The effectiveness of a new model structure behavior visualization technique:

An experimental study of the Forio Model Explorer

William Schoenberg
Supervised by: Pål Davidsen

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Philosophy in Systems Dynamics

System Dynamics Group
Department of Geography
University of Bergen, Norway

November 2011

Abstract

System dynamisists have a difficult time communicating the knowledge learned during the model building process to the general public and decision makers. (Warren and Langley 1999) In this study one of the most popular methods for communicating the results from the modeling process causal loop diagraming was tested vs. a new technique known as the model explorer. A dynamic task was created which participants had to solve. Participants were divided into two treatment groups, the first receiving a CLD the second a model explorer. There were no statistical differences between the two groups when comparing scores in each run, or the improvement of each participant from run to run. This means that the model explorer performed no worse or better then a CLD and should be considered another tool in the toolbox of system dynamisists for when they need to explain the knowledge gained from the modeling process. This research opens up questions around the effectiveness of the model explorer on large complex model where the CLD is generally weaker; might it be that the model explorer may prove to be more effective then current techniques there?

Table of Contents

Table of Figures

Table of Tables

1 INTRODUCTION

Many decision makers have a difficult time making decisions in complex dynamic systems (Brehmer 1992, Funke 1991, Jensen 2005, Moxnes 1998; Moxnes 2004, Rouwette et al. 2004, Sterman 1989a, Sterman and Booth Sweeny 2007). As a result many decision makers make sub-optimal decisions when faced with complex dynamic systems (Sterman 1989b). An example of this is global climate change where there are many types of GHGs all with varying impacts which matter to a whole host of stakeholders spread world-wide (Moxnes and Saysel 2009, Sterman and Booth Sweeny 2007). Adding to the complexity of the problem are the delays inherent in the system which separate cause from effect (Moxnes and Saysel 2009, Sterman and Booth Sweeny 2007). Then to deal with the problem there are many policy makers spread across the entire world at many levels of government, in many different governments all with different incentives and goals. So far the policy that we have seen from these decision makers has been ineffective at curbing the amount of GHGs in the atmosphere (Moxnes and Saysel 2009, Sterman and Booth Sweeny 2007).

The field of System Dynamics has developed many techniques and processes that allow them to analyze and understand complex dynamic systems. These techniques allow system dynamisists to work with stakeholders, decision makers, and members of the general public to communicate the mental models all people posses of complex dynamic systems (Sterman 2000). Three of the best examples of these techniques are:

1. Simulation
2. Stock and Flow Diagramming
3. Causal Loop Diagramming

Each of these techniques has its own inherent advantages and disadvantages which make it more or less useful for solving and communicating the knowledge gained from the study of different types of problems. In addition, very often these techniques are combined in order to have the best chance of success.

When using stock and flow diagramming, or causal loop diagraming a second behavior based visualization should also be provided if available so that the audience can understand not only the structure of the problem but its behavior. Though, there are many practitioners of Systems Thinking which make use of causal loop diagramming, or stock and flow diagramming without doing simulation therefore they lack the ability to present the behavior of the system.

Simulation is a technique that system dynamisists use communicate the behavior of the system under study under various scenarios. Simulation is an abstraction of the reality of the system under study (Sterman 2000). Constructing a simulation requires technical training, and specialized education. Simulations are difficult for untrained people to interpret and understand and are therefore generally the domain of experts whose job it is the construct them and then communicate and disseminate their results through an alternate means to a specific and trained audience (Größler et al. 2000). Building a simulation requires the specification of the exact relationships between variables via equations so that a computer can calculate them (Sterman 2000).

Stock and flow diagramming is used by system dynamisists to communicate the structure of the system under study (Sterman 2000). Now-a-days they are almost always created when constructing simulation models because the prevalence of Vensim, iThink , PowerSim and SMIA. These diagrams are technical and hard to understand without specific training (Wolstenholme 1999). Determining system behavior from these diagrams is difficult and requires much training and practice to be able to do reliably. To construct a stock and flow diagram requires knowledge about the presence or absence of links among variables, but not the actual specifics of the equations required by simulations (Sterman 2000). Stock and Flow diagrams do show feedback loops and delays, but they are not necessarily very clear. On the other hand though, they do show the stocks which is crucial for deriving the behavior, and while stocks do form the basis of all delays many modeling tools provide functions as abstractions for delays which do not use a stock symbol and therefore those delays do not show up in the stock and flow diagram. In large models stock and flow diagrams can get very complex, filled with thousands of variables connected by equally many thousands of arrows linking them. In addition the use of symbols, for example boxes for stocks, thick arrows for flows, thin arrows for causal links have opaque meaning to non trained people making the meaning of the whole diagram hard to grasp.

Causal loop diagrams (CLDs) are also used by System Dynamisists to communicate the structure of a system under study (Sterman 2000). They are used primarily to abstract away some of the complexities involved in stock and flow diagrams and simulations. Unlike stock and flow diagrams, CLDs are not automatically created when building simulations. CLDs communicate abstractions about the simulation model which they are based on, which themselves are abstractions of reality. Because CLDs are so abstract, they are generally pretty easy to understand, and they are best suited for showing the feedback loops and delays present within a system (Sterman 2000). CLDs also have the advantage of showing the polarity of links, but they generally do not differentiate between stocks and flows. Constructing a CLD only requires the knowledge of what variables are causally linked to each other and the polarity of that link.

Using these current techniques has created a problem where the full knowledge created and learned during the simulation process is not fully transferred to the decision makers and key stakeholders in the dynamic system (Jensen 2005). Take for example the global climate change example from above. Experts are currently having a lot of trouble convincing decision makers of the correct course of action. Part of the problem is that experts are retaining far more knowledge from the modeling process than they can pass on, or relate back to the policy makers, stakeholders or the general public. In part, this is a problem of communication. Too much knowledge is being lost in the transfer between experts and decision makers.

In order to solve this communication problem between experts and decisions makers, system dynamisists need a new technique, tool or process that they can rely on to help them communicate the knowledge they have gained from the modeling process in order to make their results approachable for non-experts (Warren and Langley 1999). Therefore, this thesis will study the effectiveness of a new technique developed to allow system dynamisists to portray the structure and behavior of simulation models to decision makers and the general public. The new technique revolves around the use of a tool called the model explorer developed by Forio Online Simulations.

The model explorer is a tool that combines many of the best attributes of simulation, stock and flow diagrams, and causal loop diagrams into one. It communicates the full simulation structure including equations without the use of symbols (such as stocks or flows) as well as the behavior of the model under a specific, user controlled scenario. The model explorer is weak at showing delays, and of labeling feedback loops, but is very good at combining structure and behavior into a single visualization. The model explorer can allow for the full set of simulation behaviors, including setting parameters, advancing the model through time and viewing of results.

In order to avoid some of the above mentioned problems of the stock and flow diagram, it allows for the abstraction of simulation structure through a series of partial visualization techniques. Rather then showing the full model structure at all times the model explorer is designed to show only the most relevant portions of the model structure to the viewer. The model explorer uses four techniques to accomplish this:

1. A to B diagrams
2. Nearest neighbor diagrams
3. Complexity Levels
4. Visibility Levels

An A to B diagram shows all of the causal links in a model from a starting point (A) to an ending point (B). These diagrams are most often used to show all of the relationships between a decision (A) and a key indicator (B). In this mode the model explorer will reveal all of the causal links that tie the decision to the key indicator. This diagram will not contain any variable C which is not on a direct path from A to B. When using A to B diagram the model explorer allows the user to track all of the causal pathways from A to B that include a third variable C. This feature is known as path highlighting. This allows users to visualize all of the causal links from A to B that depend on C. This feature is very useful when used in the context of a large model because it immediately highlights all of the causal links that the user is most interested in.

Nearest neighbor diagrams constitute another integral part of the model explorer. They allow the user to see only the most closely related variables to the chosen variable. In this manner they are very similar to Vensim Causal Tracing trees, but they use a parent centered radial layout algorithm as opposed to a tree layout algorithm. Also, nearest neighbor diagrams show any of the relationships that exist between the nearest neighbors in addition to the relationships between the chosen variable and the nearest neighbors. Nearest neighbor diagrams allow the user to explore the full structure of the model one variable at a time without having to become overwhelmed by the full structure of the model.

Complexity levels are used when displaying an A to B diagram to a user. In large models these diagrams can sometimes be gigantic containing well over one hundred variables with thousands of causal links. In order to make diagrams of that complexity approachable to non-experts the model explorer will show at first the least complex diagram consisting only of the most direct and shortest links possible from A to B. As the user shifts the complexity to higher and higher levels the model explorer will reveal further and more distant links between A and B. This feature allows the user to hide and reveal in steps the complexity of the full model structure.

Visibility levels are a tool similar to complexity levels that hide structure in the simulation model to make it easier to understand by non-experts. Each variable in a model can be given a particular visibility level. The level specifies how important that variable displayed is to a user. Visibility levels can then be named, and users then have the option to display any variables below a certain complexity threshold. Visibility levels are most often used to hide effect variables, or initial variables. Lets take for example the following set of relationships:

Market Share = { some equation, not relevant to this example }
Effect of Market Share on Brand value = LOOKUP(Market Share)
Brand Value = Effect of Market Share on Brand Value * Initial Brand Value

In many cases the variable Effect of Market Share on Brand value is going to be confusing to non experts. It is a non-necessary variable from a diagramming perspective separating the important link between Market Share and Brand Value. Therefore the modeler would assign to Effect of Market Share on Brand Value a lower visibility level (a higher number), meaning that it is a less important variable so that when users viewed the model explorer they would see a link directly from Market Share to Brand Value. Only those users who choose to view the model at the higher more complex visibility level would see the full relationship.

The model explorer displays variables are circles, and causal links as arrows. It makes no distinctions based on the type of variable. The selected variable(s) are always shown with a larger radius to make them easier to pick out and see. In the center of each circle is a sparkline which shows the actual behavior for that variable in the current simulation. The model explorer requires a simulation to work, specifically that simulation must be created in either, Vensim, iThink, PowerSim, Excel or the Forio modeling language. It is a machine generated tool, requiring nothing more then a model to function.

Because the model explorer is machine generated; the diagrams that it produces are not always the most optimal diagrams. Therefore, users are allowed to re-arrange and shift the location of variables on the screen. These changes are then remembered by the tool, and will in fact override the machine generated layouts that would occur when the next user views the same diagram.

The model explorer is a fully web-based tool that is meant for consumption by decision makers and the public at large. Its goals are not to be the most useful tool by which system dynamisists communicate about models amongst themselves, but rather how system dynamisists communicate models to regular people. It allows for exploration and insight discovery by non experts and packages up all model structure and behavior combinations allowing them to be discovered by users.

The potential value of the model explorer could be huge. If, in fact, the model explorer is a better technique for communicating the relationship between the structure and the behavior of system dynamics models then system dynamisists should prefer using to communicate their work. If it is truly more understandable to members of the general public and decisions makers, then it has the potential to increase the size of the field by making system dynamics more relatable and understandable to those who do not wish to invest the time in training or learning the specifics of the field. In order to grow, the system dynamics community needs to find and assess the added value of new techniques designed for sharing the lessons derived from the construction of models. Therefore, it is important to test the model explorer vs. CLDs to assess which is better in helping members of the general public succeed at decision making in a complex dynamic system.

2 THE MODEL

2.1 The Model Structure

The laundry detergent task model developed as the basis for this experiment is based heavily on the Bass diffusion model. The model was written in Vensim and tracks the flow of people from Potential Customers to Customers of any product, in this case laundry detergent. The flow of people from Potential Customers to Customers is called the Adoption Rate, and it is driven by the amount of money spent on marketing each week, as well as the number of customers relative to the number of potential customers. In addition to the basic Bass diffusion model, this model also tracks profit and cumulative profit, by assigning a profit per customer each week to calculate revenue, and subtracts the weekly marketing expenditure in order to calculate weekly profit.

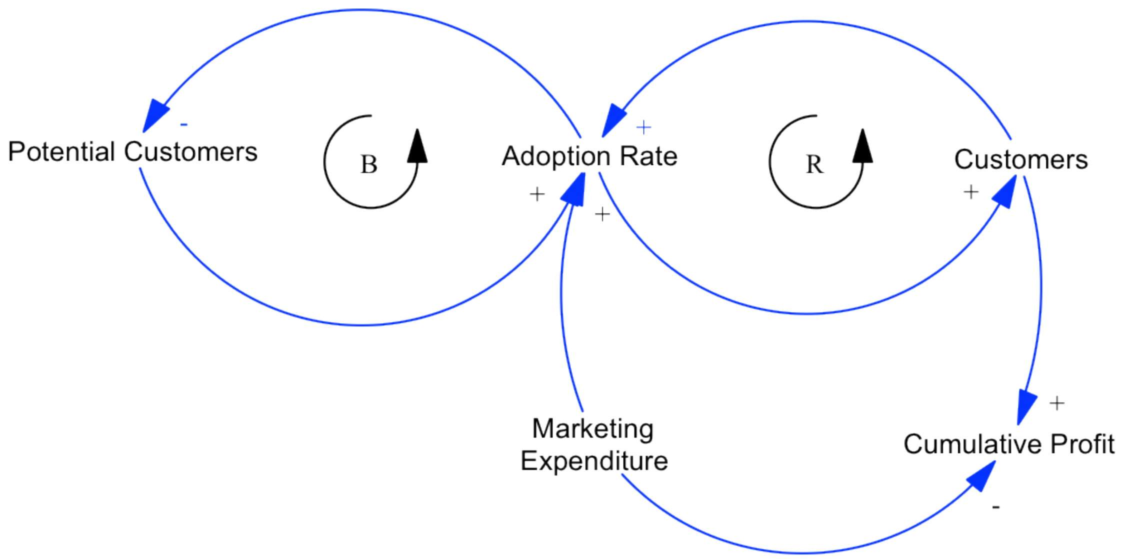

Figure 1: A CLD of the Laundry Detergent Task Model. The model contains two feedback loops which shift dominance when the Adoption Rate peaks.

The model was constructed with the Bass diffusion model as its base because it needed to be relatable to non-experts, and easily re-parameterized in a way that shifted loop dominance. The Bass diffusion model was chosen for the task because it is a well studied and easily understood model. It's structure is relatable to non-experts because it uses common and easily understood variable names and has logical causal links. In addition, because the Bass diffusion model has two feedback loops it is easy to parameterize so that either the balancing or reinforcing feedback loop is initially dominant.

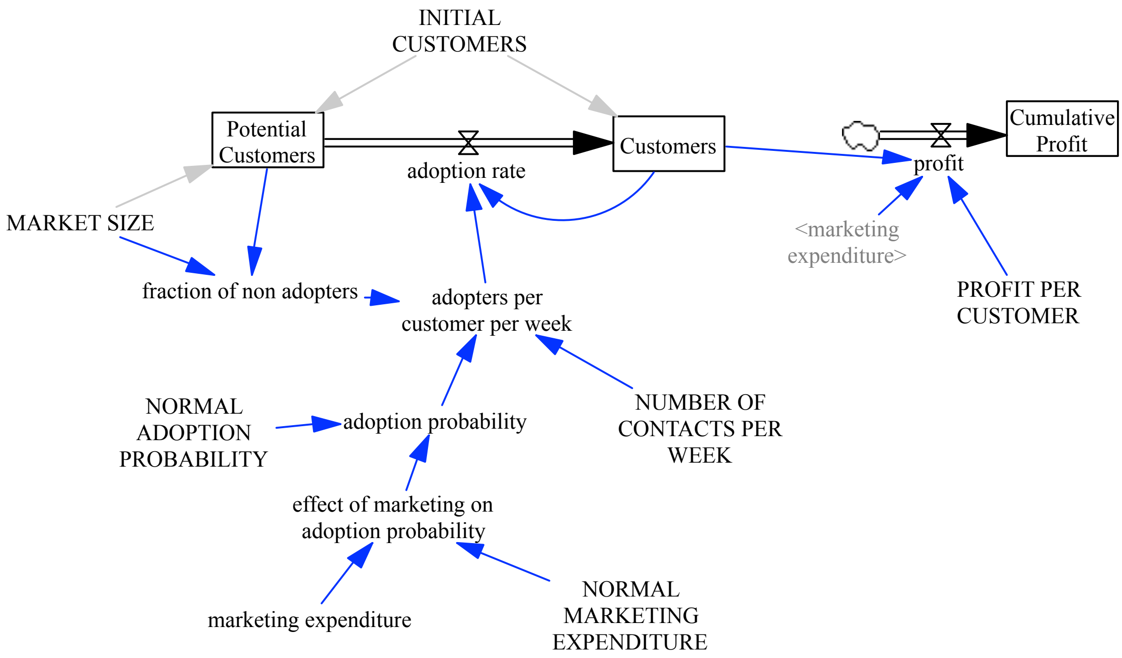

Figure 2: The Stock and Flow Diagram of the laundry detergent task model. This diagram looks very similar to that of the Bass Diffusion model, with the additional structure to track marketing expenditure and cumulative profit.

In order to fully understand the model structure lets look at the equations starting with Marketing Expenditure.

Marketing Expenditure = GAME(0) (1)

Marketing Expenditure is the decision made by the user at each step, therefore because the model was constructed using Vensim it is a gaming variable. Marketing Expenditure is then used to calculate the Effect of Marketing on Adoption Probability.

Effect of Marketing on Adoption Probability = WITH LOOKUP(Marketing Expenditure/Normal Marketing Expenditure, (0,0.05),(2.5,0.25),(5,1),(7.5,4),(10,5)) (2)

The Effect of Marketing on Adoption Probability follows an S shaped pattern. The input to the lookup table is normalized, because it divides by the Normal

Marketing Expenditure of $1,000 per week.   The output of the table is the normalized Adoption Probability. Adoption Probability is calculated as

Adoption Probability = Normal Adoption Probability * Effect of Marketing on Adoption Probability (3)

The Normal Adoption Probability is 5%, which occurs when the Marketing is $1,000. The Adoption Probability is used to calculate the Adopters per Customer per Week

Adopters per Customer per Week = Number of Contacts per Week * Adoption Probability * Fraction of Non Adopters (4)

The Number of Contacts per Week is fixed at 3, and the Fraction of Non Adopters is calculated as

Fraction of Non Adopters = Potential Customers / Market Size (5)

The Fraction of Non Adopters is used to account for the contacts between two customers each week, as opposed to a contact between a customer and a potential customer. The Market Size is a fixed constant of 2,000 people. The Potential Customers are calculated as follows

Potential Customers = STOCK(-Adoption Rate, Market Size – Initial Customers) (6)

The Potential Customers are initialized with all of the Non Customers (Market Size – Initial Customers), and the only outflow of the stock is through the Adoption Rate

Adoption Rate = Customers * Adopters per Customer per Week (7)

The Adoption Rate represents all of the new customers generated each week. The Adoption rate then drives the Customers which is calculated as

$$\text{Customers} = \text{STOCK}(\text{Adoption Rate, Initial Customers}) \tag{8}$$

The Customers and the Marketing Expenditure is then used to calculate the Profit for the week

$$\text{Profit} = (\text{Customers} * \text{Profit Per Customer}) - \text{Profit per Customer} \tag{9}$$

The Profit Per Customer is a constant fixed at $8 per customer per week. Finally, the Profit is used to calculate the Cumulative Profit

$$\text{Cumulative Profit} = \text{STOCK}(\text{Profit},0) \tag{10}$$

The Cumulative Profit uses a stock to sum up each weeks profit.

The two main feedback loops in this model are the balancing feedback loop from Potential Customers to Adoption Rate, back to Potential Customers, and the Reinforcing feedback loop from Customers to Adoption Rate back to Customers. These two feedback loops trade off dominance at the maximum of Adoption Rate. When the model is initialized with fewer then one thousand (Market Size divided by two) people the Reinforcing feedback loop is dominant, but once the number of customers is greater then one thousand the balancing loop exerts dominance. The reason for this is that Fraction of Non Adopters drops below one half, which means that each customer will contact more current customers then potential customers.

2.2 Model Parameterization and Behavior

For this experiment participant responses will be tested in two scenarios. Each scenario is a different parameterization of the model. The first scenario is the market newcomer scenario, which was created by setting the variable Initial

Customers to one hundred, which is significantly less then Market Size divided by two, which means the positive feedback loop is initially dominant. The second scenario tested is the market incumbent scenario. It was created by setting the variable Initial Customers to one thousand which is equal to Market Size divided by two which means that only the balancing feedback loop is dominant creating two very different sets of behavior.

2.2.1 The Market Newcomer Scenario

The market newcomer scenario has a variety of behavior patterns based on the market expenditure strategy undertaken by the participant. This scenario is setup such that in order to reach a high cumulative profit the participant needs to spend heavily initially causing them to start out with a negative cumulative profit. The behavior of the model under the newcomer scenario falls into three different categorizations:

1. Worse before better (Baseline)
2. Overspending and loosing money
3. Under spending and never realizing full potential

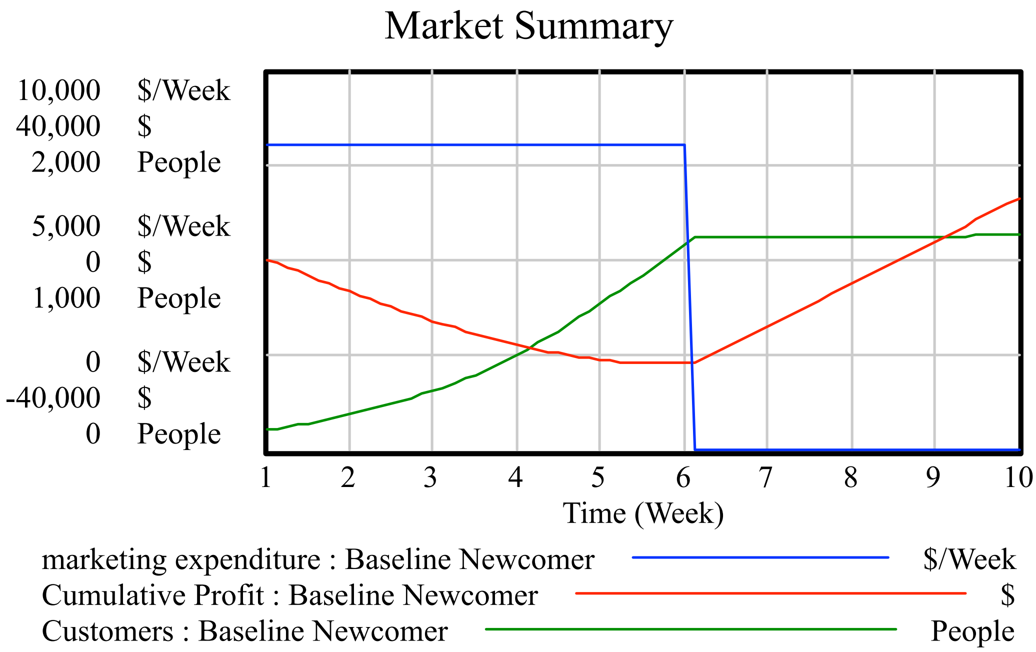

Figure 3: Result for the baseline newcomer scenario. This graph shows the marketing expenditure, number of customers and cumulative profit generated in the worse before better case.

In the worse before better behavior categorization the participant fully maximizes the utility of the reinforcing feedback loop by spending $8,000 per week for the first 6 weeks. That spending causes the participant to go approximately -$20,000 in cumulative profit. This causes the participant to capture over 50% market share ,which causes the loop dominance to shift to the balancing feedback look. At that point spending is cut off because any money spent fighting the balancing loop is wasted. Once the marketing expenditure is cut off, the cumulative profit rebounds because of the large capture of market share.

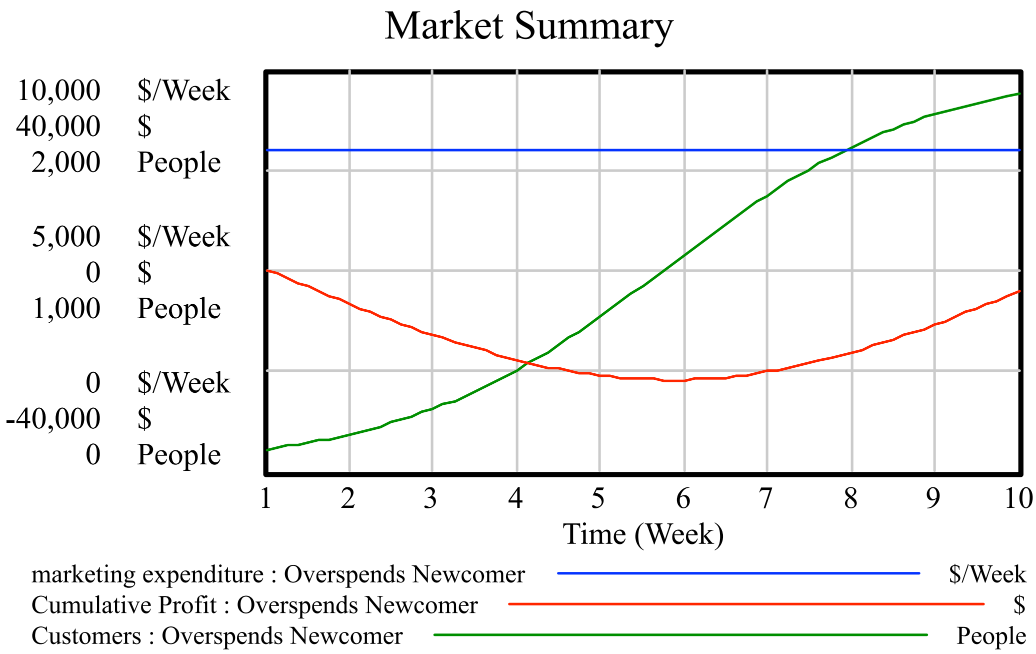

Figure 4: Results for the overspending newcomer scenario. Notice how the high marketing expenditure maximizes the number of customers, but does not maximize cumulative profit.

In the overspending market newcomer scenario the participant fully maximizes the utility of the reinforcing feedback loop by spending $8,000 per week for the first 6 weeks, but then continues to spend even after passing the inflection point in the number of customers. Therefore, even though the participants grabs almost 100% market share the cumulative profit never enters the positive region. This is because all money spent after the shift in loop dominance represents lost profit because of the effort spent combating the balancing feedback loop.

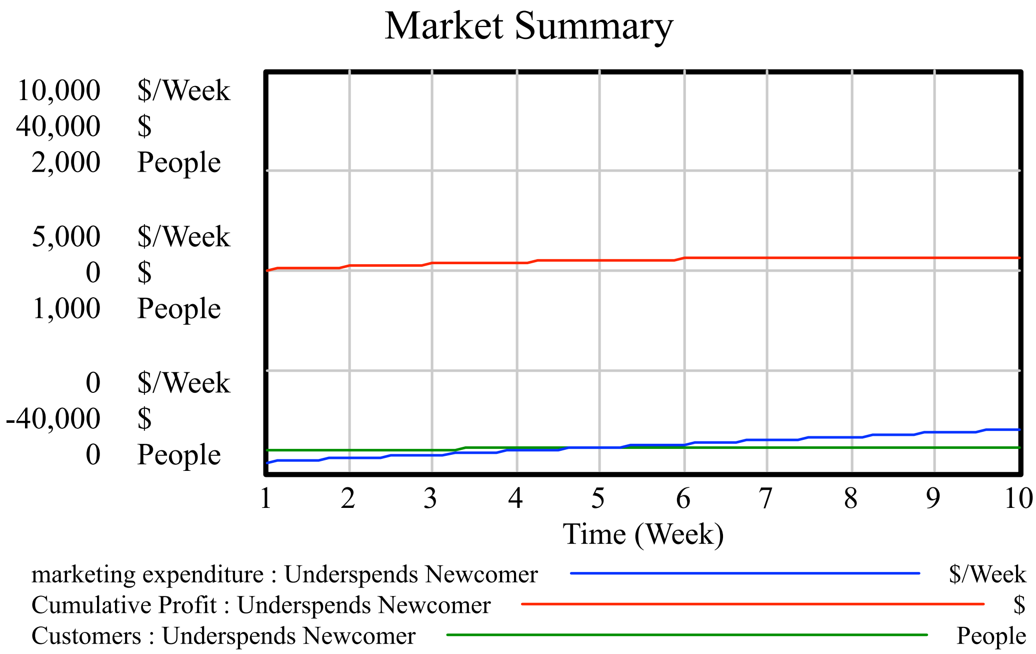

Figure 5: Results of the under spending newcomer scenario. Notice how the low marketing expenditure does not significantly increase the number of customers causing a low amount of cumulative profit.

The under spending market newcomer scenario occurs when the participant is unwilling to realize a negative cumulative profit. When this scenario occurs, almost no new customers are captured, and the only profit realized is from the initial 100 customers the participant starts with. In this scenario the reinforcing feedback loop is severely underutilized and the balancing feedback loop is never dominant.

2.2.2 The Market Incumbent Scenario

The market incumbent scenario has a variety of behavior patterns based on the market expenditure strategy undertaken by the participant. This scenario is setup such that in order to reach a high cumulative profit the participant needs to spend initially to increase their number of customers, but stop quickly so that they do not spend too much money fighting the balancing feedback loop. The behavior patterns in the market incumbent scenario can be categorized into three distinct categories

1. Quick cutoff on spending (baseline)
2. Overspending
3. Under spending

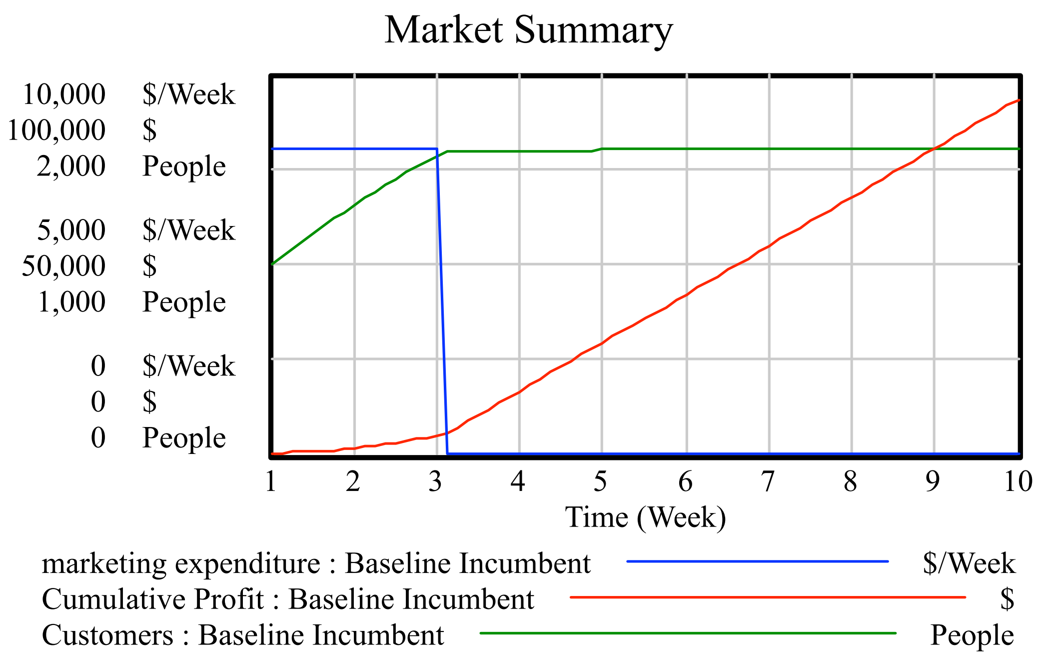

Figure 6: Results for the baseline incumbent scenario. Notice how the cumulative profit is maximized by capturing the ideal number of customers without having to combat the balancing feedback loop for too long.

In the baseline market incumbent scenario the participant maximizes the utility of the remaining strength in the reinforcing feedback loop by spending $8,000 per week for the first 3 weeks. That spending allows them to gain approximately 500 additional customers without having to loose too much profit fighting the dominant balancing feedback loop. At 3 weeks the spending is cutoff because that is the point where the balancing feedback loop becomes so much stronger then the reinforcing feedback loop that spending on marketing is wasted.

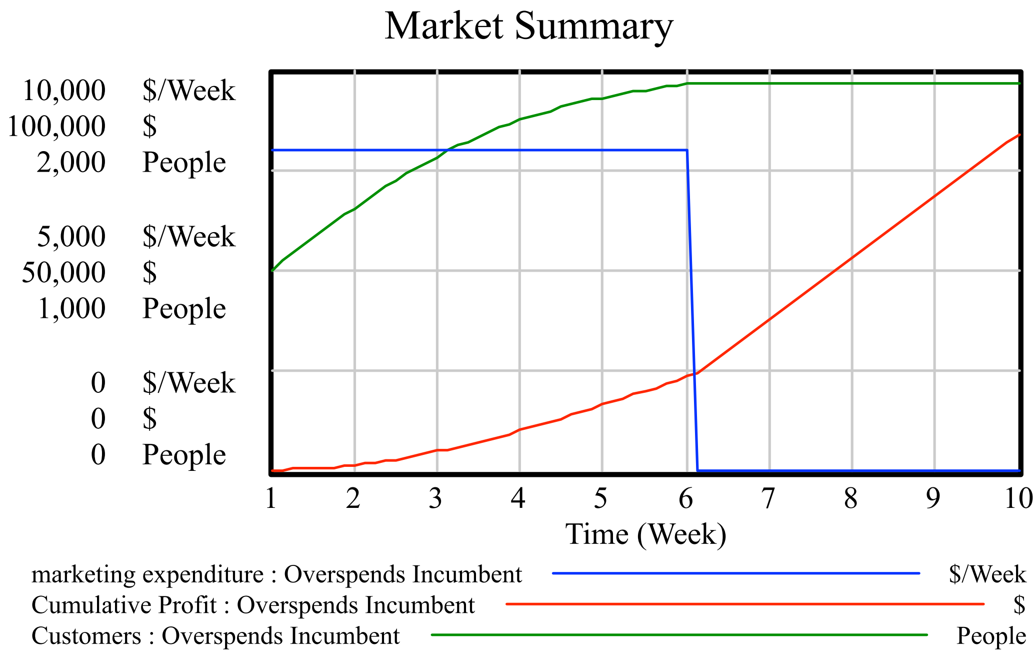

Figure 7: Results for the overspending market incumbent scenario. Notice how the high marketing expenditure maximizes the number of customer, but does not maximize cumulative profit.

In the overspending market incumbent scenario spending is not cut off at 3 weeks allowing the participant to capture a far greater market share, but at the cost of a reduced future cumulative profit. This happens because there are 3 weeks of spending ($24,000 total) which occur when the balancing feedback loop is so dominant that it would be better not to spend the money at all.

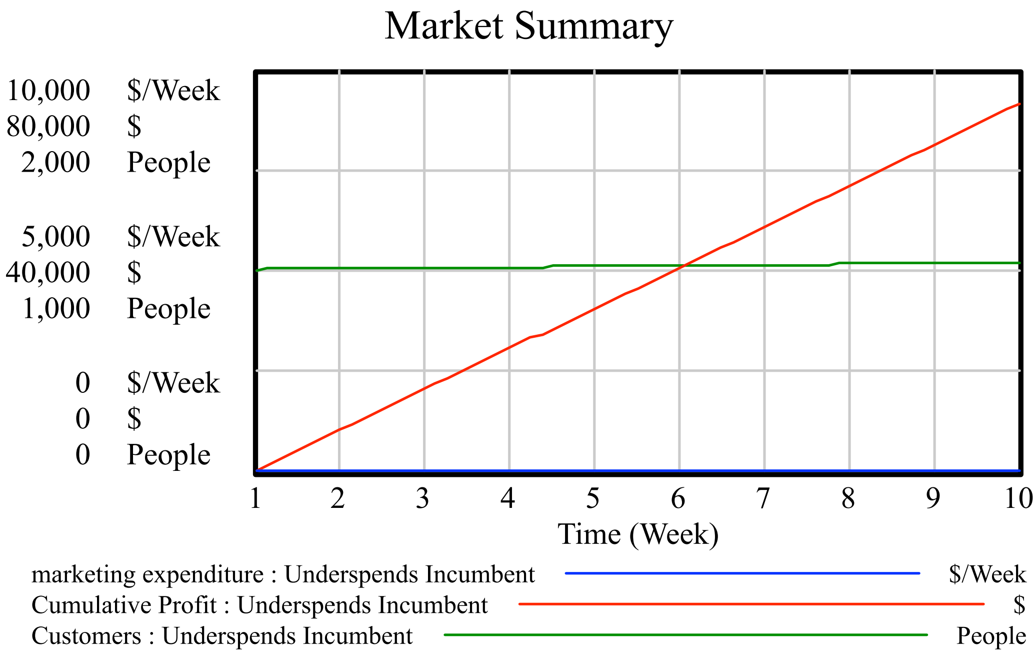

Figure 8: Results for the under spending market incumbent scenario. Notice how the non-existent marketing expenditure changes the growth of cumulative profit from exponential to linear.

Finally the under spending market incumbent scenario occurs when the participant is unwilling to spend any money during the dominant balancing feedback loop stage. Cumulative Profit is reduced in the long term because the number of customers is not raised high enough to earn the revenue required to maximize profit.

3 EXPERIMENTAL DESIGN

In order to determine whether the model explorer is a useful tool for system dynamisists to share the insights and knowledge gained from the modeling process a controlled experiment that compared the ability of a CLD versus that of a model explorer to share insights gained from the modeling process to members of the general public was constructed.

3.1 The Task

The controlled experiment required that a deterministic task based on a system dynamics model was constructed. The task developed places the participant in the role of a marketing executive at a super market. The participant has to decide how much money to spend on marketing for laundry detergent each week in order to maximize their cumulative profit.

The laundry detergent task was designed with the following requirements in order to be sure it was useful for testing the problem. The task had to:

1. Be representative of a complex dynamic system where there are decisions that need to be made by people.
2. Be simple enough to solve without knowing anything about system dynamics or having any specialized training or education.
3. Be difficult enough to solve so that the answer could not be guessed by anyone negating the need for model structure visualization.
4. Be easily re-parameterized so that it is easy to create a variety of scenarios to test whether participants understand the structure of the system or just its behavior in certain scenarios.
5. Place any bias for success onto the CLD group in order to ensure that any difference between the two groups exist because the model explorer is superior, rather then the task playing to its advantage.

These requirements ensure that the task that for this experiment did not bias the results of the experiment.

The first requirement is meant to ensure that the laundry detergent task accurately tests the participants ability to respond in a complex dynamic system. If the laundry detergent task was not representative of a complex dynamic system then it would be impossible to relate the results of the experiment back to whether or not the model explorer would help real people solve real problems.

The second requirement ensures that the participants are not overwhelmed with so much complexity that all they do is try to trial and error the problem

(Moxnes 2004). The laundry detergent task could not be so complex as to match that of a real marketing executive's job because the participants were regular people, not marketing executives. On the whole the participants lack the training required to solve a problem of that complexity regardless of the model visualization they were supplied with. Essentially, the second requirement makes sure that a regular person could learn to solve the task over the course of the experiment.

The third requirement makes sure that the participants have a reason to use and study the model visualization provided to them. If the task is too easy to solve, then the experiment would not be testing if a model explorer or CLD was more useful to solving the problem, but rather which random group of people has a better innate ability to solve this problem.

The fourth requirement ensures that the task is able to support multiple scenarios so that the experiment can confirm whether or not the participants understood the model, not just a single behavior mode of the model. In order to ensure that the fourth requirement is adequately met, the task needs to present a different shape of behavior (S shaped growth vs. goal seeking or exponential) when it is re-parameterized so that participants do not shift their behavior matching strategy from one scenario to another. Because this experiment tests whether or not people understand how the system works and function and not whether or not they can be guided into acting a certain way under a certain set of conditions.

The fifth and final requirement ensures that the results of this experiment represent the worse case scenario for the model explorer. This makes it easier to generalize these results over a wider range of tasks. If the task were biased towards the model explorer group it would be hard to determine if it were the task or the model visualization which caused any differences between the two groups.

The laundry detergent task is based on the theory behind the Bass Diffusion model which describes how potential customers transition to become actual customers based on advertising and marketing (Bass 1969). The task was based on this model in order to satisfy the above requirements.

The laundry detergent task represents a complex dynamic system for many reasons and therefore satisfies the first requirement. The first reason is that it has two key feedback loops that shift dominance over the course of the experiment. The first loop is reinforcing, the second is balancing. The task also has two stocks and it is well known that stocks are a difficult concept for regular people to understand (Diehl and Sterman 1995). It also contains a lookup table or a non linear function for translating marketing spending into adoption probability. Finally, because all of these structures are present, it has a variety of behavior modes that can all be triggered from the one decision that participants have control over.

Even though the task does represent a complex dynamic system and it contains a fair bit of complexity, it is simple enough to be solved by a regular person, - meaning that it meets the second requirement. The task does not have an overwhelming amount of complexity. In fact the Bass diffusion model, which this task is based on, is often used to introduce new students to the complexities of system dynamics models. Finally, the model is relatively small, - at 10 key equations, meaning that it can be well understood in the hour allotted for the experiment.

Because the laundry detergent task is based on a dynamic system, the answer to the task cannot be easily guessed. That means it meets the third requirement. There are many combinations of decisions, and a whole host of behavior modes which arise based on those decisions. The correct answer relies on understanding how the Bass diffusion process works, and understanding that the two feedback loops endogenously shift dominance at some point during the experiment. Because a satisfactory answer to the task requires understanding of the shifting loop dominance, the model visualization does indeed serve a purpose in the experiment, and its use (or non-use) will affect the outcomes of the participants.

Next, this task meets the fourth requirement because, by re-initializing the initial number of customers, the model produces very different behavior patterns. In fact by re-initializing the model, the dominance of the reinforcing feedback loop can be prevented from ever happening switching the behavior of the model from S shaped to a pure goal seeking. Therefore, participants who can do well across all parameterizations may be assumed to understand the model rather then just a single behavior pattern.

Finally, the task meets the final requirement because the model size is small. This ensures that any bias from the task is in favor of the CLD group. This is because the model explorer was designed to help visualize and understand large models. Hence its focuses on complexity levels and visibility levels which are completely un-used, and not useful in the laundry detergent task. This is compared to the CLD, which looses none of its advantages because of a small model and is in fact benefited because a simpler easier to understand CLD is able to show the full structure of the model and no difficult decisions have to be made about whether or not a feedback loop is worth including, given the additional complexity it would introduce in the diagram.

Also, because the two main feedback loops in the laundry detergent task model shift dominance throughout the experiment the model visualizations presented to each group should help the participants understand and learn how the model behaves when each loop is dominant so that they can transfer their knowledge from the market newcomer scenario to the market incumbent scenario. Since the treatment group has received a model explorer as a visualization tool they are potentially at a dis-advantage because the model explorer does not place the same emphasis on the feedback loops as the CLD that the control group receives. This means that the control group is at an advantage during the second scenario because their visualization highlights feedback loops which are more generally helpful to understanding behavior and may therefore learn how to perform well in the market incumbent scenario by playing the market newcomer scenario.

3.2 The Experimental Procedure

The experiment was run online over the course of 7 days. Participants were hired from Craigslist, using an ad placed in the gigs section, and paid $20 for their participation. Craigslist is an online classified ads website that allows people to post ads looking for jobs or houses, or anything. It is visited each month by approximately 63 million unique people (Site Analytics 2011). The ad was posted in the gigs / computer gigs section because the job was short , non-recurring and required knowledge of how to use a computer. The described the experiment as a fun online game that would take approximately half and hour to play and pay you $20 for the effort. The rest of the advertisement was devoted to the nuts and bolts of administrating the experiment such as how to get paid and whether or not the perspective participant had the required technology on their computer to run the experiment. The advertisement was designed in to attract the maximum number of responders (full text appendix A). Once a participant responded to the advertisement they were sent an e-mail with an anonymous user name and password, as well as a link to the simulation and were able to login at their leisure anytime during the experimental period (full text appendix B).

Participants were randomly assigned to two groups based on the order that they responded to the ad. The first user was placed in the treatment group that received a model explorer as a visual aid, the second user in the control group that received a CLD. This pattern was kept up for all participants. There were 33 in the treatment group and, 31 in the control group. Each participant who signed up was compared to all other participants using their provided e-mail address and payment details in order to prevent duplicate signups from the same person.

Each participant ran the experiment six times, three times under the newcomer scenario, three times under the market incumbent scenario. The two scenarios were required in order to test whether or not each participant understood the model structure, not just one behavior mode. The reason each participant ran each scenario three times was so that their learning could be measured as they progressed through the experiment. It also gave the participants a chance to get familiar with the UI and the concept of a simulation since the majority of the participants had never performed a task like this ever before.

3.2.1 The Experimental User Interface

The User Interface (UI) for this experiment was developed using Forio Simulate and its UI Designer tool. The simulation was administered using an interface built into the simulation which allowed the administrator to specify the number of times each participant could run the simulation as well as to see all data entered by the participant during the simulation in real-time separated by treatment group.

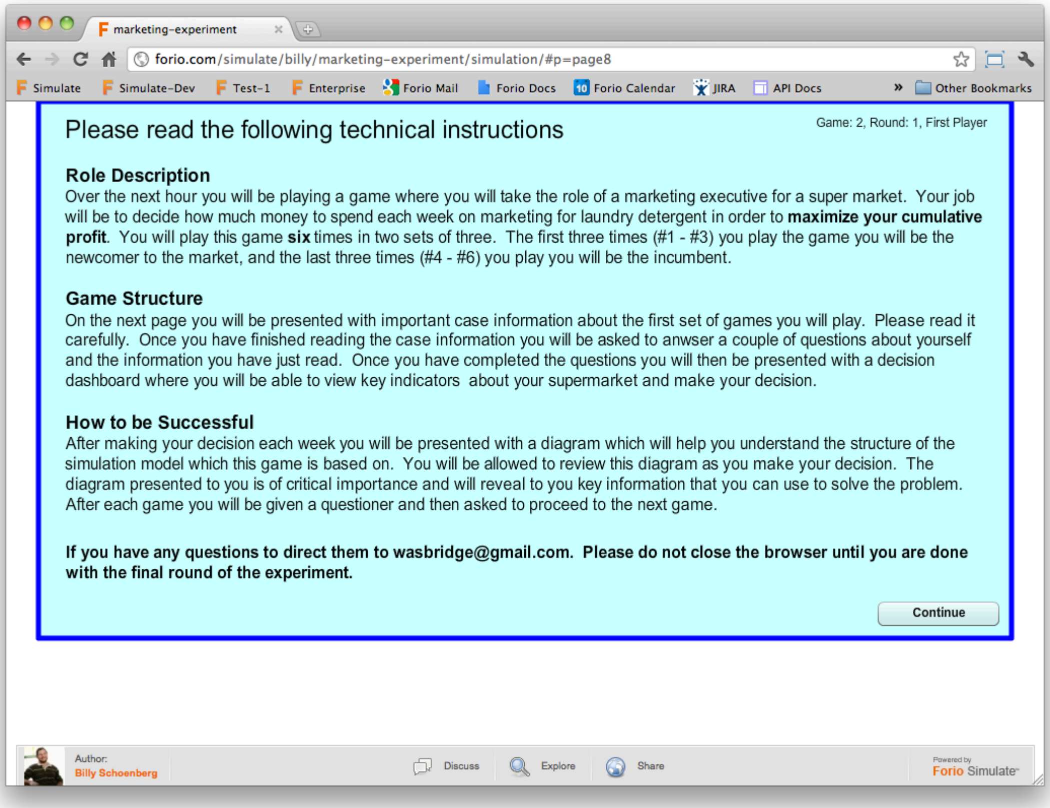

Figure 9: The Technical Instructions Screen. This screen is first seen by participants after logging in. It describes their role, the game structure and how to be successful.

The first page seen by participants after logging in is the technical instructions page. This page describes to the participant their role in the experiment, their goal for the task, an overview of the game structure, and a blurb about how to be successful in the game. This page is the first one exhibited because it gives the participant the context that they need in order to understand what they are trying to accomplish. The role description helps the participants visualize what they should be doing, and the goal description lets them know how they will be judged. The game structure information is useful to allow the participant to manage their expectations for what will happen over the course of the game, and it makes it clear to them what exactly is expected of them. Finally, the how to be successful blurb lets them know that the model visualization diagram is an important tool that they will have to use in order to be successful.

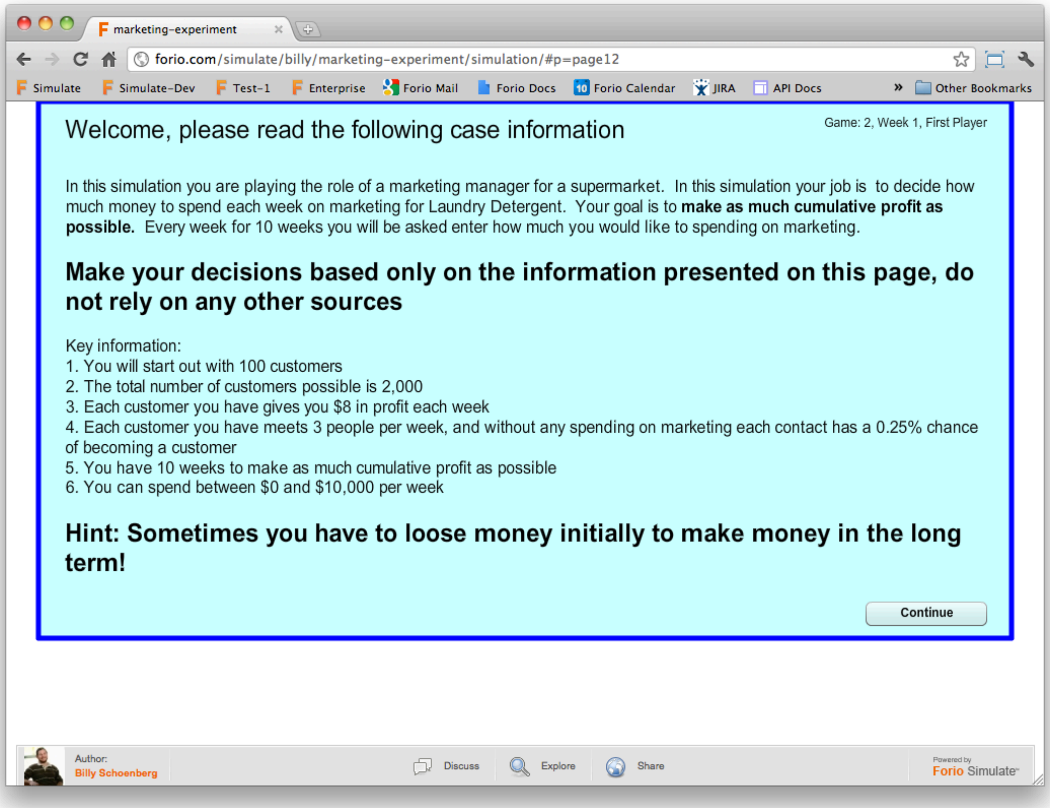

Figure 10: The Case Information Screen. This is the second screen that all participants see after logging in. It re-confirms the goal, and shows key information about the model parameterization. It also gives a helpful hint about spending into the negative.

After reading the technical instructions, the participants then went on to read the case information page. This page re-introduced the goal, warned users to rely only on information that the experiment had presented to them, showed them all of the key parameters in the model and finally gave them another hint about how to be successful. The goal information was bolded in order to make it stand out, and to ensure that it was read and understood. The blurb about relying only on information presented on the case information page was made big and bold because it was important that the participants did not assume things about the model that weren't true. The other big and bolded information was the hint that let participants know that it was okay to have a negative cumulative profit, and that in fact they may have to do that in order to be successful. The parameter information helped to set their expectations for how the model ought to have performed and set their expectations for the magnitude of spending they were allowed to make.

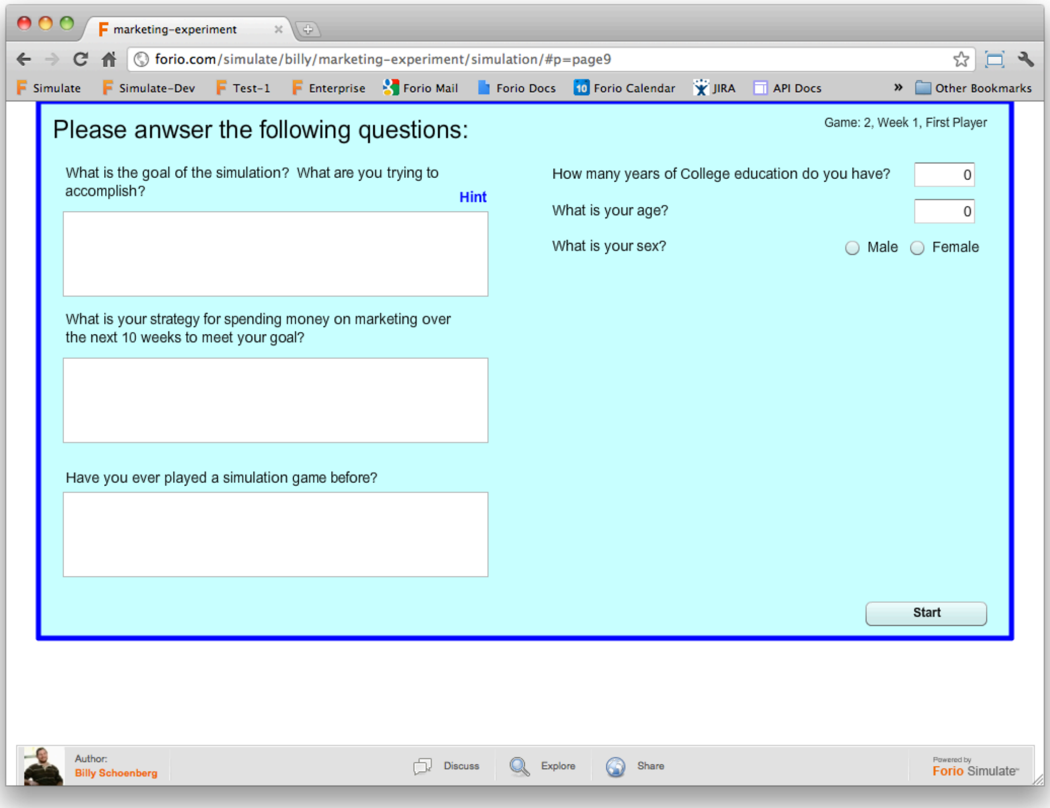

Figure 11: The Pre-Test Survey. This page is used to collected information from the participants before they play the game for the first time.

After reading the case information participants were then shown the pretest survey which asked the following questions:

1. What is the goal of the simulation? What are you trying to accomplish?
2. What is your strategy for spending money over the next 10 weeks to meet your goal?
3. Have you ever played a simulation game before?
4. How many years of College education do you have?

5. What is your age?
6. What is your sex?

The first question was asked in order to make sure they understood their purpose in the game. Since many people will skip reading instructions the experiment forced the participants to re-capitulate their goal to me, and if they hadn't read the instructions, there was a hint for them look at in order to answer the question. Any participant who cannot answer this question correctly will be removed from the analysis.

The second question was asked in order to force the participant to think about a coherent strategy for how they were going to meet their goal. It is important to make sure that each participant had in their mind a set of steps for how they were going to accomplish their goal so that when they started playing the game they were not just doing something random.

The third question was asked in order to determine whether the participant had any experience with a simulation of this type before. In order to make sure the two experimental groups shared the same approximate innate ability to solve this task, the experiment needed to check that one group was not filled with trained experts while the other was untrained.

The last set of questions was asked in order to make sure that the results of the experiment were based on the different model visualization tool rather then age, sex or level of education. In order for this experiment to speak about the model visualization all results need to be tested to make sure that they cannot also be explained by age, sex or level of education.

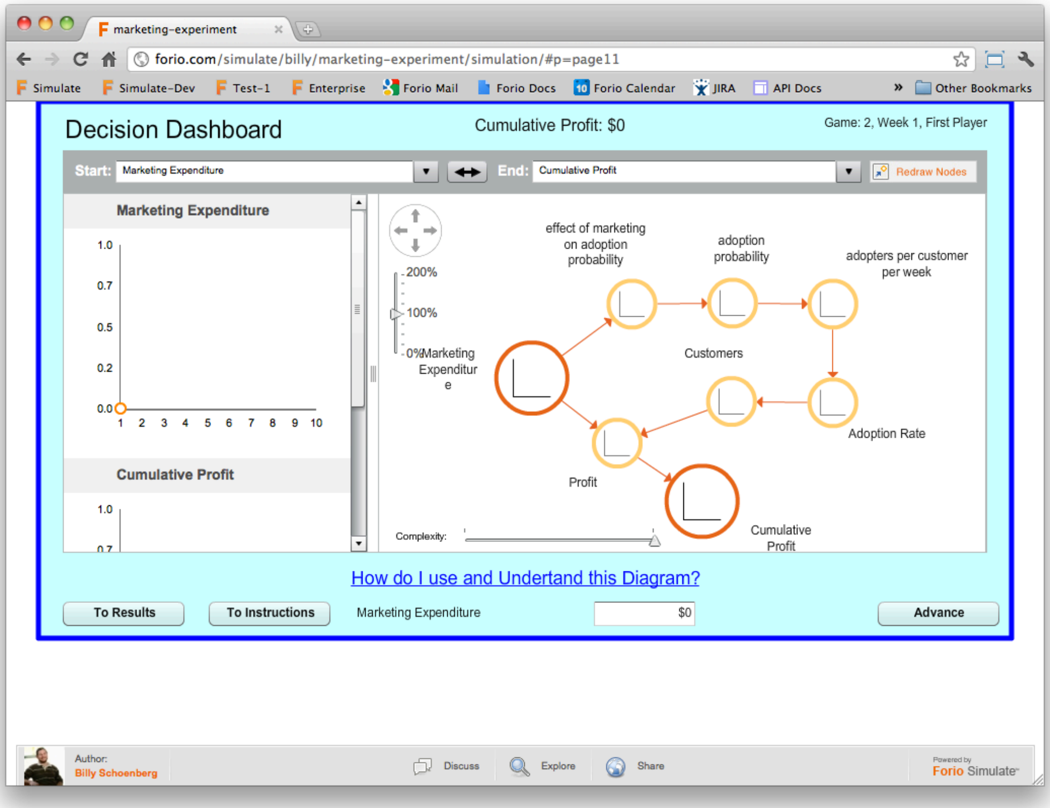

Figure 12: The Game Page - Model Explorer Group. This page is seen only by those in the model explorer group. It allows them to make their decision, advance the game, see the structure of the model and review and instructions or additional results they want.

After completing the pre-test, participants were then presented with the game page. On this page participants made their decision, advanced the simulation, viewed the model visualization, were able to review the instructions, could view their cumulative profit, or head to another page to view graphs of results. The game page had at its top a text label showing the participants current profit. This was placed on the top of the page in large text in order to make it very easy to see and check on their progress toward their goal. Directly below the cumulative profit label was the model visualization which takes up ~80% of the screen real-estate. This is done so that participants have to view the model visualization and cannot avoid seeing it. Below the model visualization was a link in big blue text that read “How do I use and Understand this Diagram?”. This link when clicked would bring them to a 1 page document that described the visualization they were looking at and how to use it. At the bottom of the page from left to right there was a button to go to graphical results, a button to view the instructions, a text entry for entering their Marketing Expenditure, and a button to advance the simulation.

For the model explorer group the model visualization presented was a model explorer. It was parameterized such that the start variable was the participants Marketing Expenditure, and that the end variable was the current Cumulative Profit. The visualization then showed a graph of all of the relationships between those two variables. Participants could clearly see that there were two ways that Marketing Expenditure affected profit, the first and most direct one was that spending reduced profit, which reduced cumulative profit, the second was that spending drove adoption probability which drove the adoption rate which increased the number of customers which increased profit. Participants where allowed to use the model explorer in any way they desired, but after each press of the advance button the model explorer reverted to displaying the above described diagram.

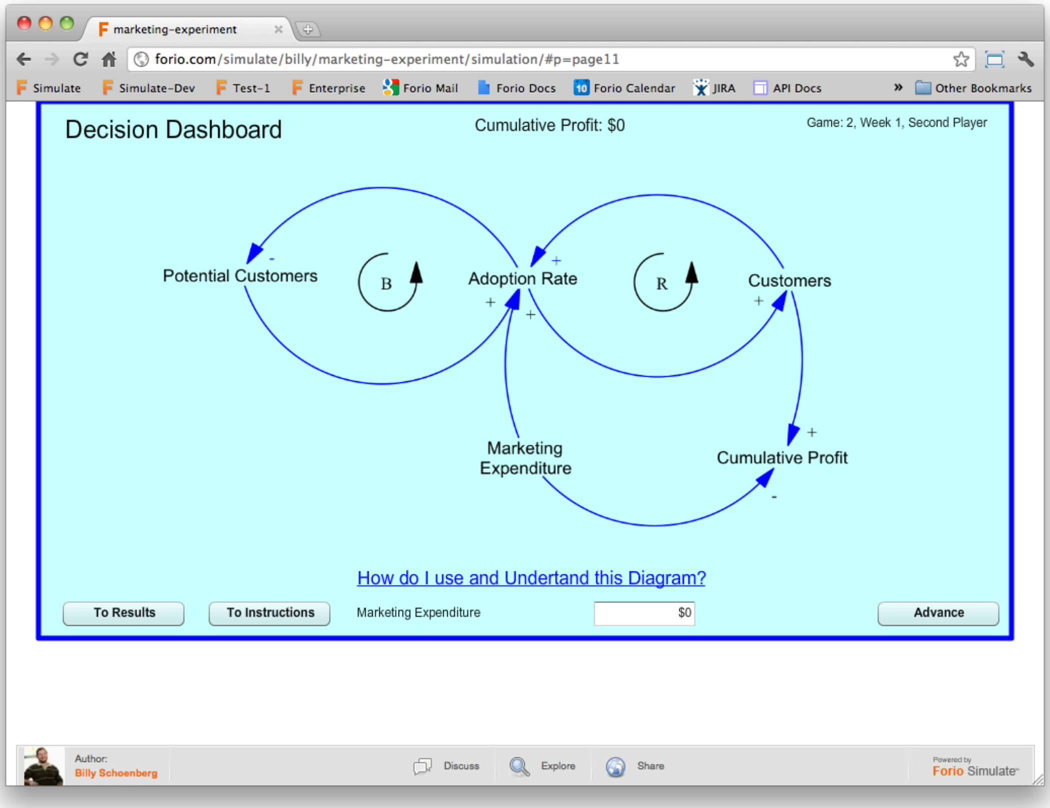

Figure 13: The Game Page - CLD Group. This page is seen only by those in the CLD group. It allows them to make their decision, advance the game, see the structure of the model and review and instructions or additional results they want.

For the CLD group, the model visualization presented was a CLD. This CLD labeled the two feedback loops in the model as well as the relationship between Marketing Expenditure and Cumulative Profit. The CLD included the appropriate link polarities and symbols to differentiate the reinforcing and balancing feedback loops.

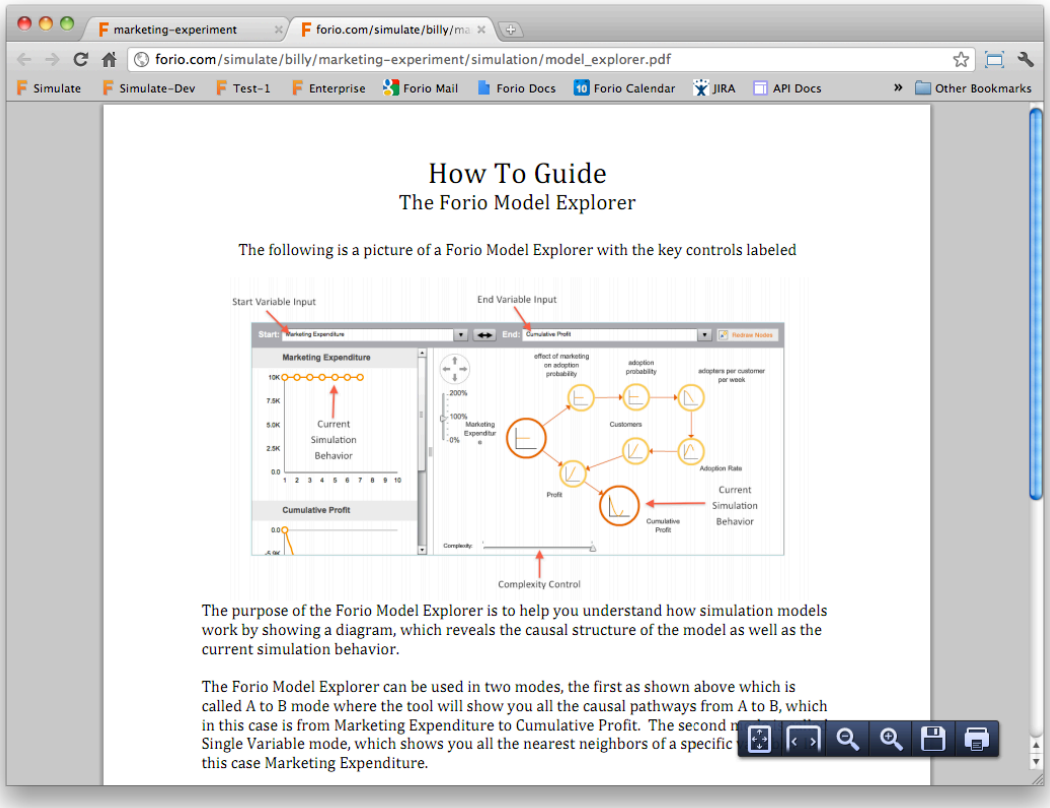

Figure 14: How To Guide: Model Explorer Group. This page opens in a new window when those in the model explorer group want to learn more about how the model explorer works. It shows a labeled diagram of the model explorer and describes how to use and understand the tool

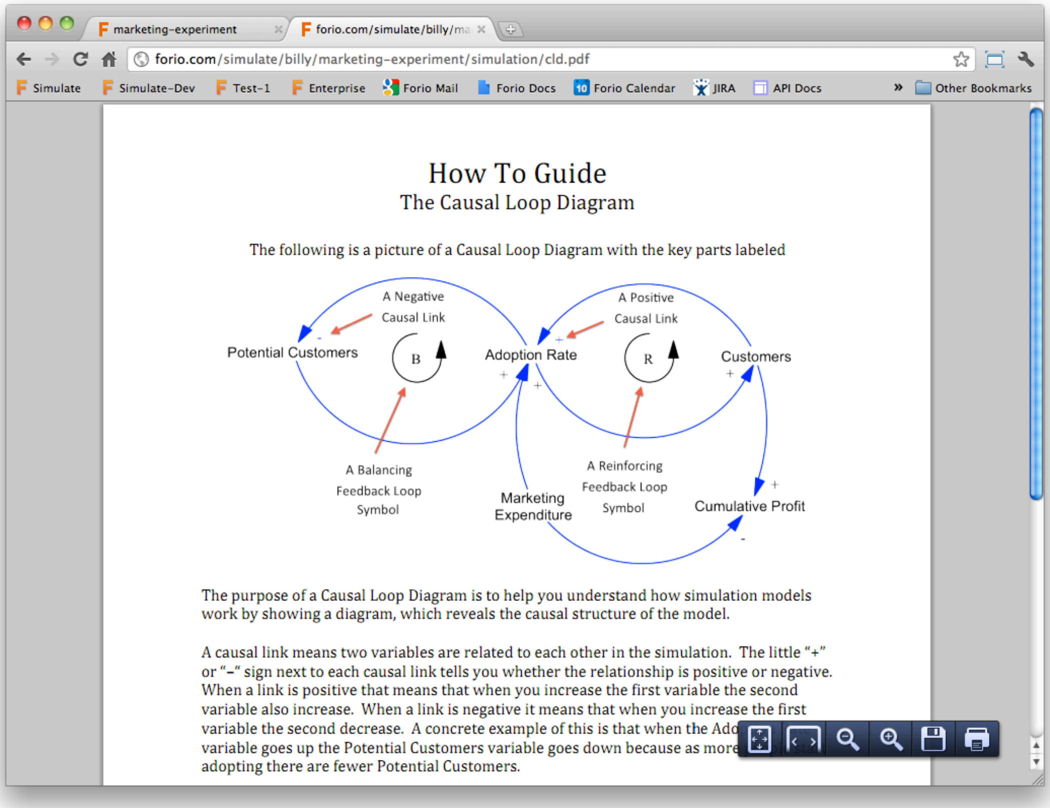

Figure 15: How To Guide - CLD Group. This page opens in a new window when those in the CLD group want to learn more about how to interpret a CLD. It shows a labeled diagram of CLD and describes how to use and understand the diagram.

As mentioned above, both treatment groups had the ability to read a 1 page document about how to use their specific model visualization tool. That link when clicked opened in a separate tab in the participants' browser with the document. For both groups the document followed the same form. First there was a labeled picture of the model visualization tool. Each key item had an arrow pointing to it along with a few word explaining what it was. Below the labeled diagram was a sentence which described how the model visualization tool could help you. For the model explorer, it explained that the tool showed both model structure and model behavior, for the CLD group it described how the CLD showed model structure. Then each document went on to explain each of the labeled items in the diagram picture and how to use each one of those items to understand the model.

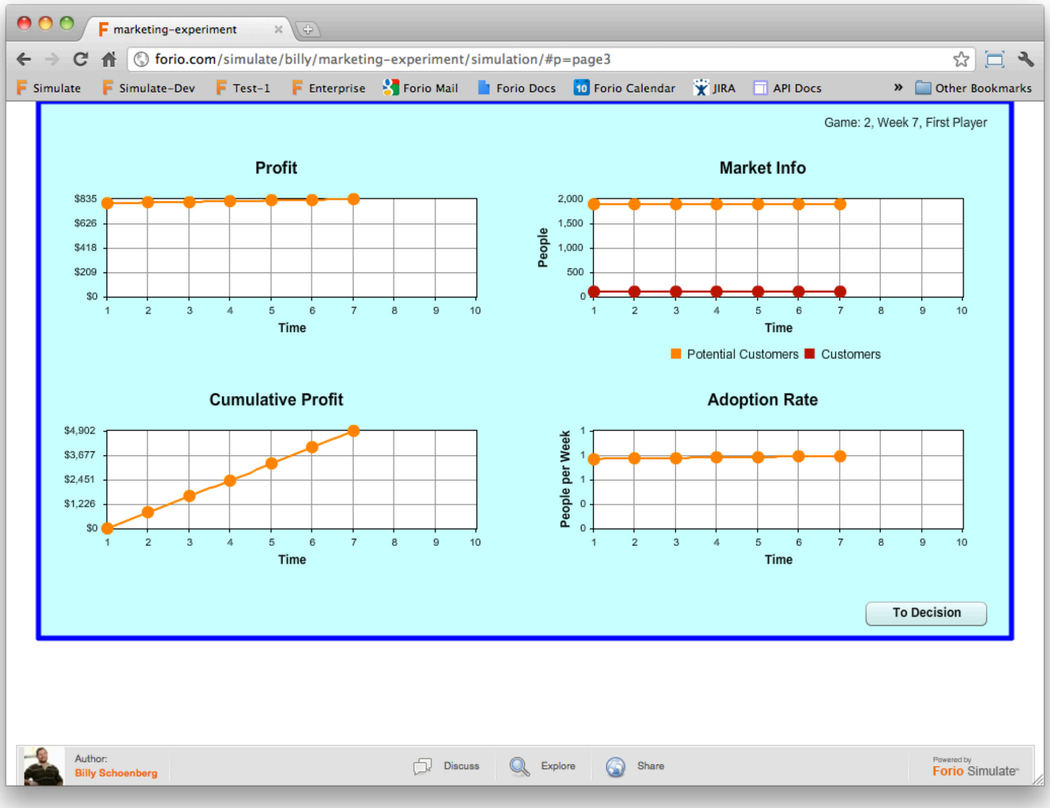

Figure 16: The Results Page. This page is visible by both groups and shows key in progress results that the participants can use to judge their progress towards their goals. Indicators for this page were picked based on their importance to solving the task.

Also available from the game page was a link to view a page of results. On this page participants from either group were able to view 4 graphs of key indicators in the model. In the upper right was a graph of profit, which was chosen to allow the participant to see how well they did each round. Next to that was a graph of Market Info which showed on the same graph the Potential Customers vs. the Customers. This graph was present in order to help the participant understand their market and to specifically see the relationship between Potential Customers and Customers. Below the Profit graph was a graph of Cumulative Profit. This graph was included because it is the goal of the simulation. Finally in the lower right-hand portion of the page was a graph of the Adoption Rate which was key to the participant understanding how much to spend on marketing. This graph was included because the optimal set of marketing decisions has the participant no longer spending money on marketing after the Adoption Rate has peaked.

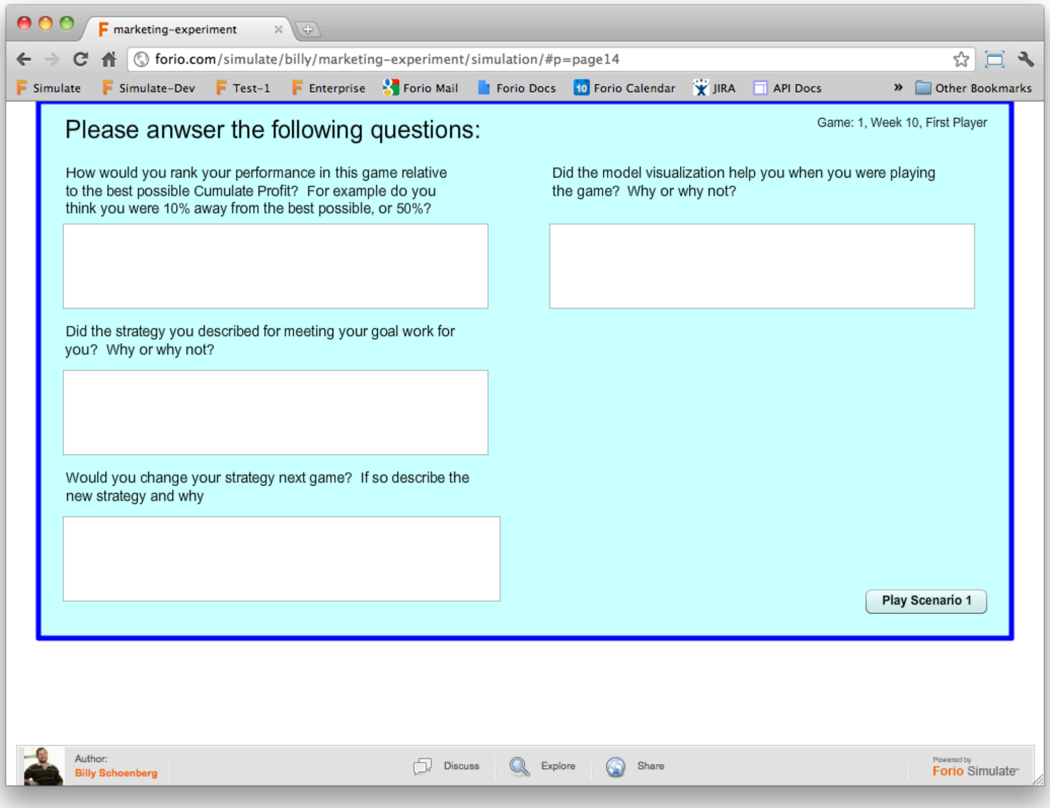

Figure 17: The Post Test Survey. This page is used to collected information from the participants after they play each game

After the participant had finished each round of the experiment they were presented with a page of post test questions:

1. How would you rank your performance in this game relative to the best possible Cumulate Profit? For example do you think you were 10% away from the best possible, or 50%?
2. Did the strategy you described for meeting your goal work for you? Why or why not?
3. Would you change your strategy next game? If so describe the new strategy and why.

4. Did the model visualization help you when you were playing the game? Why or why not?

The first question was asked in order to judge whether or not the participant thought they did well on the task. The reason for this question was to get them thinking about what the best possible score was so that they could adjust their strategy appropriately rather then just re-play with the same decisions.

The second and third questions asked were a follow-on to the first and were meant to get the participant thinking about what they were going to do in the next game. These two questions were designed to learn about any pitfalls or blocks that people experienced when trying to do well in the game. It is important to know whether or not they could textually describe their understanding of the model through a market expenditure strategy.

The fourth question was asked in order to determine whether the participant actually utilized the model visualization tool during that round. This is necessary to test for the participants preferences.

3.3 Hypotheses

Since the model explorer is a continuation and a refinement of the ideas behind the CLD and stock and flow diagrams, I believe that the treatment group (model explorer) will out perform the control group (CLD). This hypothesis is formally stated as:

H_0: *There will be no difference in the cumulative profit of any run between the control and treatment groups at an alpha of 0.05.*

H_1: *There will be a difference in the cumulative profit of at least one run between the control and treatment groups at an alpha of 0.05.*

Next, I predict that the model explorer will help participants improve more in the same number of trials as compared to the CLD. This hypothesis can be formally stated as:

> H_0: *There will be no difference in the improvement of cumulative profit between any two runs between the control and treatment groups at an alpha of 0.05.*

> H_1: *There will be a difference in the improvement of cumulative profit between any two runs between the control and treatment groups at an alpha of 0.05.*

Next, I believe that those who find the model explorer useful will perform better then those who find the CLD useful. This hypothesis can be formally stated as:

> H_0: *There will be no difference in the cumulative profit of any run between the control and treatment group participants who have identified the model visualization as helpful at an alpha of 0.05.*

> H_1: *There will be a difference in the cumulative profit of at least one run between the control and treatment groups participants who have identified the model visualization as helpful at an alpha of 0.05.*

Finally, I believe that those who find the model explorer useful will improve more over the same number of trials then those who find the CLD useful. This hypothesis can be formally stated as:

> H_0: *There will be no difference in the improvement of cumulative profit between any two runs between the control and treatment group groups participants who have identified the model visualization as helpful at an alpha of 0.05.*

H_1: There will be a difference in the improvement of cumulative profit between any two runs between the control and treatment group groups participants who have identified the model visualization as helpful at an alpha of 0.05.

4 RESULTS

4.1 The Market Newcomer Scenario

The first scenario ran by the participants was the market newcomer scenario. The average best cumulative profit in the control group (CLD) was $4,045 vs. $3,841 for the treatment group (model explorer). All averages were then tested for differences using a two-tailed Students T test in order to derive a P-value that was tested at an alpha of .05.

Table 1: Results for Scenario 1 – The best row shows the averages for the run with the highest cumulative profit.

Run	Control (N=28)	Treatment (N=31)	P-value	Result
1	-$1,621 ($8,615)	-$5,820 ($10,223)	0.0953	No Diff.
2	-$1,519 ($9,588)	-$1,354 ($10,936)	0.9512	No Diff.
3	$1,667 ($6,662)	-$14 ($9,869)	0.4516	No Diff.
Best	$4,045 ($4814)	$3,841 ($6,059)	0.8872	No Diff.

Note: standard deviation between parentheses. P-value derived from two-tailed Students T test vs. alpha of .05

The results were then filtered to not include any participant who did not find their model visualization useful in order to remove the influence of participants who did not utilize the model visualization when making decisions. With this filter applied the average best cumulative profit was $6,840 for the control group vs. $4,814 for the treatment group.

Table 2: Results for Scenario 1 filtered – Any participant who responded that the model visualization was not useful was removed and all statistics were re-calculated

Run	Control (N=4)	Treatment (N=14)	P-value	Result
1	$5,469 ($2,779)	-$3,775 ($11,857)	0.1490	No Diff.
2	$5,532 ($2,563)	$1,731 ($8,756)	0.4126	No Diff.
3	$6,786 ($176)	$407 ($11,131)	0.2787	No Diff.
Best	$6,840 ($206)	$4,814 ($5,785)	0.5031	No Diff.

Note: standard deviation between parentheses. P-value derived from two-tailed Students T test vs. alpha of .05

Next the average difference between each run for each participant in scenario 1 were compared. The average improvement in cumulative profit from run 1 to 3 in the control group was $3,288 vs. $5,806 in the treatment group.

Table 3: Scenario 1, Differences in cumulative profit – This table shows the average differences in cumulative profit between each run for each treatment group. Positive numbers show improvement while negative numbers would show a regression in ability.

Between	Control (N=28)	Treatment (N=31)	P-value	Result
1 to 3	$3,288 ($9,628)	$5,806 ($13,325)	0.4133	No Diff.
1 to 2	$101 ($11,326)	$4,466 ($15,133)	0.4109	No Diff.
2 to 3	$3,186 ($7,241)	$1,340 ($9,570)	0.2189	No Diff.

Note: standard deviation between parentheses. P-value derived from two-tailed Students T test vs. alpha of .05

Again all participants who responded that their model visualization was not helpful, were filtered out of the results. This changed the average difference in cumulative profit from run 1 to run 3 in the control group to $1,317 vs. $4,182 in the treatment group.

Table 4: Scenario 1, Differences in cumulative profit filtered – This table shows the average differences in cumulative profit between each run for each treatment group filtered to not include participants who did not find their model visualization useful.

Between	Control (N=4)	Treatment (N=14)	P-value	Result
1 to 3	$1,317 ($2,639)	$4,182 ($14,992)	0.5031	No Diff.
1 to 2	$62 ($255)	$5,506 ($13,781)	0.4509	No Diff.
2 to 3	$1,254($2,419)	-$1,324 ($8,441)	0.5621	No Diff.

Note: standard deviation between parentheses. P-value derived from two-tailed Students T test vs. alpha of .05

4.2 The Market Incumbent Scenario

The second scenario run by the participants, was the market incumbent scenario. The average best cumulative profit in the control group (CLD) was $75,242 vs. $76,781 for the treatment group (model explorer).

Table 5: Results for Scenario 2 – The best row shows the averages for the run with the highest cumulative profit.

Run	Control (N=31)	Treatment (N=33)	P-value	Result
1	$67,619 ($13,756)	$68,175 ($14,503)	0.8756	No Diff.
2	$68,802 ($14,147)	$69,996 ($14,165)	0.7371	No Diff.
3	$72,368 ($13,072)	$72,572 ($13,483)	0.9513	No Diff.
Best	$75,242 ($11,776)	$76,781 ($12,593)	0.7451	No Diff.

Note: standard deviation between parentheses. P-value derived from two-tailed Students T test vs. alpha of .05

The results were then filtered to not include any participant who did not find their model visualization useful in order to remove the influence of participants who did not utilize the model visualization when making decisions. With this filter applied the average best cumulative profit was $74,946 for the control group vs. $76,202 for the treatment group.

Table 6: Results for Scenario 2 filtered – Any participant who responded that the model visualization was not useful was removed and all statistics were re-calculated

Run	Control (N=7)	Treatment (N=20)	P-value	Result
1	$69,403 ($11,707)	$65,691 ($15,318)	0.9072	No Diff.
2	$72,533 ($8,324)	$68,966 ($16,120)	0.4590	No Diff.
3	$72,188 ($10,083)	$72,198 ($14,952)	0.9732	No Diff.
Best	$74,946 ($7,336)	$76,202 ($14,292)	0.8272	No Diff.

Note: standard deviation between parentheses. P-value derived from two-tailed Students T test vs. alpha of .05

Next we compered the average difference between each run for each participant in scenario two. The average improvement in cumulative profit from run 1 to 3 in the control group was $4,749 vs. $4,397 in the treatment group.

Table 7: Scenario 2, Differences in cumulative profit – This table shows the average differences in cumulative profit between each run for each treatment group. Positive numbers show improvement while negative numbers would show a regression in ability.

Between	Control (N=31)	Treatment (N=33)	P-value	Result
1 to 3	$4,749 ($12,666)	$4,397 ($14,738)	0.9314	No Diff.
1 to 2	$1,183 ($12,766)	$1,820 ($12,020)	0.8377	No Diff.
2 to 3	$3,566 ($13,195)	$2,576 ($10,606)	0.7719	No Diff.

Note: standard deviation between parentheses. P-value derived from two-tailed Students T test vs. alpha of .05

Again all participants who responded that their model visualization was not helpful were filtered out of the results. This changed the average difference in cumulative profit from run 1 to run 3 in the control group to $2,784 vs. $6,508 in the treatment group.

Table 8: Scenario 2, Differences in cumulative profit filtered – This table shows the average differences in cumulative profit between each run for each treatment group filtered to not include participants who did not find their model visualization useful.

Between	Control (N=7)	Treatment (N=20)	P-value	Result
1 to 3	$2,784 ($6,404)	$6,508 ($16,042)	0.5595	No Diff.
1 to 2	$3,130 ($12,375)	$3,275 ($13,245)	0.9799	No Diff.
2 to 3	-$345 ($7049)	$3,233 ($10,956)	0.4300	No Diff.

Note: standard deviation between parentheses. P-value derived from two-tailed Students T test vs. alpha of .05

5 DISCUSSION

The results of this experiment show that there are no differences between the control group (CLD) and treatment group (model explorer) in any run, or in the improvement from run to run. In the market newcomer scenario, when the results are unfiltered for model visualization usefulness, the largest difference of $4,199 between the two experimental groups is on their first run where the control group performed better. The P-value in this case is 0.0953 showing a no difference result. On run 2 the two groups show an almost identical result with a difference of only $165 in favor of the treatment group with a P-value of 0.9512. This means that there was a difference of $4,365 in favor of the treatment group between the average improvement from run 1 to 2. But this difference too was statistically insignificant with a P-value of 0.4109. Finally, in the third run the difference between the two groups was $1,653 in favor of the control group which yielded a statistically insignificant P-value of 0.1888. This means that there was a difference of $1,846 in favor of the control group between the improvement from run 2 to run 3. That difference was also statistically insignificant with a P-value of 0.2189. Even when examining the best score of each participant in runs 1 – 3, there was no significant difference with a P value of 0.7451 or when examining the total change from run 1 to 3 which had a P value of 0.4133 even with a $2,518 difference in average improvement favoring the treatment group.

After filtering out all responses from both groups in the market newcomer scenario which did not find the visualization provided useful; all results still showed no difference between the control and treatment groups. In run 1 there was a $9,424 difference between the control and treatment groups favoring the control group the P value was 0.1490. In run 2 there was a $3,801 difference in improvement between the two groups favoring the control group, but this difference was still insignificant with a P value of 0.4126. This means that there was an average difference between the two group from run 1 to run 2 of $5,444 in favor of the treatment group, but that difference was also insignificant with a P value of 0.4509. In run 3 there was a $6,379 difference between the two groups favoring the control group, but that difference too was statistically insignificant with a P value of 0.2787. The average difference in improvement from run 2 to run 3 was $2,578 which was statistically insignificant with a P value of 0.5621. Finally even the difference of $2,026 favoring the control group in the average best run generated by each participant was statistically insignificant with a P-value of 0.5031.

For the market incumbent scenario there continued to be no statistically significant differences between the two groups. In the unfiltered results for run 1, there was a difference of $556 favoring the treatment group which had a P-value of 0.8756. For run 2 that difference was $1,194 which was still insignificant with a P-value of 0.7371. This meant that there was difference of $637 in the improvement from run 1 to run 2 favoring the treatment group, but this was insignificant as well with a P-value of 0.8377. In run 3 there was a difference of $204 favoring the treatment group and that difference was insignificant with a P-value of 0.9513. This meant that there was a difference of $990 in the improvement between runs 2 and 3 favoring the control group and that difference was insignificant with a P-value of 0.7719. Finally, the difference in between the best run from each participant in each group which was $1,539 in favor of the treatment group was insignificant with a P-value of 0.7451.

The same filtering was applied, removing any participant who did not find the visualization useful and the lack of difference between the two groups still held. In the filtered results for run 1 there was a difference of $3,712 favoring the control group, but that difference was insignificant with a P-value of 0.9072. For run 2 the difference dropped to $3,567 which was insignificant with a P-value of 0.4590. This meant that the difference in improvement between the two groups was $145 which was insignificant with a P-value of 0.9799. For run 3 there was a difference of $204 in favor of the treatment group and that difference was insignificant with a P-value of 0.9513. That meant that the difference in improvement between runs 2 and 3 was $3,578 in favor of the treatment group, which was insignificant with a P-value of 0.4300. Finally, there was the difference between the best run from each participant in each group which was $1,256 in favor of the treatment group, - insignificant with a P-value of 0.8272.

The most likely potential reason that we do not see participants who have the model explorer doing statistically better then those having the CLD is because the model explorer does not do as good of a job as the CLD highlighting the feedback loops present in the system, which makes it harder for those participants to understand that shifting feedback loop dominance should drive their behavior. At the same time though, participants with the CLD do not perform better then those with the model explorer because they lack ability to link their model visualization to the behavior that they see. The combination of these reasons begins to explain the results that we observe.

In the market newcomer scenario (scenario 1) the results without filtering follow the expected trends. With each successive trial the average cumulative profit increases. From this we would expect that when the results of those who did not find the model visualization were removed from our sample, we would see the same increasing trends from run to run, but starting with a higher cumulative profit, and increasing at a higher rate. Yet when the results are filtered to remove those participants the behavior does not match expectations; instead in both groups we do see the behavior starting off at a higher point ($5,469 for the control group and $-$3,775 for the treatment group), but the constantly increasing trend from run to run does not occur in the treatment group. In run 3 for the filtered treatment group we see a dip of $1,324 in the average cumulative profit. This behavior brings up the question of why do those who find the model visualization useful fail to continually increase their profit from run to run.

It would seem that for the filtered treatment group, the results do not match the standard findings in the literature that participants start with poor performance and in general with more trials perform better (Sawicka 2005, Paich and Sterman 1993, Moxnes 1998, 1998, 2000, 2004, Jensen and Brehmer 2003). In order to determine why, a review of the scenario 1 run 2 post test questions was in order. On average, these participants rated themselves as having been in the top 67% of maximum with a standard deviation of 29%. With this level of confidence in their results, we would expect these participants to make minor adjustments to their strategies to optimize their results rather then starting from scratch with a new strategy. Yet 10 out of the 14 participants decided to make a major change to their strategy. And out of the 11 participants who did worse from run 2 to run 3, 7 were part of the sub-group that changed their strategy. Those 7 also represented the worst 7 improvements averaging a decrease in improvement of $18,271 when comparing the improvement from run 1 to run 2 to the improvement from run 2 to run 3. Although not all participants who made a major change in strategy ended up doing worse. The three participants who changed their strategy successfully were the top 3 most improved, showing an increasing average improvement of $12,786 from run 2 to run 3.

The next major difference between the filtered groups and the unfiltered groups in the market newcomer scenario is that even though the filtered group achieved a higher best cumulative profit, their rate of improvement is less than the unfiltered group average. This is interesting because we would expect those who found the model visualization useful to not only perform better, but to improve quicker (with the caveat that there actually was room to improve, which there was), - yet the results fail to back that assertion. In both the treatment and control groups, we see less overall improvement from runs 1 – 3 in the filtered groups vs. the unfiltered. We would expect those who utilize and find the model visualization tools helpful to perform better because they ought to have a better understanding of how the system works when compared to their peers who disregard the model visualization. Yet it seems that this cannot be the case because of their stunted improvement. The data seem to suggest that those who found the model visualization useful, were not able to fully understand and utilize their diagram to make improvements at a higher rate then the total population.

In the market incumbent scenario (scenario 2) we also have some interesting differences between the filtered and unfiltered groups in both the control and treatment groups. In this scenario, the average best run of the filtered groups was less then the average best run of their respective unfiltered groups. In the market incumbent scenario, finding the model visualization useful did not correspond with getting a higher score. This is unexpected behavior because the literature says that those with an understanding of the system should perform better then those without (Spector 2000). In order to test this fully, the experiment would have to be modified to add a 3rd group which received no model visualization in order to determine whether there were any statistically significant differences between the groups with a model visualization and those without.

The improvement for the filtered groups in the market incumbent scenario was not always lower than that of the unfiltered groups as it was in the market newcomer scenario. For instance, the treatment group that was filtered, showed more improvement then its unfiltered counterpart across both runs, while the same did not hold true for the control group. For the control group the filtered sub-group showed more improvement from run 1 to run 2, rather then continuing to increase from run 2 to run 3, it showed a slight decrease.

The next interesting result is the very small number of participants in the control group who found the model visualization helpful for all 3 runs in either scenario. In the market newcomer scenario only 4 of 28 (14%) found the CLD helpful when making decisions compared to 45% (14 of 31) in the treatment group. For the market incumbent scenario the ratio was 7 to 31 (23%) compared to 20 of 33 (61%) for the treatment group. This is interesting because the CLD is supposed to be a relatable form of model structure visualization in the toolkit of system dynamisists (Sterman 2000). This result could have come about for a variety of reasons, the first being that the participants in this experiment were not students who had studied system dynamics or other engineering/mathematical concepts, - rather anonymous members of the general public. For them, the complexity of even the CLD was probably too much to handle.

The second reason why a higher percentage of participants found the model explorer help as compared to the CLD is that since people tend to skip reading instructions the participants in both groups did not read, or did not fully digest the text in the How to Use this Diagram link (Moxnes 1998). This would have left the users of the CLD confused about how their diagram was supposed to help them. Whereas for those in the model explorer group, even if they did not understand the model structure visualization, may still have benefited from the behavior visualization portion of the model explorer. This means that those using the model explorer could be classified into two levels of understanding. The first would be those who understood the behavior portion of the visualization only and the second level would be those who understood the behavior and structure visualization combined. Since the instructions for the model explorer were really unnecessary to understand the behavior portion of the visualization, the average non-instruction reading participant would have found the model explorer at least partially helpful, while same participant would find the CLD to be completely opaque.

Another interesting result is that while many participants were able to articulate a coherent and viable strategy, when it came time to execute, they would waver and change approaches halfway through the game, resulting in sub-optimal results. Let us take for example participant 22 in the treatment group during run 1 of the market newcomer scenario. This user answered the pre-test strategy question saying the following:

> "Heavy advertising at the start to get the customers up, then easing it off as word of mouth helps to increase customers on its own"

This strategy, if followed, would lead to nearly optimal results, yet this user chose to spend between $4,000 and $5,000 from time 1 – 6. Subsequently, when their adoption rate peaked they changed their strategy to spend $6,000 till the end. This participant started the simulation off spending below what we would expect for someone who wanted to spend heavily initially, and then completely ignored their own strategy by increasing their spending just as the word of mouth effect was wearing off. This decision represented the worst possible decision that could be made and that decision made the user the worst performer in run 1 of the market newcomer scenario in both the treatment and control groups. This participant offered the following reason as for why they changed their mind:

> "The adoption was low enough I ended up abandoning the strategy and increased my advertising costs, which while giving me a much bigger negative profit short-term seemed to be turning itself around quicker than lower advertising."

This answer seems to indicate an understanding of the linkage between customers and profit, but a misunderstanding of how the word of mouth effect depletes as the number of customers increases. It seems that the participant saw their declining cumulative profit and was not satisfied with the number of customers they had captured. Therefore they decided to focus on capturing more customers rather then to cut their losses and start spending less. This participant then suggested the following strategy for their next run:

> "Going much heavier on the advertising out the gate next time to hopefully get the customer count up much quicker."

With this adjustment in strategy the participant saw a much larger initial growth in customers and may have felt that they had reached the maximum utility of the word of mouth effect, therefore cutting off their marketing expenditure which lead them to one of the top scores in the market newcomer scenario. Consequently, it seems that results generated by participant 22 in the treatment group was driven by an improving decision heuristic, developed upon an understanding of the simulation structure, and subsequently reinforced by the results the participant generated.

The next interesting result is the uniformly high standard deviations around the means in both scenarios and groups, filtered or unfiltered. For instance, in the market newcomer scenario one standard deviation in either direction from the mean of the third run in either the treatment or control groups represents a range of strategies from nearly optimal to terrible. This means that the range of strategies employed by the participants, were not consistent. And since the standard deviations did not shrink noticeably as the number of trials increased, this means that the participants did not come up with the same strategies at the same times. The only place where we see a significant decrease in standard deviations is between the individual runs and the best runs from any of the three runs. This suggests that while each group saw a wide array of strategies in use for each of the runs 1 through 3, in the best cumulative profit run, only a subset of those strategies were used.

While running and analyzing the results of this experiment, it has become clear that some flaws exist. For instance, because this experiment was not facilitated in person, it is hard to tell how much effort was put into solving the problems presented. To ameliorate this concern, the experiment could be improved by adding questions asking the participants to rate the effort expended in solving the task for each run. This would allow an analysis to be performed in which we try to correlate effort with performance, and to investigate whether differences may exist between the two treatment groups on that basis.

The next difficulty due to the lack of in person facilitation is that it is impossible to tell whether outside resources were utilized, or whether participants communicated with each other. Since, in order to collect participants for this experiment, the word of mouth effect was utilized, there is a high probability that participants spoke to one another about the experiment before both had completed the exercise. In order to account for this problem, the experiment could be modified to take place at a single date and time using webinar software which reports the currently selected window on the users screens. By forcing the experiment to happen at one time, we can be sure that the participants cannot inadvertently tell others about the task before they themselves have completed it. Also, by utilizing a webinar tool which reports the currently selected window on the participants screen, we could be sure that they maintained their focus on the experiment for the entire time.

Another problem related to the lack of in person facilitation is that it is possible for one person to pose as multiple participants. While this was controlled for by checking for duplicate e-mail addresses it would be best to run the experiment using webinar software mentioned above forcing each person to be present in the online webinar in order to participate and then to prevent participants who have the same IP address from participating. While limiting participation via IP address potentially prevents participants from the same network from participating it does ensure with the combination of the webinar software that each participant is indeed unique.

The next problem was associated with the amount of text presented during the simulation. Even though all of the participants who were analyzed in this study read enough of the instructions to be able to repeat the goal of the experiment, it is doubtful that they all fully utilized all the documentation and hints given to them. In order to test for a more even distribution of information about the experiment, the experiment could include web-analytics on how often each page was viewed and for how long. This would allow an analysis to be completed to show whether or not all instructions present were read. Another approach to solving this problem may be to take as many instructions as possible and provide them in both video and text form. This would allow participants who learn by watching or reading to experience the full information present in the experiment.

6 CONCLUSIONS AND FURTHER RESEARCH

For the laundry detergent task, there is no statistical difference between the control group (CLD) and treatment group (model explorer). Since the task was based on a small model which plays to the strengths of the CLD, we can be reasonably sure that the model explorer represents a viable alternative tool that can be used by system dynamisists to share the knowledge gained through the modeling process. The implications of this research suggest that system dynamisists should use the model explorer as a way to present their work to decision makers and the general public. Since the model explorer is another avenue to reach people and help them understand how to solve complex dynamic problems, this represents a new way to help the field grow and reach new wider audiences.

This study answers many questions about the effectiveness of the model explorer, but it also raises many more. Areas for future study include testing on large complex models, because the current techniques and tools system dynamisists use are weakest there. Also, how important is it to know where the stocks are? Would adding differentiation to variable types help make the model explorer a more helpful tool, or what about changing the visualization to highlight feedback loops more easily? Finally, this experiment used completely untrained random participants, but what would happen if trained system dynamisists were used? Would they find the model explorer to be more or less useful then the general public?

7 REFERENCES

Bass, F. (1969) "A new product growth model for consumer durables". Management Science **15**(5):215-227

Brehmer, B. (1992). "Dynamic decision making: Human control of complex systems." Acta Psychologica **81**: 211-241.

Diehl E, Sterman JD. 1995. "Effects of Feedback Complexity on Dynamic Decision Making." Organizational Behavior and Human Decision Processes **62**:198-215

Funke, J. (1991). "Solving complex problems: Exploration and control of complex systems." Complex problem solving: principles and mechanisms. R. Sternberg and P. Frensch. Hillsdale, NJ, Lawrence Erlbaum**:** 185-222.

Größler, A., Maier F.H., and Milling P.M. (2000) "Enhancing learning capabilities by providing transparency in business simulators." Simulation & Gaming **31**(2):257-278

Jensen, E. (2005). "Learning and transfer from a simple dynamic system." Scandinavian Journal of Psychology **46**(2): 119-131.

Jensen, E. and B. Brehmer (2003). Understanding and control of a simple dynamic system. System Dynamics Review **19**(2): 119-137.

Moxnes E. (1998). Overexploitation of renewable resources: The role of misperceptions. Journal of Economic Behavior & Organization **37**:107-27

Moxnes, E. (1998). "Not only the tragedy of the commons: Misperceptions of bioeconomics." Management Science **44**(9): 1234-1248.

Moxnes, E. (2000). Not only the tragedy of the commons: misperceptions of feedback and policies for sustainable development. System Dynamics Review **16**(4): 325-348.

Moxnes, E. (2004). "Misperceptions of basic dynamics: the case of renewable resource management." System Dynamics Review **20**(2): 139-162.

Moxnes, E. and A. K. Saysel (2009). "Misperceptions of global climate change: information policies." Climatic Change **93**(1): 15-37.

Paich M, Sterman JD. 1993. Boom, Bust, and Failures to Learn in Experimental Markets. In Management Science, p. 1439: INFORMS: Institute for Operations Research

Rouwette, E. A. J. A., A. Größler and J. A. M. Vennix (2004). "Exploring influencing factors on rationality: a literature review of dynamic decision-making studies in system dynamics." Systems Research and Behavioral Science **21**(4): 351-370.

Site Analytics. "Craigslist.org UVs for September 2011" *Compete*, Site Analytics. Web. 13 Nov. 2011. <http://siteanalytics.compete.com/craigslist.org/>

Sawicka, A., Gonzalez, J. J., Qian Y., (2005). "Managing CSIRT Capacity as a Renewable Resource Management Challenge: An Experimental Study" Proceedings of the 2005 System Dynamics Conference 133-163

Spector, J. M. (2000). “System Dynamics and Interactive Learning Environments: Lessons Learned and Implications for the Future” Simulation & Gaming **31**(2): 528-535

Sterman, J. D. (1989). "Misperceptions of feedback in dynamic decision making." Organizational Behavior and Human Decision Processes **43**(3): 301-335.

Sterman, J. D. (1989). "Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment." Management Science **35**(3): 321-339.

Sterman, J. (2000). "Business dynamics: Systems thinking and modeling for a complex world”. Boston: Irwin/McGraw-Hill.

Sterman, J. D. and L. Booth Sweeny (2007). "Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter." Climatic Change: 213-238.

Warren, K. and Langley, P. (1999). "The effective of system dynamics to improve insight and learning in management education" The Journal of the Operational Research Society **50**(4): 396-404.

Wolstenholme, E.F. (1999) "System Dynamics for Policy, Strategy and Management Education" The Journal of the Operational Research Society: 422-428

8 APPENDIX A

The following is a copy of the advertisement placed on Craigslist.

Title:

Participate in Fun Online Experiment

Body:

Would you like to get paid $20 for no more then half an hours work playing a game online?

If so please send me (wasbridge@gmail.com) an email with subject "I'd like to participate". In that e-mail please answer the following questions:

1. Do you have Adobe Flash Player installed? What version (http://kb2.adobe.com/cps/155/tn_15507.html)?
2. When would you be able to play the game (the sooner the better)?

If you answer the first two questions yes, and are able to play sometime this weekend or this week I will send you an e-mail with a URL and some login information. Once you login you will be able to play the game. The game should take no more then 1 hour to play and for many people will take less. During the game you will be asked some questions. All results will be kept 100% confidential, but you do have to answer every question. If you leave one blank, or skip one I will not be able to pay you. Once you have finished playing the game you will e-mail me and tell me that you have finished. You will tell me what your login was, I will double check that I have all the information I need from you then I will pay you right away!

Interested? If so e-mail wasbridge@gmail.com. I will be waiting!

9 APPENDIX B

Form letter replying to potential participants

Hi,

Thanks for signing up.

Your login info is:
XXXX@forio.com
XXX

The URL of the game is:
http://forio.com/simulate/billy/marketing-experiment/simulation/

When you are done please e-mail me, I will check that I have all the data that I need from you and then I will get the info I need from you to send you an Amazon gift card.

Billy

P.S. I am looking to have ~80 people run this game over the next week do you know any others who may be able to help?

10 APPENDIX C

Technical Instructions

Role Description

Over the next hour you will be playing a game where you will take the role of a marketing executive for a super market. Your job will be to decide how much money to spend each week on marketing for laundry detergent in order to maximize your cumulative profit. You will play this game six times in two sets of three. The first three times (#1 - #3) you play the game you will be the newcomer to the market, and the last three times (#4 - #6) you play you will be the incumbent.

Game Structure

On the next page you will be presented with important case information about the first set of games you will play. Please read it carefully. Once you have finished reading the case information you will be asked to anwser a couple of questions about yourself and the information you have just read. Once you have completed the questions you will then be presented with a decision dashboard where you will be able to view key indicators about your supermarket and make your decision.

How to be Successful

After making your decision each week you will be presented with a diagram which will help you understand the structure of the simulation model which this game is based on. You will be allowed to review this diagram as you make your decision. The diagram presented to you is of critical importance and will reveal to you key information that you can use to solve the problem. After each game you will be given a questioner and then asked to proceed to the next game.

11 APPENDIX D

Case Instructions. The {$X} variables are replaced by model values based on the current scenario.

In this simulation you are playing the role of a marketing manager for a supermarket. In this simulation your job is to decide how much money to spend each week on marketing for Laundry Detergent. Your goal is to make as much cumulative profit as possible. Every week for 10 weeks you will be asked enter how much you would like to spending on marketing.

Make your decisions based only on the information presented on this page, do not rely on any other sources

Key information:

1. You will start out with {$1} customers
2. The total number of customers possible is {$2}
3. Each customer you have gives you {$3} in profit each week
4. Each customer you have meets {$4} people per week, and without any spending on marketing each contact has a {$5} chance of becoming a customer
5. You have 10 weeks to make as much cumulative profit as possible

Hint: Sometimes you have to loose money initially to make money in the long term!