USING CONTROLLED EXPERIMENTS TO CALIBRATE COMPUTER MODELS: THE AIRPLANE GAME AS A LEAN SIMULATION EXERCISE

Zofia K. Rybkowski¹, John-Michael Wong², Glenn Ballard³ and Iris D. Tommelein⁴

ABSTRACT
Simulation games may be used to introduce lean principles to those who are considering implementing them. However, they can also function as controlled experiments against which to calibrate a computer model and they can even be adapted to serve as the gold standard of scientific experimentation, the randomized-controlled trial. Results generated from a live playing of the Airplane Game validate an EZStrobe computer-based simulation model representing one part of the game. Close alignment of results suggests that the computer model will likely be able to accurately predict outcomes from similarly structured, real life activities, such as those encountered in a design office or on a construction site.

KEY WORDS
lean, Airplane Game, discrete-event simulation, controlled experiment, randomized controlled trial, EZStrobe

INTRODUCTION
Lean construction methodologies, such as those implementing production schedule levelling, pull (kanban), just-in-time delivery, Last Planner™, mistake proofing (poka yoke), and continuous improvement, are gaining widespread acceptance within the construction industry (e.g., Salem et al. 2006). Case studies suggest that application of lean principles can favourably impact a construction project’s budget and schedule (e.g., Khanzode et al. 2005; Kim et al. 2007; Koerckel and Ballard 2005; Pasqualini and Zawislak 2005; Seppanen and Aalto 2005; Simonsson and Emborg 2007). Case studies are helpful because they illustrate the application of lean principles to environments for communities of interest, such as design
Using Controlled Experiments to Calibrate Computer Models: The Airplane Game as a Lean Simulation Exercise

Zofia K. Rybkowski, John-Michael Wong, Glenn Ballard and Iris D. Tommelein

People, Culture and Change

and construction. However, case studies also represent microcosms of complexity. Confounding variables exist and case studies generally lack key elements of rigorous scientific inquiry, such as the establishment of an experimental control group for testing the impact of a single variable. Therefore, it is difficult to quantify the relative magnitude of benefits of individual lean principles and assess the best order in which to introduce them. Detty and Yingling (2000) argue that adopting lean methodologies based on reported experiences of others requires relying on general rules of thumb and "faith-based justification." To systematize the quantification of lean, they used discrete-event simulation to measure the relative impact of the more easily modelled aspects of lean, such as continuous flow, just-in-time inventory management, quality at the source, and production schedule levelling. Aspects of lean applied to the architecture-engineering-construction industry (AEC), such as the impact of variation, multi-tasking, and batching on production system performance have also been simulated using computers (Arbulu et al. 2002a, 2002b; Disney et al. 1997; Farrar et al. 2004; Tommelein 1999). Pioneering computer simulations of matching problems and standardization in a production system for pipe-spool supply and installation (Tommelein 1998, 2006) demonstrate relationships between both product and process variability on buffering and productivity, and show the benefits of using real-time control, aka. "pull" (kanban), to improve performance. While computer simulations offer a level of control that case studies lack, sceptics may argue it is difficult to know if their output accurately represents magnitudes obtained on actual projects.

To overcome some of these limitations, Tommelein et al. (1999) developed a game to simulate the impact of workflow variability on the productivity of construction trades. The simulation, called the Parade of Trades Game, introduced random variation by rolling a die by hand, but the game could also be played using a computer. The Lean Construction Institute has been using a "teaching simulation" from Visionary Products Inc. (2007, 2008) and refers to it as the "Airplane Game" to introduce concepts of lean to construction project teams. Inspired by the Airplane Game, Sacks et al. (2005, 2007) subsequently developed a computer model of a live simulation game called LEAPCON™. The LEAPCON™ model helps to investigate the separate and combined influences of specific lean interventions and monitor them through time. Independent variables included batch size, multi-skilling, pull versus push; and dependent variables such as work in progress (WIP), completed units, and cash flow were measured. A number of other lean simulation games have been documented by Verma (2003). By their decision to use a simulation game as a testing ground for combinations of lean principles, the researchers implied their belief that such games can serve as models for real life scenarios. More investigations using various lean simulation games, similar to those described would enhance our understanding of lean.

This paper focuses on the Airplane Game. This game tests lean principles in four separate phases, adding new principles to each phase successively, and using the prior phase as the successive phase's control for most,
Using Controlled Experiments to Calibrate Computer Models: The Airplane Game as a Lean Simulation Exercise

Zofia K. Rybkowski, John-Michael Wong, Glenn Ballard and Iris D. Tommelein

Although experimental methodology varies by field and an investigation may be limited by real constraints, valid scientific experimentation generally follows five steps: (1) formulate a hypothesis, (2) randomly assign participants to the intervention group or to the control group, (3) measure the dependent variable(s) in one or both groups, (4) introduce the treatment or intervention, and (5) measure the dependent variable(s) again. A hypothesis is a testable, proposed explanation for an observed phenomenon that predicts a relationship between an independent variable and the dependent variable(s).

Although experimental methodology varies by field and an investigation may be limited by real constraints, valid scientific experimentation generally follows five steps: (1) formulate a hypothesis, (2) randomly assign participants to the intervention group or to the control group, (3) measure the dependent variable(s) in one or both groups, (4) introduce the treatment or intervention, and (5) measure the dependent variable(s) again. A hypothesis is a testable, proposed explanation for an observed phenomenon that predicts a relationship between an independent variable and the dependent variable(s).

The most rigorous form of experimentation, the randomized, controlled trial (RCT) is primarily used in clinical research, but its principles can be applied to social science research as well. RCTs are often considered to offer the highest degree of reliability (lowest confounding) of results because, in addition to the requisite control group and randomization of participants, there is "double blinding" of both experimenters and subjects (Leandro 2005; Sandercock 1993). During double blinding, neither experimenters...
nor subjects know whether they are part of either the experimental or control group; therefore, they cannot (sub)consciously influence the results—a phenomenon sometimes referred to as a “placebo effect.”

Promotion of the conduct of RCTs sits at the heart of The Cochrane Collaboration (Cochrane 2008), an organization dedicated to enhancing reliability of research results. While The Cochrane Collaboration's expressed mission is improve health care decision-making, their methodology can be applied to any field seeking to improve the rigor of its research methodology. For example, van der Molen et al. (2007) assessed research on injury prevention in the construction industry.

METHODOLOGY

DESCRIPTION OF EXPERIMENT

An experiment was conducted on Phase 2 of the Airplane Game (Figure 1) as described by the manufacturers, but with some modification. According to the instruction manual, players should be seated around a table in four assembly workstations, one quality control station, and one teardown station. Workstations are arranged in cellular layout, each station with a supply of specific Lego® blocks. Completed Lego®-block assemblies from each workstation are passed, sequentially, to the next workstation for further transformation, until a Lego® airplane is assembled, checked for defects, and torn down. Figure 2 shows, at the centre, the seating layout as illustrated in the instruction manual.

Because the purpose of this paper is to show how a lean game might serve as controlled experiment against which to calibrate a computer model, only two lean principles were tested: the concepts of (1) pull vs. push and (2) batching. The processes were modelled using EZStrobe (Martinez 1996, 2001; Martinez and Ioannou 1999) and calibrated against a run with actual players. To keep this paper short, only workstations 1 through 4 are shown in Figure 2. Workstations 5 and 6 for quality control and teardown were not included in the live simulation as described, nor added to the computer model. We may extend the model at a later time to include additional workstations.

Playing the game required three types of Lego® blocks: 4-pin, 8-pin, and 16-pin. Quantities for each block type were initialized to the amounts available to the researchers (the number available will affect how many units can be completed). The blocks were then assembled in respective workstations.

Figure 1: First Four Workstations from the "Airplane Game" (Reprinted with permission from Visionary Products, Inc. 2008).
Using Controlled Experiments to Calibrate Computer Models: The Airplane Game as a Lean Simulation Exercise
Zofia K. Rybkowski, John-Michael Wong, Glenn Ballard and Iris D. Tommelein

Figure 2: EZStrobe Computer Simulation of the Airplane Game
**Live Simulation**

Before beginning the game, a facilitator measured individual workstation assembly times with the players. This was done by measuring the time needed for each player to create 5 assemblies and computing an average time per assembly.

The experiment should be extended by having different players each make a greater number of assemblies, and then characterizing the assembly times by means of a probability distribution (such as a beta or PERT distribution). The researchers did not do this as they first wanted to deliver proof of concept of the methodology to be followed. They recognize that in this very issue of repetition lies one of the challenges of conducting statistically valid RCTs.

Four trials were run as follows: (1) batch size of 5 with push, (2) batch size of 5 with pull, (3) batch size of 1 with push, (4) batch size of 1 with pull. Players were instructed to assemble pieces in their workstations as evenly and systematically as possible.

*Batch size* refers to the number of assemblies that must be completed before transferring assemblies from one workstation to the next. *Push* refers to a process whereby players make and transfer assemblies to the next player, regardless of that next player's needs (i.e., regardless of how many assemblies pile up in-between players). In contrast, *pull* refers to a process whereby players make and transfer assemblies to the next player, only as needed by that player. For example, a "batch size of 5 with pull" means that (a) players cannot transfer their pieces to the next workstation until they have completed 5 assemblies, whereupon they transfer all 5 assemblies in one batch together, and (b) players do not begin to make new assemblies until their customer—the succeeding workstation—has emptied its work area of incoming assemblies, thereby implicitly placing an order (aka. submitting a kanban) to request a new batch of 5 assemblies. Here, the number of units requested to be replenished (production batch) equaled the number of units in the batch transferred (transfer batch); in general, such equality is not required. The game was played for six minutes (360 seconds).

**Computer Simulation**

A computer model was designed and implemented in EZStrobe to mimic the actions of the live simulation game (Figure 2). Measured workstation assembly times from the live simulation were input to the computer simulation. The purpose of the live simulation was to calibrate and reasonably validate the computer simulation. The same metrics were gathered for the EZStrobe simulations as for the live simulations. The model is controlled using the parameters of batch size $B$, kanban size $K$, batch transfer durations, and workstation activity durations. Here, $B = K$. The same model with different parameters can simulate a variety of lean principles.

Figure 2 shows durations as being deterministic for the sake of simplicity. However EZStrobe can model a duration by means a probability distribution, based on data availability (e.g., a beta or PERT distribution), from which the program will sample a duration each time the corresponding activity gets instantiated (Martinez 2001, Law and Kelton 2000). A stochastic model would be more realistic in nature and exhibit characteristics not observed in a
deterministic model like the one shown in this paper.

Push and pull can be simulated by changing the $K$ and $B$ parameters. The flow of assemblies through the simulation is controlled by operators $\text{OperatorW}Si$ (one at each workstation $WSi$) and supply $\text{Thex}Pin\text{Supply}$ queue sizes. Each workstation draws the number of Lego® blocks as required per the game instructions from its $\text{Supply}$ queue, and then outputs assemblies. An initial value of 1 in each $\text{OperatorW}Si$ queue ensures that only one assembly is worked on at a time.

The queues labelled $\text{KanbanFromWS}j$ restrict workstation activities. A push system exists when there is no limit to the number of items a workstation's can work on. This is modelled here by a value of $K = 1,000$ so that the resource in this queue does not constrain the assembly process: a workstation can continue making more assemblies as long as supplies are available. (i.e., $K$ is set to a number much larger than the number of 4-pin, 8-pin, or 16-pin Lego® blocks in any of the supply queues. Another way to model "push" is to remove the $\text{KanbanFromWS}j$ queue from the model altogether.). In contrast, a pull system (build-on-demand) exists when a workstation can output only up to a set number of assemblies and then has to wait for a kanban from down the line, signalling a request for more (quantity $K$) assemblies to be made. In this case, the previous workstation must wait to start making its next batch until the next workstation completes and transfers its current batch.

The effect of cellular layout can be modelled by changing the durations for the $\text{TransferBatch}$ activities. In this model, all batch transfers take 2 seconds and it is left it unspecified as to who actually performs the transfer. The transfer time is consistent with the assumption of cellular layout where there is a short, consistent piece transfer time between the workstations. A non-cellular layout would be reflected by longer times and different times for different workstations, and may include walking and other transportation time between the stations.

The first start of the $\text{PlaneDone}$ activity captures the time of the first completed airplane. The number of airplanes completed in the total simulation time is equal to the total count of the number of assemblies that have entered the $\text{OutputW}S4$ queue. Work-in-process (WIP) at each workstation $n$ is measured at the end of the simulation by summing the number of assemblies in the queue with input for workstation $n+1$, the number of assemblies in the queue with output (not yet transferred) for workstation $n$, and the contents of the workstation $n$ assembly activity. Since quality control was not modelled here, the contents of the output of workstation 4 is not included in the WIP. Nevertheless, that queue was modelled in order to control the kanban $\text{KanbanFromWS}4$.

RESULTS

Table 1 shows data collected from the live simulation versus the computer simulation. Computer vs. live results match quite well. Differences may be attributed to factors such as simplifying assumptions made in the computer simulation (incl. deterministic durations), and transitional behaviour of the players in the live simulation (e.g., players began to get bored or tired toward the end of the game and slowed down their assembly times). Clearly, further data...
collection and computer model refinement are in order to more accurately capture the live simulation.

Table 3: Results from the Airplane Game based on Computer and Live Simulation

<table>
<thead>
<tr>
<th>Transfer type</th>
<th>Planes completed</th>
<th>Time elapsed until first plane</th>
<th>WIP from WS1</th>
<th>WIP from WS2</th>
<th>WIP from WS3</th>
<th>WIP from WS4</th>
<th>Total WIP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Batch Size 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer Push</td>
<td>15</td>
<td>138</td>
<td>54</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>63</td>
</tr>
<tr>
<td>Live Push</td>
<td>12</td>
<td>150</td>
<td>30</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>Computer Pull</td>
<td>10</td>
<td>138</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Live Pull</td>
<td>10</td>
<td>145</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td><strong>Batch Size 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer Push</td>
<td>20</td>
<td>46</td>
<td>55</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>Live Push</td>
<td>20</td>
<td>43</td>
<td>51</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>57*</td>
</tr>
<tr>
<td>Computer Pull</td>
<td>12</td>
<td>46</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Live Pull</td>
<td>12</td>
<td>39</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

*WS1 ran out of pieces at 5'20"

**DISCUSSION**

A purpose of many lean production simulations is to educate players and increase receptivity to change through play (Tommelein et al. 1999, Verma 2003; Visionary Products Inc. 2007). Consultants and managers wishing to introduce lean principles to a production situation facilitate playing the game so that workers can see for themselves that application of lean principles enhances productivity, diminishes WIP, reduces the need for rework, and offers other benefits.

In addition to providing lean experience to participants, the structure of the Airplane Game embodies many of the critical features of controlled experimentation, similar to those found in psychological, sociological, and medical research (Bernard 2000). Results of this research suggest that, if administered properly, lean games can function as scientific experiments or even RCTs. In fact, the EZStrobe simulation can also be structured to mimic an RCT by providing a (1) control group, (2) randomization of participant activity times through the use of a random number generator, and (3) double blinding, since a computer program remains indifferent to the psychological forces that contribute to a placebo effect in traditional clinical and social science research. A computer model eliminates problems that may confound an experiment using human subjects, such as assembly time variations due to a learning curve or fatigue. The computer model can isolate the effect of these human factors from the effect of implementing individual lean principles. This improves the researcher's ability to accurately quantify the impact of lean principles.

Using a computer simulation increases the reliability of results since any enhancement of productivity of the experimental group vis-à-vis the control group can be said to be mathematical and capable of being generalized, and not purely due to the human variability or the particularities of the environment under examination.

The researchers found that comparing results from the live simulation— even a deterministic one— against those generated by a...
Using Controlled Experiments to Calibrate Computer Models: The Airplane Game as a Lean Simulation Exercise

Zofia K. Rybkowski, John-Michael Wong, Glenn Ballard and Iris D. Tommelein

computer (here, using EZStrobe), enhanced their understanding of lean. In fact, misaligned results during early trials revealed inaccuracies in the initial version of the computer model that the researchers might not have noticed otherwise. The modelling effort forced the researchers to spell out system characteristics and the model was subsequently adjusted until validated by results from the live game. This calibration lends confidence that the computer simulation is likely accurate as designed and can serve as a reasonable predictor of outcomes during "what if" scenario testing.

CONCLUSION

Lean simulation games offer educational benefits that cannot be found in textbooks. Additionally, relative simplicity makes lean games ideally-suited to serve as controlled scientific experiments or even as RCTs for testing lean principles.

For this research, a computer model representing parts of the airplane game was created, and then refined and tested against a live simulation. The close agreement of live game play and computer simulation validates the model, so the model can be used confidently to test outcomes that involve varying pull versus push and batch size. The simplicity and repeatability of the game makes it easy to test the computer model’s accuracy because it is unfettered by multiple confounding variables.

Additional rigor may be introduced into the validation process by undertaking multiple RCTs with experimental subjects, by characterizing assembly times using probability distributions for each workstation, and by making other refinements to the EZStrobe model. When seeking to apply the computer model to represent an actual sequence of activities, it is best for RCTs to be performed on the actual processes being studied so that standard deviations for individual activity times can be established. This being said, it must also be acknowledged that properly performing RCTs requires time and resources to reward experimental subjects. Performing such experiments also entails obtaining Institutional Review Board approval. Because of these constraints, we did not undertake to perform actual RCTs in this experiment. However, we found that, for the purpose of validating a process pathway, at least one live run of a controlled experiment helped enormously to validate the computer model.

ACKNOWLEDGEMENTS

We gratefully acknowledge the assistance Michael Deese and Ryan Golightly of Visionary Products, Inc., for supplying the instruction manual to the Airplane Game, and the assistance of Chiaowen Deng, Sebastien Humbert, Catalina Ochoa, and Andreas Schmidt for playing the Game. Research for this paper was in part supported by gifts made to the Project Production Systems Laboratory (http://p2sl.berkeley.edu/). All support is gratefully acknowledged. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Project Production Systems Laboratory.
REFERENCES


Using Controlled Experiments to Calibrate Computer Models: The Airplane Game as a Lean Simulation Exercise

Zofia K. Rybkowski, John-Michael Wong, Glenn Ballard and Iris D. Tommelein


