Simulation Assisted Learning in Statistics:

How Important are Students' Characteristics?

William Nilsson (Corresponding author)
Department of Applied Economics, University of the Balearic Islands
Ctra Valldemossa Km 7,5, E-07122 Palma de Mallorca, Spain
Phone: +34 971 17 13 77, Fax: +34 971 17 23 89, E-mail: william.nilsson@uib.es

Tomás del Barrio Castro
Department of Applied Economics, University of the Balearic Islands
Ctra Valldemossa Km 7,5, E-07122 Palma de Mallorca, Spain
Phone: +34 971 17 32 56, Fax: +34 971 17 23 89, E-mail: tomas.barrio@uib.es

Abstract: A simulation assisted learning method is introduced to deal with students' misconceptions concerning the properties of estimators; bias, efficiency and consistency. The design of the exercise is based on cognitive conflict theory, i.e. the simulations provide contradictory information to obtain a conceptual change and induce students to abandon misconceptions. The evaluation of the intervention is done by a pre-test and both a post-test and a delayed post-test. As predicted by the cognitive conflict theory, individual characteristics, such as values and attitudes towards learning and passing a prior mathematical course, are important to obtain a meaningful cognitive conflict.

Keywords: Monte-Carlo simulation, misconception, cognitive conflict.

Author note:
This work was partially supported by Spanish Ministry of Science and Innovation (grant # ECO2010-20829). The authors thank Xisca Amengual, CTI, for retrieving the students' answers in the delayed post-test, (in addition to the test score) in the final exam. The authors are also thankful to seminar participants at the Second Lisbon Research Workshop on Economics, Statistics and Econometrics of Education, January 2013 and at the University of the Balearic Islands, as well as students and teachers that participated in the project.

1. INTRODUCTION

The process of acquiring knowledge is not necessarily straight forward and depending on the topic, students can develop serious misconceptions. New information can be misunderstood and concepts can be too simplified in a way that the original idea is lost. A key feature of cognitive conflict theory is to identify the misconception and provide anomalous data or contradictory information to obtain a conceptual change. Limón (2001) provides an excellent review of the literature on cognitive conflict theory. An important finding is that, despite results of some positive effects, it seems to be difficult to achieve a strong conceptual change. In some cases, the new information is simply assimilated and a cognitive conflict is never achieved. In other cases only a partial change is achieved, but this change eventually disappears after some time of the intervention. Limón (2001) argues that earlier literature has been too focused on the individual's cognitive process, and thereby leaving out important factors related to individual characteristics. To obtain a meaningful conflict many variables should be considered. For example, students should be motivated, have a certain amount of prior knowledge and have reasoning abilities. It is possible that the students do not realize that the provided information is contradictory at all, or simply, do not find the exercise interesting. In some cases the information is considered contradictory, but the students are unable to realize its implication for their current beliefs.

Many basic concepts in statistics are highly abstract, and Watts (1991) found that this is the main reason for considering statistics relatively difficult compared to other subjects. Simulation assisted learning provides a method to overcome these difficulties and students can, for example, clearly distinguish a parameter estimate from its corresponding parameter in the population.

Liu, Lin & Kinshuk (2010) use Simulation-Assisted Learning Statistics (SALS) in a study of 72 students. The experiment is based on the cognitive conflict theory, and the purpose is to correct misconceptions concerning the correlation coefficient. The results indicate that the method is effective to correct misconceptions. Morris et al. (2002) also use computer-based learning activities to contribute to students understanding of the correlation coefficient and measures of central tendency. Their method involves direct manipulation of data and graphical displays. Pre- and post-tests indicate that the exercise had a significant effect concerning the mean, but the difference was not statistically significant for the correlation coefficient. Hodgson & Burke (2000) provide a short review of the literature on using simulations in the understanding of statistics. They conclude that simulations can promote understanding, but the exercise can also provide new misconceptions. If a student does not possess sufficient skills, it is possible that he cannot distinguish salient from non-salient features of the exercise, and, as a consequence, a student can acquire both knowledge and new misconceptions.

The purpose of this study is to evaluate a simulation assisted learning method based on the cognitive conflict theory. Misconceptions concerning the properties of estimators, i.e. bias, efficiency and consistency, are common, and a simulation exercise is proposed to obtain a cognitive conflict. We use pre-test, post-test and delayed post test to evaluate if a meaningful cognitive conflict is obtained. In addition, a sufficiently large sample is collected to analyze the importance of different characteristics related to the student. Prior knowledge, cognitive ability, values and attitudes towards learning, motivation and interest are analyzed in relation to the performance of the simulation exercise.

Individual characteristics are found to be important to explain if a meaningful cognitive conflict is achieved or not. The values and attitudes towards learning seems to

be particular important. Having a passing grade in mathematics is also important for a meaningful cognitive conflict. The characteristics of the students are, indeed, important to obtain a meaningful conceptual change. Interestingly, a characteristic that favor an instant cognitive conflict does not necessarily do so in a delayed post-test implemented in the final exam. Evaluating both short-run and long-run effects of an intervention is, accordingly, found to be important.

2. METHOD

2.1 Design

The study was implemented in three stages. The first stage was to obtain information on individual characteristics of the students. At the second session of the course, all students were asked to answer a survey in the computer-lab. The second stage was implementing the simulation study at approximately week 13-14 of the course. Before that session the students had received a theoretical class which included a discussion concerning the properties of estimators; bias, efficiency and consistency. The final stage was to include a delayed post-test as a part of the final exam.

2.2 Participants

The experiment was implemented at the University of the Balearic Islands in the course Analysis of Economic Data. The course is a mandatory introductory statistical course in the second semester of the Economics program and the Business and Administration program. Five teachers implemented the exercise in eight different groups, and a total of 186 students attended the session for the simulation exercise. This is the sample that is analyzed in this study. The average age, among the students that participated in the

session for the simulations, and also had answered the survey during the first week is 20.4 years old. About 50.6% are female students.

2.3 Individual characteristics

The idea to collect information on individual characteristics is to be able to evaluate the performance of the computer assisted learning method in relation to the cognitive conflict theory. In particularly, it is interesting to identify for which characteristics the method is effective or not. Limón (2001) includes a list of variables related to the learner that is claimed to be important to obtain a meaningful conflict. Below we have formalized the main characteristics into several different questions and measures.

Table 1. The measurements of the individual characteristics.

Concept	Measurement
Prior knowledge	- 3-questions measuring statistical knowledge [0-3] - Grade Point Average (GPA) [0-10, but truncated at 5 for university students] - Grade in mathematics [0-10]
Reasoning abilities	Cogntive Reflection Test [0-3]
Values and attitudes towards learning	*"How strongly do you agree or disagree with the following statements? (1=strongly agree to 5 = strongly disagree)"* [Simply knowing the answer, rather than understanding the reasons for the answer to a problem, is fine with me.] [When I find it too difficult to understand a problem I often try to memorize its solution, instead of making the effort to understand it.] [Simply passing a course (in general), rather than understanding the content of the course, is fine with me.] [Simply passing this course in statistics, rather than understanding the content of the course, is fine with me.]
Motivation and interests	[Doing well on this course is important to me.] [I will engage in good effort throughout this course.] [I am curious about how I do on the evaluations of this course relative to others.] [I am not concerned about the score I receive on the assessments of this course.] [This is an important course to me.]

To measure prior knowledge we propose three different measures. Firstly, we let the student answer three different statistical questions, giving one point for each correct answer. The first two of these questions have been used in (Toplak, et al. 2011) and the third was used in Hoppe & Kusterer (2011). The questions are included in Appendix A.

Secondly, we asked the students for their Grade Point Average (GPA) from secondary schooling which was used to access University studies. This measure captures a more general prior knowledge. Thirdly, we asked for the grade in a course in mathematics that they had during the first semester. This question was included in the survey in the same session as the simulation exercise, because not all students had received their final grade at the time for the first survey.

The reasoning abilities are measured with the Cognitive Reflection Test (CRT) presented by Frederick (2005). The test consists of three questions that are all constructed with an intuitive *incorrect* answer. The questions are included in Appendix A. To find the correct answer students will usually have to override their initial thought, and reflect further, to find the correct answer. The test has received a massive impact in the literature. The reasons are probably because the test is not time consuming to implement in a survey, and the test possess strong explanatory power since it measures the performance, instead of relying on self-reported characteristics of the participants.

To measure values and attitudes towards learning the students were asked: *"How strongly do you agree or disagree with the following statements? (1=strongly agree to 5 = strongly disagree)"* and several different statements were presented. The exact formulations are included in Table 1. The main idea is to capture the students' self-assessed opinion about their willingness to understand, or simply know the answer, memorize it, or even pass a course without understanding its content. These kinds of questions are also used in Ardelt (2003). Motivation and interest were measured in the same way. The students were asked to reveal their self-assessed opinion concerning the importance of the course, the effort they would engage etc. These questions were adapted from The Student Opinion Scale (SOS) included in Sundre (2007, unpublished observations). Note that the SOS actually is a post-test, i.e. it is implemented after the

students had taken their exam. In this study we preferred to ask the questions the first week, to avoid endogeneity problems, i.e. motivation is likely to be affected by events occurring during the course. If we evaluate the effect of motivation measured at the final part of the course, we would likely capture effects that actually comes from these factors affecting motivation and the relevant effect of motivation would be difficult to determine.

2.4 Pre- and post-tests

The session for the simulation exercise was structured in the follow way; All students answered a short test to measure their initial knowledge and possible misconceptions. After approximately 10 minutes, when the students had submitted their answers, the teacher introduced the simulation exercise. The students used an Excel-sheet with simulations to answer 10 questions. The questions were presented with instructions on how to use the simulations to answers the questions. The students were allowed to seek help from the teacher and/or peers to understand the exercise. No discussion was done concerning the initial questions included in the first survey. After approximately 30 minutes most students had answered and submitted their answers. Once all of the students had submitted the answer the teacher made a short review of the correct answers and students were allowed to ask questions. This summary took approximately 5 minutes, and once it was finished the students opened the post-test, which included the exact same questions as were used in the pre-test, and, in addition, a few questions concerning, for example the grade in an earlier course in mathematics.

In the final exam, approximately 4-5 weeks after the session, the same questions were repeated to obtain information on a delayed post-test. The students were never informed that the same questions were to form part of the final exam. The students did,

however, have the opportunity to review the material as it was a part of course, and obviously could be included in the final exam. There are two important reasons to include the delayed post-test in the final exam. Firstly, this is a way to maximize the sample size, since attendance is, without doubt, higher in the final exam, compared to a single class. Secondly, evaluating the delayed post-test in the final exam simulates the most common learning situation, i.e. the students receive a learning intervention, but the evaluation of the students' performance usually is in form of a scheduled exam. Accordingly, the students' own decisions and efforts are important parts of the learning process and this is relevant for the long-run effect of an intervention. Obviously, this should be kept in mind when the results are analyzed. The alternative would be to evaluate the delayed post-test as a surprise test in a class a few weeks after the intervention. The students' decisions and efforts would still matter, and the effect could, in any case, not exclusively be interpreted as due to the intervention. The effect would also be transitory as students later would review the material and the relevant effect would not necessarily be the same. We find it more interesting to evaluate the aggregate effect and see the intervention and the students' efforts, as integrated parts of the learning process. All of the pre-and post-tests included the following questions, where the correct answers are marked in bold.

Table 2. Pre- and post-tests

Questions	Multiple-choice answers	
1. It is said that an estimator is unbiased if:	a)	It is an estimator that does not make mistakes.
	b)	It is an estimator that makes very small mistakes.
	c)	**It is an estimator that does not make systematic errors.**
	d)	It is an estimator that is very reliable.
2. An estimator is more efficient than another:	a)	If the errors are smaller.
	b)	**If it is more precise.**
	c)	If it is always right.
	d)	None of the above.
3. An estimator is consistent:	a)	If it is unbiased and efficient.
	b)	If it is an optimal estimator.
	c)	If as the sample size grows the bias is smaller.
	d)	**If as the sample size grows its bias and variance is decreasing.**

Common misconceptions were deliberately included among the answers. During the session the idea is that the students will face contradictory evidence that will guide them to a conceptual change in favor of a correct answer.

2.5 Using simulations to study the properties of estimators

In the exercise, Monte-Carlo simulations are used to study the properties of four different estimators of the population mean. 100 samples were generated for the sample-sizes; 20, 40 and 60 observations. Each sample was generated based on a normal distribution, with a population average, $\mu = 5$ and a population variance, $\sigma^2 = 1$, $N(5,1)$. The function in Excel that generates a random draw, as explain above, is **=NORMINV(RAND(),5,1)**. The following estimators of the population mean were calculated for each of the samples.

Table 3. Description of the estimators used in the simulation study

Estimator	Comment
$\mu_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	The estimator is the sample mean.
$\mu_2 = \frac{1}{n-1}\sum_{i=1}^{n-1} x_i$	The estimator is the sample mean for the $n-1$ first observations. The last observation is simply not used in the calculation.
$\mu_3 = \frac{1}{n+1}\sum_{i=1}^{n} x_i$	The estimator is similar to the sample mean, but the sum is divided by $n+1$.
$\mu_4 = x_1$	The estimator is simply to use the first observation in the sample.

In addition to calculating these measures for all samples, the mean, variance and mean squared error (MSE) were calculated. All the calculations were prepared in Excel and automatically updated for new random samples each time a student pressed F9. The students did not need to do any calculation and could focus on analyzing the results.

The instructions and questions that were made available to the students are included in Appendix B.

Related to the misconceptions included for the concept “bias”, it is easy to see that all estimators commit mistakes, in the sense that the estimator is not equal to the population mean of 5. Repeating the simulations a few times makes it evident that these mistakes can be large, and still they are not a systematic mistake. Hence, estimator 4 is unbiased, but certainly not *reliable* as option d) suggest for an "unbiased" estimator. Estimator 3, systematically underestimates the population mean, and this is the example of a biased estimator.

The misconceptions concerning if an estimator is efficient is that "it is always right" and "if the errors are smaller". Clearly, an estimator that is always right would not be possible to have, given that we have a random sample from a population and that a *variable* (i.e. not a constant!) is analyzed. Efficiency does not really refer to the size of the errors, but rather the relative accuracy of the estimators, given the sample size. This is the reason to compare the variances of unbiased estimators or the MSE, if at least one of the estimators is biased. Estimator 3 has the smallest variance, but it is not found to be more efficient than estimator 1. The reason is that it has a bias that makes the MSE higher in this case.

Finally, the first misconception concerning a consistent estimator is that it is consistent if it is "unbiased and efficient". Estimator 3 was found to have a bias, but if we study the MSE for different sample sizes, we see that it is decreasing with a larger sample size. We also found that estimator 4 was unbiased, but its MSE was not decreasing with larger sample size. Hence, a biased estimator can be consistent and an unbiased estimator can be inconsistent. We can also see that estimator 2 is not efficient,

but still it is consistent. What matters is that the variance *and* the bias decreases as the sample size increases.

3. RESULTS

In order to check if the monte-carlo experiment has a positive effect, we use the test for paired proportions proposed by McNemar(1947) and Liddell (1983) (see Liddel and Armitage, Berry and Matthews (2002) for details).

Denote by n_{00} the population frequency that answer an incorrect options before and after, n_{11} the correct option before and after the monte-carlo experiment, n_{10} the correct option before and an incorrect option after the monte-carlo experiment and n_{01} an incorrect option before and the correct option after the monte-carlo experiment. McNemar (1947) introduced a statistic that follows the chi-square distribution, with one degree of freedom, that allows us to test the null hypotheses that the expected value $n_{01} - n_{10}$ is zero (that is, the difference of the frequency of students that have [incorrect-correct] and [correct-incorrect]). Additionally Liddell proposed a point estimator and a confidence interval estimator of the relative risk $R = n_{01} / n_{10}$ and an F-type statistic to test the null that $R = 1$.

The statistics were calculated to evaluate the performance of the intervention and table 4 includes the results for questions 1, 2 and 3 for both pre- and post-test and pre- and delayed post-tests.

Table 4. Results for the complete group

	Pre-and post-test								
Question	$\hat{n}_{11}$	$\hat{n}_{00}$	$\hat{n}_{01}$	$\hat{n}_{10}$	McNemar	$\hat{R}$	$\hat{R}_L$	$\hat{R}_U$	$F_{R=1}$
1	40	92	38	8	18.28***	4.75	2.42	10.16	4.22***
2	71	57	25	25	0.02	1.00	0.60	1.66	0.96
3	33	95	36	14	8.82***	2.50	1.44	4.50	2.33**
	Pre-and delayed post-test								
1	28	92	45	21	8.02***	2.14	1.35	3.46	2.05**
2	63	51	34	38	0.13	0.89	0.59	1.36	0.87
3	27	91	49	19	12.37***	2.58	1.61	4.23	2.45***

Notes: *** significant at 1% level. $\hat{n}$ refers to the absolute sample frequency, the first subindex is (0 = incorrect answer, 1 = correct answer) in the pre-test and the second subindex is used in the same way, but for the post-test. $\hat{R}$, $\hat{R}_L$ and $\hat{R}_U$ denote the point estimate of the relative risk, lower and upper confidence interval proposed by Liddell and $F_{R=1}$ is the Liddell test to test the null $R = 1$.

A meaningful cognitive conflict was obtained for both the concepts “bias” and “consistency”, where a statistically significant difference is achieved. The result also holds for the delayed post-test. Despite having a statistically significant statistic, a large proportion of the students answered incorrectly both in the pre-test and the post-tests. For many students a meaningful cognitive conflict was not achieved and it is important to identify the characteristics that opposes, or favor a conceptual change.

For question 2, “efficiency”, no statistically significant change was found. The McNemar statistic was calculated for all subgroups, but it was never statistically significant. These results are not included in the paper. The conclusion is that the cognitive conflict was poorly designed, irrespective of the students’ characteristics, for this particular question.

Table 5 includes descriptive statistics for the quantitative variables for the students that participated in the simulation study.

Table 5. Descriptive statistics

			Percentage scoring 0, 1, 2 or 3.			
	mean	std. dev.	0	1	2	3
Statistical knowledge test	0.88	0.81	36.42%	41.62%	19.08%	2.89%
CRT	0.87	0.96	45.09%	30.64%	16.18%	8.09%
Grade in mathematics	4.96	2.59				
GPA	6.57	1.31				

The average score on the prior statistical knowledge test was 0.87 with about 36% failing all three questions. The average score on the Cognitive Reflection Test (CRT) was 0.87, which seems reasonable, although fairly low compared to previous studies using the test (Frederick, 2005, Hoppe, & Kusterer, 2011, Toplak, et al., 2011, Brañas-Garza et. al., 2012). Students were allowed to use to use their own paper and pencil to solve the questions, but these tools were not handed out to the students. It is possible that the test-score could have been improved by providing these tools.

The qualitative variables captures valuations from *(1=strongly agree to 5 = strongly disagree)"*. In the analysis these variables were used to create two groups depending of the answers given. For example, the number of students that answered, 1, 2 or 3, respective 4 or 5 to the statement "Simply knowing the answer, rather than understanding the reasons for the answer to a problem, is fine with me." can be found in tables 6 and 7. These tables also include the McNemar statistic calculated for all subgroups with different characteristics. In the case of tables 6 and 7 we did not report the results about the Liddell tests in order to save space, as we obtain equivalent results with both tests.

Table 6. Evaluating exercise concerning "bias", post-test and delayed post-test, for different samples.

	Post-test	Delayed Post-test	Post-test	Delayed Post-test	
Sample based on variable;	McNemar	McNemar	McNemar	McNemar	Sample based on variable;
[Simply knowing the answer, rather than understanding the reasons for the answer to a problem, is fine with me.] 1, 2, 3	3.5* (0.064) [59]	0.41 (0.522) [61]	10.24*** (0.001) [97]	7.61*** (0.006) [103]	[Simply knowing the answer, rather than understanding the reasons for the answer to a problem, is fine with me.] 4, 5
[When I find it too difficult to understand a problem I often try to memorize its solution, instead of making the effort to understand it.] 1, 2, 3	2.4 (0.121) [56]	0.64 (0.424) [60]	12.04*** (0.001) [100]	7.31*** (0.007) [104]	[When I find it too difficult to understand a problem I often try to memorize its solution, instead of making the effort to understand it.] 4, 5
[Simply passing a course (in general), rather than understanding the content of the course, is fine with me.] 1, 2, 3	5.06** (0.024) [69]	1.53 (0.216) [74]	10.23*** (0.001) [86]	6.04** (0.014) [89]	[Simply passing a course (in general), rather than understanding the content of the course, is fine with me.] 4, 5
[Simply passing this course in statistics, rather than understanding the content of the course, is fine with me.] 1, 2, 3	1.78 (0.181) [59]	0.17 (0.677) [63]	12.96*** (0.000) [97]	8.76*** (0.003) [101]	[Simply passing this course in statistics, rather than understanding the content of the course, is fine with me.] 4, 5
[Doing well on this course is important to me.] 3, 4, 5	2.29 (0.131) [26]	0.36 (0.546) [29]	11.28*** (0.001) [130]	6.61*** (0.010) [135]	[Doing well on this course is important to me.] 1, 2
[I will engage in good effort throughout this course.] 3, 4, 5	4.17** (0.041) [26]	0.9 (0.343) [29]	9.03*** (0.003) [128]	4.69** (0.030) [133]	[I will engage in good effort throughout this course.] 1, 2
[I am curious about how I do on the evaluations of this course relative to others.] 3, 4, 5	5.06** (0.024) [72]	7.26*** (0.007) [76]	8.52*** (0.004) [83]	1.09 (0.296) [87]	[I am curious about how I do on the evaluations of this course relative to others.] 1, 2,
[I am not concerned about the score I receive on the assessments of this course.] 1, 2, 3	2.08 (0.149) [41]	0.05 (0.82) [45]	15.56*** (0.001) [114]	9.03*** (0.003) [118]	[I am not concerned about the score I receive on the assessments of this course.] 4, 5
[This is an important course to me.] 3, 4, 5	0.8 (0.371) [26]	0.00 (1.00) [30]	12.12*** (0.000) [129]	6.94*** (0.008) [133]	[This is an important course to me.] 1, 2
Prior statistical knowledge test <= 1	6.76*** (0.009) [125]	3.84** (0.050) [131]	9.09*** (0.003) [35]	3.76* (0.052) [37]	Prior statistical knowledge test > 1
Grade in mathematics < 5	0.36 (0.302) [63]	0.00 (1.00) [59]	13.79*** (0.000) [102]	5.03** (0.025) [100]	Grade in mathematics >= 5
GPA <=6.5	5.5** (0.019) [86]	3.22* (0.072) [87]	11.53*** (0.001) [67]	4.114** (0.043) [99]	GPA > 6.5
CRT <= 1	9.63*** (0.002) [120]	7.20*** (0.007) [126]	5.56*** (0.027) [40]	0.56 (0.453) [42]	CRT > 1

Notes: All statistics are calculated using a pre-test result. ***, **, * indicate significant at 1%, 5% respective 10% level. Numbers in parenthesis are p-value and sample size.

3.1 "Bias", pre-test and post-test

Table 6 includes the results for different subgroups concerning the first question, "bias", when the test was evaluated using pre-test and post-test and delayed post-test. The third

and fourth column represents the sample that is expected to perform better. The first four questions intend to capture a student's values and attitudes towards learning, in particular the willingness to understand. For the post-test, the McNemar statistic is strongly significant for the groups that express a stronger willingness to understand. For the groups that do not express the same willingness to understand, the statistic is not statistically significant for two questions and statistically significant at the 5% level for one question, and at the 10% level for another question. The questions that measure a more general self-assessed motivation and interest also identify that the simulation exercise is more suitable for the group with higher expected performance. The groups with expected lower performance are, however, often quite small. A statistical significant effect is found for the subgroups of different score on the prior test in statistical knowledge, the Cognitive Reflection Test and the GPA score. For the group that did not pass the mathematical course, the simulation exercise did not produce a significant statistic, which indicates that the intervention was not effective for this group.

Table 7. Evaluating exercise concerning "consistency", post-test and delayed post-test, for different samples.

	Post-test	Delayed Post-test	Post-test	Delayed Post-test	
Sample based on variable;	McNemar	McNemar	McNemar	McNemar	Sample based on variable;
[Simply knowing the answer, rather than understanding the reasons for the answer to a problem, is fine with me.] 1, 2, 3	0.27 (0.606) [59]	5.50** (0.019) [61]	6.76*** (0.009) [97]	5.92** (0.015) [103]	[Simply knowing the answer, rather than understanding the reasons for the answer to a problem, is fine with me.] 4, 5
[When I find it too difficult to understand a problem I often try to memorize its solution, instead of making the effort to understand it.] 1, 2, 3	0.75 (0.386) [56]	4.00** (0.046) [60]	5.28** (0.022) [100]	7.31*** (0.007) [104]	[When I find it too difficult to understand a problem I often try to memorize its solution, instead of making the effort to understand it.] 4, 5
[Simply passing a course (in general), rather than understanding the content of the course, is fine with me.] 1, 2, 3	3.04* (0.081) [69]	5.76** (0.016) [74]	2.04 (0.153) [86]	4.69** (0.030) [89]	[Simply passing a course (in general), rather than understanding the content of the course, is fine with me.] 4, 5
[Simply passing this course in statistics, rather than understanding the content of the course, is fine with me.] 1, 2, 3	1.56 (0.211) [59]	6.86*** (0.009) [63]	4.32** (0.038) [97]	5.03** (0.025) [101]	[Simply passing this course in statistics, rather than understanding the content of the course, is fine with me.] 4, 5
[Doing well on this course is important to me.] 3, 4, 5	3.20* (0.074) [26]	4.90** (0.027) [29]	3.69* (0.055) [130]	7.22*** (0.007) [135]	[Doing well on this course is important to me.] 1, 2
[I will engage in good effort throughout this course.] 3, 4, 5	3.20* (0.074) [26]	6.75*** (0.009) [29]	3.69* (0.055) [128]	4.89** (0.027) [133]	[I will engage in good effort throughout this course.] 1, 2
[I am curious about how I do on the evaluations of this course relative to others.] 3, 4, 5	5.5** (0.019) [72]	7.04*** (0.008) [76]	1.14 (0.286) [83]	4.69** (0.030) [87]	[I am curious about how I do on the evaluations of this course relative to others.] 1, 2,
[I am not concerned about the score I receive on the assessments of this course.] 1, 2, 3	5.79** (0.016) [41]	3.77* (0.052) [45]	1.241 (0.265) [114]	6.80*** (0.009) [118]	[I am not concerned about the score I receive on the assessments of this course.] 4, 5
[This is an important course to me.] 3, 4, 5	2.25 (0.134) [26]	3.125* (0.077) [30]	3.69* (0.055) [129]	7.84*** (0.005) [133]	[This is an important course to me.] 1, 2
Prior statistical knowledge test <= 1	4.36** (0.037) [125]	10.29*** (0.001) [131]	1.23 (0.267) [35]	2.12 (0.146) [37]	Prior statistical knowledge test > 1
Grade in mathematics < 5	1.56 (0.211) [63]	1.25 (0.264) [59]	6.26** (0.012) [102]	11.025*** (0.001) [100]	Grade in mathematics >= 5
GPA <=6.5	4.65** (0.031) [86]	11.17*** (0.001) [87]	0.94 (0.332) [67]	2.53 (0.112) [76]	GPA > 6.5
CRT <= 1	1.63 (0.201) [120]	13.02*** (0.000) [126]	5.06** (0.024) [40]	0.56 (0.453) [42]	CRT > 1

Notes: All statistics are calculated using a pre-test result. ***, **, * indicate significant at 1%, 5% respective 10% level. Numbers in parenthesis are p-value and sample size.

3.2 "Bias", pre-test and delayed post-test

The student's values and attitudes towards learning are found to be important also for the delayed post-test. The group that expressed a stronger willingness to understand had a significant statistic, while no such effect was found for the group that disagreed or were neutral. The results for the questions expressing self-assessed motivation and

interest are mixed. Interestingly, for the group with low score (0 or 1) on the CRT the simulations had a statistically significant statistic, but no such effect was found for the group with a score of 2 or 3. Having a passing grade in mathematics was again found to be important to achieve a meaningful cognitive conflict.

3.3 "Consistency", pre-test and post-test

The results concerning "consistency" can be found in table 7. The measures of a student's values and attitudes towards learning are again found to be important to obtain a meaningful cognitive conflict in the post-test. In fact, the difference is now even stronger. The questions used to capture self-assessed motivation and interest work poorly to identify for which group the intervention is more suitable. In fact, the results have a tendency to indicate a better performance for the group that is expected to perform worse. For example, the statistic is statistically significant for the group not being curious, or neutral, about the evaluation compared to other students, but no such effect is found for the group that answered 1 or 2 (*1=strongly agree*). The simulations were only effective for those with a score of 2 or 3 on the CRT, while it was not statistically significant for the group that had scores of 0 or 1. It seems that the cognitive skill to be able to override an intuitive incorrect answer is important for an instant effect of the simulation study. Having a passing grade in mathematics is also important for obtaining a meaningful cognitive conflict.

3.4 "Consistency", pre-test and delayed post-test

Having approved the course in mathematics is found to be an important characteristic to obtain a meaningful cognitive conflict in the delayed post-test. An interesting result is that the student's values and attitudes towards learning cannot

distinguish between in which group a meaningful cognitive conflict is obtained. The statistic is now statistically significant for both groups. Another interesting result is that the group with lower score on the CRT had a statistically significant statistic, but no such effect is found for the group with 2 or 3 on the CRT.

4. CONCLUSIONS

The first impression of the results of the delayed post-test for the concept "consistency" might seem unexpected. For the group that was less willing to understand, and in the post-test during the session did not had a statistically significant statistic, is found to have so in the delayed post-test. A plausible explanation is that the students simply have opted for their way of learning, i.e. they have had the time to "learn the answer" or "memorize the solution". A correct answer does not necessarily mean that a student has understood the reason for it. A question is of course, why this result did not occur for the concept "bias". It is possible that the students were unaware of their misconception, despite the intervention, in that case, but in the second case, the simulation exercise forced the students to look for the correct answer (to memorize). A recommendation for future studies is to use alternative formulations in an attempt to measure understanding of the concept.

In the pre-test and post-test evaluation for both "bias" and "consistency" the statistic is statistically significant for high CRT, but once we study the pre-test and delayed post-test, the simulation exercise has lost the effect. A possible interpretation is that a high cognitive reflection skill improves the instant effect of the simulation, as the cognitive conflict could be solved. The effect could, however, be superficial, and eventually disappear, because students felt that they solved the exercise, and hence did not feel that they had to review it before the final exam, i.e. the delayed post-test. It seems that the

cognitive conflict only was partial as the misconceptions came back. For the group with low CRT, the delayed post-test shows a significant effect. Maybe they had difficulties to solve the cognitive conflict and opted for reviewing the exercise before the delayed post-test. These results underline the importance of analyzing a delayed post-test and doing so at the time of a final exam is reasonable.

Having approved the course in mathematics is found to be an important characteristic to obtain a meaningful cognitive conflict in both instant and delayed post-test. It should be clarified that the exercise in itself does not require mathematical calculations. Simply comparing the values is enough. The grade in math could, however, capture abstract skills that could be useful for the exercise. Another possibility is that it is not this particular skill that matters, simply that the students *think* that they will be more able, and hence, it is more a motivational trigger that is functioning. Gottfried (1990) found, for example, that early achievement was correlated with later motivation. The prior statistical knowledge test did not provide such clear effect. The correct answers on these questions were never revealed to the students and it is unlikely that this measure would affect the motivation.

It is important to remember that attitudes and values towards learning, motivation and interests, are variables that are possible to change to improve the efficiency of simulation assisted learning methods. Methods to affect these variables are, however, beyond the scope of this study.

APPENDIX A

A.1 Prior statistical knowledge test

Base rate fallacy

1. In a city with 100 criminals and 100,000 innocent citizens there is a surveillance camera with an automatic face recognition software. If the camera sees a known criminal, it will trigger the alarm with a 99% probability; if the camera sees an innocent citizen, it will trigger the alarm with a probability of 1%. What is the probability that indeed a criminal was filmed when the alarm is triggered?

Sample size

2. A game of squash can be played to either 9 or 15 points. Holding all other rules of the game constant, if A is a better player than B, which scoring scheme would give player A a better chance of winning? (.) 9 points, (.) equal, (.) 15 points.

Covariation detection

3. A doctor had been working on a cure for a mysterious disease. Finally, he created a drug that he thinks will cure people of the disease. Before he can begin to use it regularly, he has to test the drug. He selected 300 people who had the disease and gave them the drug to see what happened; 200 were cured and 100 were not. He selected 100 people who had the disease and did not give them the drug in order to see what happened; 75 were cured and 25 were not.

Was the treatment positively, neutral, or negatively associated with the cure for this disease? ( ) positively, ( ) neutrally, ( ) negatively.

Correct answers; 1) ≈ 0.09008 **, which rounded about to 0.09 was also marked as correct**, 2) **15 points** and 3) **negatively.**

A.2 Cognitive Reflection Test (CRT)

Below are several problems that vary in difficulty. Try to answer as many as you can.

1. A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost? …..cents.

2. If it takes five machines five minutes to make five widgets, how long does it take 100 machines to make 100 widgets? ….. minutes.

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ….. days.

Correct answers; 1) **5 cents**, 2) **5 minutes** and 3) **47 days.**

APPENDIX B

B.1 Questionnaire that students filled in during the simulation exercise.

How do we use simulations to detect if an estimator is unbiased?

In the simulations we have specified that the mean (the parameter of interest) of the population is 5. We look at the average of the estimates obtained in each replicate. A bias implies that the estimator systematically over-estimated or under-estimated the result. An unbiased estimator does NOT mean that we will always have a point estimate equal to the population parameter. We can have unbiased estimators in which the point estimates may be far from the population parameter. The important thing is to identify whether the error is systematic or not.

1. Is estimator 1 unbiased? ( ) ***Yes****, ( ) No*

2. Is estimator 2 unbiased? ( ) ***Yes****, ( ) No*

3. Is estimator 3 unbiased? ( ) Yes, ( ) ***No***

4. Is estimator 4 unbiased? ( ) ***Yes****, ( ) No*

How do we use simulations to detect if an estimator is more efficient than another estimator?

To determine whether an estimator is more efficient than another it is necessary to compare the mean square error of the estimators. Just to compare the variance is not sufficient if one of the estimators is biased. A biased estimator, but with a very small variance can be more precise than an unbiased estimator with large variance.

5. Which estimator has the smallest variance? ***3***

6. Which is the most efficient estimator? ***1***

How do we use simulations to detect if an estimator is consistent?

To identify whether an estimator is consistent we look if the mean square error (MSE) decreases with increasing sample size. If the mean square error decreases this implies that its variance and possible bias decrease when the sample size increases. This indicates that the MSE tends to zero as the sample size goes to infinity, i.e., the estimator would eventually coincide with the parameter value.

7. Is estimator 1 consistent? *( )* ***Yes****, ( ) No*

8. Is estimator 2 consistent? *( )* ***Yes****, ( ) No*

9. Is estimator 3 consistent? *( )* ***Yes****, ( ) No*

10. Is estimator 4 consistent? *( ) Yes, ( )* ***No***

REFERENCES

Ardelt M. (2003), "Empirical Assessment of a Three-Dimensional Wisdom Scale," *Research on Aging*, 25(3), 275-324. DOI: 10.1177/0164027503251764.

Armitage P., Berry G. & Matthews J.N.S. (2002), *Statistical Methods in Medical Research*. Blackwell Science, Cornwall.

Brañas-Garza P., García-Muñoz T. & Hernán González R. (2012), "Cognitive effort in the Beauty Contest Game," *Journal of Economic Behavior & Organization*, 83, 254-260. DOI: 10.1016/j.jebo.2012.05.018.

Frederick S. (2005), "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, 19(4), 25-42. DOI: 10.1257/089533005775196732.

Gottfried A.E. (1990), "Academic intrinsic motivation in young elementary school children," *Journal of Educational Psychology*, 82(3), 525-538. DOI: 10.1037/0022-0663.82.3.525.

Hodgson T. & Burke M. (2000), "On simulations and the teaching of statistics," *Teaching Statistics*, 22, 91-96. DOI: 10.1111/1467-9639.00033.

Hoppe E.I. & Kusterer D.J. (2011), "Behavioral biases and cognitive reflection," *Economics Letters*, 110, 97-100. DOI: 10.1016/j.econlet.2010.11.015.

Limón M. (2001), "On the cognitive conflict as an instructional strategy for conceptual change: a critical appraisal," *Learning and Instruction*, 11, 357-380. DOI: 10.1016/S0959-4752(00)00037-2.

Liddell F.D.K. (1983), "Simplified Exact Analysis of cas-referent matched pairs; Dichotomus Exposure Studies," *Journal of Epidemiology and Community Health*, 37, 82-84. DOI: 10.1136/jech.37.1.82.

Liu T.C., Lin Y.-C. & Kinshuk. (2010), "The application of Simulation-Assisted Learning Statistics (SALS) for correcting misconceptions and improving

understanding of correlation," *Journal of Computer Assisted Learning*, 26, 143-158. DOI: 10.1111/j.1365-2729.2009.00330.x.

McNemar Q. (1947), "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, 12, 153-157. DOI: 10.1007/BF02295996.

Morris E.J., Joiner R. & Scanlon E. (2002), "The contribution of computer-based activities to understanding statistics," *Journal of Computer Assisted Learning*, 18, 114-124. DOI: 0.1046/j.0266-4909.2002.00219.x.

Toplak M.E., West R.F. & Stanovich K.E. (2011), "The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks," *Memory Cognition*, 39(7), 1275-1289. DOI: 10.3758/s13421-011-0104-1.

Watts D.G. (1991), "Why is Introductory Statistics difficult to learn? And what can we do to make it easier?" *The American Statistician*, 45, 290-291. DOI: 10.1080/00031305.1991.10475824.