group appeared to be more unexpected in the light of the literature reviewed in the initial part of this work. The performance of participants in the data analysis part of the pre, posttest was virtually unaffected by the data analysis intervention; this result contradicts claims, made in the first part of this work, namely that computer-based data analysis activity should have positive effects on participants data analysis skill. Additionally, sampling knowledge performance of participants in the data analysis condition improved significantly, although not to the same extent as for participants in the sampling condition. Gains in the data analysis condition were important and this was not predicted in the theoretical review conducted for this study. These two facts require further scrutiny of alternative evidence.

In the next section, the results of three different types of evidence are presented in order to shed light on the quantitative results presented to this point. The first section reviews the learning process in a sub sample of students from the beginning of statistical instruction to the intervention point. The second section presents a short protocol analysis study of a different sub sample of students answering pre, posttest items while thinking aloud. In the third section, students' written explanations to pretest and posttest items are classified and analyzed in relationship to the quantitative results presented before, and to the theory presented in the second chapter. Of particular interest is the fact that change in sampling knowledge, and no change in data analysis skill was found in participants in the data analysis condition. Two ideas were fundamental in the explanation of these results: first, that, given the high levels of initial data analysis knowledge, it made sense to suppose that the courses provided prior to intervention point much of the practice required to develop data analysis knowledge; and, second, that the improvement on sampling results was produced by the acquisition of decision rules that did not suggest a deep understanding of probability.

4.5 THE DEVELOPMENT OF DATA ANALYSIS KNOWLEDGE

An informed evaluation of students changes during the data analysis intervention requires understanding the context in which interventions took place. Students in this study participated in three statistics courses, all with an emphasis in data analysis. A central claim in explaining the small gains in data analysis knowledge is that students had undergone intense prior practice in data analysis tasks, and that, in some way, this prior practice overshadowed the intervention effects. The underlying idea is that data analysis is a complex ability whose underpinnings cannot be acquired in a short period of time. Data analysis, as conceived here, requires practice sustained for a time period long enough to allow students to master coordination of theory, representation, and data.

To support this idea, I describe the development of students' knowledge in the political science course from the beginning of instruction to the intervention point. Understanding this trajectory requires acknowledging that there is a representational specificity in statistical training and that substantial instructional activity is required in order for students to transfer what they have learned in one representation to another. To support this idea, students' transition from boxplots to histograms is described. A second layer of explanation delves into the assertion that data analysis knowledge is more than using statistical packages. This section shows that even when students were trained to produce representations using statistical packages, they did not display adequate representational interpretation.

4.5.1 Representational vs Interpretative skill

Instruction about boxplots relied on the concepts of center and spread measures that had been taught before week 6. Evaluations in week 5 indicated that students were able to identify the mean, median, and standard deviation of a small data set, and therefore that they had the notions of spread and center basic for the understanding of boxplots. In week 6, students in the political science course were trained to produce boxplots using SPSS and to interpret the results of this type of representation in terms of spread and measures of central tendency. At the end of the class, the results of the class exercises sent by students to the instructor via email showed that most students were able to graph variables and factor groups results as boxplots using SPSS. However, when asked to interpret boxplots in the quiz at the end of the class (Figure 8), more than half of the students were unable to produce an unified interpretation of a graph displaying four boxplots.

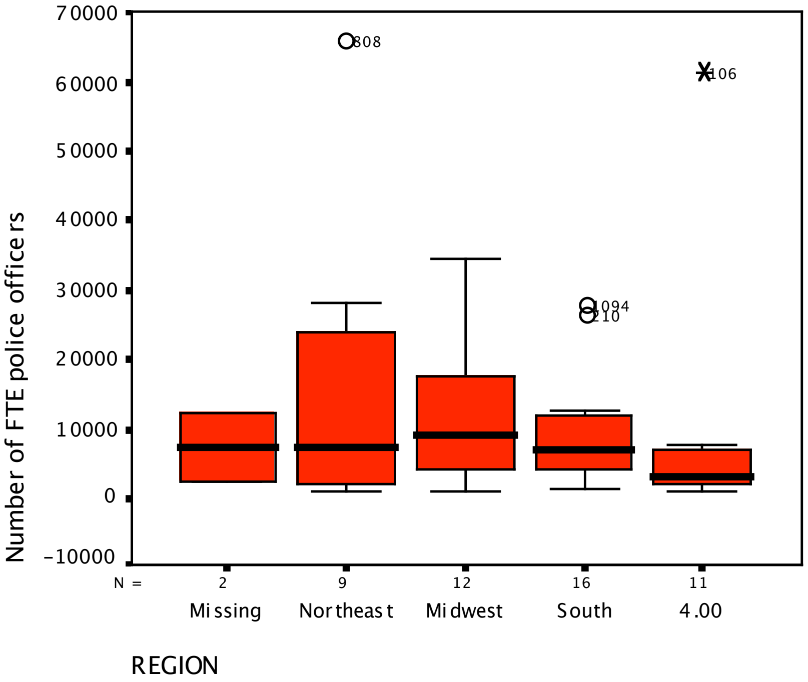


Figure 8. Quiz Week 6

They were able to list information coming from the Boxplots, but they could not produce a consistent comparison of the different groups. For example, student ER compared the number of police officers in the Midwest and in the Northeast by listing the central values of those states in the FTE police officers variable:

> *"The northeast counts with 23000 FTE police officers while the Midwest has 18000 police officers".*

"El northeast cuenta con 23000 FTE police officers mientras el Midwest cuenta con 18000 efectivos de policia".

This answer did not include any mention of variability and treated the upper quartile in the graph as raw data (upper quartile=total number of police officers), and it did not include any explicit comparison among groups, despite the fact that it was explicitly asked in the quiz. Student TY showed the same pattern when in the same comparison listed the information contained in the boxplot:

"In the south the great majority of states have less than 20000 policemen and only two states have…have more than 20000 policmen, the majority of states in the south are very homogeneous in the number of policemen, since the higher 25% is not really significant, that is, there are very few states that are different in the # of policemen in the superior quartile. In the Midwest, the superior 25% is broader than the inferior (25%) but there are many states with more than 20000 policemen".

"en el sur la gran mayoria de stados tienen menos de 20000 policias y solo 2 estados tienen mas de 20000 policias. la mayoria de estados del sur son muy homogeneso en cuantoa numero de plicias, pues el 25% mayor es muy poco significativo, es decir, hay muy pocos estados que se alejan en cuanto a # de policies enel cuartil superior. En el Midwest el 25% superior es mas amplio que el inferior pero hay muchos estados con mas de 20000 policias".

This answer is more complete than the one provided by ER because it presents an interpretation of spread and divides the interpretation into quartiles as boxplots do. However, it stills avoids comparison. Overall, 4 students out 14 provided comparisons based on adequate boxplots interpretations; 2 presented comparisons based on inadequate boxplot interpretations; 3 students presented adequate interpretations but no comparisons among boxplots; and 5 students did not produce adequate interpretations, nor comparisons during the exercise.

From this result, the instructor decided to spend the beginning of the following class (week 7) reviewing several comparison examples and portraying the characteristics of a global interpretation of boxplots. The purpose of this instructional sequence was to go beyond the use of SPSS and articulate representation and interpretation around cases (e.g. differences between States where Carter won, and States were Ford won). Exercises avoided the use of SPSS and presented several graphs containing boxplots elaborated by the instructor: the general sequence implied presenting a graph, modeling the comparisons, and telling the students explicitly that no individual boxplot interpretations "would suffice" without an integrative interpretation. Then, the sequence continued by asking students to solve similar exercises and by providing feedback. In the quiz given two weeks later (week 8), students showed some improvement, but it took them two weeks of graph interpretations to go from partial interpretations to global comparisons. Week 8 quiz required students to interpret boxplots and histograms that displayed divorce rates from four USA regions. In the first part of the quiz, from 13 students that solved the quiz, 7 students presented adequate boxplot comparisons; 4 made mistakes while trying to elaborate boxplot interpretations but tried to produced comparisons among boxplots; and 3 did not try to compare boxplots, and made mistakes in the individual boxplots interpretations. A prototypical right answer in this quiz was (FS):

"The graph shows that the West is the region with the highest divorce rate on average. It is followed by the south, the Midwest, the northeast, and the west. It is important to note that the 25% with the highest divorce rate is very disperse. The Midwest presented an atypical data that is 359. In the last quartile the Midwest is very concentrated around the 50% around the median. In the last region, west presents the highest divorce rate and has the highest divorce rate of all the graph".

El grafico muestra que el west es la region con una tasa mayor de divorcios en promedio. Le siguen el south , el midwest y el north west. Cabe resaltar que el 25% con la tasa mas alta de divorcios en el west es muy disperso. El midwest presenta un dato atipico que es 359. El ultimo cuartil de midwest esta muy concentrado alrededor del 50%, alr ededor de la mediana . En utlimas la region west presenta la mas alta tasa de divorcios y tien los casos con la tasa mas alta de divorcios de toda la grafica".

In the same line of improvement, student WU, who in the first quiz did not produce a comparison, nor an adequate boxplot interpretation, wrote the following answer that presented not only a correct boxplot interpretation but also included a explicit global comparison:

"West-South: between west and south despite there are higher extreme data points in the west, the median is not that different, and on average the (divorce) rates are not that different. Between the Midwest and the south, the rate is clearly different since the mean of the south is equivalent to the highest data points in the midwest (or it is proximate). General Interpretation, the west and the south for some reason not established in the

graph present mean divorce rates much higher to the extent than their averages are the extreme cases of other regions".

Oeste-Sur: entre el oeste y el sur si bien hay datso extremos mas altos, la meidana no es tan distinta y en promedio las tasas no son disimiles. Entre medioeste y sur, la tasa es bien ditinta pues la media del sur es uno de los dtos mas altos (o se aproxima). Interpretacion general: el oeste y el sur, por una razon no establecida en el grafico presentan tasas medias mucho mas altas de divorcio tanto que sus promedios son casos estresmos de otras regions".

The evolution from week 6 to week 8 shows that the process of representational interpretation, particularly but not exclusively of boxplots, is a skill that does not arise automatically from using SPSS, or from being able to follow a procedural routine, finding the way through computer windows or filling dialog boxes. It requires building an interpretative framework in which to place the results of the computer-based processing. This skill, as shown by week 6 quiz, requires specific training on interpretation and comparison of specific types of representations, and not general instruction on the nature the use of SPSS. This training needs to be directed to specific tasks (e.g, comparing distributions) and needs to promote global graph interpretations; otherwise students will tend to elaborate concrete interpretations of data and to focus mainly on procedural activity.

4.5.2 Representational Specificity

The second aspect of statistical instruction affecting the outcome of the data analysis intervention is Representational Specificity. Representational Specificity was identified during the political science course when instruction on Box plots did not transfer to solving problems that included Histograms. The qualitative nature of this finding does not allow asserting that this phenomenon generalizes to other types of representation; however, this finding presents a plausible explanation of why instruction in the data analysis intervention did not produce significant gains from pretest to posttest in the data analysis dimension. The fact that the data analysis exercises presented all exercises but one as Venn Diagrams could decrease the intervention effect on the posttest exercises that were presented as histograms. The identification of representational specificity took place in week 7 when students had undergone the instruction on boxplots, had solved the first quiz, and were starting the initial instruction on histograms.

At first, students were told that the x axis displayed the variable's magnitude (value) and the y axis the counts within certain ranges of variable X. Then, concepts of spread and center in this type of representation were introduced and connected to some examples. Then, SPSS routines were presented to students using a data base that provided information about USA regions and States. After the instructor had demonstrated the procedure, students were asked to construct histograms of some variables and to send the results via email to the instructor. After several practice exercises, students were asked to solve a quiz that required them to find similarities and differences between a variable represented as a histogram (expenditure in education) and another variable represented as a boxplot (number of high school graduates). It was expected that students easily would draw parallels between both types of representations because they had received instruction on boxplots during week 6 (e.g., y axis variable value; quartiles as presenting spread) and instruction on histograms during week 7 (Figure 9).

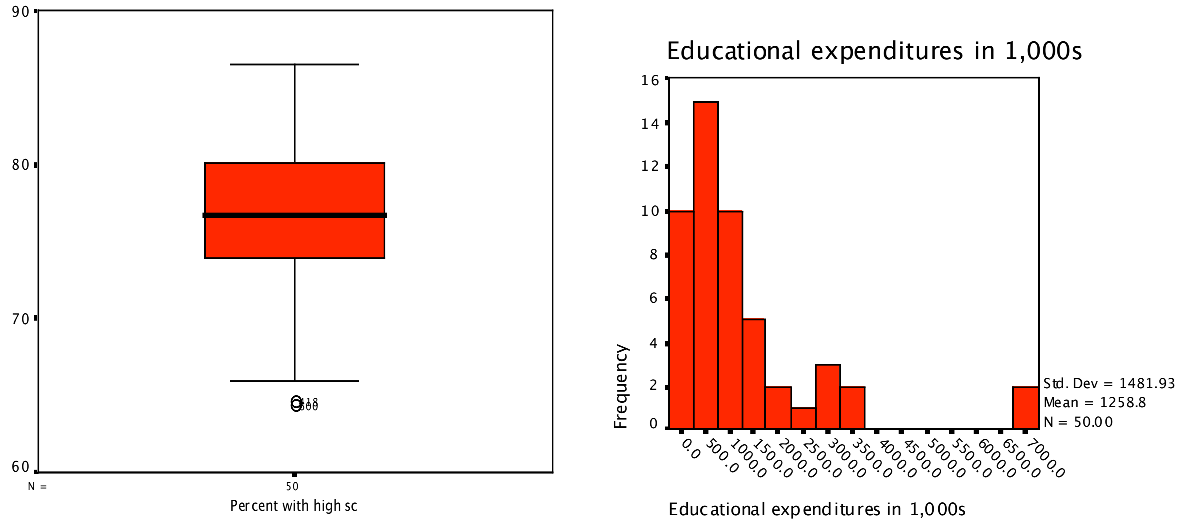


Figure 9. Quiz Week 7

The results of the quiz however showed a different picture. Students were unable to evaluate the differences between box plots and histograms. On the contrary, students produced two typical types of answers. The first type of answer consisted of lists of unconnected information pieces extracted from the boxplot or the histogram, without any explicit comparison between them. This finding was surprising because students were told explicitly in the initial part of the class that comparisons were part of a complete answer in problems that required to compare box plots. However, students were unable to transfer that instruction to the solution of this question requiring compare boxplots and histograms. For example, student WU provided this answer:

> "*The Boxplot shows that the four quartiles are found until approximately 3000 (educational expenditure). But it shows also that not all the states are in the four quartiles because there are 5 observations (states) that are below these quartiles. The*

histogram shows that 45 states, that are the majority, are found until 3000, and that there are 5 states above 3000. The histograms show also that there are 15 states in 500".

El boxplot muestra como hasta aproximadamente 3000 (gasto educativo en miles) se encuentra los 4 cuartiles, esto es todos los datos. Pero tambien meustra que no son todos los datos los que estan en los 4 cuartiles pues hay 5 observaciones (estados) es estan por debajo de eestos cuartiles. El histograma muestra que hasta 6000 se euncentra 45 estados que es la mayoria que hay 5 estados por encima de 3000. El histograma muestra tambien que hay 15 estados en 500".

The second type of answer was a definition of boxplot, side by side with a definition of histograms, or a comparison of general characteristics of both types of representations without connection to the particular problem or data set. This type of answer represent some tendency to produce mechanical answers from text definitions, without understanding the problem, or application conditions. The answer of student AC is an example of this type of response.

"The histogram and the boxplot are utilized to represent the behavior of continuous variables. The boxplot groups by group, as for example by region or position, and it permits to calculate the mean and standard distribution. The histogram represents the amount of data within the groups of a continuous variable for example how the population is distributed in relation with the mean and what type of normal distribution it takes. That is, if it has bias to right or to the left of the mean".

El histograma y el boxplot se utilizan para representar el comportamiento de variables continuas. El boxpot agrupa por grupos, como por ejemplo por region o posicion, y permite calcular media y distribucion estandar. El histograma representa la cantidad de datos dentro de los grupos de una variable continua por ejemplo como se distribuye la poblacion en relacion con la me dia y que tipo de distribucion normal toma la tendencya, es ecir, sies normal o presenta sesgos hacia laderecha o la izquierda de la media".

The first type of answer was given by 30.7% students; the second type of answer was given by 38.4% students. Full interpretations or adequate comparisons were presented by 30.7% Students. Full interpretations were usually shorter and went directly to the point. They define the whole comparison in two or three lines. The first and second type of answers were longer because students could not give an unified account of the relationship between both variables. This difference can be seen in student BA that gave a full interpretation as answer for week 7 quiz:

"These diagrams are very different, which shows that there is not relationship between States educational expenditure, and the percentage of the population with high school diploma".

Estos diagramas son muy diferentes, lo que muestra que no hay una relación entre el gasto en educación de los Estados y el porcentaje de poblacion con diploma de bachillerato".

For week-8 quiz, students improved in their ability to compare boxplots and histograms. This quiz asked students to evaluate the difference among 4 boxplots and then to evaluate the differences among 4 histograms, and to establish the relationship between both representations. In the quiz, histograms and boxplots were presented as representing different variables but they were actually built from the same data sets. Ten out of 14 students believed that the boxplots and histograms displayed the same information.

The two phenomena depicted here explain to some extent the low change in the data analysis condition. Understanding of data analysis is only produced by sustained activity and feedback on interpretative mistakes, as well as, by strong and clear presentation of what an adequate interpretation requires. Activity and feedback on SPSS routines does not creates improvement on interpretative skills, at least that accompanied by contingent interpretation in context. In addition, the other phenomenon, representational specificity, suggests that training in the data analysis intervention could not produce change from pretest to posttest, because this intervention presented examples as abstracts Venn diagrams, but not as histograms or boxplots. A Venn diagram is the same for any data distribution and that makes hard to develop specific interpretative skills for deciphering boxplots and histograms in this type of instruction.

4.6 PROTOCOL ANALYSIS: THE COGNITIVE PROCESS OF SOLVING STATISTICAL ITEMS

Studying the evolution of knowledge during the political science statistics course explains to some extent the interventions effects on students knowledge. However, some points require the consideration of a different sort of evidence. Protocol analysis provides additional insight on the micro processes that allow or difficult knowledge acquisition in both data analysis and sampling knowledge. Particularly, this type of evidence offers important insight on the reasons why the interventions produced significant changes in sampling knowledge of students in both the sampling and data analysis conditions.

This section presents think aloud protocols obtained from 12 students answering 3 data analysis and 3 sampling items in the pre, posttest. The examination protocols of students solving data analysis questions is designed to describe the reasoning process of solving prototypical data analysis items and to explain why this process was untouched by the interventions. The examination of the protocols of students solving the sampling part of the test is aimed at understanding the changes in the reasoning process that made the interventions effective in improving students answers. In this regard, an important goal of this analysis was discerning between gains produced by the acquisition of propositional information, from gains produced by more complex modifications in the reasoning process, such as the elaboration of complex representations of probability theory, or chance models.

4.6.1 Protocol Analysis of Data Analysis Items

The three questions selected for this analysis represented prototypical tasks of distributions comparison. These items (Data1, Data2 and Data3 from the first section of the pre, posttest) required students to compare several pairs of distribution graphs with different mean differences, spreads and sample sizes. The items were coded according with the system described in the methods section.

4.6.1.1 The General Process of Decision Making in Data Analysis Items

The cognitive process depicted by the protocol analysis suggests that the solution of data analysis problems goes through one or several rounds of description, before participants start comparing and trying to answer the items. This process was evident in the analysis of participants' answering sequences in which descriptions of spread and center preceded the production of answers. Participant MC for example conducted one round of describing and comparing before starting to infer in order to producing an answer to Data2:

> *In this, in the B (P)/, eeeh,no, so,/this has like the sample more compacted (DS)/, this has the sample more disperse (DS)/, but if I, I don't know, compare this two segments, I think they are like the same (CC)…/ then the differences should be larger in this (I)/because there is a larger dispersion of the… the samples (E)/. Maybe, here, they are a little bit… lies!, je, no.(O)/ in this there is a larger differences of the samples because in this they overlap , they over…lap more the samples (A)/, therefore, in this there is a larger difference (A).*

> *En esta en el B (P)/eeeeh no pues /este tiene como la meustra mas recogida(DS)/, este tiene la muestra mas dispersa (DS)/ pero si yo no se comparo estos dos segmentos creo que son como igual(CC)…. /entonces las diferencias deberian ser mayor en esea (I)/ porque hay una dispersion mas grande de la… las muestras(A)/. De pronto aqui estan como un poquito… mentiras… je… n (O)/ . en esta hay una mayor diferencia de las muestras (A)/ porque en esta se superponen se superp..ponen mas las muestras (E)/, entonces en esta hay mayor diferencia.*

Once the spread and center description and comparison were made, participants produced answers based on decision rules that will be described later. The percentage of students producing answers after at least one round of describing and comparing are presented in table 17. This type of behavior was observed for no less than 66% of the cases in any item. In most cases, when the describing-comparing-answering sequence was not observed, it was due to the fact that some students produced short integrated answers that came almost automatically. In the codification system of this study, when answering codes were accompanied by explaining or inferring codes, they were coded as a single answering code.

Table 17. Percentage of Protocols containing Describing and Comparing prior to Answering

Item	Data1 (%)	Data2 (%)	Data3 (%)
Pretest	83.3	75	66.6
Posttest	75	75	66.6

The same phenomenon was seeing when answering codes were analyzed. All protocols for data analysis items contained references to center and spread as justification for a given answer. Clearly, the justification varied as a function of the item being answered (Table 18). Center was core for Data1 and spread for Data2. This fact indicates that students were aware of the core elements of the problem, and they had been taught to identify center and spread differences in graphical representations.

Table 18. Answering Codes containing Justifications in terms of Center and Spread

	Pretest Items (%)			Posttest Items (%)		
	Data1	Data2	Data3	Data1	Data2	Data3
Center	83.3	33.3	41.3	91.6	33.3	33.3
Spread	16.6	66.6	58.3	8.3	66.6	66.6
Sample	0	0	0	0	0	0

4.6.1.2 Accuracy of Descriptions

When the accuracy of descriptions was analyzed, it was found that most participants were able to produce adequate descriptions of the information being depicted by the items' graphs. Center differences were described adequately by most participants in all items. That is, in all items most participants were able to identify the cases with largest center differences. The few errors observed in students answers were related to perceptual mistakes in items were the distributions had different spreads. Spread differences were described adequately in all items for the pretest and posttest. No evident differences between pretest and posttest were found (Table 19).

Table 19. Accuracy of Descriptions

	Pretest Items (%)			Posttest Items (%)		
	Data1	Data2	Data3	Data1	Data2	Data3
Center	100	83.3	83.3	100	83	83
Spread	100	100	100	100	100	100

4.6.1.3 Change in Decision Rules for Data Analysis Items

The natural question from the above presented results is what factor made the difference on students that produced right answers. Most students conducted a process of description and comparison, and most of them were able to describe correctly the elements of the problems, therefore, the difference between right and wrong answers could not lie in those aspects. The analysis of students answers shows that the critical element in the adequate solution of data analysis items is the presence of certain decision rules. A decision rule was defined as a combination of center and spread parameters that would determine the significance of a difference. The decision rules were identified using either the information coming from the answering code, or the combined information of the cognitive actions produced immediately before the answer, and the information provided by the answer itself.

The decision rules identified in the pre, posttest answers were compared with the final answer provided by each participant. This analysis show that there was positive relationship between decision rules and right answers in the pre and posttest data analysis items. Four main decision rules were found. The first rule (1): the larger the center difference, the more significant the mean difference (and vice versa); the second rule (2): the larger the difference between the extremes of the distribution, the larger the mean difference; the third rule (3): the larger the spread of both distributions, the more significant the difference. Finally the fourth rule (4): the larger the spread of both distributions, the less significant the difference. To simplify the presentation when decision rules appeared in different ways in participants' protocols but they had the same meaning, they were coded as the same rule. For example, decision rule 1 was expressed by participant 4 in the following way: "the gray are.. is smaller, therefore the difference is less", implying that there was less space between the centers of both distributions in the graphs.

In Data1, it was clear that decision rules 1 and 2 were associated with right answers (Table 20). Most mistakes in this item were caused by erroneous description of basic features of the representation, by erroneous inferences (explicit or implicit) from the representation (e.g. the height of the curve as the mean value) or by comparisons based on erroneous descriptions or inferences from basic features. In Data2 and Data3, decision rule 3 was associated with wrong answers and decision rule 4 was associated to right answers. In some cases, students used decision rules 1 and 2 because they considered erroneously that there were center differences between the distribution pairs being compared.

Table 20. Percentage of Right and Wrong Answers by Decision Rules

		Pretest		Posttest	
Item	Decision Rule	Right Answer (%)	Wrong Answer (%)	Right Answer (%)	Wrong Answer (%)
Data1	A larger centers' difference implies a higher significance	41.6	8.3	50	0
	A larger difference between extremes implies a higher significance	33.3	16.6	41.6	8.3
	A larger distributions' spread implies higher significance	0	0	0	0
	A larger distributions' spread implies a lower significance	0	0	0	0
Data2	A larger centers' difference implies a higher significance	0	8.3	0	8.3
	A larger difference between extremes implies a higher significance	0	8.3	0	0
	A larger distributions' spread implies higher significance	0	58.3	0	58.3
	A larger distributions' spread implies a lower significance	25	0	33.3	0
Data3	A larger centers' difference implies a higher significance	0	8.3	0	0
	A larger difference between extremes implies a higher significance	0	8.3	0	0
	A larger distributions' spread implies higher significance	0	50	0	66.6
	A larger distributions' spread implies a lower significance	33.3	0	33.3	0

4.6.1.4 Long and Short Answers in Data Analysis Items

Additional analysis of participants answers pointed out that there were two general kinds of answering strategies. The first strategy was to solve items after a few rounds of exploration. In this type of strategy, definitive answers were given after little exploration of the problem space, when the more relevant elements of the problem had been identified (e.g., distributions' spread, mean differences). The second strategy implied a deeper exploration; in this strategy, participants not only assessed the main elements of the problem, but they tried to uncover relationships not evident in those elements. In strategy 1, participants did not try to discover "the decision rule" for solving a problem, but they guessed from the information they had; they either had or did not have the answer. In the second strategy, students tried to discover the decision rule through the exploration of relationships hidden in the graphs (e.g., "the curves overlap more when there is less spread). This difference was evident in the higher presence of inferring codes in strategy 2 protocols, but no other difference could be identified in the coding of both strategies, except the length of the protocol. The reason why some students produced type 1 and type 2 strategies might be related to participants' global conceptions of mathematics as a discipline (Boaler & Greeno, 2000), to the epistemological beliefs held by students (Shoenfeld, 1983), or to individual differences in the tendency to produce self-explanations (Chi & Bassok, 1989).

In the pretest, only 3 students produced strategy 2 responses for at least one item; the other 9 participants produced strategy 1 responses. In the posttest, all students produced strategy 1 answers, in part because they evaluated the problem as redundant and they felt that the interventions did not provide new information. Strategy 2 was included because it permits one to illustrate complex use of information in data analysis items. Strategy 1 permits one to illustrate more standard processes on data analysis items. Students that used strategy 2 arrived at correct decision rules even when they did not have those decision rules at the beginning. Students in strategy 1 trusted decision rules in the absence of checking mechanisms, arriving quickly to conclusions that not necessarily implied deep comprehension of the rule.

4.6.2 Protocol Analysis of Sampling Items

The answers for sampling items (Sam1, Sam3 and Sam5) were usually shorter and the distinction between short and long answer was not made for sampling items. This fact suggested that fewer transformations were necessary to convert the items text into pieces of information. In many cases answers were given right after reading the items with only short explanations as support. In this section, only protocols to Sam3 and Sam5 are reported. Protocols for Sam3 were usually short and uninformative; in the pretest, most students expressed that they did not know the answer; in the posttest, most students reported in very short statements that “larger samples implied more significance”. Protocols of students answering Sam3 and Sam5 were classified according to the coding system presented in the methods section. The coding system classified the cognitive actions of students into different categories. Describing and comparing were coded when students referred one or several elements of the problem directly, indicating or comparing either its spread, central tendency or sample size. Inferring and explaining codes were used when students concluded or justified an answer. Answering was coded when students presented direct answer to the items’ questions.

4.6.2.1 Decision Rules in Sampling Items

The protocols analysis of sampling items showed some change from probabilistic misconceptions in the pretest to ideas that relate correctly sample size to mean parameters in the posttest. This change was, however, associated to the acquisition of decision rules, without evidence of gains in the understanding of probabilistic behavior. Answering codes were analyzed an classified. In the pretest, four main types of answers were identified. Causal explanations (1); equiprobability bias (2); larger samples produce lower significance (3); and larger samples produce higher significance (4). Causal explanations (1) referred to answers in which the probability of some event was evaluated in terms of causal mechanisms. For example in Sam3, student 10 answered that increasing the number of times you weigh a rock does not increase the precision of the weight estimation, because *"you have to be weighing it wrong, what you have to change is the instrument, no how many times you weigh it"*

"*Osea, la esta pesando mal, lo que tiene que cambiar es el instrumento y no cuantas veces la pesa".*

Equiprobability bias (2) was coded when participants asserted that all events regardless the sample size had the same probability. Student 8 answer in Sam3:

"*I'd say that both have the same probability, thus, because I don't think that there is an influence that you weighted 20 times, I imagine that, after some point it is going to weight the same the rock".*

"Yo diria que ambos tienen la misma probabilidad pues porque no creo que influya que la pese 20 veces me imagino que despues de un punto siempre va a pesar igual la roca".

Larger sample sizes produce higher significance (3), and larger sample sizes produce lower significance (4) were coded when participants expressed the same propositional idea in any way. For Sam5, larger samples, lower significance was coded also when students expressed that small samples tended to produce more extreme data; larger samples, higher significance was coded when students affirmed that larger samples produce less extreme sample means (Table 21).

Table 21. Percentage of Misconceptions and Decision Rules in Sampling Items

Sampling Items	Sam3 (%)		Sam5 (%)	
	Pre	Post	Pre	Post
Causal	16.6	8.3	33.3	16.6
Equiprobability	41.6	0	16.6	0
Larger sample sizes imply higher significance	0	0	33.3	25
Larger sample sizes imply lower significance	25	66.6	16.6	50
Other	16.6	25	0	8.3

4.6.2.2 Change from Pretest to Posttest in Sampling Items

No clear tendency was found in the pretest results but the presence of probability misconceptions was clear (Table 21). In the posttest, results showed that students used decision rule 4 (Larger samples, higher significance) more frequently than any other rule. This result indicated some effect of the interventions on statistical reasoning. However, the acquisition of decision rule 4 did not seem to be associated with deep changes in the reasoning structure. There were not many explanations associated with students answers in the pretest, nor in the posttest. Only 16.6 percent of students in the pretest and 25 percent of students in the posttest gave an explanation for the use of decision rule 4 in any sampling item. The sense-making function of having a complex understanding of decision rule 4 can be observed in the answer of student 6 to Sam3.

> *"The probability of having extreme values is higher in the sample of 10 students because a very high or very low value can pull the sample much more significantly than in a sample of 50 students, I mean, in the largest sample"*

> "*La probabilidad de tener valores extremos es mas alta en las muestras de 10 estudiantes porque un valor muy alto o muy bajo puede jalar la muestra mucho mas significativamente que en la meustra de 50 estudiantes, osea en la muestra mas grande".*

This type of answer shows an adequate level of command of probabilistic concepts and some sort of model of what sampling means. The same cannot be said about student 4's answer to the same item that exemplify the acquisition of decision rule 4 without a deep understanding of it:

> *"So, the larger the sample, more probability of.. no? thus the one that weights the rock 20 times"*
>
> *"pues entre mas grande sea la muestra mas probabilidad de no? entonces el que la pesa 20 veces."*

The fact that participation in the study gave students new information but did not create in most cases new ways of representing sampling phenomena at least for the protocol sample explains the low change observed for Sam5. In Sam5, students could not use directly the rule that larger samples increased the significance of the mean difference, or the trust of researchers on the observed differences. This item required students to predict given two sample sizes in which there was a larger probability of finding extreme values. Some students assume that if you have more cases, there will be a larger probability of finding extreme values. For example student 7 answered:

> *"So, here I'd choose B because within 50 students is more probable to have like the, like the highest and the lowest, because the sample is larger".*

> "*Pues aqui tambien escogeria B porque dentro de 50 estudiantes es mas probable tener como los como mas altos y mas bajos porque es mas grande la muestra".*

This answer exemplifies how a mechanical translation of decision rule 4 can lead to incorrect answers. In the same line, the protocol analysis study presented here points out that the gains caused by the interventions were due to a large extent to the acquisition of decision rules, but not to a deep transformation in the way students understood probability. The analysis of students written explanations presented in the next section will come back to this point.

4.7 EXPLANATIONS OF STUDENTS TO QUESTIONS IN THE MEASUREMENT SYSTEM

Written explanations to items were collected for the whole sample of students. Many students answered with short answers or they did not provide any explanation to their answers. The analysis of the explanations give further insight on the reasoning process underlying students' answers, and permit to evaluate the actual effects of the interventions on statistical thinking.

4.7.1 Explanations: Data Analysis Items

In this section, two types of items are considered. First, items in the first section of the pre, posttest requiring students to write some explanation, and, second, open-ended questions evaluating data analysis aspects. Specifically, the explanations included in this analysis are the explanations to items Data1, Data2 and Data4, the answer to the first open-ended question, that asked students to explain the factors that influenced the significance of a distribution difference, and the answer to the second open-ended question that asked students to define between and within groups variability.

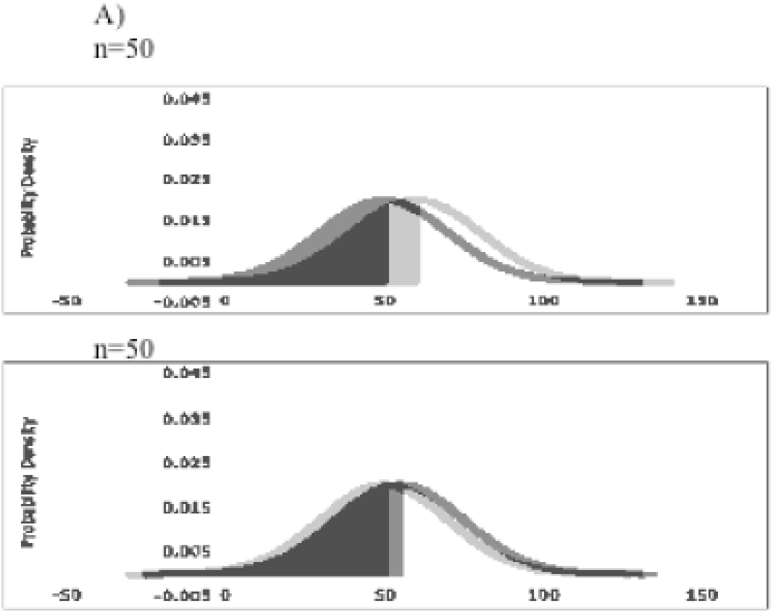


Figure 10. Data1: Comparing Distributions' Pairs with Equal Spread

For Data1, Data2 and Data4, the explanations were divided into explanations related to the center, related to the spread, related simultaneously to center and spread, tautological explanations and other. Tautological explanations are explanations based on self-evident definitions (e.g. *"because the difference is more significant"*). The "other" category was created for other types of explanations not included in the above mentioned categories. For item Data1, inter-coder reliability was 86% (Kappa=.76) in the pretest, and 82% (Kappa=.72)in the posttest. For Data2, inter-coder reliability was 76% (Kappa=.62) for the pretest, and 78% (Kappa=.64) for the posttest. For Data4, inter-coder reliability was 80% (Kappa=.66) in the pretest, and 84% (Kappa=.69) in the posttest.

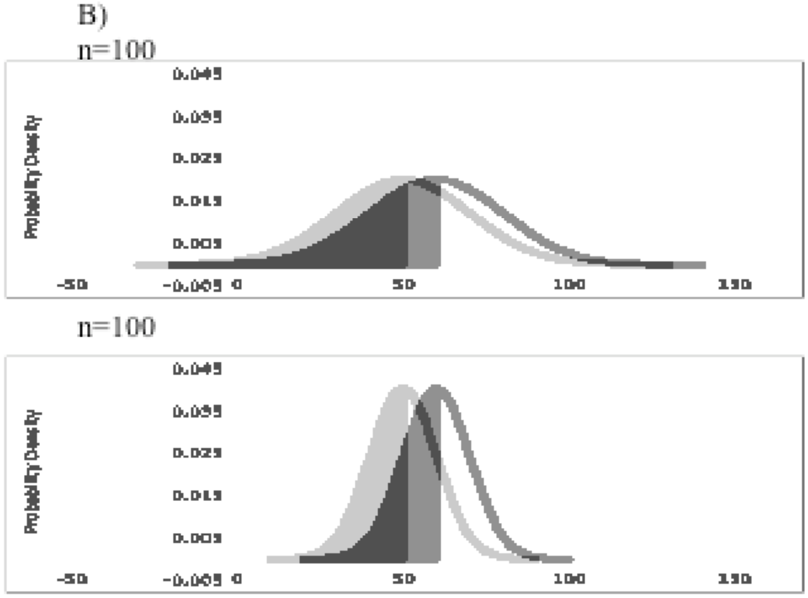


Figure 11. Data2: Comparing Distributions' Pairs with Different Spread

Results show no evident change from pretest to posttest in the distribution of categories (Table 22). Explanations based on center were central for Data1 and explanation on spread were central for Data2. This results makes sense since in Data1 the only difference between the two pairs of distributions was the central value; and the only difference between the two pairs of distributions in Data2 was the distributions spread.

Table 22. Classification of Students to Written Explanations in Percentages*(n*=84)

Items	Data1		Data2		Data4	
Explanation	Pre (%)	Post (%)	Pre (%)	Post (%)	Pre (%)	Post (%)
Center	39.3	42.9	7.1	7.1	16.7	13.1
Spread	6.0	3.6	31.0	33.3	2.4	3.6
Center and Spread	14.3	11.9	10.7	6.0	45.2	48.8
Tautological and Other	17.9	11.9	16.7	15.5	9.5	9.5
Missing	22.6	29.8	34.5	38.1	26.2	25

For Data4, the question was more interesting: participants were asked to describe the difference between two distributions in the context of a case. The distributions varied in their central values and in the variability surrounding it. Most students evaluated the difference in terms of central values, but they mentioned spread in their evaluations (pre=45%,post 49%). Taken as a whole, results show that participants had even in the pretest awareness of central values and variability as parameters in the comparison of distribution pairs. However, the use of spread was ambiguous. Students were able to identify spread but they did not know to use it. In many cases, students used spread in inadequate ways. For example, several participants assert that *"the difference between the distributions is larger because the spread is larger" (73).* Another interesting phenomenon was the confusion shown by some students between the curve displaying the data distribution and a sampling distribution curve. Some students tried to extract directly the significance of the difference from the curve:

> *"I think that in the two levels the difference is equal because the levels of significance are equal (0.02))".*

> *"Me parece que los dos niveles de diferencia igual ya que los niveles de significancia son iguales (0.02)".*

For the open-ended questions, results showed that students were able to differentiate the concepts of between and within variability, but they did not use that concept when determining significance. In the second open-ended question, 61% in the pretest (24 missing) and 60% in posttest (23 missing) of the students defined correctly within and between variability. However, when asked to mention the factors that determine the significance of mean differences in the first open-ended question, students mentioned center differences, that is, between groups variability as the main factor (Table 23). The use of spread in this type of comparisons was limited, which to some extend explains why Data2 and Data4 (that required evaluating spread in the comparison) had the lowest answer rates among the data analysis items.

Table 23. Classification of Written Answers to the Open-ended Question 1

	Pre (%)	Post (%)
Center	35	32
Spread	7	7
Center and Spread	17	19
Causal	5	6
Tautological and Other	5	5
Missing	32	31

4.7.2 Explanations: Sampling Items

The analysis presented in this section was elaborated based on students explanations to one item in the pre, posttest questionnaire, and to three open-ended questions devoted to sampling. The item from the pretest is Sam1 that asked students to evaluate the significance of mean differences given different samples sizes. The open-ended questions asked whether or not and why two samples obtained from the same population have the same mean (3); what factors explained the mean differences between two groups of people in a hypothetical case (4); and what effects sample size had in the sample mean (5).

The analysis of open-ended question number 3 shows that students had intuitive notions of the sampling process (Table 24). They understand that sample parameters varied within certain range, and that sample means from the same population are different but similar. When asked whether or not two samples extracted from the same population would have the same central values, most students answered that the samples should be different but similar (44% pretest; 52% posttest).

Table 24. Written Answers to Open-ended Question 3

	Pre (%)	Post (%)
Equal	6	4.8
Different but similar	44	52.4
Different	20	15.5
Missing	29	27.4

However, the analysis of Sam1 and open-ended question number 5 shows that students cannot specify the particular effects of sample size on the sampling process (Table 25). Additionally, the analysis of Sam1 in the pre, posttest, and open-ended question number 5 shows that the participation in the interventions produced positive changes on students answers (Table 25). In explanations for Sam1 and open-ended question number 5, students went from stating that mean differences had the same significance independent of their sample sizes to associate large sample sizes with larger significances of mean differences.

Table 25. Written Explanations for Sam1 and Open-ended Question 5 for participants in the Sampling Condition

	Sam1		Open-ended Question 5	
	Pre (%)	Post (%)	Pre (%)	Post (%)
Equal	19	8.3	21.4	7.1
Different because larger samples produce higher significance	9.5	25	10.7	17.9
Different because larger sample sizes produce lower significance	3.6	6.0	6	3.6
Missing	13.1	15.5	16.7	16.7

This effect was present in both interventions and was larger for the sampling intervention. The gains in the data analysis condition were not predicted by the theoretical review. An alternative hypothesis is that gains in declarative knowledge have a lot to do with the observed change; that is why, both the data analysis and the sampling intervention, produced important change in students never exposed to probabilistic training. The analysis of open-ended question number 4 that asked participants to explain the factors associated to the significance of sample mean differences is consistent with this hypothesis. It shows that there seems not to be a deep change in students understanding of probability (Table 26).

Table 26. Written Answer to Open-ended Question 4 (n=84)

Explanation	Pre (%)	Post (%)
Causal Factors	66.5	48.8
Random Sampling	2.6	7.7
Causal Factors + Random Sampling	2.6	5.2
Missing	28.2	38.5

In open-ended question 4, participants used mainly causal factors to explain the mean differences between two groups in a hypothetical situation. There was no important difference between both pretest and posttest. Students did not incorporate what they learned about sample size and significance to evaluative situations requiring explanation of the reasons producing mean differences. In other words, students were able to assert that larger samples produced larger significance, but they did not connect this knowledge with the way they explained observed phenomena. In spite of the pervasiveness of causal explanations, probabilistic misconceptions decreased from pretest to posttest. For example, student 34 answered in the pretest that sample size did not affect the significance of mean differences "*because the number of people is not what is being measured" (No porque no se esta midiendo el numero de personas);* this conception was modified in the posttest where the same student gave a better, yet simplistic answer: *"the more sample is obtained, the more significant are the results" ("entre mas muestra se tenga mas significantes son los resultados").* In another example of the way change happened during the interventions, student 48 answered in the posttest that two samples coming from the same population should be different because *"they are not the same students and differences can be found in the obtained results" ("porque no son los mismos estudiantes y pueden presentarse diferencias en los resultados obtenidos").* The same student had written in the pretest "*if is the* *same variable, they (the means) should be the same because it's the same data ("si es de la misma variable deberian ser la misma porque son los mismos datos").* Finally, the change can be observed in student 79 that asserted in the pretest that "*in a large sample, there are larger possibilities of finding atypical data that alter the mean" ("en una muestra grande hay mas posibilidades de encontrar datos atipicos que alteren la media"),* and stated in the posttest that *"the larger the sample, the less the variation among samples and the more trustworthy the result" ("entre mas grande la muestra menos varian las muestras entre si y mas confiable el resultado").*

5.0 CONCLUSIONS

The results presented in the previous chapter portray a complex picture regarding statistical reasoning and the effects of computer-based interventions on this process. This picture includes quantitative evidence that indicates that computer-based interventions improve student's ability to think about probability and repair some important probabilistic misconceptions. This sort of evidence shows also that the gains in probabilistic knowledge are stronger for students exposed to simulations than for students exposed to data analysis exercises and worked-out examples. These results are consistent with predictions made in the literature review that indicated that the use of simulations improves the representation of random process and therefore it increases the levels of statistical reasoning. However, qualitative analysis conducted on open-ended explanations and protocols of students solving statistical problems show that no deep change is occurring in the students' representation of random processes. While participants seem to incorporate the notion that larger samples increase significance, they do not seem to modify the conceptions of statistical explanation that underlie the coordination of theory and evidence in the statistical process. This lack of deep effect can be explained by the short time of the interventions and by the absence of additional mechanisms supporting learning such as feedback systems, natural language tutoring systems, and blackboards. In a more complete version of online instruction, these mechanisms would be available to support students' process of learning.

5.1 SUMMARY

Interventions reduce the prevalence of statistical misconceptions such as the equiprobability bias, and the law of large numbers, but the effects on data analysis knowledge are very soft. Accounting for these results requires considering different factors contributing to this outcome. In the first place, participants knew more about data analysis than about probability before the intervention as evidenced by pretest scores. At the moment of the intervention, students in the three groups had undergone intensive training in data analysis as part of the basic sequence of the statistics courses. In part, this training was a necessary prior to ANOVA and other mean comparison procedures; at some other level, this training was a personal and pedagogical choice of instructors that consider data analysis a central part of social sciences' use of statistics. The intensive training fostered in students skills for data representation and interpretation, as shown by the evolution of students in the political science course. The sub-study in the political science course showed that shortening the distance between representation and interpretation implied data-based dialog and substantive feedback on students' answers. This study showed also that connecting different types of representations in order to correct representational specificity implied pedagogical support beyond the use of statistical packages.

These results considered altogether suggest that the data analysis intervention was less effective perhaps because students had developed their data analysis skills to a considerable level prior to the intervention. Additionally, the same nature of the data analysis intervention contributed to the lack of effect on students' data analysis knowledge. As shown by the comparison of the interventions made in the methods section, the data analysis intervention had lower levels of engagement than the sampling intervention; authentic data analysis is time consuming and, even with computers, it requires students to spend considerable time and resources on one task in order to produce an interpretable result. Low levels of engagement are not a characteristic of online instruction, but a defect in this particular design. In the context of this study, this constraint led to the presentation of content through worked-out examples instead of through student exercises. This difference between interventions combined with the high base line for data analysis knowledge affected the magnitude of students gains in data analysis skill.

A complementary issue explored in this work was the relationship between inference, data analysis, and sampling knowledge as constituents of statistical thinking. The theoretical framework suggested that students' conducting statistical inference combined data analysis and sampling knowledge to produce plausible interpretations in statistical cases. This point was supported by the results of this study showing that change in inference knowledge is predicted significantly by change in sampling and data analysis, and by the participation in the sampling intervention.

5.2 CHALLENGES FOR ONLINE STATISTICAL INSTRUCTION

This study has shown that computer-based interventions can be to some extent effective to generate changes on students' reasoning. However, this effect is constrained to areas where students possess low knowledge (e.g., sampling knowledge in the pretest). In many regards, the results of this study suggest that the efficiency of computer-based tools needs to be mediated by pedagogical practices either human or computer-supported. Emphasis on the necessity of adequate scaffolding of statistical argumentation is made here. Scaffolding can be provided in the form of blended teaching or mediated through different kinds of computer tools from automatic feedback based on task analysis to participation in discussion boards. Learning statistics means coordinating theory, data and representation in the context of random effects and variability. While representation of data and randomness can be facilitated through computers, understanding probability and commanding statistical argumentation do not grow automatically from the use of data analysis statistical packages. Computers provide several tools that can help this process; feedback systems and discussion boards are some of them.

5.3 RESEARCH QUESTIONS

At this point is convenient to go back to the research questions proposed in the methods section. Research question 1 inquired whether or not it was possible to build a measurement system able to capture the core ideas of data analysis and sampling in a reliable and valid way. The results of this study show that such system is possible, and that it can capture students gains in statistical knowledge. The posttest results show that the coordination of distribution graphs, test results and conclusions in context that measured inferential knowledge in the items designed by the researcher correlated to some extent with standardized measures of inference and data analysis knowledge, but they did not correlate with sampling measures. Additionally, the results of this study show that the coordination of data analysis and probability is related to students performance in inferential statistics. Specifically, gains in data analysis and sampling knowledge predicted students gains in inference knowledge.

Research question 2 inquired about the real potential of online education to teach data analysis, sampling and inference knowledge in short time frames. The answer provided by this study to this question is that computer-based simulations favor the learning of probability and formal aspects of statistics. On the other side of the question, the results of this study show that a limited number of computer-based data analysis exercises does not produce gains in data analysis knowledge when students have been exposed previously to activities of representation and comparison of data sets. Additionally, the sub study of the evolution of the political science course shows that data analysis knowledge grows over time.

Research question 3 reviewed the possible tradeoffs between data analysis and simulations as tools to teach statistical inference. Regarding research question 3, this study shows that there is an important tradeoff between the authenticity of data analysis exercises, and the sustained activity of sampling simulations in the teaching of statistical content, particularly, of probability. Simulations are more effective to teach probability than data analysis exercises. Additionally, as shown by the codification of the interventions, the number of sampling questions in the data analysis intervention is lower than the number of sampling questions in the sampling intervention. A possible explanation for this fact is that conducting data analysis requires more instructional time than using simulations. This tradeoff makes simulations a better tool to teach probability.

5.4 IN HINDSIGHT

The results of this study present a promising line of research. However, it is necessary at this point to indicate important flaws in this design that need to be avoided in future research. The first element that needs to be considered is the no effect observed in data analysis knowledge. The data analysis intervention presented few exercises, and some features of the OLI course were excluded in order to increase the comparability of the two conditions of this study. These features such as learning checks and videos could change the outcome of comparisons similar to those presented in this study. Additionally, the time of instruction used here is below the standard time of a naturalistic experiment. Several weeks of instruction may be necessary to observe change in data analysis activity.

More important, the design of the study created a systematic bias in favor of the learning of sampling knowledge. Sampling knowledge was better represented in this study than data analysis and inference knowledge. Understanding sampling size effects and observing simulations of sampling distributions are tasks that are close to the core of probability. Comparisons of center and spread among distributions are just a marginal subset of what data analysis is. The items evaluating data analysis in the pre, posttest questionnaire focused on tasks that are secondary to data analysis. The same can be said about the gains in inference knowledge. ANOVA and understanding of ANOVA tables are just a minimal part of what statistical inference is. In further studies, it is necessary to correct this flaw using data analysis tasks that capture the complexity of the statistics.

5.5 LIMITATIONS AND FURTHER DEVELOPMENTS

This dissertation suggests a research agenda on statistical learning and educational technology. This agenda includes a randomized study comparing two perspectives in statistical instruction; a protocol analysis study on the reasoning process underlying sampling and data analysis exercises; and a classroom study on the instructional process of statistical data analysis from the beginning of the class to the intervention point. This agenda is a work in progress.

Regarding the randomized study, the first point to note is that the comparison between the sampling and the data analysis perspectives is not transparent. In order to produce an ecologically-valid comparison, this study mixed several factors in what constituted two opposed perspectives. In the quest for ecological validity, several facets were varied across the lines of the interventions. For example, the sampling intervention contained more sampling exercises than the data analysis intervention. This variation seemed natural because of the emphasis on sampling in the sampling perspective, but it introduced confounding effects in the design. The comparison here proposed contrasted two broad views of statistical instruction. However, fine-grained studies that provide analytical evidence regarding two aspects of the interventions are necessary: first, it is necessary to prove that higher engagement produce higher levels of learning in all domains and not just on the sampling dimension of statistical thinking; and second, it is necessary to test if worked-out examples are effective in learners at initial stages of data analysis training. The studies required to fill this gaps are experimental in nature and entail controlling all aspects of a particular comparison even if that implies sacrificing ecological validity.

The think aloud protocol analysis study needs a larger sample and maybe one or two additional studies. One additional study on the instructional conditions that foster complex and larger sequences of description and comparison in the solution of data analysis items. This study needs to delve into the specific characteristics of those sequences, connecting in one hand the statistical reasoning research and in the other the literature about self-explanations. In addition, this complementary study needs to identify the factors producing long and deep exploratory behavior, from the presence of certain declarative statements in instruction to differences in the disciplinary epistemologies held by students. Another protocol analysis study is necessary regarding the particularities of the knowledge gains in the sampling condition; particularly it is necessary a detailed description of the effects that a working model of probability has on students answering complex probability problems. In other words, it is necessary to identify what advantages a working model of probability has over a declarative representation of the effects of sample size. Protocols of some students in this dissertation permit to anticipate that a mental representation of sampling allows students to apply probability to a widest range of situations, including for example item Sam5 where students had to revert the decision rule to produce a correct answer. In the decision rule taught in the interventions, larger sample size increased significance; in item sam5, larger sample sizes decreased the probability of finding extreme values. This reversion was difficult for some students holding just declarative knowledge.

Regarding the findings of the classroom study, experimental evidence in a controlled setting is necessary to confirm the existence of representational specificity and to evaluate the actual effects of the distance between interpretation and representation in students' learning. For example, representational specificity could be tested using a design in which students with some knowledge about either histograms or box plots would be asked to pair graphs of both types representing the same and dissimilar data. Finally, a large controlled classroom study is required to capture in greater detail the teacher's instructional moves and the changes in classroom discourse. This study must include sessions video recording and open-ended interviews with students at different points of the course.

APPENDIX A

ADDENDUM TO AMERICAN STATISTICIAN ARTICLE (08/2005)

June 6, 2006

Javier Corredor and Gaea Leinhardt

This report follows up on the article that appeared in the American Statistician (Larreamendy-Joerns, Leinhardt, Corredor, 2005) in which we compared six on-line statistics courses. In this report we add to that two additional courses: the OLI statistics course and ActivStats. We add OLI because it is the focus of our work and we add ActivStats because although it is a CD and not an online course it is one that is highly used and deeply respected by the statistics community itself.

To review briefly, in the American Statistician article we engaged in the following activities: we compared the overall content and approach of the courses; we focused on the types, frequency, and breadth of instructional examples; we compared the types of exercises students were asked to complete; and finally we estimated the cognitive complexity or demands of the exercises. In this report we add two additional courses to the original coding and discuss the results.

Content Analysis

To compare the courses in terms of the content they presented, we built a list of the topics covered by each course. This list of topics was initially conceived as a mechanism to compare the content of the courses with the AP Central Statistics Exam content. We started by listing the topics that were covered by the AP exams as a base line. Next, we looked for the topics covered by instruction in the online courses. Appendix A shows the details of the results of this analysis. OLI and ActivStats covered 70% of the AP content. Cyberstat covered the 88%; and Seeing statistics covered the 51% of the AP contents. The online courses also covered topics that were not part of the AP list, but, as it is shown in Appendix A, those topics were few in all the courses. The exception was Cyberstats which is by far the most comprehensive course.

The topic list (Appendix A) presents the subjects that are included in the four online courses that appeared most complete: Seeing Statistics and Cyber stat from the AS article, and OLI and ActivStats in this report. All courses cover more or less the same topics. However, some courses are larger than others and that makes a difference in the depth in which topics are taught. Cyberstats is an especially large course. This feature has the advantage of permitting a deeper exploration of the topics. In some sense a larger course is more akin to provide more and more authentic examples and exercises, and to give more information in general.

From the list of topics, we noticed that the four courses organized their content in three sub areas: Examining data, Probability, and Inference. Examining data refers to methods that permit one to find, summarize and represent patterns in data. Probability refers to content devoted specifically to the laws of chance, and to the characteristics of sampling distributions. And Inference refers to procedures of hypothesis testing that contrast the predictions of probability theory with the patterns in actual data. All the four courses make explicit distinctions between these three content sub areas when they presented the overview of the course in the introduction. Additionally, all the courses contained content on “producing data’, that is, on methods of collection of data, and on theory in the sources of data and their characteristics. However, this last area of content was less large and presented in a more implicit way. The distinction between the four sub areas was present in the AP content and appears to be a standard distinction within statistics.

Courses Description and Resources

OLI. The OLI statistics course is one of several courses included in the Online Learning Initiative authored by CMU faculty and available for free in the Internet. The course has been designed with attention to cognitive principles of learning (“Cognitively-informed education”) that encourage the active use of declarative knowledge, and the provision of timed feedback. The design of the course also avoids instructional situations that create high working memory loads. This course presents content through different means: text, video, interactive exercises (that provide feedback on student answers). Additionally, it presents an Intelligent tutoring system (Stat Tutor). The course checks student learning constantly with “did I get this?” questions and it provides “learning by doing” that presents complex situations to students to work on. Additionally, the course includes a section called “many students wonder” that gathers questions provided by students using the course. The goals of each unit are presented and they are easily accessible by clicking in an icon in the navigation bar.

The course content follows a structure that attends to what they called “the big picture” of statistics. The big picture of statistics refers to the process of “converting data in useful information”. This process includes three basic steps: collecting data, summarizing data, and interpreting data. At the end of each unit, the OLI course provides a section with resources that include data sets, and examples, but the examples are not enough to compensate for the small number of examples provided in the content.

In terms of resources (Table 1), the OLI course is very complete. It is at the same level of the top online course available today (CyberStat, ActivStats), and it is far better than most of the other courses including SeeingStatistics. The OLI course, however, lacks of online active interactive and simulations, and rather it relies on links to external applets for that purpose. Some simulations of random behavior are conducted using the random generators functions from statistical packages, but this strategy can create problems for students not familiar with programming, and is not nearly as natural or easy as having such simulations built in and easily available.

Table 1. Available Online Resources per Course

Resources	CyberStats	SeeingStat	OLI	ActivStats
Applets	✔	✔	✔	✔
Videos			✔	✔
Statistical Software	✔		✔	✔
Virtaul labs				✔
Note-taking facilities	✔		✔	
Course Map	✔	✔	✔	✔
Glossary	✔	✔		✔
Search engine	✔	✔		
Course management system	✔		✔	✔
Links to external sources	✔		✔	
Electronic forums	✔			
Multiple-choice questions	✔		✔	✔
Short answer questions	✔	✔	✔	✔
Feedback	✔	✔	✔	✔

ActivStats. ActivStats is an online course developed by Paul Velleman a highly respected statistician and statistics educator. This course is not available for free in the Internet. It is released in a CD that can be bought online for about $50. The course is very complete, it provides a course management system that keeps track of students' actions and helps them to follow the course sequence. ActivStats has the characteristic of presenting an example in each unit –usually in the form of a video- and developing that example carefully through the unit. One strength of this course is that it provides active simulations that illustrate important statistical ideas. Additionally, the course provides drag and drop quizzes, and a system of data management that contains data sets and acts as a statistical package. In this data management system, the students can conduct authentic research by following the instructions for each exercise that appeared in boxes when the students are using the system.

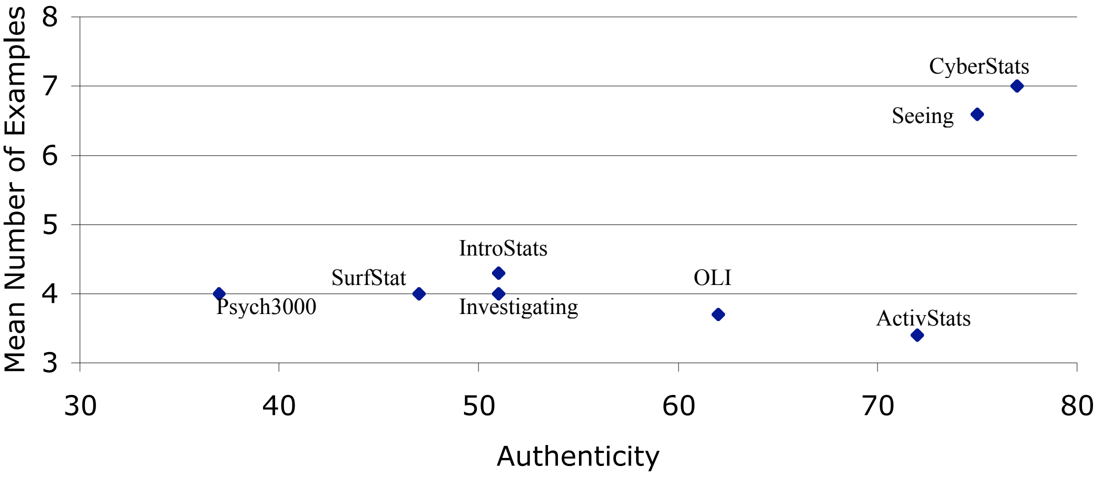


Figure 1. Authenticity vs Mean number of examples

Examples

Examples are a key element in any mathematical explanation (Leinhardt, 2002; Renkl, 1997; Zhu & Simon, 1987). Examples are also a critical mechanism for learning as shown by Simon and his colleagues. But examples must be carefully selected to show the range of conditions of use, the subtleties of differing circumstances, and authenticity of the domains core questions (Risland 1991). In addition, examples must be unpacked or explicated in a variety of levels of depth to assure that students engage with them (Zhu and Simon, 1987). To code examples we selected three target topics in introductory statistics (Central Tendency, ANOVA, and Regression) and identified all of the material that constituted an example (for details see Larreamendy-Joerns et al, 2005). For each example we then categorized a number of features. Figure 1 shows the new results for the additional courses.

With respect to frequency CyberStat and Seeing have the most in terms of number and variety of examples while the remaining six courses all cluster around 3.5 or 4 examples per unit. OLI and ActivStats both have fewer than these two leaders, however, they do unpack their examples in greater depth. Frequency is important not only because of issues of range but because it appears that the actual cognitive arrangement of concepts is expanded if there is greater frequency. With respect to issues of authenticity OLI and ActivStats are comparable and in the upper half of the reviewed courses. The authenticity index was developed by scoring each example as to whether it had no cover story, a cryptic cover story, or an authentic cover story. Authenticity reflects the uses of statistics and the field in and of itself. The majority of examples in OLI and ActivStats had cover stories and some of the stories were developed in considerable depth. There is an inherent trade off between depth and frequency when the total length is fixed. The number of examples in Seeing is very high because the course is designed to allow for additional examples on demand that are suitable for different content areas (economics, psychology, sociology).

Figure 2 shows the quality of examples in terms of the variability. There are two basic indicators of variability. One is the mean number of different topics covered by the examples within the different units evaluated. The other is the mean number of different covers that are included to explain each topic. In both indicators ActivStats and the OLI appear low. One explanation for this is that both courses present few examples but they organize a large set of activities around them. In the case of ActivStats, it is true that there is an important set of random simulation generators that produce data each time that the user requests it. For the case of OLI, a factor that could affect this result is that the structure of the course divides areas the

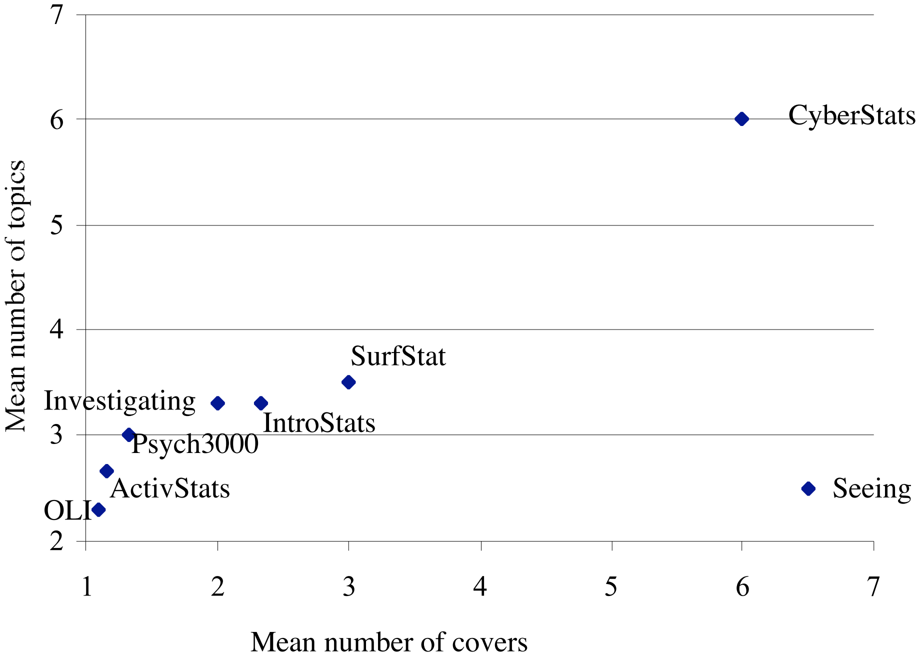


Figure 2: Mean number of different example covers and conceptually distinct examples per unit and course.

content in a way that is not typical of the content distribution of most courses. OLI course have big units of content that represent the structure of statistics (e.g. examining data), and subdivide those units in specific procedures (e.g. central tendency measures). Most other courses use specific procedures as thematic units. How does it affect the counting of examples? Being the studied units in the case of OLI subordinates to other larger thematic units make this units smaller, and thus, the number of examples and their variability naturally lower.

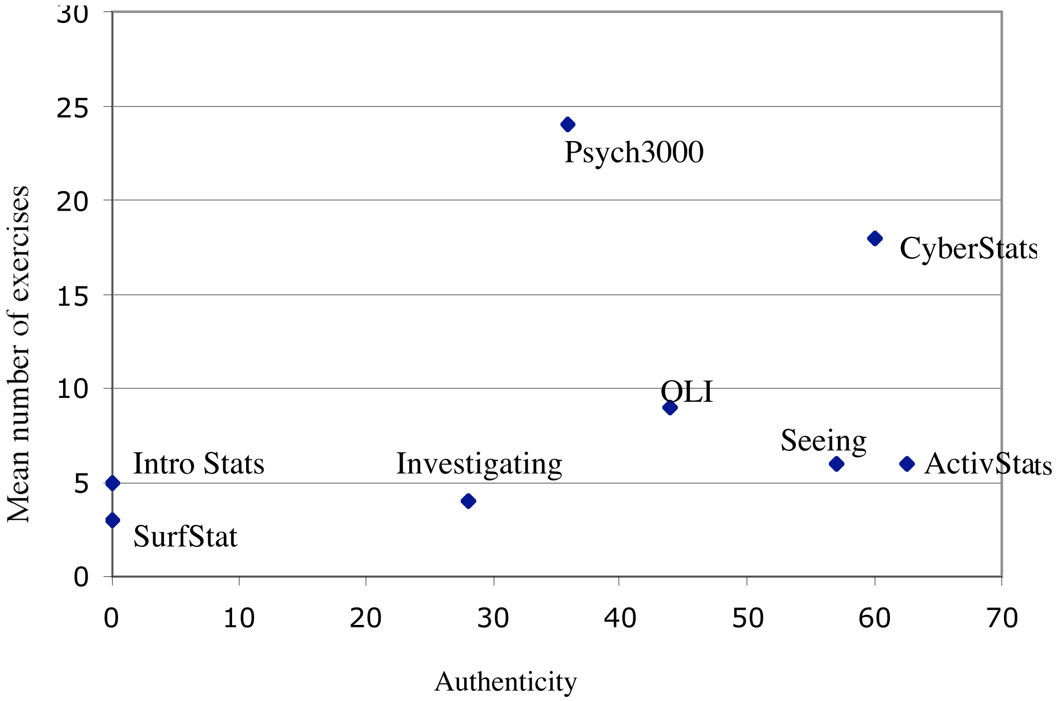


Figure 3. Mean number of Exercises and Authenticity

Exercises

Figure 3 shows a comparable analysis of the courses for exercises. In earlier evaluations of statistics texts exercises along were the main criteria for evaluation of material because they are so central to the learning process in quantitative domains. Here we compared the frequency and authenticity of the exercises in which students were expected to engage. Other than CyberStat, OLI has a reasonable authenticity index and a reasonable frequency of exercises for the students. In the case of ActivStats, the presence of random simulation generators in the exercises can compensate the low number of exercises. Random simulations produce multiple data sets on which the students can work. It is important to note that a tradeoff associated with the use of random generators is the lack of authenticity inherent to them. While using sets of real data creates both a natural data structure and a credible cover story, using random generators creates a large of set of cases, but the data structure is fixed (although it is possible to use probability functions to build them), and the cover story usually false and poor. The authenticity score for ActivStats was lowered by the fact that this course contains a large number of this type of exercises.

In the case of OLI, the authenticity score for the exercises can be explained in the following way. The OLI contains two types of exercises: Learning by doing exercises and checking knowledge exercises (“Did I get this?”). The first type of exercises has a high level of authenticity because it refers to complete situations where learners are confronted with authentic problems. The second type of exercises (“Did I get this’) has a low level of authenticity because the purpose of the exercise is not to face students with a complex situation, but to evaluate how they had learned basic chunks of information.

In terms of the cognitive demand of the exercises, both OLI and ActivStats show a reasonable level of demand. In the case of OLI more than 75% of the exercises demand from students the complex use of procedures. In the case of ActivStats complex use of procedures was also the more common category. This was also the case of CyberStats and Seeing. What this means is that the students face complex tasks and are required to use the knowledge they have in a flexible way. Students have to coordinate sequences of goals and subgoals, and divide the task into different steps. It means also that they work through ill-defined tasks in which they define the framework of the task. Another characteristic of these courses is that they have a variety of cognitive demand levels.

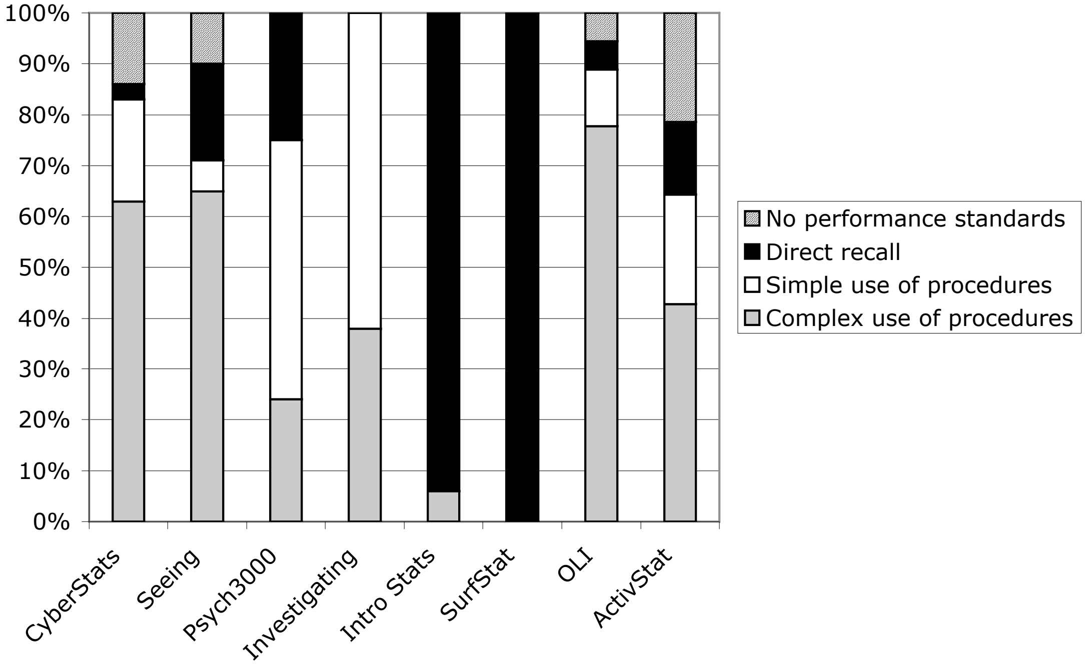


Figure 4. Cognitive Demand of Exercises.

Conclusions and Possible contrast for OLI

From this report we can conclude that OLI and ActivStats are solid courses and seem to be among the upper set of courses available. They present students with authentic examples and exercises. They present students with exercises that have high cognitive demand. The resources available in the courses are as good as or better than the resources provided by other online courses. The main weakness of both courses is the low number of examples. In part, this is due to the fact that these courses are relatively short. A solution that could work in this case is to create pop-up windows that present additional examples at the request of the learner.

A characteristic that negatively differentiates OLI from all the other courses is the lack of interactive simulations. The simulations provided by OLI must be created using the random number generator functions of statistical packages. Interactive simulations in the form of Applets have advantages and disadvantages. A disadvantage is that they give the students quick visualization of statistical ideas. A disadvantage is that students do not engage in modeling tasks by themselves. These modeling tasks can be a core for the understanding of statistical ideas.

The results of this study suggest some points of contrast for research. presence or absence of simulations; increase or decrease in quantity and range of statistical examples; increase or decrease in range of covers and breadth of examples . These three features suggest contrasts between OLI, ActivStats, and Cyberstats might be informative.

REFERENCES

Larreamendy-Joerns, J., Leinhardt, G. & Corredor, J. (2005). Six Online Courses: Examination and Review. *The American Statistician*. 59 (3). 240-251.

Leinhardt, G. (2001). Instructional explanations: A commonplace for teaching and location for contrast. In V. Richardson (Ed.), *Handbook of research on teaching* (4 th Ed., pp. 333-357). Washington, DC: American Educational Research Association.

Renkl , A. (1997). Learning from worked- examples: A study on individual differences. *Cognitive Science*. 21(1): 1-30.

Risland, E. L. (1991). Example-based reasoning. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds), *Informal Reasoning in Education* (pp. 187-208). Hillsdale, N.J.:LEA.

Zhu , X., & Simon , HA (1987). Learning mathematics from examples and by doing. *Cognition and Instruction*, 4, 137-166.

Appendix A: Topics' List

Content	AP	OLI	ActivStats	CyberStat	Seeing Statistics
The big picture		Introduction			Introduction
EDA		Examining	Examining	Examining	Examining
Dotplot	Examining		Examining		Examining
Pie Chart		Examining		Examining	

Bar-chart		Examining	Examining	Examining	
Histogram	Examining	Examining	Examining	Examining	Examining
Simple plot					Examining
Stem-plot	Examining	Examining	Examining	Examining	
Cumulative Frequency plot	Examining				
Measures Center	Examining	Examining	Examining	Examining	Examining
Mean	Examining	Examining	Examining	Examining	Examining
Median	Examining	Examining	Examining	Examining	Examining
Typical values					Examining
Measures of Spread	Examining	Examining	Examining	Examining	Examining
Range	Examining	Examining	Examining	Examining	Examining
Median absolute deviation					Examining
Inter-quartile range	Examining	Examining	Examining	Examining	
Standard Deviation	Examining	Examining	Examining	Examining	Examining
Box Plot	Examining	Examining	Examining	Examining	Examining
Independent Variable		Examining	Examining	Examining	
Dependent Variable		Examining	Examining	Examining	
Context of data			Examining		
Backtobackstem Plots	Examining	Examining		Examining	
Comparing boxplots		Examining	Examining	Examining	
Frequency tables bar charts	Examining				
Marginal and joint in freq.tables	Examining				
Conditional relative frequencies	Examining				
Scatterplot	Examining	Examining	Examining	Examining	Examining
Linear transformations	Examining		Examining	Examining	Examining
Linear Relationships		Examining	Examining	Examining	Examining
Correlation Coefficient-r		Examining	Examining	Examining	Examining
RegressionI	Examining	Examining	Examining	Examining	Examining
Causation		Examining		Examining	
Sampling	Producing	Producing	Producing	Producing	
Study design		Producing		Producing	
Randomization	Producing	Producing		Producing	Producing
Methods of data collection	Producing			Producing	Producing
Census	Producing				

Experiment	Producing		Producing	Producing	
Observational study	Producing			Producing	
Survey design	Producing	Producing	Producing	Producing	
Control group	Producing			Producing	
Block designs	Producing			Producing	
Simulation	Others				
randomness			Probability	Probability	
Probability (definition)	Probability	Probability	Probability	Probability	Probability
Probability (calculation)	Probability	Probability	Probability	Probability	
Relative frequency of events	Probability	Probability		Probability	Probability
Equaly likely events	Probability	probability		Probability	
Probability rules	Probability	Probability	Probability	Probability	
Conditional probability	Probability	Probability	Probability	Probability	
Large numbers	Probability		Probability	Probability	
Bayes Rule		Probability		Probability	
Binomial distribution					Probability
Random Variable	Probability	Probability	Probability	Probability	
Discrete random variable	Probability	Probability	Probability	Probability	
Continuous random variables	Probability	Probability	Probability	Probability	
Normal random variables	Probability	probability	Probability	Probability	Probability
Sampling distribution	Probability	Probability	Probability	Probability	Probability
Central limit theorem	Probability		Probability	Probability	Probability
Sample vs population	Probability	Probability	Probability	Probability	Probability
Sampling distribution proportion	Probability	Probability		Probability	
Sampling distribution mean	Probability	Probability	Probability	Probability	Probability
t-distribution	Probability		Probability	Probability	Probability
Chi square distribution	Probability			Probability	Probability
Confidence			Inference	Inference	Inference
Point Estimation	Inference	Inference	Inference	Inference	

decision making					Inference
Interval Estimation	Inference	Inference	Inference	Inference	
Interval Estimation slope and differences	Inference			Inference	
Hypothesis Testing	Inference	Inference	Inference	Inference	Inference
Sample Size	Inference	Inference	Inference	Inference	
Inference for means	Inference	Inference	Inference	inference	
Inference for proportions	Inference	Inference	Inference	inference	Inference
Ht for proportions	Inference	Inference	Inference	inference	Inference
Ht for mean	Inference	Inference	Inference	Inference	
Z test for one variable	Inference	Inference	Inference	Inference	Inference
t test for one variable	Inference	Inference	Inference	Inference	Inference
Two sample t test (independent)	Inference	Inference	Inference	Inference	Inference
Two sample t tests (paired)	Inference	Inference	Inference	Inference	Inference
ANOVA		Inference	Inference	Inference	Inference
Chi	Inference	Inference	Inference	Inference	Inference
Regression linear		Inference	Inference	Inference	Inference
F distribution			Inference	Inference	Inference
Multiple regression			Inference	Inference	Inference
Time series				Inference	
Treatment				Producing	
tree diagram				Examining	
Binomial distribution				Probability	
Bivariate (data)				Examining	
Blind				Producing	
Cause-and-effect diagram				Inference	
Composite event				Examining	
Compound event				Examining	
Ecological correlation				Inference	
Elliptical point cloud				Examining	
Flowchart				Examining	
Jitter				Other	
Long-run mean				Examining	
Lurking variable	Examining			Examining	
Memoryless property				Examining	
Pareto diagram				Examining	
Placebo				Producing	

QQ-plot				Inference	

OLI 18/61 (70%)
Active stats 18/61 (70%)
CyberStats 7/61 (88%)
Seeing Statistics 31/61 (51%)

.

APPENDIX B

SIMULATIONS

The VESTAC project (Darius et al, 2002) presents a collection of applets devoted to different statistical concepts. VESTAC contains applets devoted to visualize diverse distributions (Univariate and Bivariate normal) and demonstrate the Central Limit Theorem by showing how the mean of samples from “different distributions gradually approaches a normal distribution”. There are also a group of applets devoted to show the relation between population and sample for the particular cases of regression and ANOVA: that is, to show how the values of the statistics (e.g. regression slope) vary within certain limits when repeated sampling from a population is conducted.

Similar projects have presented by Nicholson et al (2000), Cramer & Neslehova (2003), and by Harner & Hengi Xue (2003). Nicholson et al (2000) present a system that creates samples based on the selection of parameters made by the learners. Neslehova and Cramer (2003) present a collection of applets that integrates a textbook type of course with active simulations (EMILeA-stat) and other tools for teaching statistics; and Harner and Hengi Xue (2003) (http://www.8-mobius.com) present a java-based environment (myJavaStat) to teach statistics, that includes, among other modules, a section for simulation of probabilistic processes. The design of myJavaStat's module for probabilistic simulations is more sophisticated than a stand-alone applet. It permits through five steps to perform some built-up modeling of the probabilistic process. This feature permits students to make choices about the structure of the task. First, students have to decide what kind of distribution they are dealing with (e.g. normal, binomial, etc). Second the students have to select a sampling method and a sample size. Third, the statistics of interest and the event are introduced in the system. After that, the simulation is conducted and presented graphically and analytically to the students. This type of program permits to simulate complex processes and conducting pseudo-experiments. This approach is an interesting alternative that has the functionality of stand-alone simulations but that has the flexibility and open task structure of programmed simulations.

Wood (2005) and Drier (2000) describe simulations that add a new characteristic: they not only permit learners to operate in abstract distributions, but they reproduce micro worlds where the sampling happens. In this case, students are not operating on the properties of abstract distributions, but on the structure of objects in the virtual environment. Wood (2005) presents a program that samples "balls" from hypothetical "buckets" to illustrate how typical values distribute within confidence intervals when you aggregate multiple samples. Drier (2000) describes a similar program: the probability explorer. The probability explorer is a computer program that permits students to simulate random events (e.g. tossing coins) by operating in virtual icons in a micro-world. The program provides also several graphical displays of the virtual situation that permit students to access several representations of the events (e.g. from cumulative frequencies to bar graphs). Students can decide many facets of the experiment, as for example, the elements to be sampled, the sample size and the outcomes' probabilities.

Shaughnessy and Ciancetta (2002) present a type of simulation that is even more authentic. They simulate the behavior of two “fair spinners” to help students understand how the probability of an event is calculated. This type of simulation is an alternative to teaching “the laws of probability” in a formal manner. It permits to exemplify how probability works without proving it mathematically. Shaughnessy & Ciancetta (2002) report also that working with this simulation improves students’ probabilistic reasoning as measured by some NAEP items. It is important to note, that the fair spinner is a virtual version of a question in the NAEP exam and then it shares many superficial features with the questions used to assess the effectiveness of the simulation.

Another approach to present simulations in an authentic situation is the S.A.M.P.L.E.R – Statistics As Multi-Participant Learning-Environment-(Wilensky & Stroup, 1999). S.A.M.P.L.E.R is a participatory simulation that gathers information from several terminals that permit students to participate as individual agents in a complex phenomenon. The terminals transmit the individual decisions of each student to the S.A.M.P.L.E.R server that, according to some parameters controlled by the instructor, creates the population attributes. Then, the program permits the students to take samples from different parts of this population.

Pedagogical experiences using random simulations. DelMas, Garfield and Chance (1999) present a class intervention that uses a sampling distribution program. This program permits students to explore the process of sampling by using an interactive tool that allows them to select the sample size and “the shape of a population” from which samples are going to be obtained. The program conducts repeated sampling and then creates diverse representations of the simulation results (e.g. Histograms). DelMas, Garfield and Chance (1999) not only present the program but they explain how the simulations are used in a pedagogical model to teach statistical concepts. The pedagogical model asks the students to solve problems using the simulation program. Students explore questions like “What is the relationship between sample size and the spread of the sampling distribution?”. After working with the simulations, the students are evaluated. The authors found poorer results when the intervention activities and the assessment task were different than when activities were similar to the assessment items. This result shows that transfer between tasks with different superficial features is difficult for students in statistics.

Sanchez (2002) emphasizes the pedagogical approach than to the tool. He describes how the use of programmed simulations using Fathom (a data management program) promotes the learning of probability. This project belongs to the second type of simulation and focused on problem solving by simulation. Instead of taking a formal approach, learners could use random number generators to simulate problem situations, and solve questions related with the characteristics and results of random distributions. In subsequent interviews, Sanchez found that participants considered the simulations useful to facilitate the solving problem process around core probabilistic concepts (e.g. the concept of relative frequencies).

Blejec (2002) generates simulated data and use it to clarify statistical concepts during instruction. It is unclear, however, how the students can see the connection between the simulated data and the statistical properties they are observing since it is the teacher, not the students, who program the simulation. Students cannot witness the actual sampling process or the programming of the sampling process when studying data produced by the simulation, so, why should they regard the data as representing real sampling situations?

APPENDIX C

DYNAMIC VISUALIZATIONS

Several examples of uses of dynaic visualizations to teach statistical data analysis can be found in the literature. The VESTAC project (Darius et al, 2002) provides dynamic visualizations and random simulations. The random simulations were described in the section devoted to computer tools in the space of chance. The dynamic visualizations of VESTAC are Applets devoted to several statistical topics. Some of these topics are descriptive statistics concepts, such as histograms, box-plots, QQ-plots, and correlation. Other topics are part of inferential statistics, for example, confidence intervals, one- and two-sample tests, the types of errors in hypothesis testing, the least-square method, the regression line, and the relationship between explained and unexplained variability in ANOVA. The dynamic visualizations devoted to regression and ANOVA combine both random generation of data, like simulations, and dynamic visualization of the characteristics of these inferential procedures.

West and Ogden (1998) present a summary of the "interactive demonstrations" available on the Internet and show how these demonstrations can be used to teach the representational features of histograms, (e.g., the effects of the width or number of bins). The authors describe also how these interactive demonstrations can be used to illustrate the characteristics of samples generated randomly and the behavior of the regression line in relationship to different data sets created by students. Their program also provides simulations that represent probability concepts (central limit theorem, confidence interval) and inferential procedures (power and hypothesis testing).

Godino et al (2003) describe how applets can be used in problem solving situations. They use specifically two applets: one that works as a spreadsheet and permits users to perform calculations and to do graphing of data sets; and another that permits users to compare the mean and the median as central tendency measures of two small artificial samples (n=7).

Biehler (2003) describes the use of Fathom, a statistical package that permits users to create visualizations in order to explore statistical concepts in the context of an introductory statistics course. In New Zealand, Stirling (2002) reports the use of CASTS, a computer-based application, to conduct simulations, and explore data in descriptive and inferential statistics. Marasinghe (2002) presents a series of modules to teach statistics (programmed in lisp-stat) that permit students to study confidence intervals and to visualize least squares fitting and variation among experimental units.

A similar project has been presented by Cramer and Camps (2003). Their project, called EMILeA, is devoted to statistics instruction. EMILeA-stat covers a variety of topics in statistics and probability. It contains interactive visualizations and java applets that help students to comprehend concepts in data analysis, probability, and inference. In the same vein, Mittag (2003) describes a project called “new statistics” that offers 60 applets to teach diverse statistical concepts. Applets in this project relate to diverse topics and they contain descriptive simulations and other types of tools that permit students to visualize statistical concepts. Razi and Hiemenz (2003) refer to the same project as being part of a statistic learning sequence with diverse scenarios for learning.

APPENDIX D

PRE, POSTTEST

Pre-test

Section 1.

Task 1

In this section, you will find several distribution graphs. Each graph presents two data distributions. One in black, the other in gray. From the eight distribution pairs displayed below, five have significant differences at p=0.01. You have to identify which are the significant pairs and to explain which criteria you used to identify a distribution as significant.

Please, write here the letter corresponding to the distribution pairs you consider have significant differences.

______

______

______

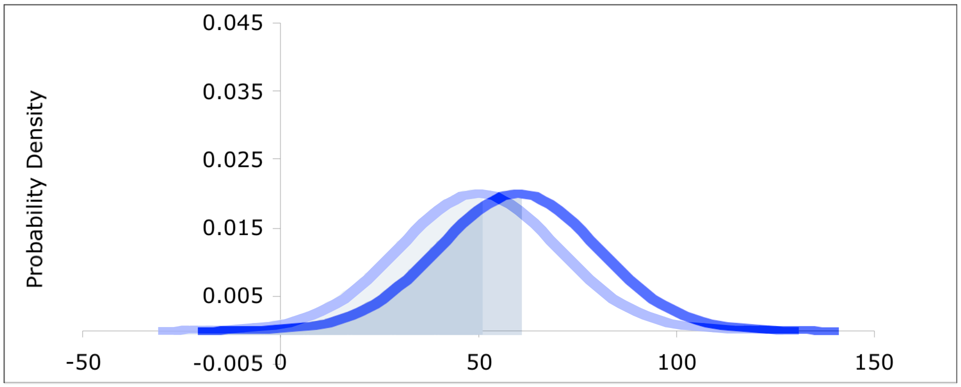
______

______

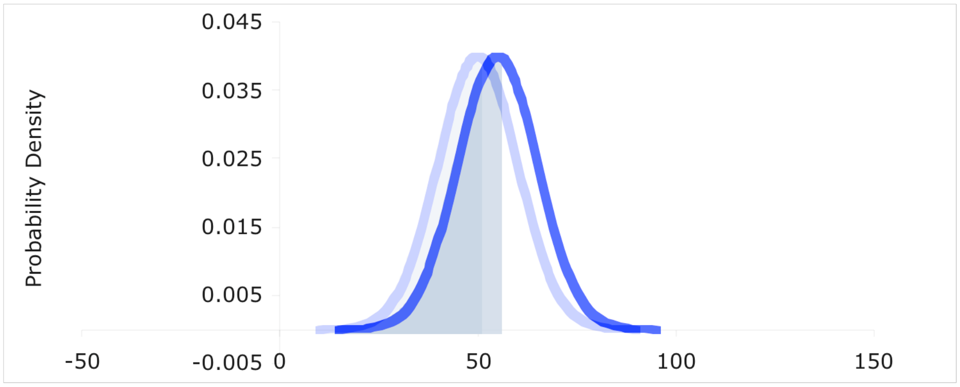
Explain the criteria you used to identify the significant differences:

Distributions

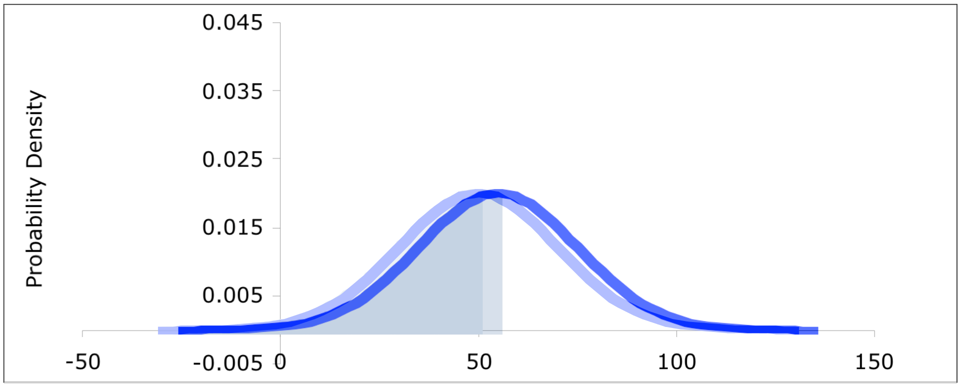
a) *n*=50



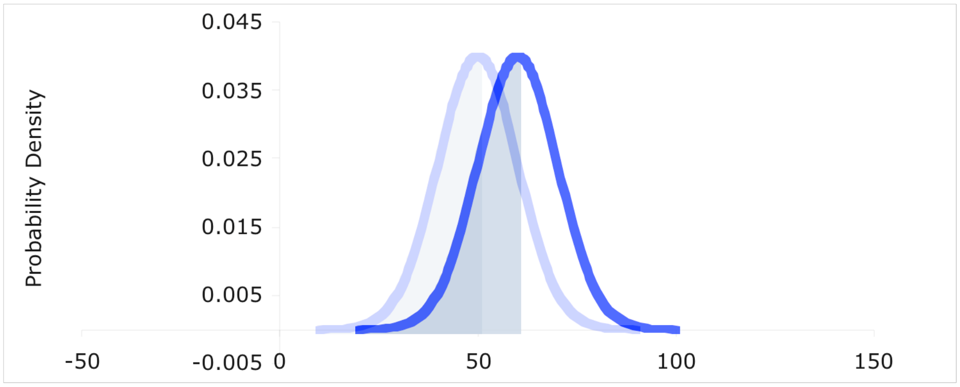
b) *n*=100



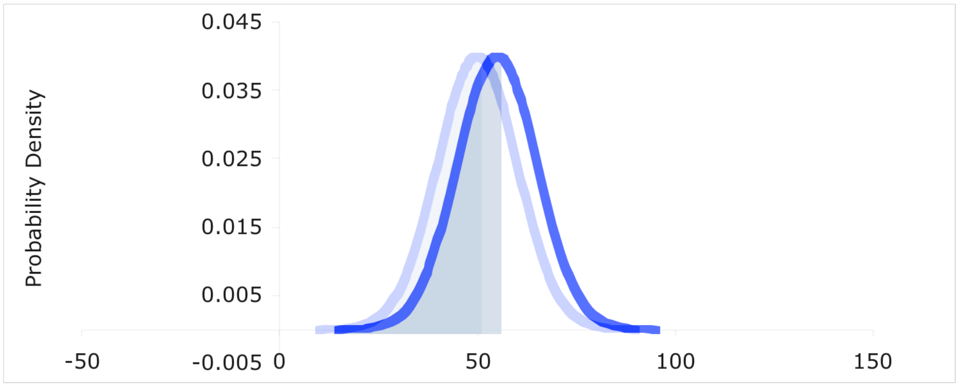
c) *n*=50


Probability Density
0.045
0.035
0.025
0.015
0.005
-0.005
-50
0
50
100
150

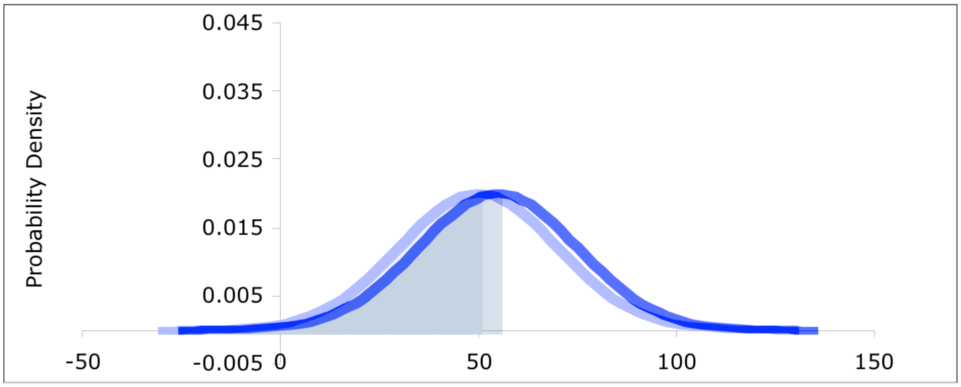
d) 100


Probability Density
0.045
0.035
0.025
0.015
0.005
-0.005
-50
0
50
100
150

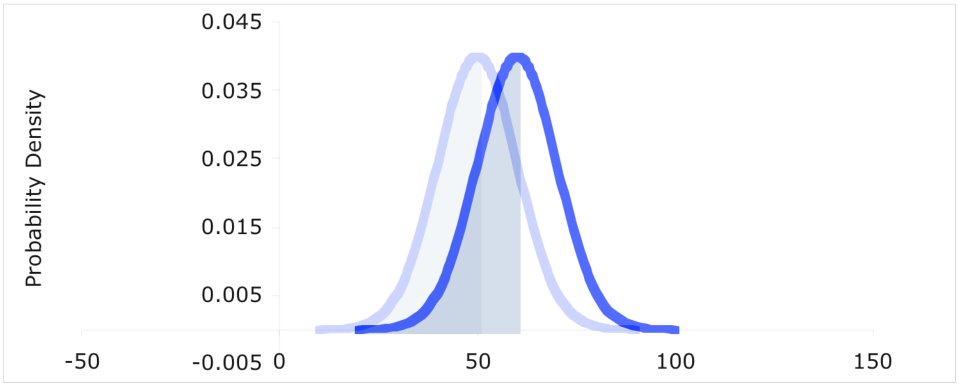
e) n=50


Probability Density
0.045
0.035
0.025
0.015
0.005
-0.005
-50
0
50
100
150

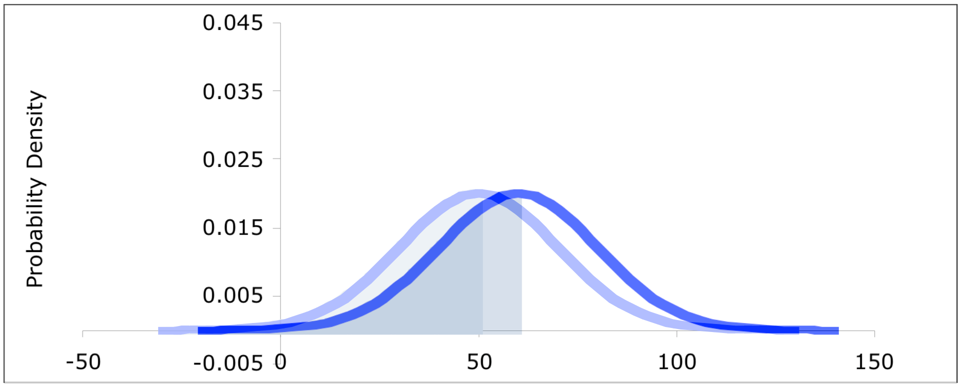
f) *n*=100


Probability Density
0.045
0.035
0.025
0.015
0.005
-0.005
-50
0
50
100
150

g) n=50


Probability Density
0.045
0.035
0.025
0.015
0.005
-0.005
-50
0
50
100
150

h) n=100


Probability Density
0.045
0.035
0.025
0.015
0.005
-0.005
-50
0
50
100
150

Pre-test

Section 1

Task 2

In the next section you will find the tables displaying the results of eight ANOVAs calculated using the same raw data using to build the distributions in task 1. Please, pair them, trying to identify which distribution corresponds to which ANOVA table. Second, identify in the distribution graph were the between groups and within group indexes were obtained from.

ANOVA	Distribution Pair
a)	
b)	
c)	
d)	
e)	
f)	
g)	
h)	

a)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	962.908	1	962.908	2.389	.125
Within Groups	39098.718	97	403.080		
Total	40061.626	98			

b)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	726.128	1	726.128	5.888	.017
Within Groups	11961.750	97	123.317		
Total	12687.879	98			

c)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2813.604	1	2813.604	23.670	.000
Within Groups	11530.017	97	118.866		
Total	14343.621	98			

d)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2463.032	1	2463.032	6.098	.015
Within Groups	39177.714	97	403.894		
Total	41640.746	98			

e)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4086.461	1	4086.461	43.880	.000
Within Groups	18439.568	198	93.129		
Total	22526.029	199			

f)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5037.576	1	5037.576	13.883	.000
Within Groups	71845.268	198	362.855		
Total	76882.844	199			

g)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	950.020	1	950.020	8.943	.003
Within Groups	21034.761	198	106.236		
Total	21984.781	199			

h)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	525.998	1	525.998	1.421	.235
Within Groups	73286.964	198	370.136		
Total	73812.963	199			

Section 1

Task 3

Case 1.

Searching for gender differences in the use of online and offline resources, a group of Colombian researchers replicated the Roy et al (2003) study on web and library search patterns. The study assigned students randomly to one of two search conditions (web or library). The students then have to find information related to a target question ("How do mosquitoes find their prey?"). The Colombian study showed different results than the original study (Roy et al, 2003). The original study found that the web was superior to the library for searching target-specific information but this difference was not significant, and that in the web condition boys learned more target-specific information than girls. In the Colombian case, results showed that the web was superior to the library in learning gains with a significant difference, and that boys learn more target-specific information than girls in both the library and web conditions. (Sig=0.01)

Identify a graph and the ANOVA result for the next situations and explain why having the same differences in each pair, in some cases the different is significant and in other it is not.

- A difference of 5 points between boys and girls in the web condition with a significant difference (Colombian study).
- A difference of 5 points between boys and girls in the library condition with a significant difference (Colombian study).

- A difference of 10 points between the web and library condition with a significant difference (Original study).
- A difference of 10 points between boys and girls in the library condition without a significant difference (Original study).
- A difference of 10 points between the web and library condition without a significant differences (Original study).
- A difference of 10 points between the web and library condition with a significant difference (Colombian Study).

Roy, M., Taylor, R. & Chi, M.T.H. (2003). Searching for information on-line and off-line: Gender differences among middle school students. Journal of Educational Computing Research, 29(2), 229-252.

Case 2.

A group of biology students wants to replicate Creasy et al (1997) study on the population and biology of the majid spider crab. They are particularly interested in finding out if there are significant size differences between male and female specimens. They collected specimens from two populations –one found at 150 m and the other found at 650m depth- on the coast of Oman. In both populations, females were consistently larger than male by the same amount. However, in the 150 meters population the different was significant, and in the 650 population the difference was not significant, when they conducted one way ANOVAs. Due to this fact, they collected a larger sample from both populations, this time the observed differences were smaller than in the first version of the study but they were found significant this time.

Identify a graph and the ANOVA result for the next situations and explain why having the same differences in each pair, in some cases the difference is significant and in other it is not.

- A difference of 10 cm between females and males in the 150 m condition with a significant difference (Original study).
- A difference of 10 cm between females and males in the 650 m condition without a significant difference (Original study).
- A difference of 5 cm between females and males in the 150 m condition with a significant difference (replication).
- A difference of 5 cm between females and males in the 650 m condition without a significant difference (replication).
- A difference of 10 cm between the mean of crabs found at 150 m and at 650 m without a significant difference in the original study.
- A difference of 10 cm between the average size of crabs found at 150 m and at

650 m with a significant difference in the replication.

Simon, R. A, Tyler, P., Young, C., &, Gage, J. (1997). The Population Biology and Genetics of the Deep-Sea Spider Crab, Encephaloides armstrongi Wood-Mason 1891 (Decapoda: Majidae). Philosophical Transactions: Biological Sciences, Volume 352, Issue 1351, pp. 365-379.

Section 2

Open-ended Questionnaire

What the standard deviation means conceptually? What does it indicate you? ____________

____

What is the difference between group variance and within group variance? ______________

____

What within group variance means conceptually? What does indicate you? ______________

____

What between group variance means conceptually? What does indicate you? ____________

____

When you get samples for two groups and you find differences, those differences can be due to which factors? ____

____

If you sample twice from a population, the results are going to be the same for both samples?

____

____

If you have several pairs of data distributions, and all they have the same mean difference, the results of the ANOVA table are going to be the same for all the pairs?

____

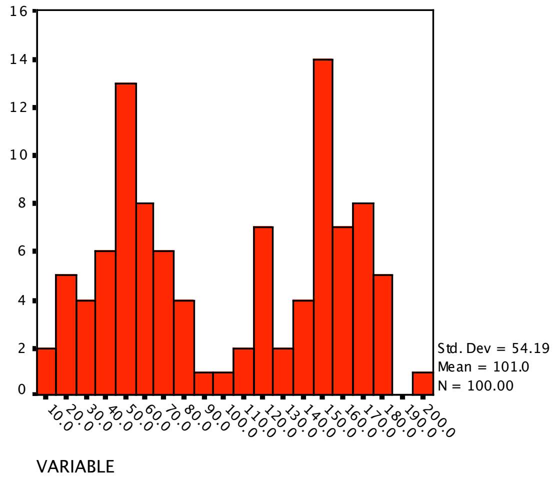
Open-ended Questionnaire 2

ANOVA

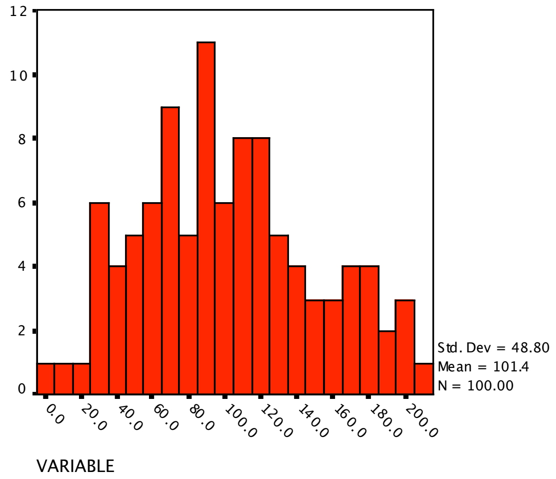
VARIABLE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7.311	1	7.311	.003	.958
Within Groups	526506.800	198	2659.125		
Total	526514.111	199			

a)



b)



The above displayed graphs correspond to two the data distribution for two different schools. Graph a) corresponds to the distribution of data of co-ed school in mathematical competency for fourth grade; distribution b) corresponds to the results of a masculine school in the same test. 100 students out of 400 were tested in each school due to budget limitations. The ANOVA table was obtained by comparing both data distributions.

Is there any difference between the schools? ____________________________________

____

Is this difference significant? ____

____

Explain the results in your own words. ____

____

Explain the reason for these results.____

____

How would you change this study to make it better? ___________________________

____

If you repeat the study, will you find the same results? _________________________

____

Why? ____

____

Section 3

Data Analysis

In this exercise, you will receive a set of data that compares the levels of performance of two groups of therapists. One group has less than 3 years of experience; the other group has more than 20 years of experience. They have been evaluated in their skill for producing adequate evaluations of psychological cases and suggesting appropriate treatments for those cases.

Acording to Ericsson (2000)

"Among investigators of expertise, it has generally been assumed that the performance of experts improved as a direct function of increases in their knowledge through training and extended experience. However, recent studies show that there are, at least, some domains where "experts" perform no better then less trained individuals (cf. outcomes of therapy by clinical psychologists, Dawes, 1994) and that sometimes experts' decisions are no more accurate than beginners' decisions and simple decision aids (Camerer & Johnson, 1991; Bolger & Wright, 1992)".

Expert Performance and Deliberate Practice: An updated excerpt from Ericsson (2000)

http://www.psy.fsu.edu/faculty/ericsson/ericsson.exp.perf.html

Conduct the necessary calculations using SPSS and write a paragraph stating your conclusion with respect to Ericsson's (2000) assertion.

Section 4

Standardized Items

The standardized items used in the final version of the main questionnaire were taken from the CAOS test and the ARTIST Scale.

Data5, Data6, Data7 are items 11, 12, and 13 in the CAOS TEST 4 Version 31, September, 8, 2005.

Inf2, and Inf3 are items 23 and 24 in the CAOS TEST 4 Version 31, September, 8, 2005.

Sam3 is Item 6 is the Artist Scale (Measures of Spread), April, 2006.

Sam4, Sam5, and Sam6 are Item 1, Item 2, and Item 5 in the Artist Scale (Sampling Variability) April, 2006.

These materials can be requested at:

https://app.gen.umn.edu/artist/tests/index.html

APPENDIX E

INTERVENTION 1: DATA ANALYSIS

The data analysis intervention was taken from the OLI course.

Module 9 (Sampling Distributions) from Unit 4 (Probability) was used.

Module 10 (Introduction) and Case I of Module 12 (Inference for relationships) from Unit 5 (Inference) were used.

Feedback systems included in the OLI course were not used in this study. Only text and exercises were kept.

This course can be visited at.

http://www.cmu.edu/oli/courses/enter_statistics.html

APPENDIX F

INTERVENTION 2: SAMPLING

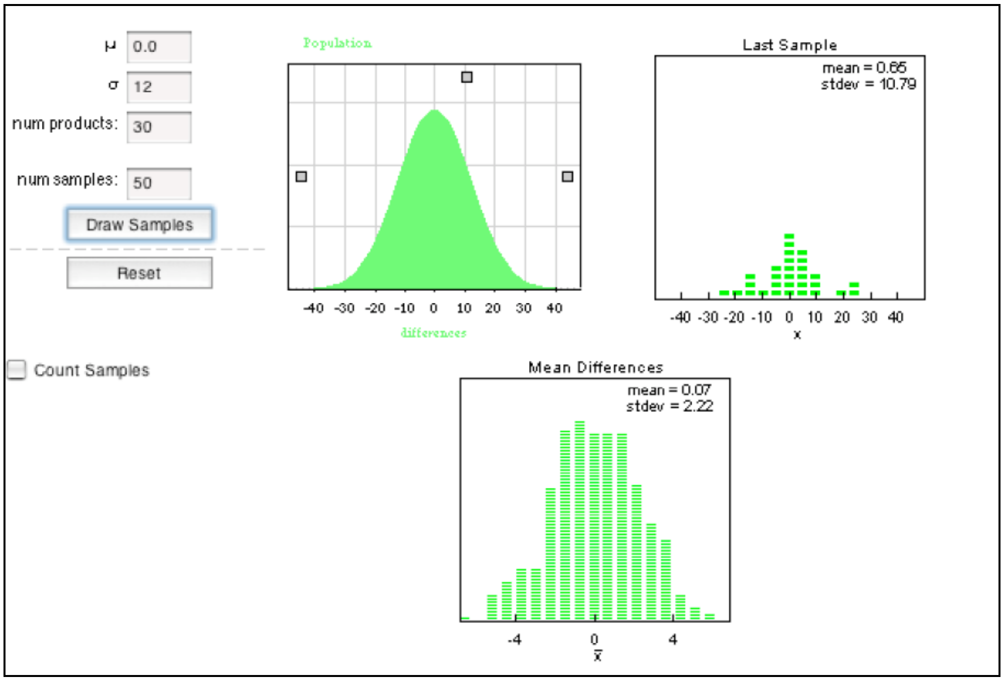
INTERVENTION 2

Until now you've read the general idea of ANOVA: Evaluating the relationship between a independent categorical variable and a dependent continuous variable by comparing group means. But, do you know what that really means? Why isn't it enough to see the plain means and compare their values? After all, the numbers are pretty clear; if a group mean is higher than other, well, it is higher, right? Well, in this section we will delve into the reasons for which we need a procedure to evaluate differences. Part of the reason we need a procedure to evaluate mean differences is that mean differences are not transparent. A mean does not speak by itself, it needs to be contrasted with the variability that surrounds it, and with the conditions under which the sample was obtained. For example, if you hear someone saying that all Latin-Americans party a lot, you can object that statement by saying that it is not true for everyone: some people here and there party more than others. In this section we explore the factors that affect the interpretation of mean differences, and some of the solutions to the problem of group comparisons.

Applet

In this section you'll be using the next Applet. This applet allows you to draw random samples while controlling the population parameters. The Applet allows you to control four parameters: μ (the mean of the population), σ (the standard deviation of the population), number of products (the sample size), and number of samples (the number of samples you obtain each time you click "draw"). The Applet displays three diagrams. The diagram under "population" displays the population characteristics you've chosen. The diagram under "last sample" displays the results of the last sample you draw. The diagram under mean differences displays the cumulative result for all the samples you have drawn. For example, if you are evaluating 30 students from the total population of a school, the results of that sample will be displayed under “last sample” and the characteristics of the population will be displayed under “population”. In this Applet you can control the mean and standard deviation of the school by changing μ (the mean of the population), or σ (the standard deviation of the population). If you draw a second sample, then the results of that sample will be displayed under “last sample” and the combined results of the first and second sample will be displayed under “mean differences”. Do you notice that in the first trial, the “last sample” results and the “mean differences” results are the same. You can also draw several samples at the same time by changing “num samples” or changing the size of the sample by modifying “num products”.

http://statweb.calpoly.edu/chance/applets /Shopping/Shopping.html.



μ 0.0

σ 12

num products: 30

num samples: 50

Last Sample: mean = 0.65 stdev = 10.79

Mean Differences: mean = 0.07 stdev = 2.22

Please, set the indicator “num samples” to 1, and sample a couple of times.

What does a point (dot) in the bottom center diagram means? ______________________

Why does a point (dot) in the up-right diagram means? ___________________________

What is the difference between the three diagrams? ________________________

If you want to get a sample of *n* = 50, what do you have to do? _______________

Exercise 1

In this exercise, we explore the reasons why we need to test for mean differences. The basic goal of this exercise is to show that mean differences can be produced by factors different to the existence of a real mean difference. Suppose you are trying to find out the true mean of verbal reasoning for the students of a given school. Suppose also that you can't test all the students in the school. So, you draw a sample of students.

Draw a sample using the tool. Suppose that the average score of the students in the school, not in the sample, is 20 and establish that average for the population in the average window (marked with miu); suppose that the standard deviation (marked with sigma) is 7.5. Then, draw a sample.

What is the mean of this sample:_____________________________________

Is the mean of this sample equal to the mean of all the students in the school?

____

Now, draw another sample from the population with the same parameters.

What is the mean of this second sample? _______________________________

Is the mean of this sample equal to the mean of the population (all students in the school)?

____

Is the mean of this sample equal to the mean of the first sample?

____

If you have several samples from a population, how can you know what it is the true mean of the population. Or for the case of the example, if you can't test all the students in the school, how can you know what is the true average score of the school's students in verbal reasoning. The only real way to know it is testing all the students in the school. If you can't test them all, there is no way to know the exact mean of the school. However, we can know that the majority of the sample means fall within a certain range. To prove this, we conduct the next exercise.

Draw 10 samples using the Applet. Compare the diagrams in the Applet.

What the diagram in the bottom of the Applet represents? ____________________

____

What the diagram in the upper right corner of the Applet represents? ___________

____

What the diagram in the upper left corner of the Applet represents?_____________

____

Then, draw 40 samples more using the automatic function. You can observe how the sample means (in the diagram at the bottom of the applet) accumulate in certain zones of the distribution.

Try to identify the value at highest point of this distribution (the distribution of sample means at the bottom of the applet):

____

To which number is this score is similar:

a) to the mean of the population b) to the mean of the last sample

Now, that we know that the samples accumulate close to the population mean, we want to know how far from the population mean, the samples means fall. Using the counting function of the Applet, please build the next table…

Table 1.

Less than 5	Between 5 and 12.5	Between 12.5 and 20	Mean = 20	Between 20 and 27.5	Between 27.5 and 35.	More than 35.

As a conclusion, we cannot know if the mean we obtained from a sample is the exact mean of the population, but we can know that those samples are more likely to fall within certain limits. We do not go to the specific details of this calculation but its possible to obtain an interval around the sample means that indicates, with a certain probability, where the population mean should be. For example, if we obtain a sample that has a mean of 30, we cannot be sure that 30 is the mean of the school, but we can assert that the mean of the school is between 25 and 35 with 80% probability, meaning that if we draw many samples they will fall between 25 and 35 80% of the times.

Can you explain this idea using the example of the school? Suppose you're trying to find the mean of the school on verbal reasoning but you only have a sample of 50 students. You obtain the sample using the Applet. What mean did you obtain? ______

Where the mean of the population must be? _________________________________

____

Exercise 2

Now we go to a more interesting question. If we are testing for a mean difference, how can we know really exists. For example, you are comparing to high schools in knowledge about ecology. You go to school 1 get a sample and evaluate it. Then, you go to the school 2 and evaluate another sample. You analyze the data you have and see that school 1's average is higher than school's 2. What if you were just lucky and got the higher sample mean from school 1 vs the lower sample mean of school 2. What if there is not any real different? The next exercise will help to explore the possible ways to solve this problem.

Let’s simulate the situation described above. You have two schools. School 1 has an average of 20; school 2 has an average of 40. The standard deviation is 7.5. Draw 50 samples from each school and fill the following table.

<5	>X>10	10>X>15	0>X>25	5>X>30	0>X>35	5>X>40	0>X>45	5>X>50	0>X>55	5>X>60	0<X

Table 2.

Can you observe a difference between the mean distributions produced by the two schools? ____

____

What is the most common value within this sample from school 1?[2] ______________

____

You can see that there is a common range where means from both schools fall? What is this common range? ____

Are there more or less sample means in this range than in the other intervals? _______________

Well, as the last exercise shows, there is a range of error when comparing two groups, but normally the sample means from a higher mean population tend to be higher than the means from a population with a lower mean. Statistical tools that test for mean differences evaluate the strength of you conclusion regarding the existence of mean differences.

[2] This question does not have an actual answer; it is just to check they are able to differentiate between sample mean and sample distribution of the mean.

A complementary question here is if mean differences created but sampling look the same that mean differences created by systematic effects in the model. In terms of our example, if we take two samples from school 1 and get a mean difference for them, does that mean difference tend to be higher or lower than if we take two mean differences from the same school and then we calculate the difference between both of them.

To find out, let's make the next exercise. Draw 10 samples from the school 1 (mean 20, sd 7.5) and 10 samples from school 2 (mean 40 sd 7.5), pair them and calculate the difference. Write the differences you find: ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

Now, draw 20 samples from school 2 and pair them in the order you are obtaining them (e.g. pair the first with the second, the third with the fourth). Then, calculate the difference for each pair. Write the differences: ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

What differences are larger in average? ____

____

Other way to see this is use the information that we collected in table 2.

Table 3

School1/ school2	10>X>15	15>X>20	20>X>25	25>X>30	30>X>35	35>X>40	40>X>45	45>X>50
10>x>15	0	5	10	15	20	25	30	35
15>x>20	-5	0	5	10	15	20	25	30
20>x>25	-10	-5	0	5	10	15	20	25
25>x>30	-15	-10	-5	0	5	10	15	20
30>x>35	-20	-15	-10	-5	0	5	10	15
35>x>40	-25	-20	-15	-10	-5	0	5	10
40>x>45	-30	-25	-20	-15	-10	-5	0	5
45>x>50	-35	-30	-25	-20	-15	-10	-5	0

This table displays the possible mean differences for school 1 and school 2. For example, if school 1 has a mean that is between 10 and 15, and school to has a mean that is between 15 and 20, the maximum mean difference in this group is 10 and the minimum mean difference is 0. The value displayed in each box is in the middle of these two values. The idea of this exercise is to see approximately where there are more likelihood of finding mean differences. To do this, you have to take the mean distribution of school 1 and the mean distribution of school 2 and put the values side by side in the table, and then multiply them. That will indicate you where more mean differences can be found. For example, if you have two means in school 1 (20, 21) and two means in school 2 (25, 27), you have four possible mean differences (25-20, 25-21, 27-20, 27-25).

In which box there are more possible mean differences? ___________________

Now do the same exercise but instead of using the data from schools, use the data from school 1, (like if you were taking two samples from school 1).

Table 4.

School1/ School1	10>X>15	15>X>20	20>X>25	25>X>30	30>X>35	35>X>40	40>X>45	45>X>50
10>x>15	0	5	10	15	20	25	30	35
15>x>20	-5	0	5	10	15	20	25	30
20>x>25	-10	-5	0	5	10	15	20	25
25>x>30	-15	-10	-5	0	5	10	15	20
30>x>35	-20	-15	-10	-5	0	5	10	15
35>x>40	-25	-20	-15	-10	-5	0	5	10
40>x>45	-30	-25	-20	-15	-10	-5	0	5
45>x>50	-35	-30	-25	-20	-15	-10	-5	0

In which box, there are more possible mean differences? ________________________________

Are the results from table 3 and table 4 different? _____ How? ____________________

____

Well, the question to solve here is this: if we take two samples from the same school, we'll find differences due to the sampling process, however, there is not any real mean difference, because the mean of the population/school is the same. How can we differentiate this situation from a situation where we have two samples from different schools and there is, in fact, different mean for the schools? ____

____

Well, the answer to the question is this. It is not possible to differentiate fully between both situations, but it is possible to know how likely is a mean difference to come from random sampling. That is, if a sample is likely to come from sampling, we assume that there is not enough evidence to assume there is a real difference between the schools. For example, look for the most common value in table 3, and write here ____: Is that value the most common value in table 4? ____

The theoretical implications of the idea presented before are really important. For example, imagine that you are comparing men and women in IQ and that there is not any difference between both populations. That is that the population has the same mean (like if you were sampling from school 1 all the time). You can find differences when evaluating a sample, however, when evaluating the likelihood of those differences you should find that they're most similar to sample produced by populations with the same mean, than to be produced by populations with different means.

Write here another example of this type of comparison (use drawings if you need it).

____

____

____

There are two factors that influence the distribution of means coming from a population: the sample size and the standard deviation of the population. The smaller the sample size, the larger the variability of the sample means found. You can see this by obtaining 50 samples with a sample size of 50 and 50 samples with a sample of size of 5.

Fill this table.

Table 5.

	X<10	10>X>15	20>X>25	25>X>30	30>X>35	35>X>40	40>X>45	45>X>50	50>X>55	55<X
Size=50										
Size=4										

Where do you get more extreme sample means? _____________________

The fact that a smaller samples size creates larger variability is very important for analyzing mean differences because the reliability of your conclusions decreases. For example, in the case of the school, you can be sampling from only school 1, and finding large differences caused just by random sampling. In the case of the men and women comparison, you can find large differences but without the existence of any real IQ difference between both population. Statistical test like ANOVA evaluate the mean differences accounting for the differences in sample size.

The second factor affecting the distribution of mean samples is the standard deviation. The larger the standard deviation of the population, the larger the variability among samples and the larger the standard deviation of the samples themselves…and of course, the lower our confidence in the mean difference we are observing. To see this idea, you can do next exercise.

Draw (paint) two samples with a normal distribution.

Figure 1

Now, draw two samples with the same mean difference but make them longer like if they had a larger standard deviation. If you can’t figure this out, you can’t use the simulator and put the standard deviation at different levels, so you can see how samples with small and large distribution look.

Figure 2

Compare the graphs. In which case, do the samples overlap more? ____________

In both cases the mean difference is the same, but, in which case you think that difference is clearer? ____

Now, using the random sampling generator, draw 50 samples from a population with a large standard deviation and 50 from a population with a small standard deviation. Fill the following table.

Table 6

	X<10	10>X>15	20>X>25	25>X>30	30>X>35	35>X>40	40>X>45	45>X>50	50>X>55	55<X
Size=50										
Size=4										

In which case, do you find more extreme values? ____________________

In which case, do you think larger mean differences can be produced just by the sampling process? ____

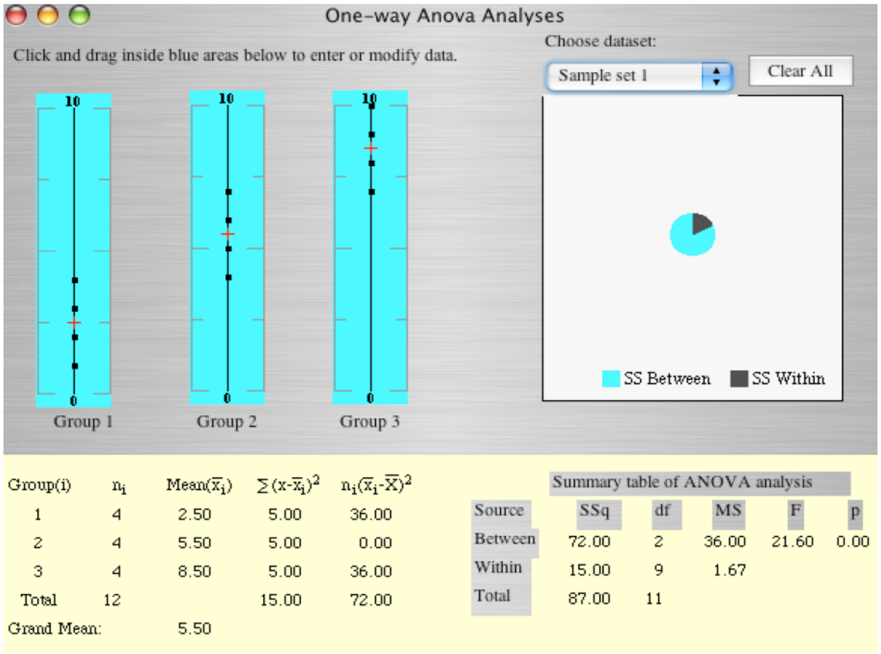
____

The idea behind ANOVA

As you have seen, mean random sampling produces mean differences. These mean differences however fall within certain limits and means from samples obtained from two populations differ more than means from samples obtained from the same population. We need a reliable method to differentiate both cases. ANOVA is the procedure that permits us to establish whether or not we have differences produced by random sampling. To do so, ANOVA compares within and between group variability. Within group variability is the variability among the members of a group that is not explain by the categorical variable that determines the groups used in the ANOVA. For example, in a study that compares men and women, within group variability is the variability among men, and among women. Between group variability is the variability between men and women, that is, the mean difference between both groups. Remember that in this example, gender is the categorical variable of the study, and men and women the two groups of this variable. Gender can explain the difference between men and women, but it cannot explain the differences within each group.

Now you have read the idea behind ANOVA. In the next pages you will find some exercises that help you to clarify the meaning of this idea. The first exercise requires you to use the following Applet. This Applet visualizes the relationship between explained and unexplained variance in a pie diagram. In the right half of the Applet, three vertical lines represent three different groups in an ANOVA (See figure). In each line, black dots represent the data points (individual scores) and a red line represents the mean of the group. Students have seven data sets to work with. Once the data set is uploaded, you can move the points (black slots) in each line. The Applet automatically modifies the numerical indicators presented in an ANOVA table in the lower part of the diagram, and the relationship between explained and unexplained variance in the pie diagram.

http://www.ruf.rice.edu/~lane/stat_sim/one_way/index.html



Click and drag inside blue areas below to enter or modify data.

Choose dataset: Sample set 1

Group(i)	n_i	Mean($\bar{x}_i$)	$\sum(x-\bar{x}_i)^2$	$n_i(\bar{x}_i-\bar{X})^2$
1	4	2.50	5.00	36.00
2	4	5.50	5.00	0.00
3	4	8.50	5.00	36.00
Total	12		15.00	72.00
Grand Mean:		5.50		

Summary table of ANOVA analysis

Source	SSq	df	MS	F	p
Between	72.00	2	36.00	21.60	0.00
Within	15.00	9	1.67		
Total	87.00	11			

In this applet, yo u can move the dots to create different configurations of data. If you have equal means, the between variance is large. If you have equal means and the dots are separated from each other, the between variance is small and the within variance is large. If you have large mean differences, the between variance grows. If you separate the dots, even if the means do not change, the within variance grows. You can see the explained and unexplained variance in the pie diagram at the right of the Applet. SS Between (Sum of Squares Between groups) and SS Within (Sum of Squares Within groups) represent the explained and unexplained variances respectively.

Tell us in your own words, what between group sums of squares represents

____

You can see the same indicators in the ANOVA table in the lower part of the Applet. In that table, you can see the sum of squares and the resulting F and p-value.

Create two samples with large within variance and describe it here: ____________________

____

Create two samples with small within variance and describe it here: ________________________

____

Create two samples with small between variance and large within variance, and describe it here:

____

____

Create two samples with small within variance and large between variance, and describe it here:

____

____

How do the F and the p value change when between SS increases?

____

____

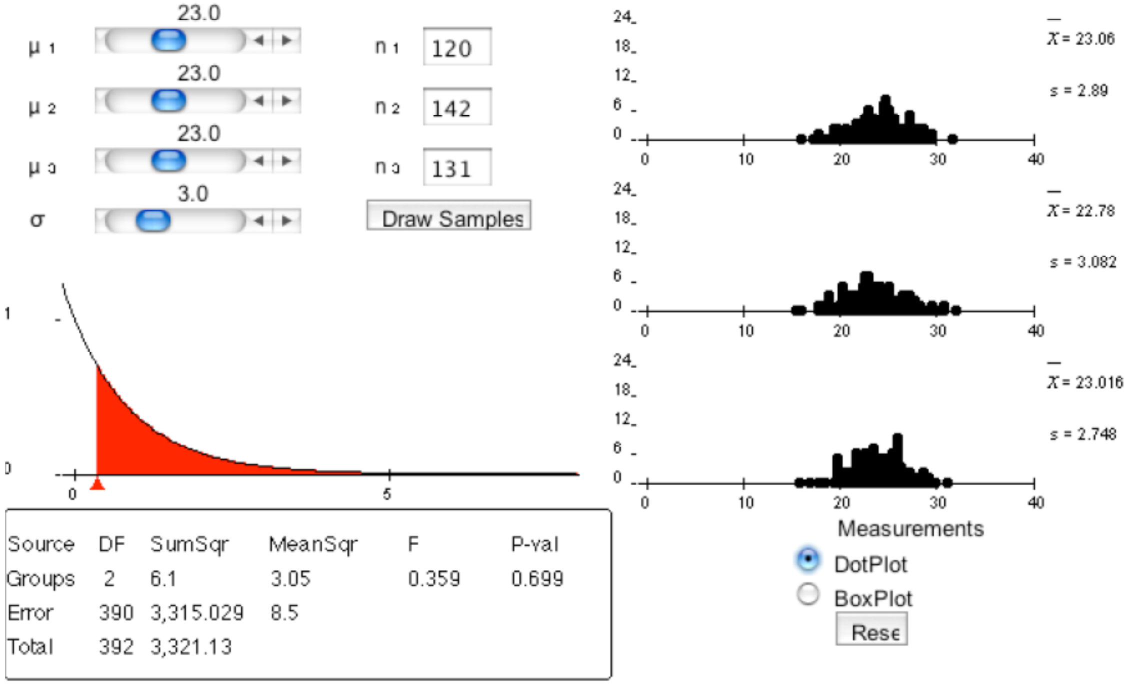
How do the F and the p-value change when within SS increases?

____

So, what the F and the p-value indicate us? ____

The fundamental mechanism of ANOVA is comparing within and between variance. This is so because when randomly sampling from a single population (like in the school example), the distance (the mean difference) between groups belonging to the same population tend to be the same than the distances (the score differences) between the cases in each group. No example of this situation was given in the section 1 of this intervention because it is very tedious to do, but it follows the same logic. When you sample randomly from a single population, the scores you obtain tend to have the same between variance (group mean difference) than within variance (group standard deviation). Of course, this is not true all the time. Some time you have samples where the within variance is smaller than the between variance. But probabilistically, samples with large between variance and small within variance tend not to happen when you are sampling from a single population. They only happen when you sample from populations that have different means. In this way ANOVA tests for mean differences. ANOVA's reasoning follows a logic like this: if the ratio between "between" variance and "within" variance is similar to the ratio should come out of random sampling, the procedures assume that there is not a effect of the categorical variable, and that the differences are the product of random sampling. If this ratio is more similar to the ratio we should expect when you are sampling from populations with different means, then the procedure assumes that the differences come from the effect of the categorical variable.

For the following exercises, you will use the following Applet.



μ 1 23.0
μ 2 23.0
μ 3 23.0
σ 3.0

n 1 120
n 2 142
n 3 131

$\bar{X}$ = 23.06
s = 2.89

$\bar{X}$ = 22.78
s = 3.082

$\bar{X}$ = 23.016
s = 2.748

● DotPlot
○ BoxPlot

Source	DF	SumSqr	MeanSqr	F	P-val
Groups	2	6.1	3.05	0.359	0.699
Error	390	3,315.029	8.5		
Total	392	3,321.13			

.

This Applet permits you to sample from three groups while controlling population and sampling parameters. You can control the means of the three populations, and the size and standard deviation of the samples. The results for each group can be displayed as histograms (dot plots) or as box plots (see figure). Numerical results are presented as an ANOVA table in the bottom of the Applet. In this case, between groups is called "groups"; and within groups is called error (because it is not explained by the model). The probability of obtaining each sample is displayed as a red band in an ANOVA distribution graph

In this exercise you will observe how random sampling produce variation in the ratio between "between" and "within" variance, but overall the relationship stays constant among different samples. Please, set both Applets with a small mean difference –all the mean groups at the same level- and draw several samples.

Fill the next tables.

Table 7

Applet 1.

Sample	1	2	3	4	5	6	7	8
Between SOS								
Within SOS								
F								
p-value								

What you did when you set all the groups at a similar level was to establish the population means. You can see that the real means you obtain vary from sample to sample –in the same fashion that they did in the part 1 of this intervention-. For this same reason, the between and within variance, and the F value, vary among trials. What we want to show however is that, even if this values vary, they vary in a different way than in a situation where there is a large means difference. To see this, you have to set the Applets with a large mean difference that is with the group means at very different levels. Please fill these table.

Table 8

Sample	1	2	3	4	5	6	7	8
Between SOS								
Within SOS								
F								

p-value								

Do you find a difference between the results of table 7 and 8, and the results of table 9 and 10?

____

____

____

When you have large mean differences in the population, the sample between variance tends to be larger than the within variance. That is how ANOVA establish if a sample permits to suppose that there is a mean difference among the different groups in the sample.

For the case of ANOVA, the results are influenced by the sample size of the sample and standard deviation of the population, for the reasons explained in the first part of this intervention.

How does sample size influence the behavior of the sample means when sampling from a population? ____

____

____

How does the standard deviation of the population influence the behavior of the sample means when sampling from a population? ___________________________

____

____

How does it affect our reasoning when comparing group means? ____________

____

____

____

In the next exercise, we will see that this influence happens also when comparing means using ANOVA. We are going to vary sample size and standard deviation, while keeping the mean difference constant. Set the group means at a similar level that you did when filling tables 9 and 10. Now, increase the standard deviation and fill the following tables.

Table 9

Applet.

Sample	1	2	3	4	5	6	7	8
Between SOS								
Within SOS								
F								
p-value								

How the values in these tables look different that the values in tables 9 and 10? ____

____

____

Increasing the standard deviation increases the within groups variance because cases in each group vary more. For the same reasons observed in figure 1 and 2, the significance of the differences decreases (the p value). Now, we are going to evaluate the influence of sample size. Take the same parameter used in the last exercise, but use a smaller sample size. Fill the following tables.

Table 10

Applet.

Sample	1	2	3	4	5	6	7	8
Between SOS								
Within SOS								
F								
p-value								

How the values in these tables look different that the values in tables 11 and 12?

____

____

____

Why?

____

____

____

Observe the behavior of the p-value. What can you conclude about the meaning of this indicator?

____

APPENDIX G

INTERVENTIONS' CODING AND ARGUMENT

1. Coding of the Interventions

In the initial stage of coding, the whole interventions' text was parsed into three types of larger units: exercises, examples, and paragraphs. An exercise was defined as a fragment of text that required students to solve one or more questions referring a single situation or data set. An example was defined as a fragment of text that illustrated a concept or procedure using a situation with at least a cover a story and a set of data results and/or representations. No distinction was made between worked-out examples and simple examples. Isolated sentences that exemplified concepts were not included in the list of examples. No mathematical pure examples were found in any of the intervention; in other words, sets of data and results were always presented with a cover story. All remaining text was classified as pure text and parsed into paragraphs; that is, units of minimum 3 sentences referring a single concept (e.g., sampling variability).

In the second stage of coding, exercises, examples and paragraphs were divided into smaller units. Exercises were divided into questions; a question was a part of an exercise that required a student to produce a single theoretical answer or mathematical result. Examples and paragraphs were divided into idea units. An idea unit was defined as a non-redundant proposition (subject + predicate) (Chafe, 1985; Steffensen et al., 1979). When two or more consecutive sentences referred to the same proposition, they were coded as a single idea unit. When an idea unit appeared two or more non consecutive times, each appearance of the idea unit was kept as an independent element in the idea units list.

2. Construction of the Idea Units Map

The idea unit map (Table 1) was built based on the data analysis intervention. This intervention's text and exercises were taken from the OLI course and served as an external source of the interventions' content. Once the map was finished, the sampling intervention was coded according to this map. The process of building the data analysis map started by classifying the idea units into redundant and non-redundant ideas. Redundant ideas referred the idea units that were repeated in the idea unit list; non-redundant ideas referred all the other idea units. As explained before, when an idea unit appeared two non-consecutive times in the text, it was kept as two separate idea units in the list. Therefore, in the list there were some idea units that were repeated two or more times. Repetition was considered as an indicator of the importance of the idea unit within the argument of the data analysis intervention. Even though, this criteria was arbitrary and did not include the meaning of the idea units, it provided an unambiguous way to rate the idea units' saliency.

After that, redundant idea units were sorted out to produce the intervention content map (Table 1). Specifically, redundant idea units were classified according to the core ideas they support. Supporting a core idea meant basically three things: paraphrasing, serving as an argument or explaining an aspect or detail of the core idea. Some of the redundant idea units had a unique relationship with the core ideas, other participated in two or more parts of the argument and therefore they had a relationship with several core ideas. This situation was more evident in the last part of the intervention where most core ideas required information from the first parts of the intervention to build an argument. To represent this situation, redundant idea units related to more than one core idea appear as complete sentences in front of the first core idea they support, and as a number in front of any other core idea they are related to.

Table 1. Idea Unit Map

Idea	**Sub-idea**	**Sub-idea**	**Sub-idea**	**Sub-idea**
1.ANOVA evaluates the relationship between a categorical and a	1.1.ANOVA tests the relationship between a categorical and a	1.2.Variables can be represented in different forms.		

continuous variable.	continuous variable.			
2.ANOVA is necessary to evaluate the relationship between categorical and continuous variables.	2.1.Sampling creates variation	2.2.Sampling creates variation within certain limits.	2.3.Parameters and Statistics are different (mean and SD)	2.4.ANOVA is useful to establish whether or not the observed differences are due to systematic differences among groups or to other factors.
3.ANOVA compares within and between group variances to elaborate conclusions on the sample means	3.1.ANOVA compares between and within variances to produce conclusions.	3.2.Between and within variances are defined in different ways (or representation).	3.3.SD influences the comparison of group means because it increases within variance.	3.3.SD influences the comparison of group means because it increases within variance.
4. The variation among group means is considered negligible when within and between variances are similar.	4.1.The F test permits to decide about the ha and the null hypothesis, that is about the group differences.	4.2.The F test permits to decide between the null and the alternative hypothesis, that is about the group differences, based on the mean differences and in the group variation.	2.1 2.2 2.4	2.4
5. The degrees of freedom affect ANOVA's interpretation.	5.1.Sample size influence sampling variation.	2.1 2.2		
6. The elements of ANOVA tables 7. Interpretation of p-value.	6.1.ANOVA Tables have sum of squares, F and p-values that are the result of 4.	4.1	7.1.Meaning of the p-value.	4.2 6.1
8. Interpretation of the results of ANOVA	8.1.Sampling Variation happens in real research and needs to be account for.	8.2.The question of a study connects with the mean differences and the method of ANOVA to test for true mean differences (ha y ho).		

The next step was situating the remaining idea units in this map. To do so, the remaining idea units (non-redundant idea units) were arranged in groups corresponding to the redundant ideas in the table. The relationship between the non-redundant idea units and the redundant idea units in the map were of three types: the non-redundant idea unit could be paraphrasing the redundant idea in a way that was not similar enough to considered them the same idea unit; the non-redundant idea could be explaining a detail or aspect of the redundant idea; or the non-redundant idea unit could be connecting a detail in the explanation to the redundant idea.

3. Counting the Number of Idea Units supporting each Redundant Idea

Once the groups were formed, the number of non-redundant idea units aligned with each redundant idea unit was counted. This procedure was conducted for the general text and for the examples in both interventions. The number of questions in the exercises devoted to each redundant idea was added up also. Table 2 presents six numbers below each redundant-idea. The three numbers in the first row present the counting results for the sampling intervention; the three numbers in the second row represent the same results for data analysis intervention. In both rows, the first number represents the number of idea units in the general text; the second number presents the number of idea units in the examples; the third number presents the number of questions in the exercises devoted to each idea.

4. Classification of the Redundant Ideas Units

The redundant idea units in table 2 were classified according to the spaces of inferential statistics: Data Analysis, sampling or statistical inference. Data analysis idea units were defined as idea units that participate in the comparison of two or more data distributions from a graphical or numerical point of view. Idea units in this category explained the characteristics of data representations, the idea and types of variance, the specific influence of explained and unexplained variance in drawing of conclusions about mean differences. Sampling ideas were defined as idea units that explained the relationship between sample and population characteristics. This type of idea focused on sampling variability, particularly on the effects of population variance and sample size on sample mean differences. Inference idea units were defined as the idea units that connected both data analysis and sampling ideas to inference in real situations. This definition included connections from ANOVA results to conclusions in context, explanations about the meaning of ANOVA tables, and the variables' structure on which ANOVA operates. The classification given to each redundant idea appears in Table 2 after the redundant idea number in the form of the first letter of the statistical space assigned to each one (e.g., D for Data Analysis; S for Sampling; I for Inference).

Table 2. Redundant Idea Units Counting and Classification

Sub-idea (line)	**Sub-idea**	**Sub-idea**	**Sub-idea**

1.1. (I) ANOVA tests the relationship between a categorical and a continuous variable.
S= (1)(0)(0)
D=(2)(22)(1)
	1.2. (D) Variables can be represented in different forms.
	S=(0)(0)(0)
	D=(0)(6)(0)		

2.1. (S) Sampling creates variation
S=(5)(0)(8)
D=(4)(13)(0)
	2.2. (S) Sampling creates variation within certain limits.
	S=(7)(2)(10)
	D=(0)(13)(0)
		2.3. (S) Parameters and Statistics are different (mean and SD)
		S=(3)(0)(4)
		D=(22)(7) (3)
			(I) 2.4.ANOVA is useful to establish whether or not the observed differences are due to systematic differences among groups or to other factors.
			S=(6)(1)(3)
			D=(20)(0)(3)

3.1. (D) ANOVA compares between and within variances to produce conclusions.
S=(2)(0)(0)
D=(9)(8)(0)
	3.2. (D) Between and within variances are defined in different ways (or representation).
	S=(2)(0)(3)
	D=(3)(3)(3)
		3.3. (D) SD influences the comparison of group means because it increases within variance.
		S=(3)(0) (5)
		D=(6)(5)(3)	

4.1. (I) The F test permits to decide about the ha and the null hypothesis, that is about the group differences.
S=(1)(0)(2)
D=(10)(1)(2)
	4.2. (D) The F test permits to decide between the null and the alternative hypothesis, that is about the group differences, based on the mean differences and in the group variation.
	S=(3)(0)(1)
	D=(3)(21)(1)
		2.1
		2.2
		2.4
			2.4

5.1. (S) Sample size influence sampling variation.
S=(3)(2)(11)
D=(8)(0)(0)
	2.1
	2.2		

6.1. (I) ANOVA Tables have sum of squares, F and p-values that are the result of 4.
S=(2)(0)(9)
D=(1)(8)(2)
	4.1	7.1. (I) Meaning of the p-value.
		S=(2)(0)(3)
		D=(3)(1) (10)
			4.2
			6.1

7.1. (S) Sampling Variation happens in real research and needs to be account for.
S=(1)(1)(2)
D=(0)(9)(0)
	7.2. (I) The question of a study connects with the mean differences and the method of ANOVA to test for true mean differences (ha y ho).
	S=(4)(0)(8)
	D=(2)(11)(3)		

5. Items Coding

After the coding of the interventions was complete, 16 questionnaire items were coded in the three statistical spaces: Data analysis, sampling, and inference (Table 3). 6 items came from the first section of the pre, posttest task, 1 item came from the open-ended questionnaire, and 9 items came from the collection of standardized items used to evaluate the students. For coding purposes, data analysis items required participant to compare two or more distributions represented in different ways; sampling items required students to understand the relationship between population and sample characteristics; and inference items required participants to interpret ANOVA results. Side by side with this process, items were classified according to the redundant idea units that were necessary to solve them (right column of table 3). This list was produced by comparing the goal of the exercise, the elements of the problem that needed to be account for and the input information in each exercise. For example, if a item required students to interpret a significant result using the information in ANOVA table, idea units related to interpretation of ANOVA results and the ANOVA table elements were included in the list; if the item additionally required student to account for sample size, then, an idea unit related to sample variability was included in the list. In a different example, if the item used as input information a graph representing two distributions, then, idea units related to data representation and comparison were included. Finally, each item was classified according to the most common type of idea required to solve it. For instance, if an item required participants to use more sampling idea units than any other type of idea units, then the item was classified as a sampling item. The classification of items using this method and the direct classification coincided.

Table 3. Item Classification

Item	Goal	Accounting for	Space	Idea Units
1	Identify the more significant difference	Different central values	D	1.2, 3.1,3.2,4.2,
2	Identify the more significant difference	different spreads	D	1.2. 3.1,3.2,3.3,4.2,
3	Identify the more significant difference	different sample size	S	2.1,5.1,8.1,
4	Produce a conclusion	different spreads	D	1.2,2.4,3.1,3.2,3.3,4.2,8.2
5	Produce a conclusion	Different sample size	S	2.1,2.4,5.1,8.1,8.2
6	Produce a conclusion	Different p-values	I	2.4,4.1,6.1, 7.1,8.2
7	Produce a conclusion in Context	Different central values	D	1.1,1.2,3.1,3.2,4.2,8.2
8	Evaluate a conclusion	Different central values	D	1.2,2.4,3.1,3.2,4.2,8.2
9	Evaluate a conclusion	Different central values	D	1.2,2.4,3.1,3.2,4.2,8.2
10	Evaluate a conclusion	Different central values	D	1.2,2.4,3.1,3.2,4.2,8.2
11	Interpret a significant result	Sample size	I	1.1,2.4,2.3,4.1,5.1,7.1,
12	Interpret a significant result	Different central values	I	1.1,2.4,3.1,3.2,4.1,4.2,7.1,8.2
13	Connect population and sample characteristics.	Sample size	S	2.1,2.2,2.3,5.1,
14	Connect population and sample characteristics	Different spreads	S	2.1,2.2,2.3,3.1,3.2,3.3,4.2,
15	Connect population and sample characteristics	Sample size	S	2.1,2.2,2.3,5.1,
16	Different sampling distribution and data distribution	Different spreads	S	1.2,2.3,3.1,3.2,3.3,4.2,

6. Reliability

The purpose of the above described process was establishing which questions in exercises, and which idea units in examples and general text connected to each statistical space. However, once that connection was established, it was possible to obviate all the other sub-steps in the exploratory process and to focus on the relationship between idea units, or questions, and the statistical spaces. Reliability indexes were obtained only for the classification of idea units from general text and examples, and for the classification of questions from the exercises because the comparison between the conditions of this study was based exclusively on that classification. No reliability was calculated for the relationship between redundant and non-redundant idea units, nor for the relationship between redundant ideas and statistical spaces.

Two independent coders classified 59.2% of the idea units in general text and examples, and 49.5% of the exercises' questions. Both coders classified the items into data analysis, sampling and inference using for that the idea units and questions list and the definitions used to classify the redundant idea units in section 3.4.2.4. The agreement rate for both coders was 85.7% for the idea units, and 92.7% for the exercises' questions. A similar process was conducted for the item classification. The classification of items (Forth column from left to right in table 5) in the three statistical spaces was conducted independently by two coders using the definitions in section 3.4.2.5. The agreement rate was 81.2%.

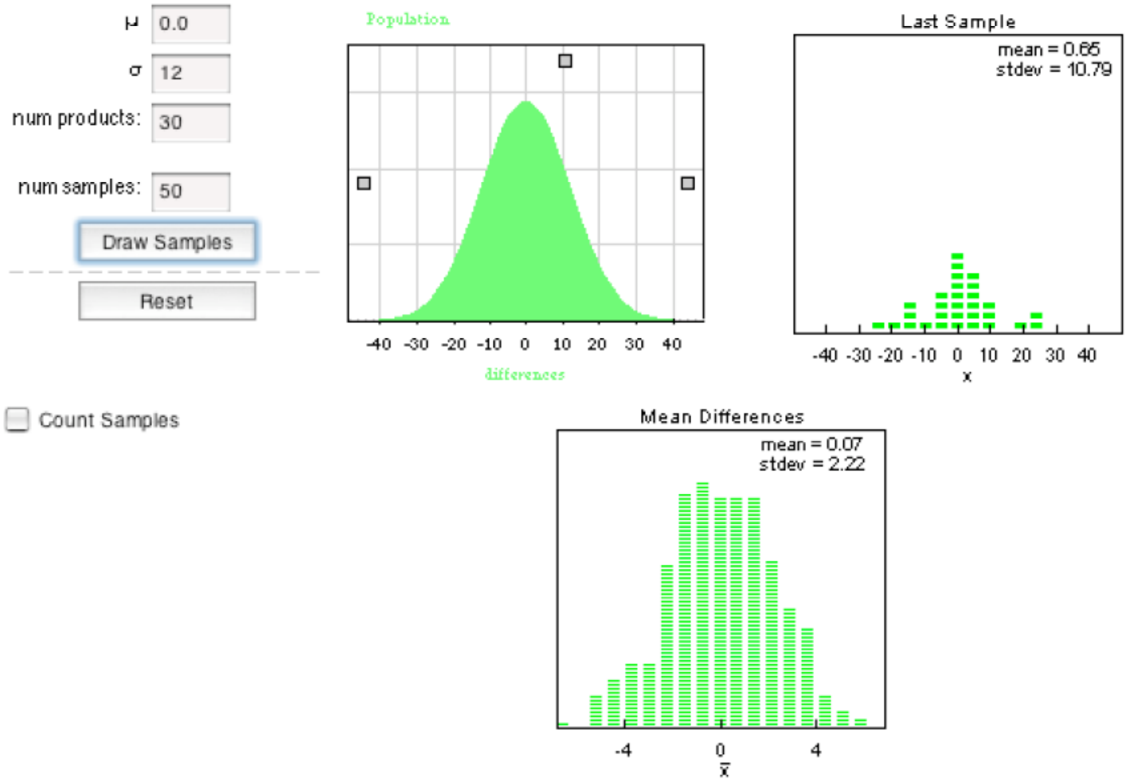
7. Description of Intervention 1: Data Analysis

The Data Analysis condition required students to go through an instructional experience that combined the sampling distribution and ANOVA sections of the OLI course (Open Learning Initiative) (Appendix E). The OLI course is part of the Open Learning Initiative funded by the Hewlett Foundation. This initiative aims to produce several online courses, accessible through Internet, that innovate both in the content and in the instructional means used to teach statistics. The OLI course shows the "big picture" of statistics, that is the relationship among exploratory data analysis (EDA), probability, and statistical inference. The OLI course claims to provide students with several opportunities to explore data sets using computer packages. Both the use of the programs and the interpretation of the results are modeled in the course through several examples.

In the first pages of the sampling distribution section of the OLI course included in this intervention, students learn the concept of sampling distribution and explore the relationship between sample and population in an authentic example. After that, students go through the ANOVA section of the OLI course. This section has seven pages that explain ANOVA's logic and process of hypothesis testing. The first page presents the introduction to ANOVA. In this introduction, ANOVA is regarded as a procedure to evaluate the relationship between a categorical independent variable with two or more groups and a dependent continuous variable. This introduction indicates also that evaluating this relationship implies comparing the group means of the categories defined by the independent variable. Included in the introduction is an example that organizes the explanation across the chapter; the example is a study of the relationship between academic major and frustration scores. The second page explains that ANOVA's F-test works in a different way than other inferential tests because the hypothesis used in ANOVA is not directional. The third page explains the "idea behind the ANOVA F-test" using two pseudo-authentic scenarios (that is, two possible configurations of data for the example mentioned above). The idea of ANOVA, according to this text, is that it compares within- and between- group variances to draw conclusions about the sample mean differences (e.g., "when the variation within group is large (like in scenario #1), the variation (differences) among the sample means could become negligible and the data have very little evidence against Ho"). On the fourth page, the same idea is restated and a short quiz is given. After that, the text continues with an explanation of how the degrees of freedom affect ANOVA's interpretation. Alongside this text, the assumptions of ANOVA are presented and instantiated in an example. On the fifth page, the meaning and location of the *p*-value are explained. Finally, the text provides a very short explanation about how *p*-values can be interpreted in context. The sixth page presents a worked-out example of ANOVA use; this example deals specifically with the relationship between the educational level of a journal and the number of words in its ads. Included in this page is a “learning by doing” exercise that asks students to conduct a complete ANOVA analysis using Excel. The seventh page presents some final comments on the interpretation of the results of ANOVA, specifically on the fact that a significant ANOVA F-test does not specify which groups in particular are producing the differences detected by the ANOVA. Alongside this explanation, the text presents a visual method to detect those differences by comparing the confidence intervals of the group means involved in the ANOVA..

8. Description Intervention 2: Sampling with Simulations

In this condition, students went through an instructional experience built on the same ideas of the OLI course. This intervention asked students to use simulations instead of data analysis to learn ANOVA (Appendix F). Initially, students were asked to use a random-sampling simulator available at http://statweb.calpoly.edu/chance/applets /Shopping/Shopping.html. This random-sampling program allowed them to draw samples from a population while controlling the sample mean and the standard deviation (Figure 1). Students could see the results of a particular sample and the cumulative results of repeated sampling; they could also control the scale of the histograms that presented the results. Numerical results were presented in the top right corner of the histograms. The first activity was to identify the elements in the diagram. A brief explanation was provided in order to map the elements in the graphical display of the Applet onto statistical concepts.



μ 0.0

σ 12

num products: 30

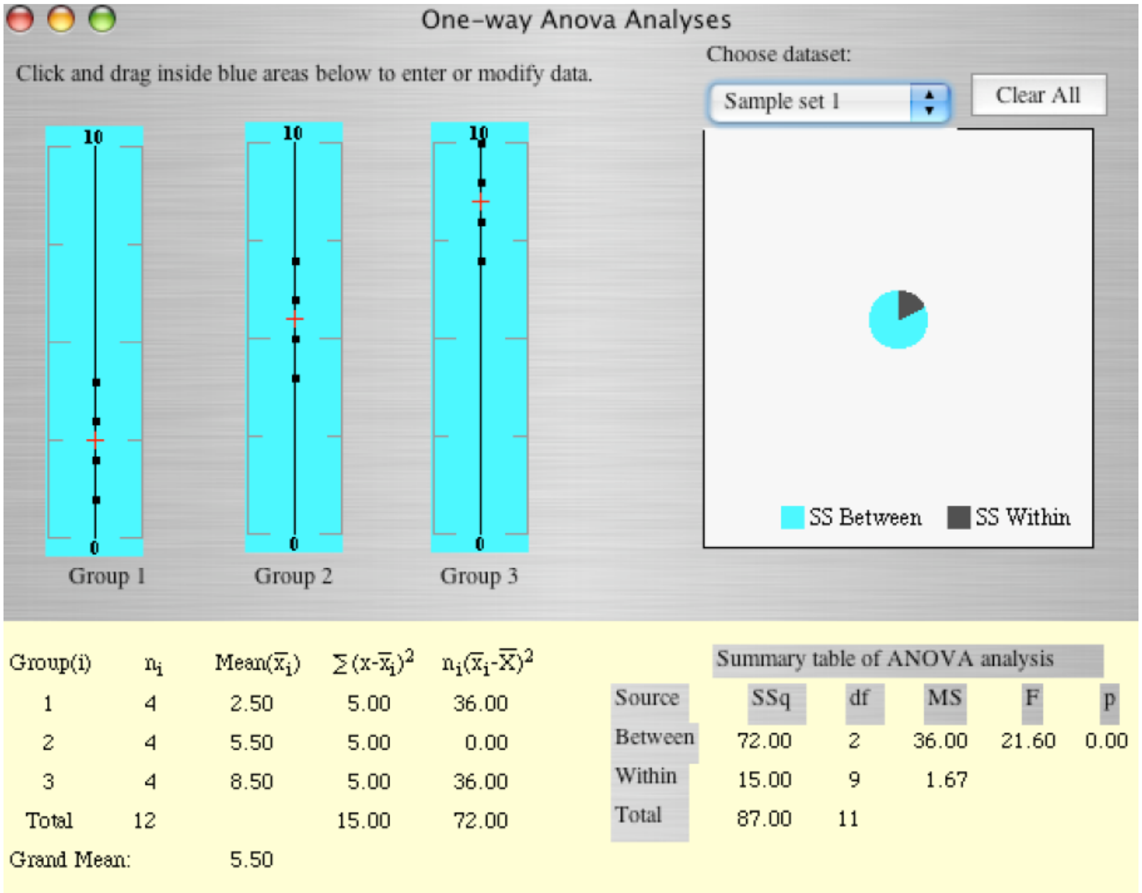
num samples: 50

Figure 1. Applet of the Sampling Distribution of the Mean

To do so, a pseudo authentic example was presented (e.g., "you are trying to find the mean in mathematical reasoning of tenth grade students, but you are unable to evaluate them all. You sample 30 of them; their individual results are presented in the top right diagram") and some questions were asked (e.g. "What does the point in the bottom center diagram mean?") to be sure the students understood and could use the Applet diagram. In order to explain why ANOVA is necessary to evaluate mean differences, students were asked to draw several samples from the population and to write down the mean of some of them. Then they were asked to think about what process is drawing a research sample analogous. Students had to realize that drawing a sample is equivalent to collecting data on a given phenomenon. The instruction provided in this part of the intervention suggested that it is impossible to observe all the possible instances of a given phenomenon, and that therefore sampling is a necessary part of scientific research. This was the point where the sampling intervention introduced questions on the variability of the samples. Students were asked basically why samples obtained from the same population have different sample means, and then they were asked to consider how researchers can be sure of the accuracy of their research conclusions. The instruction then suggested that sampling creates variability but that sampling variability is not unpredictable. Samples vary within certain ranges that depend on the parameters of the population. To prove that, students were asked to draw a large number of samples (*n*=40) and use a sample counter – available in the same Applet-- to quantify the number of samples under or below certain limits. The next step was to extend these conclusions to the case of two populations with different means. Students sampled from each population and had to find out how many sample means fell within a common range for both groups. They, then, repeated the same exercise but this time they sampled from the same population. At this point students were asked to compare the means from the one-population sampling and the means from the two-population sampling. Then, students repeated the same exercises but they played with different parameters (sample size and standard deviation) to establish how those parameters affect the confidence in the observed mean difference. Finally, there was an explanation about how ANOVA helps researchers to identify whether a difference is the result of either systematic effects or random sampling.

At this point, the intervention explained that the idea behind ANOVA is the comparison of within- and between- group variability. To explain this idea, the intervention used two new applets. The first Applet visualized the relationship between explained and unexplained variance in a pie diagram (http://www.ruf.rice.edu/~lane/stat_sim/one_way/index.html). In the right half of the Applet, three vertical lines represented three different groups in an ANOVA (see figure 2).

In each line, black dots represented the data points (individual scores) and a red line represented the mean of the group. Students had seven data sets to work with. Once the data set was uploaded, students could move the points (black slots) in each line. The Applet automatically modified the numerical indicators presented in an ANOVA table in the lower part of the diagram, and also modified the relationship between explained and unexplained variance depicted in the pie diagram.



Click and drag inside blue areas below to enter or modify data.

Choose dataset: Sample set 1

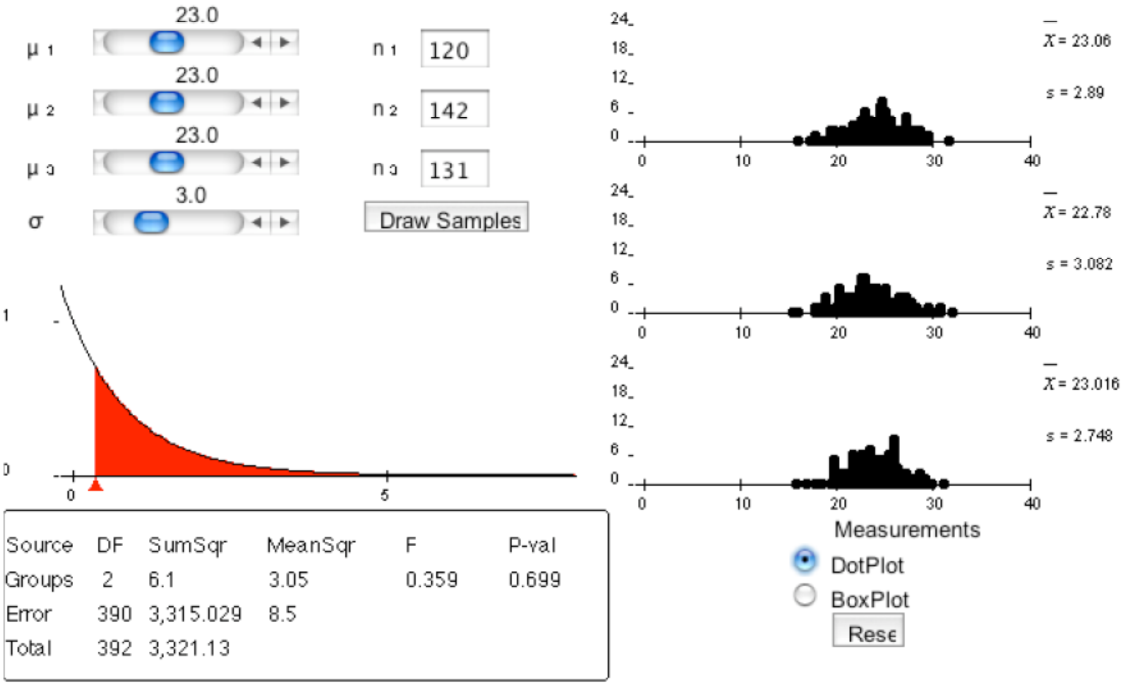
Group(i)	n_i	Mean($\bar{x}_i$)	$\sum(x-\bar{x}_i)^2$	$n_i(\bar{x}_i-\bar{X})^2$
1	4	2.50	5.00	36.00
2	4	5.50	5.00	0.00
3	4	8.50	5.00	36.00
Total	12		15.00	72.00
Grand Mean:		5.50		

Summary table of ANOVA analysis

Source	SSq	df	MS	F	p
Between	72.00	2	36.00	21.60	0.00
Within	15.00	9	1.67		
Total	87.00	11			

Figure 2. First ANOVA Applet

The second Applet used in the intervention to illustrate the idea behind ANOVA came from the Rossman and Chance Applet collection. This collection is available at http://www.rossmanchance.com/applets/Anova/Anova.html. This Applet permitted students to sample from three groups while controlling population and sample parameters. Students controlled the means of the three populations, and the size and standard deviation of the samples. The results for each group could be displayed as histograms (dotplots) or as boxplots (see Figure 3). Numerical results were presented as an ANOVA table in the bottom of the Applet. The probability of obtaining each sample was displayed as a red band in an ANOVA distribution graph. Sampling results were not accumulated in any graph of this Applet.



μ1 23.0
μ2 23.0
μ3 23.0
σ 3.0

n1 120
n2 142
n3 131

$\bar{X}$ = 23.06 s = 2.89
$\bar{X}$ = 22.78 s = 3.082
$\bar{X}$ = 23.016 s = 2.748

Source	DF	SumSqr	MeanSqr	F	P-val
Groups	2	6.1	3.05	0.359	0.699
Error	390	3,315.029	8.5		
Total	392	3,321.13			

Measurements
● DotPlot
○ BoxPlot

Figure 3. Second ANOVA Applet

The second part of this intervention asked students to use the first Applet to visualize the relationship between explained and unexplained variance in a small data set. Students were asked to move the points in the graph to obtain different amounts of explained variability. They had to find out which data configuration produced larger explained variance results. After that, students had to explain the meaning of the indicators displayed in the ANOVA table by finding the relationship between different configurations of data in the graph and different results in the F and *p*-values. Students had also to find out the effects of sample size and standard deviation on F and *p*-values, by changing them while keeping constant the relationship between explained and unexplained variance.

After that, students were asked to use the third Applet. First they had to draw 10 samples from a situation where the populations had small mean differences. Then, they had to draw 10 samples from a situation where the populations had large mean differences. Participants should explore the difference between the configurations of data in those two situations and find out that large mean differences produced higher levels of explained variance in the Applet. The next step asked students to play with sample size and variability (standard deviation) to see how those parameters affected the obtained samples. The final question in this exercise was about the meaning of the *p*-value in this context according to the result of the simulations.

APPENDIX H

Motivation

Motivation was defined as the disposition to act and persevere towards a goal (Svinicki, 1994). This definition was used to elaborate a small questionnaire of six items that asked students to evaluate their own effort during the interventions, and the perceived usefulness of the activity. This variable was introduced to control for the possible effects of motivation on learning during the study and it was obtained at the end of the intervention. Previous research indicates that motivation plays an important role in the process of learning and in the performance of students during testing situations, face-to-face instruction, and online learning (Pintrich & Schunk, 1996). The items in this instrument were of three types. The first three items evaluated students' concentration, effort and dedication. These items measured the perceived performance of students, under the assumption that students with low levels of dedication to the task (e.g., skipping exercises; writing random answers) would report that situation, under no academic or social pressure. This type of item has been used before in questionnaires evaluating motivation in testing situations (Sundre & Moore, 2002). The second type of item assessed students' opinion of the task; that is, whether or not students considered the task worthy. Task value is a common measure of students' motivation, and it relates to performance in face-to-face and online activities (Artino, 2008; Schunck, 2005). The last item of the scale asked students to evaluate their learning during the study. Self-efficacy is another common measure of motivation, and it predicts engagement and learning in online environments (Pintrich & De Groot, 1990).

HOLISTIC MODELS FOR STATISTICAL KNOWLEDGE PRE, POSTTEST CHANGE

The same analysis was conducted on the sampling scores (Table 2). Four different models were tested. The more complex model (Model 1) included pre-post change as within subjects factor, setting and intervention as between subjects factor, and motivation and completion as covariates. Two intermediate levels were built by introducing motivation or completion as sole covariates (Model 2 and Model 3). The simplest model included just intervention and setting as explanatory variables (Model 4).

Table 2. Holistic Models for Sampling Knowledge

	Model 1			Model 2			Model 3			Model 4		
Source	SS	F	Sig.	SS	F	Sig.	SS	F	Sig	SS	F	Sig
Occasion	.05	1.20	.27	.02	.42	.51	.02	.54	.46	2.03	40.52	.00
Occasion*Completion	.03	.68	.41	.07	1.45	.23	NA	NA	NA	NA	NA	NA
Occasion*Motivation	.32	7.01	.01	NA	NA	NA	.36	7.93	.00	NA	NA	NA
Occasion*Intervention	.41	9.00	.00	.41	8.38	.00	.42	9.12	.00	.42	8.45	.00
Occasion*Setting	.10	1.14	.32	.02	.24	.78	.12	1.33	.27	.02	.28	.75
Occasion*Intervention* Setting	.03	.42	.65	.02	.22	.80	.05	.54	.58	.00	.09	.91
Error	3.51			3.84			3.55			3.91		

The results for the sampling scores behaved similar to the results for the global scores. In the simplest model, pre-post change and the interaction of pre-post change with intervention were significant. The introduction of motivation or completion as covariates made pre-post change non significant; in the case of motivation, this effect was caused by the significant relationship of motivation with pre-posttest change. In the case of completion, this effect was produced by the various effects of this variable with the other predictors.

3. Holistic Models for Data Analysis Knowledge

Similar models were calculated for the data analysis score (Table 3). These models showed that the pre-post test data change was not significant under any circumstance. Neither the introduction of completion, nor motivation as covariates created significant effects in the pre-post change. These variables did not have significant effects on the trajectories of the different groups of participants from pretest to posttest. This set of results points out that the change in data analysis scores was small and no significant interaction existed with pre- post change.

Table 3. Holistic Models for Data Analysis Knowledge

	Model 1			Model 2			Model 3			Model 4		
Source	SS	F	Sig.	SS	F	Sig.	SS	F	Sig	SS	F	Sig
Occasion	.14	2.22	.14	.05	.77	.38	.06	.93	.33	.15	2.26	.13
Occasion*Completion	.08	1.33	.25	.12	1.86	.17	NA	NA	NA	NA	NA	NA
Occasion*Motivation	.10	1.65	.20	NA	NA	NA	.14	2.19	.14	NA	NA	NA
Occasion*Intervention	.12	1.97	.16	.13	1.96	.16	.13	2.02	.15	.13	2.01	.16
Occasion*Setting	.02	.17	.84	.00	.01	.98	.04	.31	.73	.00	.07	.93
Occasion*Intervention* Setting	.13	1.06	.34	.17	1.32	.27	.18	1.37	.26	.20	1.49	.23
Error	4.96			5.07			5.04			5.19		

4. **Holistic Models for Inference Scores**

Models explaining the change from pretest to posttest in the inference scores were difficult to interpret (Table 4). This situation was in part due to the weak effects of the explanatory variables on the change of inference scores, and also to the crossed relationships among predictor variables in the model. The results can be summarized the in the following way: In the simplest model (model 4) the pre, posttest change was highly significant, the effect of setting on this change was moderately significant, and the effect of intervention was non-significant. When motivation was introduced only the interaction of motivation with pre-post change was significant. When completion was introduced and motivation taken out, no variable had a significant effect.

Table 4. Holistic Models for Inference Knowledge

	Model 1			Model 2			Model 3			Model 4		
Source	SS	F	Sig.	SS	F	Sig.	SS	F	Sig	SS	F	Sig
Occasion	.00	.01	.89	.23	3.62	.06	.08	1.39	.24	1.17	18.03	.00
Occasion*Completion	.08	1.38	.24	.03	.53	.46	NA	NA	NA	NA	NA	NA
Occasion*Motivation	.45	7.56	.00	NA	NA	NA	.40	6.72	.01	NA	NA	NA
Occasion*Intervention	.24	4.03	.04	.24	3.73	.05	.23	3.93	.05	.24	3.71	.05
Occasion*Setting	.13	1.08	.34	.38	2.91	.06	.16	1.35	.26	.40	3.13	.04
Occasion*Intervention* Setting	.28	2.32	.10	.28	2.19	.11	.22	1.86	.16	.26	2.05	.13
Error	4.58			5.04			4.67			5.08		

Finally, when the complete model was evaluated the interaction between pre, posttest change and motivation was highly significant, and the interaction between pre-post change and intervention was moderately significant. These outcomes indicate that the interaction between motivation and pre-post change was consistently significant for inference scores, and that a moderate relationship between intervention and the trajectories of change was only visible when motivation and completion were introduced in the model. The introduction of the covariates redistributed variance from the pre-post change and put it on the interaction between intervention and pre-post change; at the same time, the introduction of the covariates subtracted unexplained variance from the model. These two facts combined increased the significance of the interaction between intervention and pre-post change. The opposite effect was observed in the relationship between setting variable and pre-post change: the covariates reduced the influence of the interaction between setting and occasion because pre-posttest had a slightly negative relationship with setting. That was so, because students in settings with high pre-post change in inference had slightly lower values in motivation and completion. Overall, only a highly significant relationship between motivation and pre-post change, and a moderate relationship between intervention and change seem to be supported by these results.

APPENDIX I

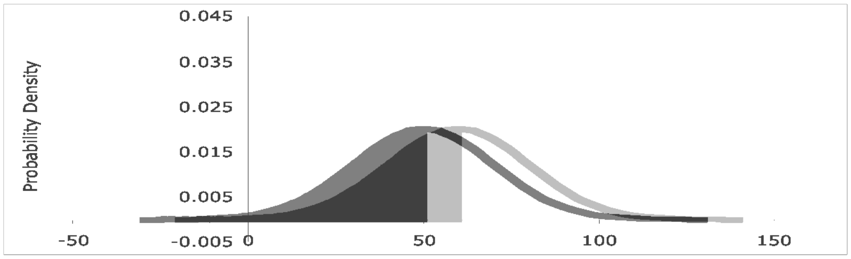
INTERVENTIONS IN SPANISH

Actividades Iniciales

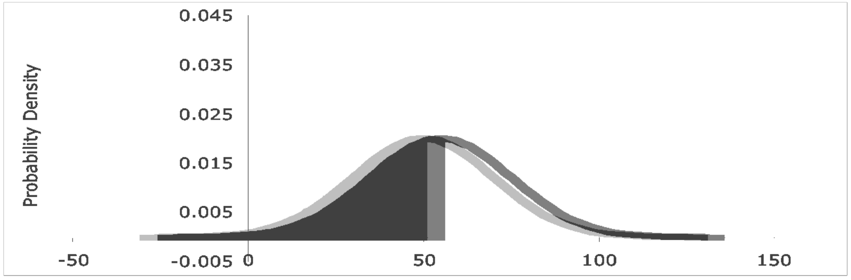
1. En cada uno de los siguientes pares de distribuciones indique con una X cual cree usted presenta una diferencia más significativa? Y explique brevemente su respuesta a la derecha de las distribuciones.

n= el tamaño de los grupos comparados.

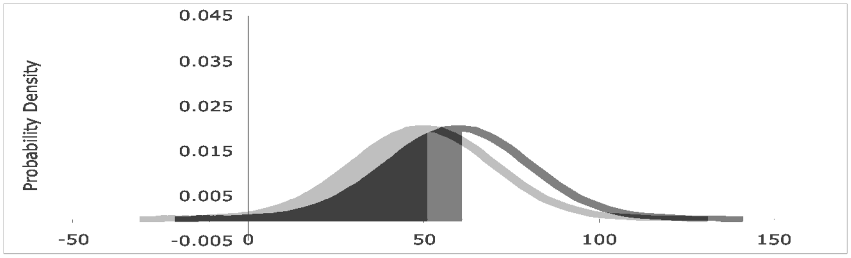
A)
n=50



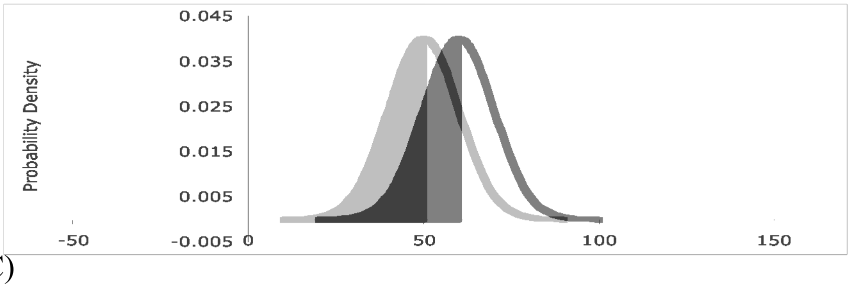
n=50



B)
n=100

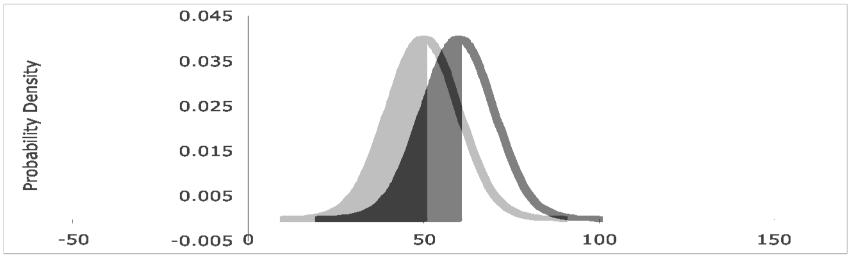


n=100

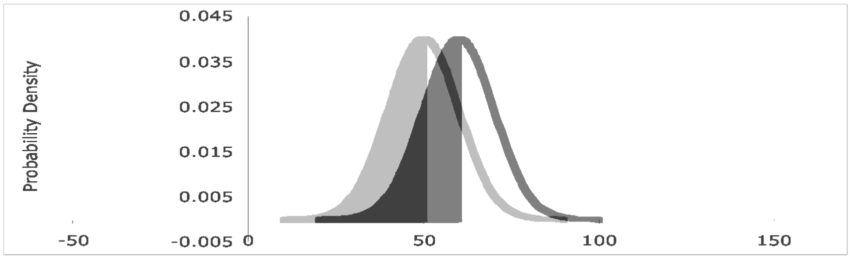


C)

n=100



n=50

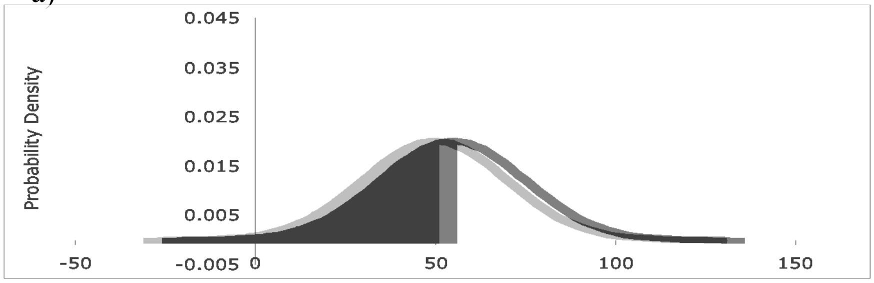


2- A continuación se presenta un caso. Usted debe identificar que gráficos corresponden a los resultados del caso.

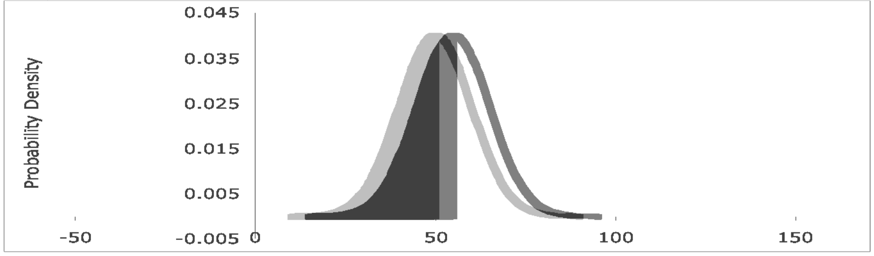
En busca de diferencias significativas en razonamiento matemático entre los estudiantes de noveno y décimo grado, un grupo de investigadores evaluó una muestra de estudiantes en dos colegios. La diferencia entre los estudiantes de noveno y décimo fue la misma para los dos colegios. Sin embargo, los tests estadísticos mostraron que la diferencia era significativa para el colegio A pero no para el colegio B.

A) A continuación usted encuentra dos pares de distribuciones. Uno corresponde al colegio A y el otro corresponde al colegio B. Su tarea es identificar cual es cual.

a)



b)



B) Otra posibilidad para explicar los resultados es que:

a) El colegio A tenga una muestra el doble de grande que el colegio B.
b) El colegio B tenga una muestra el doble de grande que el colegio A.

C) Cual de las siguientes dos tablas de ANOVA corresponde al colegio A?
a)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2813.604	1	2813.604	23.670	.000
Within Groups	11530.017	97	118.866		
Total	14343.621	98			

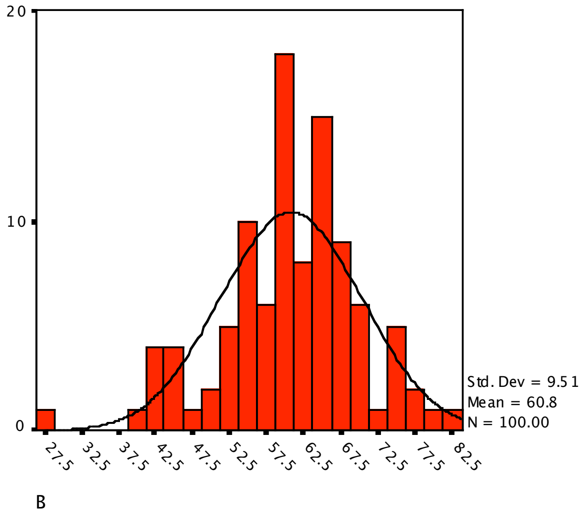
b)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	962.908	1	962.908	2.389	.125
Within Groups	39098.718	97	403.080		
Total	40061.626	98			

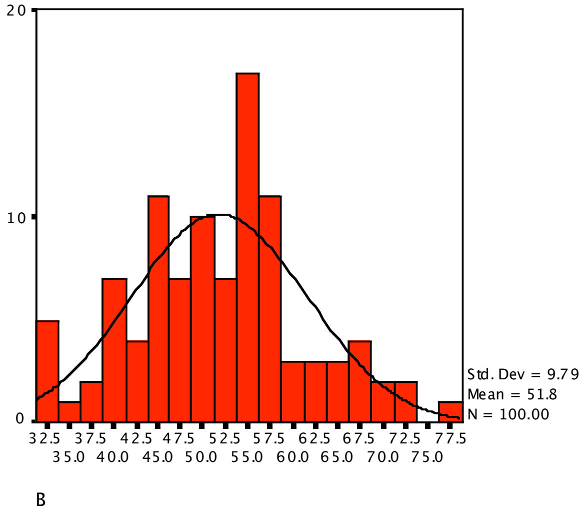
3. Lea cuidadosamente el siguiente caso y responda las preguntas al final.

Un grupo de estudiantes, como parte de su formación en métodos, decidió recolectar datos y realizar un macroestudio. Ellos eligieron comparar el numero de pulsaciones por minuto para hombres y mujeres. Ellos consiguieron medir las pulsaciones por minuto de 100 hombres y 100 mujeres. A continuación usted encuentra las graficas representando las distribuciones de datos para hombres y para mujeres.

Hombres



Mujeres



Escriba una breve interpretación de los resultados obtenidos.

____

____

____

____

____

Si usted utiliza un test estadístico cree que encontrará diferencias significativas? ______
Porque? ____
____

Preguntas Abiertas

A continuación usted encontrará una serie de preguntas respóndalas brevemente en el espacio indicado.
a) Cual es la diferencia entre variabilidad entre grupos (explicada) y variabilidad dentro de los grupos (no explicada)?

____
____
____

b) Cuando usted toma muestras de dos grupos –por ejemplo, toma una muestra de 30 hombres y 30 mujeres para compararlos en algún atributo- las diferencias que usted encuentra pueden ser debidas a que factores?

____
____
____

c) Si usted toma una muestra de estudiantes de un curso de estadística y obtiene una media para cierta variable, y después toma otra muestra al azar del mismo curso, las medias de las dos muestras van a ser las mismas? Porque?

____
____
____

d) Cual es el efecto del tamaño muestral sobre la significancía de una diferencia de medias?

____
____
____

e) Cual es el efecto de la variabilidad interna de dos grupos en el proceso de comparar las medias de estos grupos?

____
____
____

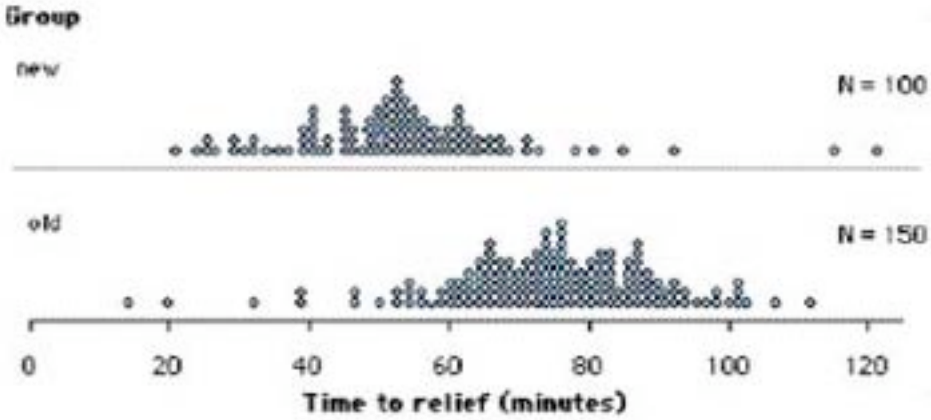
Dibuje dos distribuciones de datos y señale donde esta la variabilidad entre grupos, la variabilidad dentro de los grupos y la variabilidad producto del proceso de muestreo?

____
____
____

Selección Múltiple

Responda los siguientes ítems.

Una compañía productora de drogas desarrolló una nueva fórmula para su remedio contra el dolor de cabeza. Para testear la efectividad de esta nueva fórmula, 250 personas fueron seleccionadas aleatoriamente de la población general de pacientes con dolor de cabeza. 100 personas fueron asignadas al azar al grupo que recibiría la nueva formula (new). Las otras 150 personas recibieron la medicación que contiene la formula antigua (old). El tiempo que tardo la droga en eliminar los síntomas del dolor de cabeza fue anotado para cada paciente. Los resultados de estas pruebas se presentan en la parte inferior. Las preguntas 1, 2 y 3 presentan las conclusiones de tres estudiantes sobre los resultados del estudio. Usted debe indicar si la conclusión es valida o no.



1*. La formula antigua funciona mejor. Dos personas que tomaron la formula antigua se sintieron mejor en menos de 20 minutos; por el contrario, ninguna persona que tomo la nueva formula sintió alivio en menos de 20 minutos. También, el peor resultado –casi 120 minutos- sucedió con la nueva formula.
a) Valida
b) No valida.

2*. El tiempo promedio que requiere la nueva formula para producir efecto es menor que el tiempo de la formula antigua. Yo concluiría que las personas tomando la nueva formula tienden a sentir alivio aproximadamente 20 minutos más rápido que aquellos tomando la formula antigua.
a) Valida.
b) No valida.

3*. Yo no concluiría nada porque el numero de personas en los dos grupos no es igual, entonces no tiene sentido comparar los dos grupos.
a) Valida.
b) No valida.

Un investigador en una ciencia ambiental está realizando un estudio para investigar el impacto de un herbicida particular en una especie de peces. El tiene 60 peces saludables y los asigna aleatoriamente a dos condiciones: un grupo control o un grupo experimental. El grupo experimental estuvo en contacto con el herbicida por un periodo de 10 días. El grupo experimental mostró niveles mas altos de una encima maligna.

4*. Suponga que un test de signficancia fue realizado correctamente en los datos y mostró que no había diferencias significativas entre las medias de los grupos experimental y control en la encima indicada. Que conclusión se puede obtener de estos resultados?

a. El investigador no debe estar interpretando los resultados correctamente; debe haber una diferencia significativa.
b. El tamaño de la muestra puede ser muy pequeño para detectar diferencias estadísticas significativas.
c. Debe ser verdad que el herbicida no causa niveles más altos de la encima maligna.

5*. Suponga que un test de significancia fue realizado correctamente y mostró una diferencia significativa en el promedio de la enzima entre el grupo control y el grupo experimental. Que conclusión se puede obtener?
a. Hay una asociación evidente, pero no una relación causal entre el herbicida y los niveles de la encima.
b. El tamaño de la muestra era muy pequeño como para obtener una conclusión valida.
c. El investigador ha probado que el herbicida causa niveles más altos de la encima maligna.
d. Hay evidencia que el herbicida causa niveles más altos de la encima para esta especie de peces.

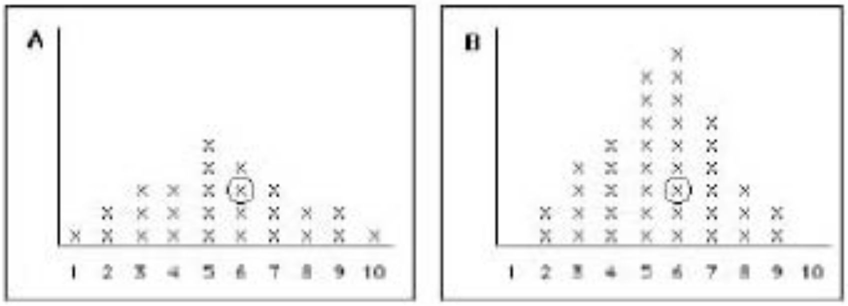
6*. En un curso de geología, los estudiantes aprendían como usar una balanza para hacer predicciones precisas de los pesos de diferentes rocas. Un estudiante planea pesar una roca veinte veces y calcular el promedio de las 20 medidas para calcular el verdadero peso de la roca. Otro estudiante planea pesar la roca 5 veces y calcular el promedio de las 5 medidas para estimar el peso verdadero de la roca. Que estudiante tiene mas probabilidad de obtener un estimativo muy cercano al verdadero peso de la roca?
a. El estudiante que peso la roca 20 veces.
b. El estudiante que peso la roca 5 veces.
c. Ambos estudiantes tiene la misma probabilidad de encontrar el verdadero peso de la roca.

7*. Considere dos poblaciones en el mismo estado. Ambas poblaciones tienen el mismo tamaño muestral (22.000). La población 1 está conformada por estudiantes de una universidad estatal. La población 2 está conformada por todos los residentes de un pueblo pequeño. Considere la variable Edad. Que población debería tener la distribución estándar más alta?
a. La población 1 debe tener una distribución estandar más alta.
b. La población 2 debe tener una distribución estandar más alta.
c. Ambas poblaciones deben tener la misma desviación estandar porque ambas tienen el mismo tamaño muestral.
d. No hay suficiente información para resolver esta pregunta.

8*. Usted intenta evaluar el nivel de motivación por el aprendizaje en los estudiantes de una universidad. Primero obtiene 30 muestras de 10 estudiantes cada una y calcula los promedios de esos grupos. Después obtiene 30 muestras de 50 estudiantes y calcula los promedios de los

grupos. En cual de los dos grupos de muestras, usted tiene más probabilidad de obtener el valor mas alto y el valor mas bajo?
a) en las muestras de 10 estudiantes
b) en las muestras de 30 estudiantes
c) la probabilidad es igual en ambos casos
d) no se puede saber

9*. La figura A representa los pesos de una población de 26 canicas. La figura B representa los pesos promedio de varias muestras de 3 canicas obtenidas de la población representada en la figura A. Un valor está encerrado en un circulo en cada distribución. Hay alguna diferencia entre lo que es representado por el X encerrado en la figura A, y lo que es representado por el X encerrado en el circulo en la figura B. Seleccione la mejor respuesta.

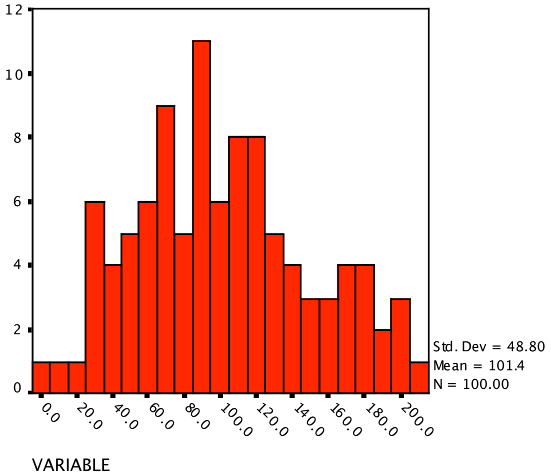


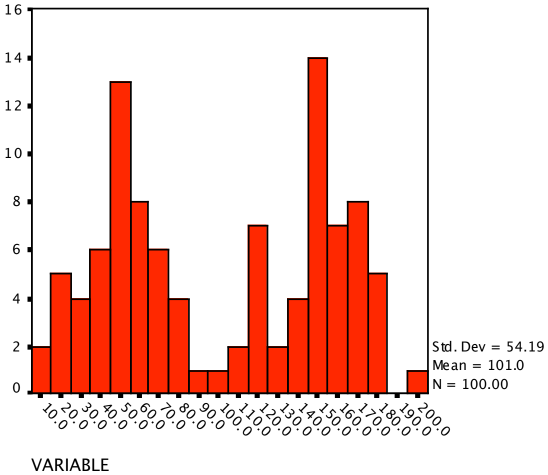
a. No, en la figura A y en la figura B, el X representa una canica que pesa 6 gramos.
b. Si, en la figura A hay un rango mas amplio de valores que en la figura B.
c. Si, la X en la figura A es el peso de una sola canica, mientras el X en la figura B representa el peso promedio de 3 canicas.

*Ítems 1 to 6 were taken from the CAOS test as explained in Appendix D.

Análisis de Datos

1. A continuación usted encuentra dos graficas.





Esas graficas corresponden a una comparación en la estatura de los estudiantes de dos colegios. La grafica de la parte superior corresponde a un colegio femenino, la grafica de la parte inferior corresponde a un colegio mixto. Ambas graficas representan submuestras de 100 estudiantes seleccionados al azar en esos colegios.

Una ANOVA fue llevada a cabo y produjo los siguientes resultados.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7.311	1	7.311	.003	.958
Within Groups	526506.800	198	2659.125		
Total	526514.111	199			

Escriba una breve interpretación de los resultados:

Usted cree que hay alguna forma de mejorar la comparación y producir resultados más significativos en la comparación?

____

____

Si usted toma otras muestras de estudiantes al azar de estos colegios, cree usted que obtendrá los mismos resultados?

____

____

____

____

2.

Evalúe su nivel de concentración en este ejercicio de 1 a 10 (siendo 1 bajo, 10 alto): ___
Evalué su nivel de esfuerzo en este ejercicio de 1 a 10 (siendo1 bajo, 10 alto): ___
Evalúe su nivel de dedicación a esta actividad de 1 a 10 (siendo 1 bajo, 10 alto): ___
Evalué de 1 a 10 que tan interesante encuentra usted esta actividad: ___
Evalué de 1 a 10 la metodología de esta actividad ___
Evalué de 1 a 10 que tanto cree usted que aprendió en esta actividad: ___

BIBLIOGRAPHY

Abelson , P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence. Erlbaum Associates.

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of Ph.D. programs in North America. *American Psychologist*, 45, 721-734.

Albert, J. H. (1993). Teaching Bayesian Statistics using sampling methods and MINITAB. *The American Statistician*, 47, 182-191.

Alldredge, R. & Som, N. (2002). Comparison of multimedia educational materials used in an introductory statistical methods course. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Artino, A. R. (2008). Motivational beliefs and perceptions of instructional quality: predicting satisfaction with online training*.* *Journal of Computes Assisted Learning*, 24(3).

Bakker, A. (2002). Route-type and landscape-type software for learning statistical data analysis. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Batanero, C. Godino, J. Vallecillos, A. Green, D. & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics Education in Science and Technology*, 25(4), 527-547.

Ben-Zvi, D. (2000). Towards Understanding the Role of Technological Tools in Statistical Learning. *Mathematical Thinking and Learning*, 2 (1), 127-155.

Ben-Zvi, D. (2002). Seventh Grade Student's sense making of data and data representations. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D. (2004). Reasoning about Variability in Comparing Distributions. *Statistics Educational Research Journal.* 3(2), 42-63.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics*, 45, 35–65.

Belli, G . (2003). *Finding, evaluating, and organizing Internet resources: issues for statistics instruction*. Paper presented at the International Association of Statistical Education Conference on Statistics Education and the Internet, Satellite Conference to the 54th Session of the International Statistical Institute (ISI), Berlin, Germany. Proceedings at http://www.ph-ludwigsburg.de/iase/proceedings/

Biehler, R. (1993). Software tools and mathematics education: The case of statistics, In C. Keitel & K. Ruthven (Eds.), *Learning from computers: Mathematics education and technology*. Berlin:Springer.

Biehler , R. (1995). Probabilistic thinking, statistical reasoning, and the search of causes: Do we need a probabilistic revolution after we have taught data analysis? *Newsletter of the international study group for research on learning probability and statistics*, 8(1).

Biehler, R. (2003). Interrelated learning and working environments for supporting the use of computer tools in introductory courses. In: International Statistical Institute (Ed.) CD-ROM *Proceedingsof IASE Satellite conference on Teaching Statistics and the Internet*, Berlin, Max-Planck-Institute for Human Development.

Blejec, A. (2002), Teaching statistical concepts with simulated data. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Boaler, J., & Greeno, J. G. (2000). Identity, agency, and knowing in mathematical worlds. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning* (pp. 171-200). Stanmford, CT, Ablex.

Braun, W. J. (1995). An illustration of bootstrapping using video lottery terminal data. *Journal of Statistics Education.* 3(2).

Brooks, D. W., Schraw, G., & Crippen, K.J. (2005) Performance -related feedback: The hallmark of efficient instruction . *Journal of Chemical Education*. 82, 641-644.

Bryc, W. (1999). Decoding a scrambled text: A hands-on project to illustrate sampling and variability. *Journal of Statistics Education*. 7(2).

Burgess, T. (2002). Investigating the 'data sense' of pre-service teachers. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Burrill, G. (2002). Simulation as a tool to develop statistical understanding. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Cai, J. & Gorowara, C. C. (2002). Teachers' conception and constructions of pedagogical representations in teaching arithmetic average. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education*, 29(1), 3-20.

Carr, R. (2002). A data analysis tool that organizes analysis by variable types. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Chafe, W. L. (1985). Linguistic differences produced by differences between speaking and writing. In D.R. Olson, N. Torrace, & A. Hildyard, A. (Eds.), *Literacy, Language and Learning. The Nature and Consequences of Reading and Writing.* Cambridge: Cambridge University Press, p 105-123.

Chance, B., DelMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi and J.Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer.

Chi, M. T. H. (1990). Memory development. In M. W. Eysenck, A. Ellis, E. Hunt, &P. Johnson-Laird (Eds.), *The Blackwell dictionary of cognitive psychology* (pp. 218-222). Oxford, England: Basil Blackwell.

Chi, M. T. H., & Bassok, M. (1989). Learning from examples via self-explanations. In L. B. Resnick (Ed.), *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser* (pp.251-282). Hillsdale, NJ: Erlbaum.

Cohen ,D. K., Raudenbush, S.W., & Ball, D.L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, 25 (2),. 119-142.

Cobb, G., and Moore, D. S. (1997). Mathematics, statistics and teaching. *The American Mathematical Monthly,* 104 (11), 801-823.

Connor, D. (2002). CENSUSATSCHOOL 2000: Creation to collation to classroom, UK. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Conti, C. Lombardo, E. (2002). The Italian census at school. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Corredor and Leinhardt (2006). Evaluation of statistics: Addendum to american statistician article. *Technical Report Hewlett Foundation*. LRDC.

Cramer, K. & Kamps, U. (2003). *Interactive graphics for elementary statistical education*. Paper presented at the International Association of Statistical Education Conference on Statistics Education and the Internet, Satellite Conference to the 54th Session of the International Statistical Institute (ISI), Berlin, Germany.

Cramer, E. & Neslehova, J. (2003). *(e)Learning the Basics of Probability*. Paper presented at the International Association of Statistical Education Conference on Statistics Education and the Internet, Satellite Conference to the 54th Session of the International Statistical Institute (ISI), Berlin, Germany.

Crandall, J. Dale, T.C. Rhodes, N.C. Spanos, G.A. (1990). The Language of Mathematics: The English Barrier in Labarca, A. & Bailey, L. (eds.) *Issues in L2: Theory as Practice/Practice as theory. Proceedings of the 7th Delaware Symposium, 1985*. Norwood, N.J: Ablex Publishing Co. 129 -150.

Cobb, P. & Hodge, L. (2002). A relational perspective on issues of cultural diversity and equity as they play out in the mathematics classroom. *Mathematical Thinking and Learning*. 4(2&3) 249-284.

Cobb, P., & Hodge, L. (2002b). *Learning, identity, and statistical data analysis*. Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS6), Cape Town, South Africa.

Cumming, G. (2002). Live figures :interactive diagrams for statistical understanding . In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Curtis, D. A., & Harwell, M. (1998). Training doctoral students in educational statistics in the United States: A national survey. *Journal of Statistics Education*. 6 (1).

Darius, P., Schrevens, E., van der Knaap, H., Portier, K., Massonnet, G., Lievens, L., Duchateau, L. and Thas, O (2003). *Using web-based tools for teaching statistical concepts and experimentation skills.* Paper presented at the International Association of Statistical Education Conference on Statistics Education and the Internet, Satellite Conference to the 54th Session of the International Statistical Institute (ISI), Berlin, Germany.

Dede, C. and Lewis. M. (1995). *Assessment of emerging educational technologies that might assist and enhance school-to-work transitions*. Washington, DC: National Technical Information Service.

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).

DelMas, R., Ooms, A., & Garfield, J., (2006). Assessing students’ statistical reasoning. Paper presented at the *Seventh International Conference on Teaching Statistics*. Salvador, Bahia, Brazil: 2-9 July.

DeMoivre, A. (1756). *The Doctrine of chances*. 3rd edition. www.ibiblio.org/chance

Dreyfus, T., & Eisenberg , T. (1996). On different facets of mathematical thinking. In R.J. Sternberg and T. Ben-Zeev (Eds.), *The Nature of Mathematical Thinking (*pp. 253-284). Hillsdale, NJ: Lawrence Erlbaum Associates.

Drier, H. S. (2000). The Probability Explorer: A research-based micro world to enhance children's intuitive understandings of chance and data. *Focus on Learning Problems in Mathematics.* 22(3-4), 165-178

Ebel, R. L. (1954). Procedures of the analysis of classroom tests. *Educational and Psychological Measurement*, 14, 352-364.

Elmore, P. B., and Woehlke, P. L. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987, *Educational Researcher,* 19-20.

Engelhart, M. D. (1965). A comparison of several item discrimination indices. *Journal of Educational Measurement*, 2(1), 69-76

Evans, K., Leinhardt, G., Karabinos, M., & Yaron, D. (2006). Chemistry in the field and chemistry in the classroom: A cognitive disconnect? *Journal of Chemical Education*. 83, 655-661.

Fields, P., & Collins, P. (2005). *An assessment of computer-based learning methodology in teaching and introductory statistics hybrid course*. Paper presented at the 55th session of the International Statistical Institute (ISI). Sidney, Australia, April 5-12, 2005.

Finney, S. J., & Schraw, G. J. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology,* 28, 161 – 186

Finzer , W. & Erickson , T. (2005). Curriculum innovations using census microdata: a meeting of statistics, mathematics and social science. In G. Burrill and M. Camden (Eds.), *Curriculum Development in Statistics Education: International Association for Statistics Education 2004 Roundtable*. Voorberg, the Netherlands: International Statistics Institute

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Boston: D. Reidel Publishing Company.

Ford, M. J. & Forman, E. A. (2006). Learning and instruction in science: Elaborating the design approach. In C. Conrad & R. C. Serlin (Eds), *Sage handbook for research in education: Engaging ideas and enriching inquiry* (pp. 139-155). Thousand Oaks, CA: Sage Publications.

Friedman, L. & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*. 59(2), 127-136.

Friel, S., Bright, G., Frierson, and Kader, G. (1997). A framework for assessing knowledge and learning in statistics (K-8). In I. Gal and J. Garfield (eds.), *The assessment challenge in statistics education* (pp. 55-63). Amsterdam: IOS and Press International Statistical Institute.

Gal, I., Rothschild, K., and Wagner, D. A. (1990). *Statistical concepts and statistical heuristics in school children: Convergence or divergence?* Paper Presented at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts, U.S.A.

Garfield, J. (1995). How students learn statistics. *International Statistical Review*, 63, 25-34.

Garfield, J. (1997). Preface. In. J. Garfield & G. Burrill, (Eds). Research on the role of technology in teaching and learning statistics. *Proceddings of the 1996 International Asociation for Statistical Education Round Table Conference.* International Statistical Institute, Voorburg, The Netherlands.

Garfield, J. (1998). The statistical reasoning assessment: development and validation of a research tool. In Pereira-Mendoza, L. (Ed.) *Proceedings of the Fifth International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute, 781-786.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*. 10 (3).

Garfield, J. and Ahlgren, A. (1988).Difficulties in learning basic concepts in probability and statistics: implications for research. *Journal for Research in Mathematics Education*, 19(1), 44-61.

Garfield, J. & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, 4(1), 92-99.

Gobbo, C., & Chi, M. T. H. (1986). How knowledge is structured and used by expert and novice children. *Cognitive Development*, 1, 221-237.

Godino, J. D., Ruiz, F., Roa, R.. Pareja, J. L. & Recio, A. M. (2003). Analysis of Two Internet Interactive Applets for Teaching Statistics in Schools. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Harner , E. J. & Xue, H. (2003). MyJavaStat: an Environment for Teaching Statistics. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Hawkins, A., Jolliffe, F. and Glickman, L. (1992). *Teaching Statistical Concepts* London: Longman.

Hawkins, A. (1997). Myth-conceptions. In. J. Garfield & G. Burrill, (Eds). *Research on the Role of Technology in Teaching and Learning Statistics*. Proceedings of the 1996 International

Association of Statistical Education Round Table Conference. Voorburg, The Netherlands: International Statistical Institute.

Heid, M. K. (1988). Resequencing skills and concepts in applied calculus in applied calculus using the computer as a tool. *Journal for Research in Mathematics Education*, 19 (1), 3-25.

Hooper, L. (2002). Making Census Count in the Classroom. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*. 19, 265-88.

Jun, L. & Pereira-Mendoza, L. (2003). Misconceptions in probability. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Kahneman, D., Slovic, P. and Tversky, A. (1982), *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology,* 3(3), 430-453.

Kaput, J. (1992). Technology and mathematics education. In Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning*. Macmillan, New York, pp. 515-556.

Kazak, S. & Confrey, J. (2004). *Investigating educational practitioners' statistical reasoning in analysis of student outcome data*. Paper presented at the Tenth international conference of mathematical education. Copenhagen, Denmark. Retrieved at http://www.stat.auckland.ac.nz/~iase/publications.php?show=11

Kersten, T. (1983). Computer simulations to clarify key ideas of statistics. *Two-Year College Mathematics Journal*, 14, 416-421.

Konold, C. (1989), Informal conceptions of probability. *Cognition and Instruction*, 6, 59-98.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education* 3(1).

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Konold, C., Polltsek, A., & Well, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield and G Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 151-168). Voorburgh, The Netherlands: ISD.

Konold , C., Pollatsek, A., Well, A.D., Lohmeier, J., and Lipson, A. (1993). Inconsistencies in Students' Reasoning About Probability. *Journal for Research in Mathematics Education,* 24, 392-414.

Konold , C., Well, A.D., Pollatsek, A., and Lohmeier, J. (1993). Teaching the law of large numbers via dynamic sampling. In J.R. Becker and B. J. Pence (Eds.) *Proceedings of the Fifteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. San Jose: San Jose State University, pp. 299-305.

Larreamendy -Joerns, J.,Leinhardt G. and Corredor J. (2005). Six Online Statistics Courses: Examination and. Review. *The American Statistician*, 59, 240-251.

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics*, 23, 557-568.

Lehrer, R. & Schauble, L. (2002). Distribution: A resource for understanding error and natural variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. Lovett & P. Shah (Eds.), *Thinking with Data.* Mahwah, NJ; Lawrence Erlbaum Associates.

Leinhardt, G. (2001). Instructional explanations: A commonplace for teaching and location for contrast. In V. Richardson (Ed.), *Handbook of research on teaching* (4th Ed., pp. 333-357). Washington, DC: American Educational Research Association.

Leinhardt , G., & Larreamendy -Joerns, J. (2007). Variation in the meaning and learning of variation: In M. Lovett & P. Shah (Eds.), *33rd Carnegie Symposium on Cognition: Thinking with data.* Mawhaw, NJ: Lawrence Erlbaum.

Leinhardt, G., & Leinhardt, S. (1980). Exploratory data analysis: new tools for the analysis of empirical data. In D. Berliner (Ed.) *Review of Research in Education,* Vol. 8, Washington, DC: AERA. 85-157.

Leinhardt, G., & Leinhardt, S. (1983). Exploratory data analysis. In T. Husen, & N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies* (pp. 1294-1303). Oxford: Pergamon Press.

Leinhardt ,G., Zaslavsky ,O. and Stein ,MK.(1990). Functions, graphs and graphing: Tasks, learning and teaching. *Review of Educational Research,* 60 (1).

Lesgold, A. (1999). Multiple representations and their implications for learning. In M. van Someren, E. Boshuizen, P. Reimann & T. de Jong (Eds.), *Learning with multiple representations.* Oxford: Elsevier.

Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician*, 54, 196-206.

Marasinghe, M. (2003). Computer modules for teaching statistical concepts. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

McClain, (2002). Uses of interactive minitools to explore authentic data with teachers to create individual and collective thinking. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Masnick, A. M., Klahr , D., & Morris, B. J. (in press) Separating signal from noise: Children's understanding of error and variability in experimental outcomes. In, Lovett & Shaw (Eds) *Thinking With Data*. Mawah, NJ: Erlbaum.

Meyer, O., & Lovett, M. C. (2002). Implementing a cognitive tutor in a statistical reasoning course: Getting the big picture. In *Proceedings of the Sixth Annual International Conference on the Teaching of Statistics*.

Mills, J. (2002). Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. *Journal of Statistics Education*, 10 (1).

Mittag , H.J. (2003). Interactive visualization for statistics education and for official statistics, in CD-ROM of the *Proceedings of the 53rd ISI session*.

Moore, D. S. (1992), Teaching statistics as a respectable subject. In Gordon, F & Gordon, S. (Eds.)*Statistics for the Twenty-First Century, MAA Notes No. 26*, Washington: Mathematical Association of America, pp. 14-25.

Mori, Y. Yamamoto Y. and Yadohisa, H. (2003): Data -oriented Learning System of Statistics based on Analysis Scenario/Story (DoLStat). In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Moschkovich, J. (in press). Bilingual mathematics learners: How views of language, bilingual learners, and mathematical communication impact instruction. To appear in N. Nassir and P. Cobb (Eds.), *Diversity, equity, and access to mathematical ideas.* Teachers College Press

Nickerson, R. S. (2004). *Cognition and chance. The psychology of probabilistic reasoning.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Nicholson , J. R., Mulhern, G. & Hunt, D. N. (2002). Wizardry or pedagogy? What is the driving force in the use of the new technology in teaching statistics? In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Nisbett, R., Krantz, D., & Jepson, C. (1993). The use of statistical heuristics in everyday inductive reasoning. In Nisbett, R (Ed.), *Rules for reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Oberschall, A. (1987). The two empirical roots of social theory and the probability revolution. In L. Kruger, G. Gigerenzer, & M. S. Morgan (Eds.), *The Probabilistic Revolution: Ideas in the sciences.* (pp 103-131). Cambridge, MA: MIT Press.

Parker, M., & Leinhardt, G. (1995). Percent: A privileged proportion. *Review of Educational Research,* 65 (4), 421-48.

Pesek, D. D., & Kirshner, D. (2000). Interference of instrumental instruction in subsequent relational learning. *Journal for Research in Mathematics Education*, 31(5), 524-540.

Perkins, D. N. (1992). Technology meets constructivism: do they make a marriage? In Duffy, T. M. & Jonassen, D. H. (eds*), Constructivism and the Technology of Instruction: a Conversation.* pp 45-55. New Jersey: Lawrence Erlbaum Associates.

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning*, 5(2&3), 131-156.

Phillips, B. (2003). *Overview of online teaching and internet resources for statistics education.* Paper presented at the International Association of Statistical Education Conference on Statistics Education and the Internet, Satellite Conference to the 54th Session of the International Statistical Institute (ISI), Berlin, Germany. Proceedings at http://www.ph-ludwigsburg.de/iase/proceedings/

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning component of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33-40.

Razi , N., and Hiemenz, B. (2003), *New statistics: Elements for blended learning.* Paper presented at the International Association of Statistical Education Conference on Statistics Education and the Internet, Satellite Conference to the 54th Session of the International Statistical Institute (ISI), Berlin, Germany. Proceedings at http://www.ph-ludwigsburg.de/iase/proceedings/.

Regnier, J. C. (2003). Interrelated learning and working environments for supporting the use of computer tools in introductory courses. International Statistical Institute (Ed.) CD-ROM *Proceedings of IASE Satellite conference on Teaching Statistics and the Internet*, Berlin, Max-Planck-Institute for Human Development.

Renkl , A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology.* 93 (2) 346-362.

Rubin, A. (2002). Interactive visualization of statistical relationships. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Saldanha, L & Thompson, P. (2003). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*. 51: 257-270.

Sanchez, E. (2002). Teachers' Beliefs about usefulness of simulations with the educational software Fathom for developing probability concepts in statistics classroom. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction,* 22(2), 129-184.

Shafer, G. (1993). The significance of Jacob Bernoulli's Ars Conjectandi for the philosophy of probability today. *Bayesian Statistics and Econometrics Conferences,* Basel, Switzerland.

Shaughnessy, J.M. (1992). Research in probability and statistics: Reflections and directions. In Grouws, D. A. (Ed.). *Handbook of Research on Mathematics Teaching and Learning* (pp.465-494). New York: Michigan Publishing Company.

Shaughnessy, M. & Ciancetta, M. (2002). Students' understanding of variability in a probability environments. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Shaughnessy , J. M., Garfield , J., & Greer , B. (1996). Data handling, in A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick & C. Laborde (Eds.), *International Handbook of Mathematics Education*,1, 205-237. Dordrecht, Netherlands: Kluwer

Scheaffer, R., (1995). *Introduction to probability and its applications*. Belmont, CA: Wadsworth Publishing Co.

Schoenfeld, A. H. (1983). Beyon the purely cognitive: Beliefs systems, social cognitions, and metacognitions as driving forces in intellectual performance. *Cognitive Science*, 7, 329-363.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science*, 23 (3), 337-370.

Schwarz, C. & Sutherland, J. (1997). An On-line workshop using a simple capture-recapture experiment to illustrate the concepts of a sampling distribution. *Journal of Statistics Education*. 5(1).

Shunck, D. H. (2005). Self-regulated learning: The educational legacy of Paul R. Pintrich. *Educational Psychologist*, 40, 85-94.

Singley, M. & Anderson, J. (1989). *The transfer of cognitive skill.* Cambridge: Harvard Univ. Press.

Snee , R. (1993). What's missing in statistical education? *The American Statistician*. 47, 149-154.

Snir, J. Smith, C. and Grosslight, L (1995) Conceptually enhanced simulations: A computer tool for science teaching. In Perkins, D. (ed.) *Software goes to school*. Oxford U. Press.

Stanton, J. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, 9(3).

Steffensen, M. S., Joag-dev, C., & Anderson, R. C. (1979). A cross-cultural perspective on reading comprehension. *Reading Research Quarterly*, 15, 10-29.

Stein, M.K., and Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, 2(1), 50–80.

Stigler, S. (1986). *The History of Statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stirling, D. (2003), Interactive content in web pages teaching statistics. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update,* 14(1), 8-9.

Svinicki, M. D. (1994). Research on college student learning and motivation: will it affect college instruction? In P. Pintrich, D. Brown, & C. E. Weinstein (Eds.) Student Motivation, Cognition and Learning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tall, D. (2001). Cognitive development in advanced mathematics using technology. *Mathematics Education Research Journal*. 12 (3), 196–218.

Tappin, L. A. (2000). Statistics in a nutshell. *Journal of Statistic Education*. 8 (1).

Taylor, R. P. (1980). Introduction. In R. P. Taylor (Ed.), *The computer in school: Tutor, tool, tutee* (pp. 1-10). New York: Teachers College Press.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Vargas, E. & Larreamendy, J. (2006). *Implementation of the OLI Statistics courseware at Universidad de los Andes (Colombia*). Technical Report Hewlett Foundation. LRDC.

Velleman , P. F. and Moore, D. S. (1996). Multimedia for teaching statistics: Promises and pitfalls. *The American Statistician*, 50, 217-225.

Watson, J. M. (2002) Creating cognitive conflict in a controlled research setting: sampling. In: Phillips, Brian. (ed.). *ICOTS 6* : 7-12 July 2002 Cape Town, South Africa.

Watson, J. M. (2002). Inferential Reasoning and the Influence of Cognitive Conflict. *Educational Studies in Mathematics*, 225-256.

Watson, J. M. &, Moritz, J. B. (1999). The Beginning of Statistical Inference: Comparing Two Data Sets. *Educational Studies in Mathematics*, 37: 145-168.

Watson, J. & Moritz, J. (2000) The longitudinal development of understanding of average. *Mathematical Thinking and Learning*, 2, 11-50.

West, R. W., and Ogden, R. T. (1998). Interactive demonstrations for statistics education on the World Wide Web. *Journal of Statistics Education*, 6(3).

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67 (3), 223-265.

Wilensky, U. & Stroup, W. (1999). Participatory simulations: Network-based design for systems learning in classsrooms. *Proceeding of the Conference on Computer-Supported Collaborative Learning*, Stanford University.

Williams, A. N. (1998). Students' understanding of the significance level concept. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee & W. Wong (Eds.) ,*Proceedings of the ICOTS5*, 743-749. Nanyang Technological University, Singapure.

Wisenbaker, J. (2002). A Personal journey toward a virtual introductory statistics course: Not (quite) ready for prime time. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Wisenbaker, J. (2003). Extending the journey toward a virtual introductory statistics course. In *Proceedings of IASE Satellite: statistics & the Internet [CD-ROM],* Berlin, Germany. International Statistics Institute.

Woolley, K. K. (1997). How variables uncorrelated with the dependent variable can actually make excellent predictors: The important suppressor variable case. Paper presented at the *Annual Meeting of the Southwest Educational Research Association*, Austin, January, 1997.

Wood, M. (2005). The role of simulation approaches in statistics. *Journal of Statistics Education*, 13 (3).

Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction*, 4, 137-166.