Experimentation with a System Dynamics Based Interactive Learning Environment:

A Case Study of Accessibility of Norwegian Municipalities Websites

Ahmed Abdelgawad, Jaziar Radianti, Mikael Snaprud
Department of ICT
University of Agder
Grimstad, Norway
{ahmedg|jaziar.radianti|mikael.snaprud}@uia.no

John Krogstie
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
krogstie@idi.ntnu.no

Abstract — **A System Dynamics (SD) simulation model including factors affecting the accessibility of Norwegian municipal websites was encapsulated in an Interactive Learning Environment (ILE). Accessibility is an important aspect of websites generally and public websites particularly. Many ways could be proposed to enhance accessibility, however the impact of selected actions is hard to predict due to diversification and contradiction, in addition to the existence of the time factor. As the SD model promised to be able to change how users think and take decisions, this ILE was tested by users in an experiment. We have conducted α, β, and γ change analysis on the results of this experiment. Results showed that the ILE was successful in changing 50% of its users' understanding and perceptions about the system's causal relations and policy options, and helping 30% redefining the standards they use to assess or evaluate these relations and policy options.**

Keywords - Experimental Design; Alpha, Beta, Gamma Analysis; Accessibility; System Dynamics; Municipal Websites; Interactive Learning Environment

I. Introduction

Accessibility of a website refers to the ability of all people to use a website irrespective of their disabilities or the client devices they use to access the Internet [1]. Accessibility is an important aspect of websites in general and of public websites in particular, to be able to serve all citizens equally. For example if a webpage is designed to retrieve user input solely via mouse clicks, people with disabilities preventing them from using a mouse, or people using mobile phones will not be able to use this webpage.

Accessibility of a website can be assessed in terms of compliance with a set of accessibility metrics defined by guidelines like WCAG 2.0 [1] and ISO 9241-20 [2]. It may be evaluated quantitatively such as site score defined in [3]. For more information see [4]. Many studies and projects have addressed evaluation of public websites accessibility, for instance [5], [6]. In addition, governments carry out benchmarking and pass laws to increase public websites accessibility. In Norway, the Agency for Public Management and eGovernment (DIFI) used to evaluate governmental websites annually [7]. For a long period, accessibility was one of three indicators used to be measured by DIFI to encourage governmental agencies to enhance the accessibility of their websites.

A public website, as an Information System, consists of people (like web-masters, editors, developers, etc.) and procedures, besides telecommunications, hardware, software, and data [8]–[10]. When considering the management process and the people component, many ways could be proposed to enhance its accessibility like consulting experts, replace the Content Management System (CMS), recruiting, and training. The expected impact of these measures ranges from slow to fast, cheap to expensive, and short-term to long-term. These diversifications and contradictions, in addition to the existence of the time factor make the decision of adopting only one way to do the task a challenge. More challenging is how to prioritise limited resources to achieve the best effect on accessibility. A simulation model is a very efficient tool that can be employed to deal with such a situation [11]–[15].

Based on the results retrieved from a set of semi-structured personal interviews with web-masters and editors from different Norwegian municipalities, Abdelgawad, Snaprud, and Krogstie [16] identified various factors and causal relations governing the processes having an impact on accessibility of Norwegian municipal websites, and compiled these relations into a System Dynamics (SD) simulation model titled "eAccessibility of Norwegian Municipalities Websites".[1] This model is intended to work as a decision support tool by helping eGovernment websites' managers to take informed decisions, and decision-makers to find policies that enable governmental organisations to enhance their websites' accessibility.

From another angle, the model is supposed to be able to change how its users think and take decisions. It is allegedly capable of changing their understanding and perceptions about the system's causal relations and policy options, in

[1] The model is licensed under a Creative Commons Attribution Share Alike license, and available at: http://forio.com/simulate/ahmedg/accessibility-of-norwegian-municipalities-websites-a-decision-support-tool/model/

other words changing their mental models which are the ways they perceive the system. It is more likely that mental models are the basis of the decision taken by managers or decision makers more than the reality of their system [15].

In this paper, the major problem and consequently the research question we are interested in answering is whether or not this model is really capable of changing its users' understanding and perceptions about the system's causal relations and policy options as it promises. We hypothesise that if the model is capable of doing this, then we can expect that the model can serve as a mean for the websites' managers and decision-makers to take more informed decisions.

To answer this question, we have updated the SD model, developed an Interactive Learning Environment (ILE) to be an interface for the model. Further, we have prepared an online questionnaire tool, and conducted an experiment with users to understand the effect of using the ILE and accordingly the model on them.

The remainder of this paper is organised as follows: the next section will describe the ILE developed for this research. In addition, it will provide a detailed description of the procedure followed to conduct the experiment, including the analysis method. The section that follows will explore and discuss the results of the experiment. The last section concludes the paper.

II. Research Methodology

A. System Dynamics based ILE

An ILE is "software for educational purposes, for supporting the process of learning, where the focus is on learning through the interaction with the computer (human-computer interactivity)" [15]. Our ILE is based on an SD model as its core. Different literature refers to an ILE as management flight-simulator, microworld, business simulator, management simulator, etc. based on certain differences [17], [18]. However, in this paper, we do not differentiate between these terms.

A mental model, in general, could be defined as "a construct of cognitive psychology. Mental models are internal representations of conceptual and causal interrelations among elements that people use to understand phenomena" [19]. The mental model in the SD context is a special case of the mental model, distinguished by the closed-loop concept and being more comprehensive [19]. Davidsen in [20] identified two purposes for developing SD based ILE in general. They aim at either changing their users' mental models for educational purposes, or identifying their users' mental models for research and validation purposes. Our ILE is developed mainly for educational purposes. We have developed it focusing on enhancing its users' mental models.

We have developed a generic reusable ILE framework as a base to develop our ILE. This generic framework can be used by others to develop their ILEs. The ILE framework is web-based. It was built using *Forio.com Epicenter.*[2] The selection of Epicenter, in addition to providing a free service plan, was its capability to directly interface our SD model format,[3] and build a Graphical User Interface (GUI) for the ILE using web technologies HTML, CSS, and JavaScript.

Furthermore, we have used PHP and MySQL to log the users' interactions with the ILE. We wanted to record all policy options chosen by the users while using the ILE, in addition to the outcomes resulting from these choices. Figure 1 shows the full ILE framework, including an online questionnaire tool that was prepared to deploy a set of questionnaires we prepared to test the ILE effect, as we will show later. We opted for an open source PHP-based tool called Limesurvey.[4] The following two subsections give more details about the client-side and the server-side of the framework.

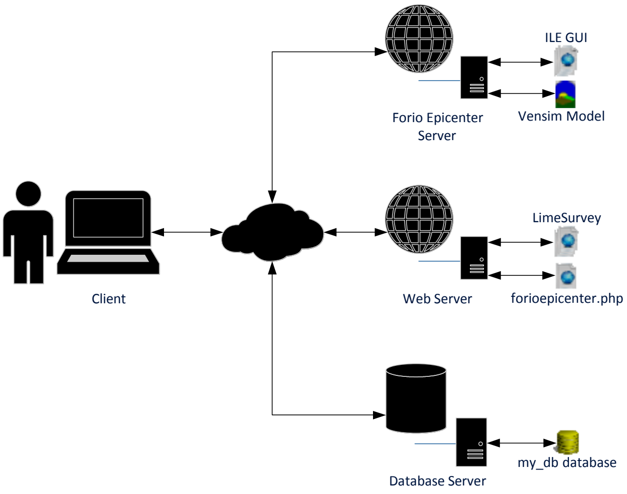


Figure 1: ILE Framework

1) Client-Side

The GUI of our ILE was coded using web technologies[5] based on Epicenter as mentioned before. Epicenter is a very powerful tool, having all what is needed to build an ILE. Nevertheless for our interface charts, we have replaced Forio's Polymer-based[6] charts with our JavaScript charts. Our JavaScript code for charts is still based on Forio's charts code, and uses the same powerful open source Forio's Contour library,[7] but in addition, it is capable of showing the results of several scenarios on the same chart.[8] This way the users are able to compare the consequences of different policies. Our JavaScript code for charts is generic, so that others can use it in building their ILEs.

[2] Forio's 3rd generation platform for simulation, modelling, and analytics. Available at: http://forio.com/

[3] We use Vensim SD modelling software package (http://vensim.com/).

[4] https://www.limesurvey.org

[5] The GUI code is based on Bootstrap (http://v4-alpha.getbootstrap.com/getting-started/accessibility/), and has been adapted to assure accessibility as far as possible.

[6] https://www.polymer-project.org

[7] https://github.com/forio/contour

[8] The JavaScript file is available at: https://forio.com/app/ahmedg/eaccessibility/elements/contour-chart.js

We have used the best practices presented in Sterman [21], [22] to design our ILE's GUI. This GUI has four tabs: Home, Instructions, Control Panel, and Dashboard shown in Figure 2, Figure 3, Figure 4, and Figure 5 respectively. The Home tab gives a brief introduction to the topic of the ILE, including basic knowledge about website's accessibility and policy options. The Instructions tab puts the user in the context of using the ILE, including specific instructions to guide her/him through the simulation or gameplay. The Control Panel tab has all policy options provided by the ILE to control the simulation, in addition to simulation time progress buttons.

The simulation starts at year 0 and can be progressed year by year or to the end of the simulation at year 6. In the Control Panel, the user can reset the simulation and start a new scenario from the beginning, whether the current scenario reached the sixth year or not. Policy options available are represented by graphical control elements for managing workforce, managing workforce time, training workforce, consulting vendor and upgrading website technology (CMS), for further information on policy options see [23], [16]. The Dashboard tab has charts showing over time behaviour of important simulation variables, needed by the user to stand on the results reflected by her/his policies entered to the Control Panel.[9]

In the first version of the model which we described in [16], we used the "Unified Web Evaluation Methodology (UWEM) score" as an indicator of website accessibility. However, in the second version, which we used inside this ILE described in this paper, the "UWEM score" was replaced with "Site Score" as defined in [3]. This "Site Score" was implemented in the model, and shown on the charts and ILE interface under the name "Website Accessibility Indicator".

2) *Server-Side*

The ILE is fully functioning by using solely the client-side, yet, as mentioned above, we wanted to log users' interactions with the ILE, i.e. record the decisions they take, and their results. To accomplish such a task, Epicenter uses Node.js [10] for client-server communications, which then could be logged to a database; however this is limited to paid subscribers. We wanted to have a generic framework that could be used by everyone.

To log users' interactions, we have developed reusable JavaScript snippets[11] and added them to all decision control elements (representing the policy options available by the SD model) and charts available on the client-side GUI. These JavaScript snippets communicate with a PHP file called "forioepicenter.php".[12] We have developed this PHP code to save the values sent by the GUI to MySQL database.[13] Both "forioepicenter.php" and MySQL database were hosted on one of the university servers. [14]

B. *Experimental Procedure*

We have conducted our experiment with volunteer students at the University of Agder, on the 8th of September 2015. A couple of weeks earlier we started spreading the invitation for a gameplay session with free pizza in Grimstad campus. In the day of the experiment 17 students showed up. Some faced technical troubles with the experimentation system, and by the end we could extract 12 useful finished surveys. Properties of the participants are presented in Table I.

TABLE I: PARTICIPANTS WHOSE RESULTS WERE ACCEPTED

Property	Value	%
Age Group	18-24	67%
	25-34	25%
	35-44	8%
Gender	Male	67%
	Female	33%
Field of Study/Work (specialisation)	ICT	92%
	Mechatronics	8%
Knowledge of Math Modelling	Yes	25%
Knowledge of System Thinking/Dynamics	Yes	17%

The whole experiment session took one hour. For 20 minutes, the experiment supervisor gave a presentation to introduce the topic and the ILE to the participants. The presentation included the terms the participants would experience during the intervention using the ILE. By the end of the presentation, the participants were instructed to connect to the Limesurvey server prepared earlier via their web browsers. In addition to the free pizza that was promised to everyone, the two highest-performing participants were promised a piece of Egyptian pharaonic collectable each.

The testing session started for everyone by answering a pre-test questionnaire. The pre-test questionnaire consists of 10 Likert 5-point scale items (strongly disagree, disagree, neutral, agree, or strongly agree). These Likert items constitute multiple-item scale unidimensional construct as described in [24], designed to test the participants' knowledge about the dynamics of the system in terms of it causal relations and possible policy options.

To build the statements of these Likert items, we have enumerated all model variables affecting the municipal website accessibility. Possible changes in the values of these variables (increase in, decrease in) were combined with different possible resulted changes on the municipal website accessibility (for example: increase after short delay, no effect, immediate decrease, etc.). The final combinations were compiled into full statements about the system, and then ordered according to their importance based on our knowledge of the system, what we wanted to show and test, and how

[9] The ILE is available at: https://forio.com/app/ahmedg/eaccessibility/eaccessibility.html

[10] https://nodejs.org

[11] The snippets are available inside HTML of the ILE. It could be shown by viewing the page source using any web browser. Furthermore, we have made these snippets generic, and marked them by HTML comment "<!--begin " and "<!--end ", to be easily copied to any other ILE.

[12] forioepicenter.php can be deployed to any server/web hotel supporting PHP. We made it available at: https://forio.com/app/ahmedg/eparticipation/helper/forioepicenter.php

[13] MySQL database tables needed by forioepicenter.php, can be reproduced at any MySQL using my_db.sql which we made available at: https://forio.com/app/ahmedg/eaccessibility/helper/my_db.sql

[14] The experiment was conducted at University of Agder.

much they are clear while using the ILE. Further, to suit the experiment duration, only 10 of these statements were selected for our Likert items, keeping a balance between reversed and non-reversed statements, and mostly following the recommendations stated in [25]. Finally we polished the wording of the final statements, for example, participants were asked to report their level of agreement or disagreement with this statement "Upgrading CMS takes long time to show an effect on the value of website accessibility".[15]

The Pre-test questionnaire was supposed to take no more than five minutes; nevertheless it was left to the participants to take as much time as they need. The participants were informed that they can ask the supervisor for help at any time; however we abstained from providing any help that could lead to biases in their answers.

The intervention using the ILE or the gameplay started as the participant ended the Pre-test questionnaire, without the option of going back to the Pre-test questionnaire. The gameplay was limited to 25 minutes. Afterwards all participants were directed automatically to the post-test questionnaire, without the option of going back to the intervention session. This way we were sure that all participants had not used the ILE for more than the designated duration.

The Post-test questionnaire contained exactly the same Likert items used in the Pre-test questionnaire, however after answering how she/he thinks now about each statement after the gameplay, the participant was asked to think back in time before the gameplay, and report how much she/he agreed or disagreed with the same statement based on her/his new understanding. This is called the Retrospective Pre-test or the Then-test [26]. It is very common that participants change their understanding between the Pre-test and the Post-test [27]. The Then-test gives the participant the opportunity to re-answer the Pre-test based on her/his new understanding/ perception after the intervention. In this case, the Post-test and the Then-test have the same base frame of reference [27].

1) α, β, and γ Change

Different SD literature presented methods to measure changes in mental models, or to compare them, for example [28]–[32]. These methods either need a human rater, or require the test subjects to have prior knowledge about certain knowledge elicitation tools, for example, the Causal Loop Diagram. We were interested in a method free from these requirements. Human raters might cause experimenter bias [33], while there was no guarantee that our experiment subjects would have enough knowledge about any knowledge elicitation tools.

In 1976, Golembiewski *et al.* in [34] distinguished between three different types of attitude change as a result of an intervention, namely **α**, **β**, and **γ**. **α** change refers to an absolute quantitative change [35]. For example, a website manager might "agree" that "Upgrading CMS takes long time to show an effect on the value of website accessibility". After the intervention, this person's level of agreement about the same statement increases to "strongly agree". This is a real change in her/his opinion on a fixed measurement scale, or **α** change.

β change refers to a measurement scale intervals recalibration, i.e. a redefinition in the measurement standards. For example, a website manager has certain understanding of the values of different agreement levels (strongly disagree, disagree … etc.) regarding the same claim mentioned above "Upgrading CMS takes long time to show an effect on the value of website accessibility". Based on her/his understanding, this website manager indicates that she/he "strongly agrees" with that claim. After the intervention, this website manager finds out that what she/he used to interpret as "strongly agree" means just "agree". Consequently his answer to such a claim would change, although this does not reflect any change in her/his opinion. This is a change in the measurement continuum, or a change in the measurement standard/scale, or **β** change.

γ change refers to a conceptual change, i.e. a redefinition of the measurement construct [34]. For example, a website manager might have no idea that "Upgrading CMS takes long time to show an effect on the value of website quality" or about that upgrading CMS has any effect on website accessibility at all. After the intervention, this website manager was provided with an understanding of the importance and possible effect of upgrading CMS, i.e. a new conceptual frame of reference, which causes a meaningful answer based on this new understanding, or **γ** change.

Many methods to assess **α**, **β**, and **γ** changes appeared since 1976, including the method suggested by Golembiewski and his colleagues. According to a comprehensive literature review conducted by Riordan *et al.* [35], there are five major methods to detect **α**, **β**, and **γ** changes:

- Ahmavaara's technique [34]
- Actual-ideal difference measures [36]
- Retrospective accounts [37]
- Confirmatory factor analysis [38]
- Latent growth modelling [39]

Our expectation of the number of the participants volunteering to our experiment was very modest, because of limited participants' availability as well as financial support. Based on that, we have opted for using the Retrospective accounts method, as it is the only method that doesn't require a large sample, in addition to that it can test for **α**, **β**, and **γ** change independently [35]. It is worth mentioning that although we are not aware of any application of this method in assessing SD based ILE effect, using this specific method for that purpose was suggested by Friedman, Cavaleri, and Raphael [40].

2) Retrospective Accounts

As the Post-test and the Then-test questionnaires are answered based on the same understanding/perception as mentioned above; the Retrospective accounts method detects **α** change by detecting the change between the Post-test and the Then-test. Furthermore, since the Pre-test and the Then-test are basically measuring the same thing based on either two different understandings/perceptions or two differently calibrated measurement scales, **γ** and **β** change are detected by

[15] All the Likert statements are available at: https://forio.com/app/ahmedg/eaccessibility/helper/questionnaires.txt

detecting the change between the Pre-test and the Then-test [37].

Even though the Retrospective accounts method supports analysis on both group and individual levels, we have chosen to focus solely on the individual level analysis, because of the limited number of participants. After all, group change is the sum of its individuals' change. Sometimes certain individual change could be overlooked by detecting only group changes [24], [41]. Furthermore, "a large amount of change exhibited by only a few individuals may be taken as evidence that the intervention had a group effect" [24].

To apply the Retrospective accounts method to our data, we have followed the practice of Birkenbach [41] in general. Nevertheless, we have opted for following Brodersen and Thornton [24] in detecting **γ** change first, then remove the participants showing **γ** change from the process of detecting **α** and **β**. According to Porras and Singh [42], when **γ** change is detected, **α** and **β** detection becomes problematic.

Answers to the questionnaire items from the Pre-, the Post-, and the Then-tests of each participant were used as raw data/basic data points [42]. Consequently for every participant, we have compiled three paired samples Pre, Post, and Then. The first step in the analysis is to try detecting **γ** change per participant. Terborg and his colleagues [37] suggested two methods:

1. Using Correlation:

For every participant, correlations between Pre & Then ($r_{\text{Pre Then}}$), Post & Pre ($r_{\text{Post Pre}}$), and Post & Then ($r_{\text{Post Then}}$) are calculated. To test for differences between $r_{\text{Post Then}}$ & $r_{\text{Pre Then}}$ and $r_{\text{Post Then}}$ & $r_{\text{Post Pre}}$, Williams's test[16] for comparing correlations of two paired/dependant samples is used to calculate $t_{(r_{\text{Post Then}})(r_{\text{Pre Then}})}$ and $t_{(r_{\text{Post Then}})(r_{\text{Pre Post}})}$ respectively [24]. **γ** change exists if the following conditions are met:

a) $r_{\text{Post Then}}$ is substantially greater than $r_{\text{Post Pre}}$

b) $r_{\text{Post Then}}$ is substantially greater than $r_{\text{Pre Then}}$

2. Using Standard Deviation:

Standard Deviations for Pre (s_{Pre}), Post (s_{Post}), and Then (s_{Then}) are calculated for every participant. Morgan-Pitman test[17] for comparing variances of two paired/dependant samples is used to calculate $t_{(s_{\text{Post}})(s_{\text{Then}})}$, $t_{(s_{\text{Pre}})(s_{\text{Then}})}$, and $t_{(s_{\text{Pre}})(s_{\text{Post}})}$. **γ** change exists if the following conditions are met:

a) s_{Post} is not different from s_{Then}

b) s_{Post} is different from s_{Pre}

c) s_{Then} is different from s_{Pre}

The highest level of **γ** change happens when both correlation and standard deviation methods to detect **γ** change occur concurrently [37]. If a participant doesn't show any signs of **γ** change, we start detecting **β** or **α** change.

To test for **β** or **α** change, mean values of Pre, Post, and Then are calculated for every participant, yielding $\overline{x}_{\text{Pre}}$, $\overline{x}_{\text{Post}}$, and $\overline{x}_{\text{Then}}$ respectively. Student's t-test[18] to compare means of two paired/dependant samples is used to calculate $t_{(\text{Then})(\text{Pre})}$ and $t_{(\text{Then})(\text{Post})}$. If $t_{(\text{Then})(\text{Post})} > t_{(\text{Then})(\text{Pre})}$, descriptively speaking, then there is more evidence of **α** change than **β** change, and vice versa [37]. Following the practice of [41], we have focused only on the size to compare $t_{(\text{Then})(\text{Pre})}$ to $t_{(\text{Then})(\text{Post})}$.

Terborg and his colleagues [37] emphasised on that t-statistics on the individual level analysis should in general be judged descriptively. Although the tests used to compute these statistics are for dependant/paired samples, which is the case, the inter-independency or independency condition inside each participant's Pre, Post, and Then samples is not met. Simply, inside each of them all data points come from the same participant [37]. For the reader's convenience, a summary of the symbols and their definitions is shown in Table V.

III. Results and Discussion

Participant P6 and P11 were removed from the analysis because of showing no variance in their Pre, Post, or Then samples. We started the analysis by detecting **γ** change. This was done via correlation and standard deviation comparisons as mentioned above. The left half of Table II shows the needed correlation values in addition to the t-statistics calculated to compare them. The t-statistic columns at the left half of the table prove that $r_{\text{Post Then}}$ is substantially greater than both $r_{\text{Post Pre}}$ and $r_{\text{Pre Then}}$ for participants P2, P4, P5, P9, and P12, consequently detecting **γ** change. The t-statistics columns on the right half of the same table cannot at all prove that s_{Post} is not different from s_{Then}, while both are different from s_{Pre} for any participant, and consequently no **γ** change was detected based on standard deviation.

After eliminating participants showing **γ** change, from **β** and **α** change detection procedure, Table III shows that P1, P3, and P7 have smaller values of $t_{(\text{Then})(\text{Post})}$ compared to $t_{(\text{Then})(\text{Pre})}$ denoting **β** change for these participants. Accordingly participants P8, and P10 have exhibited **α** change. Table IV shows the overall **α**, **β**, and **γ** changes detected for all

[16] To apply Williams's test to test the difference between two dependent correlations sharing one variable/two "paired" correlations, we used the R (The R project for statistical computing software environment https://www.r-project.org/) command: r.test {package: psych}. More information available at: http://www.personality-project.org/r/html/r.test.html

[17] To apply Morgan-Pitman test to test for equal variance of two dependent samples, we used R command: var.test {package: PairedData}. More information available at: http://artax.karlin.mff.cuni.cz/r-help/library/PairedData/html/var.test.html

[18] To apply Student's t-test to compare means of 2 paired samples, we used R command: t.test {package: stats}. More information available at: https://stat.ethz.ch/R-manual/R-patched/library/stats/html/t.test.html

participants in comparison to their answers about mathematical modelling and system dynamics knowledge, in addition to their field of work or specialisation. We could not find any association between these variables and the detected **α**, **β**, or **γ** changes. The highest phi coefficient calculated was 0.29 between the mathematical modelling and the detected **γ** [43].

50% of the participants who were included the analysis have shown **γ** change, reflecting a change in their understanding and perceptions about the system's causal relations and policy options. 30% have redefined/recalibrated the standards they use to assess or evaluate these causal relations and policy options exhibiting **β** change. In total, 80% of the participants have redefined certain knowledge as a result of using the ILE, achieving the ILE's intended goals.

From an internal validity [33] point of view, to minimise testing validity threat, we have kept the questionnaires as merely Likert-scale items, and emphasised that there is no right or wrong answers, and participants needed to report what they think/believe. Furthermore, we made sure that all participants have fully understood questionnaire items since the pre-test, to account for any misunderstanding that could be automatically clarified during the post-test solely because of repetition. The same questionnaire was administered during pre- and post-test sessions to account for any instrumentation validity threat. Moreover, to eliminate experimenter bias, we have chosen self-report questionnaire type, and kept the whole experiment computerised without any human rater interactions, except when help to clarify any vagueness was needed.

To account for possible history validity threat, participants were asked to report their prior knowledge of mathematical modelling and system thinking/dynamics. Furthermore, the experiment time was limited to almost one hour, eliminating maturation or mortality validity threats. We have to admit that the research suffered from selection validity threat due to the availability of participants, as previously mentioned merely 17 students participated in the experiment. Nevertheless, this was to a certain extent mitigated by the fact that participation was totally voluntary.

From external validity [44] perspective, participants were few, and limited to university students, yet they are mostly ICT students, who are expected –to some extent– to fill positions like website managers and decision-makers in the future, which are the users' positions originally targeted by the model. Other experiments with different samples are necessary. Longer periods between the pre-test, the treatment, and the post-test should be examined. Other sets of questionnaire items describing the model's causal relations and policy option should be used in other experiments.

IV. Conclusion

Accessibility is core for delivering usable public websites. Many ways could be proposed to enhance accessibility; however the expected impact of selected actions is hard to predict due to diversification and contradiction, in addition to the existence of the time factor, which makes decision-making a challenge. An SD model includes factors affecting accessibility of eGovernment websites was encapsulated in an ILE. The model is allegedly capable of changing its users' understanding and perceptions about the system's causal relations and policy options, in other words changing their mental models. In this paper, we have answered our research question of whether or not our SD model is really capable of changing its users' understanding and perceptions. Based on the experiment results the model was capable of achieving it promised goals.

In an experimental setting, the ILE was tested with users. We have applied **α**, **β**, and **γ** change analysis on the individual level to the results of this experiment. The results were that the ILE/model was successful in changing its users' understanding and perceptions about the system's causal relations and policy options 50% of the time, and helping them in redefining the standards they use to assess or evaluate these relations and policy options 30% of the time. In total, 80% of ILE users have redefined certain knowledge as a result of using it, achieving the ILE's intended goals, and answering our research question, provided that we could not find any evidence of the effect of the ILE users' prior knowledge or backgrounds on their ILE results. Based on these results, we recommend using the ILE/model in educating websites managers and decision-makers about their systems.

In this paper we have also provided a methodological contribution. We have developed a generic reusable ILE framework, and provided instructions on how it could be used by others in creating their ILEs. Furthermore, we have adapted the **α**, **β**, and **γ** change typology and the retrospective accounts method to test the effect of using an ILE on its users. We have also introduced our suggested approach to create the questionnaires needed to apply the **α**, **β**, and **γ** change and the retrospective accounts method in testing an ILE effect on its users, as well as our suggested steps and statistical tests needed in conducting the statistical analysis for the retrospective accounts method.

Finally, applying **α**, **β**, and **γ** change analysis to test the effect of using SD based ILE was easy and straight forward. However, more experimentation with larger samples, ideally including control groups, to test for group changes in addition to individual changes, over longer time spans, and longer questionnaire seems to be a very promising and highly recommended future research. Furthermore, comparing the **α**, **β**, and **γ** change results with results from other mental model change measurement methods more common among SD practitioners is a very important validation requirement for the method in the SD field.

Acknowledgement

We thank Kamil Gryga and Dr. Morten Goodwin for their support in facilitating the experiment. Likewise, we thank all the students who volunteered to participate in the experiment.

References

[1] 'Introduction to Web Accessibility', *W3C*, Sep-2005. [Online]. Available: http://www.w3.org/WAI/intro/accessibility. [Accessed: 06-Oct-2011].

[2] 'ISO 9241-20:2008', *ISO*, 2008. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=40727. [Accessed: 11-Oct-2011].

[3] A. Nietzio, M. Eibegger, M. Goodwin, and M. Snaprud, 'Following the WCAG 2.0 Techniques: Experiences from Designing a WCAG 2.0 Checking Tool', in *Computers Helping People with Special Needs*, K. Miesenberger, A. Karshmer, P. Penaz, and W. Zagler, Eds. Springer Berlin Heidelberg, 2012, pp. 417–424.

[4] W3C WAI Research and Development Working Group (RDWG), 'Research Report on Web Accessibility Metrics', in *W3C WAI Symposium on Website Accessibility Metrics*, First Public Working Draft., M. Vigo, G. Brajnik, and J. O. C. eds, Eds. W3C Web Accessibility Initiative (WAI), 2012.

[5] M. H. Snaprud, M. G. Olsen, and F. Aslaksen, 'Automatic Benchmarking and Presentation of the First Results from the European Internet Accessibility Observatory', in *Exploiting the Knowledge Economy: Issues, Applications and Case Studies*, M. Cunningham, Ed. IOS Press, 2006.

[6] M. G. Olsen, A. Nietzio, M. Snaprud, and F. Fardal, 'Benchmarking and improving the quality of Norwegian municipality web sites', in *Proceedings of the 5th Int'l Workshop on Automated Specification and Verification of Web Systems*, Castle of Hagenberg, Austria, 2009.

[7] 'Norge.no-Kvalitet', *DIFI*, 2009. [Online]. Available: http://kvalitet.difi.no/resultat/. [Accessed: 04-Oct-2011].

[8] E. Oz, *Management Information Systems*. Cengage Learning, 2008.

[9] E. Turban, E. McLean, and J. Wetherbe, *Information Technology for Management - Transforming Business in the Digital Economy 3rd Edition*, Third Edition edition. New York: WILEY, 2001.

[10] N. C. Georgantzas and E. G. Katsamakas, 'Information systems research with system dynamics', *System Dynamics Review*, vol. 24, no. 3, pp. 247–264, 2008.

[11] J. W. Forrester, *Industrial Dynamics*. The MIT Press, 1961.

[12] J. M. Lyneis, *Corporate Planning and Policy Design: A System Dynamics Approach*. Pugh-Roberts Ass., Incorporated, 1980.

[13] E. B. Roberts, *Managerial Applications of System Dynamics*. Productivity Press, 1981.

[14] G. P. Richardson, Ed., *Modelling for Management: Simulation in Support of Systems Thinking*. Dartmouth Publishing Group, 1996.

[15] J. Sterman, *Business Dynamics: Systems Thinking and Modeling for a Complex World*. McGraw Hill Higher Education, 2000.

[16] A. Abdelgawad, M. Snaprud, and J. Krogstie, 'Accessibility of Norwegian Municipalities Websites: A Decision Support Tool', in *UKSim 5th European Modelling Symposium on Computer Modelling and Simulation*, Madrid, Spain, 2011, pp. 225–230.

[17] F. H. Maier and A. Größler, 'What are we talking about?—A taxonomy of computer simulations to support learning', *System Dynamics Review*, vol. 16, no. 2, pp. 135–148, 2000.

[18] H. Qudrat-Ullah, *Better Decision Making in Complex, Dynamic Tasks: Training with Human-Facilitated Interactive Learning Environments*. Springer, 2014.

[19] S. N. Groesser, 'Mental Model of Dynamic Systems', in *Encyclopedia of the Sciences of Learning*, P. D. N. M. Seel, Ed. Springer US, 2012, pp. 2195–2200.

[20] P. I. Davidsen, 'Issues in the Design and Use of System-Dynamics-Based Interactive Learning Environments', *Simulation Gaming*, vol. 31, no. 2, pp. 170–177, Jun. 2000.

[21] J. Sterman, 'Interactive web-based simulations for strategy and sustainability: The MIT Sloan LearningEdge management flight simulators, Part I', *Syst. Dyn. Rev.*, vol. 30, no. 1–2, pp. 89–121, 2014.

[22] J. Sterman, 'Interactive web-based simulations for strategy and sustainability: The MIT Sloan LearningEdge management flight simulators, Part II', *Syst. Dyn. Rev.*, vol. 30, no. 3, pp. 206–231, 2014.

[23] A. Abdelgawad, M. Snaprud, and J. Krogstie, 'Accessibility of Norwegian Municipalities Websites: A Qualitative System Dynamics Approach', in *Proceedings of the 28th International Conference of the System Dynamics Society*, Seoul, Korea, 2010.

[24] D. A. Brodersen and G. C. Thornton, 'An investigation of alpha, beta, and gamma change in developmental assessment center participants', *Perf. Improvement Qrtly*, vol. 24, no. 2, pp. 25–48, Jan. 2011.

[25] B. Weijters and H. Baumgartner, 'Misresponse to Reversed and Negated Items in Surveys: A Review', *Journal of Marketing Research*, vol. 49, no. 5, pp. 737–747, Oct. 2012.

[26] G. S. Howard, 'Response-Shift Bias A Problem in Evaluating Interventions with Pre/Post Self-Reports', *Eval Rev*, vol. 4, no. 1, pp. 93–106, Feb. 1980.

[27] S. K. Rockwell and H. Kohn, 'Post-Then-Pre Evaluation', *Journal of Extension*, vol. 27, no. 2, 1989.

[28] L. Markíczy and J. Goldberg, 'A Method for Eliciting and Comparing Causal Maps', *Journal of Management*, vol. 21, no. 2, pp. 305–333, Apr. 1995.

[29] J. K. Doyle, M. J. Radzicki, and W. S. Trees, 'Measuring Change in Mental Models of Complex Dynamic Systems', in *Complex Decision Making*, H. Qudrat-Ullah, J. M. Spector, and P. I. Davidsen, Eds. Springer Berlin Heidelberg, 2008, pp. 269–294.

[30] M. Schaffernicht and S. N. Groesser, 'What's in a mental model of a dynamic system? Conceptual structure and model comparison', in *Proceedings of the 27th International Conference of the System Dynamics Society*, Albuquerque, New Mexico, USA, 2009.

[31] M. Schaffernicht and S. N. Groesser, 'A comprehensive method for comparing mental models of dynamic systems', *European journal of operational research*, vol. 210, no. 1, pp. 57–67, 2011.

[32] S. N. Groesser and M. Schaffernicht, 'Mental models of dynamic systems: taking stock and looking ahead', *Syst. Dyn. Rev.*, vol. 28, no. 1, pp. 46–68, Jan. 2012.

[33] D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, 1 edition. Belmont, CA: Wadsworth Publishing, 1963.

[34] R. T. Golembiewski, K. Billingsley, and S. Yeager, 'Measuring Change and Persistence in Human Affairs: Types of Change Generated by OD Designs', *Journal of Applied Behavioral Science*, vol. 12, no. 2, pp. 133–157, Apr. 1976.

[35] C. M. Riordan, H. A. Richardson, B. S. Schaffer, and R. J. Vandenberg, 'Alpha, beta, and gamma change: A review of past research with recommendations for new directions', in *Equivalence in Measurement*, L. L. Neider, Ed. IAP, 2001.

[36] R. W. Zmud and A. A. Armenakis, 'Understanding the Measurement of Change', *ACAD MANAGE REV*, vol. 3, no. 3, pp. 661–669, Jul. 1978.

[37] J. R. Terborg, G. S. Howard, and S. E. Maxwell, 'Evaluating Planned Organizational Change: A Method for Assessing Alpha, Beta, and Gamma Change', *The Academy of Management Review*, vol. 5, no. 1, pp. 109–121, Jan. 1980.

[38] N. Schmitt, 'The use of analysis of covariance structures to assess beta and gamma change', *Multivariate Behavioral Research*, vol. 17, no. 3, pp. 343–358, 1982.

[39] D. Chan, 'The Conceptualization and Analysis of Change Over Time: An Integrative Approach Incorporating Longitudinal Mean and Covariance Structures Analysis (LMACS) and Multiple Indicator Latent Growth Modeling (MLGM)', *Organizational Research Methods*, vol. 1, no. 4, pp. 421–483, Oct. 1998.

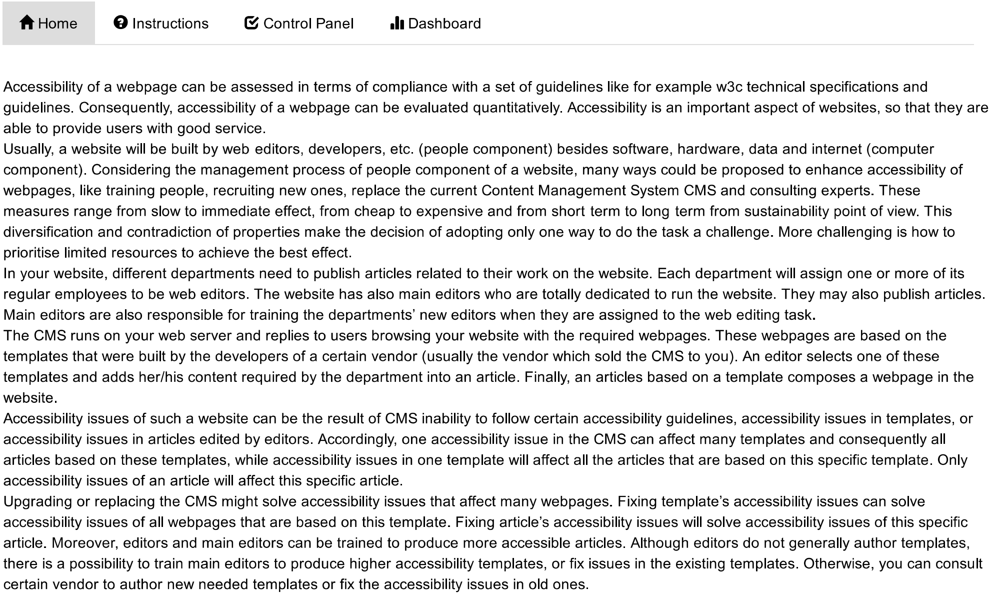
[40] S. Friedman, S. Cavaleri, and M. Raphael, 'Evaluating Changes in Systems Thinking Capacity: A Methodology Based on Alpha, Beta, Gamma Analysis', in *Proceedings of the 21st International Conference of the System Dynamics Society*, New York City, New York, U.S.A., 2003.

[41] X. C. Birkenbach, 'Self-Report Evaluations of Training Effectiveness: Measuring Alpha, Beta, and Gamma Change', *South African Journal of Psychology*, vol. 16, no. 1, pp. 1–7, Mar. 1986.

[42] J. I. Porras and J. V. Singh, 'Alpha, beta, and gamma change in modelling-based organization development', *J. Organiz. Behav.*, vol. 7, no. 1, pp. 9–24, Jan. 1986.

[43] S. Simon, ‘Stats: What is a phi coefficient?’ [Online]. Available: http://www.pmean.com/definitions/phi.htm. [Accessed: 21-Mar-2016].

[44] G. H. Bracht and G. V. Glass, ‘The External Validity of Experiments’, *American Educational Research Journal*, vol. 5, no. 4, pp. 437–474, Nov. 1968.



Accessibility of a webpage can be assessed in terms of compliance with a set of guidelines like for example w3c technical specifications and guidelines. Consequently, accessibility of a webpage can be evaluated quantitatively. Accessibility is an important aspect of websites, so that they are able to provide users with good service.

Usually, a website will be built by web editors, developers, etc. (people component) besides software, hardware, data and internet (computer component). Considering the management process of people component of a website, many ways could be proposed to enhance accessibility of webpages, like training people, recruiting new ones, replace the current Content Management System CMS and consulting experts. These measures range from slow to immediate effect, from cheap to expensive and from short term to long term from sustainability point of view. This diversification and contradiction of properties make the decision of adopting only one way to do the task a challenge. More challenging is how to prioritise limited resources to achieve the best effect.

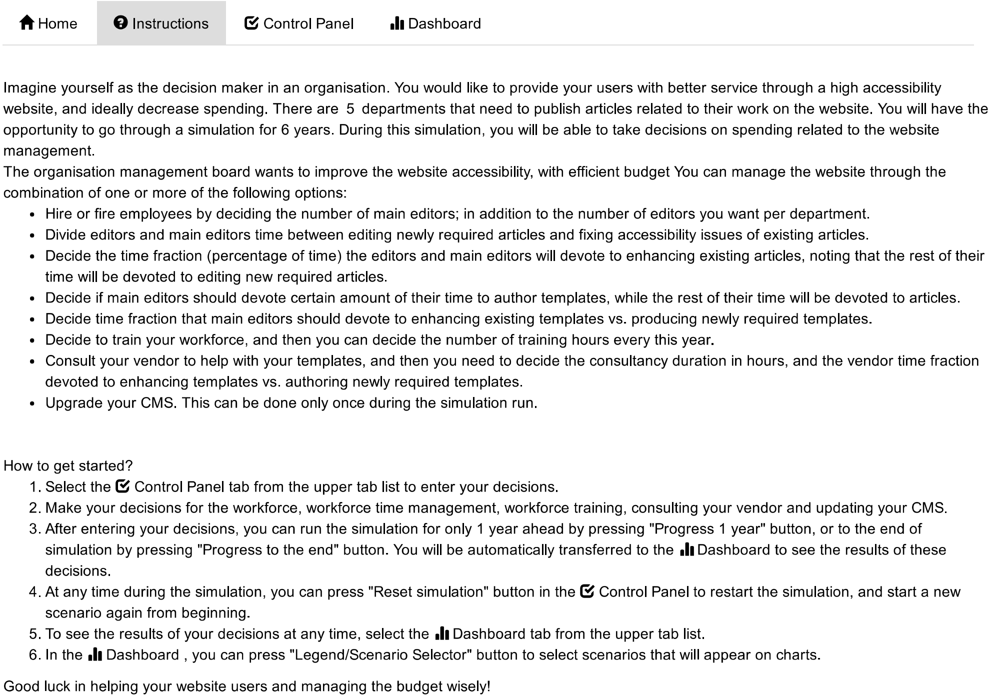
In your website, different departments need to publish articles related to their work on the website. Each department will assign one or more of its regular employees to be web editors. The website has also main editors who are totally dedicated to run the website. They may also publish articles. Main editors are also responsible for training the departments’ new editors when they are assigned to the web editing task.

The CMS runs on your web server and replies to users browsing your website with the required webpages. These webpages are based on the templates that were built by the developers of a certain vendor (usually the vendor which sold the CMS to you). An editor selects one of these templates and adds her/his content required by the department into an article. Finally, an articles based on a template composes a webpage in the website.

Accessibility issues of such a website can be the result of CMS inability to follow certain accessibility guidelines, accessibility issues in templates, or accessibility issues in articles edited by editors. Accordingly, one accessibility issue in the CMS can affect many templates and consequently all articles based on these templates, while accessibility issues in one template will affect all the articles that are based on this specific template. Only accessibility issues of an article will affect this specific article.

Upgrading or replacing the CMS might solve accessibility issues that affect many webpages. Fixing template’s accessibility issues can solve accessibility issues of all webpages that are based on this template. Fixing article’s accessibility issues will solve accessibility issues of this specific article. Moreover, editors and main editors can be trained to produce more accessible articles. Although editors do not generally author templates, there is a possibility to train main editors to produce higher accessibility templates, or fix issues in the existing templates. Otherwise, you can consult certain vendor to author new needed templates or fix the accessibility issues in old ones.

Figure 2: ILE Home tab



Imagine yourself as the decision maker in an organisation. You would like to provide your users with better service through a high accessibility website, and ideally decrease spending. There are 5 departments that need to publish articles related to their work on the website. You will have the opportunity to go through a simulation for 6 years. During this simulation, you will be able to take decisions on spending related to the website management.

The organisation management board wants to improve the website accessibility, with efficient budget You can manage the website through the combination of one or more of the following options:

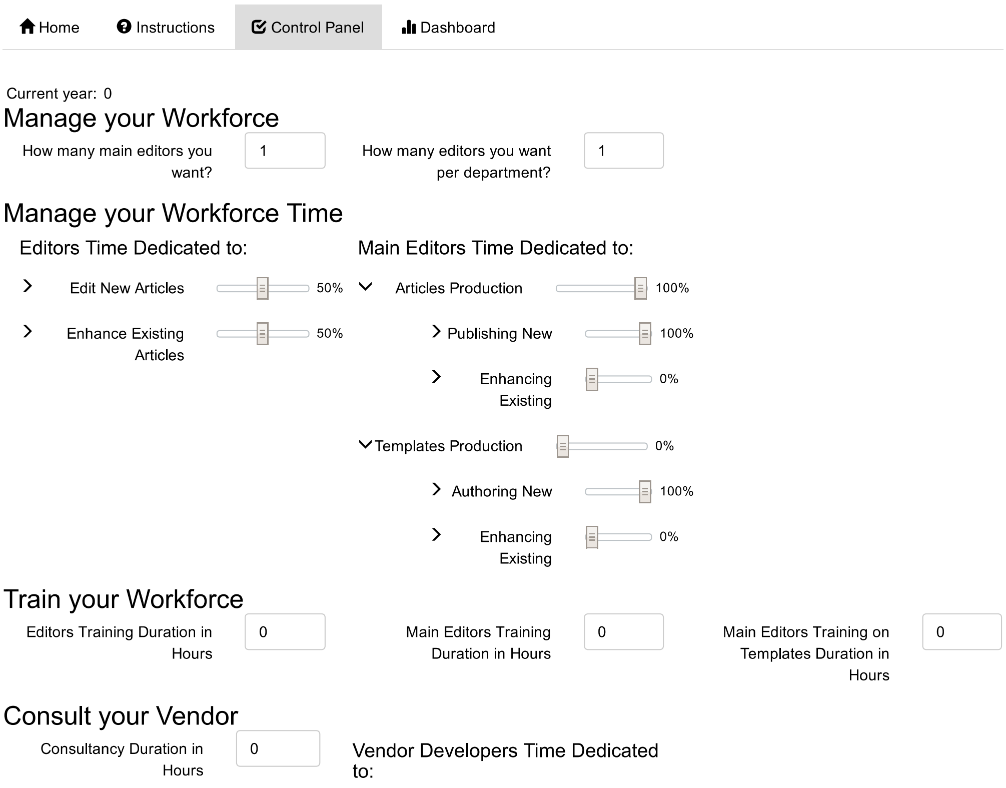
- Hire or fire employees by deciding the number of main editors; in addition to the number of editors you want per department.
- Divide editors and main editors time between editing newly required articles and fixing accessibility issues of existing articles.
- Decide the time fraction (percentage of time) the editors and main editors will devote to enhancing existing articles, noting that the rest of their time will be devoted to editing new required articles.
- Decide if main editors should devote certain amount of their time to author templates, while the rest of their time will be devoted to articles.
- Decide time fraction that main editors should devote to enhancing existing templates vs. producing newly required templates.
- Decide to train your workforce, and then you can decide the number of training hours every this year.
- Consult your vendor to help with your templates, and then you need to decide the consultancy duration in hours, and the vendor time fraction devoted to enhancing templates vs. authoring newly required templates.
- Upgrade your CMS. This can be done only once during the simulation run.

How to get started?

1. Select the ☑ Control Panel tab from the upper tab list to enter your decisions.
2. Make your decisions for the workforce, workforce time management, workforce training, consulting your vendor and updating your CMS.
3. After entering your decisions, you can run the simulation for only 1 year ahead by pressing "Progress 1 year" button, or to the end of simulation by pressing "Progress to the end" button. You will be automatically transferred to the 📊Dashboard to see the results of these decisions.
4. At any time during the simulation, you can press "Reset simulation" button in the ☑ Control Panel to restart the simulation, and start a new scenario again from beginning.
5. To see the results of your decisions at any time, select the 📊 Dashboard tab from the upper tab list.
6. In the 📊 Dashboard , you can press "Legend/Scenario Selector" button to select scenarios that will appear on charts.

Good luck in helping your website users and managing the budget wisely!

Figure 3: ILE Instructions tab



Current year: 0

Manage your Workforce

How many main editors you want? 1

How many editors you want per department? 1

Manage your Workforce Time

Editors Time Dedicated to:

- Edit New Articles 50%
- Enhance Existing Articles 50%

Main Editors Time Dedicated to:

- Articles Production 100%
  - Publishing New 100%
  - Enhancing Existing 0%
- Templates Production 0%
  - Authoring New 100%
  - Enhancing Existing 0%

Train your Workforce

Editors Training Duration in Hours 0

Main Editors Training Duration in Hours 0

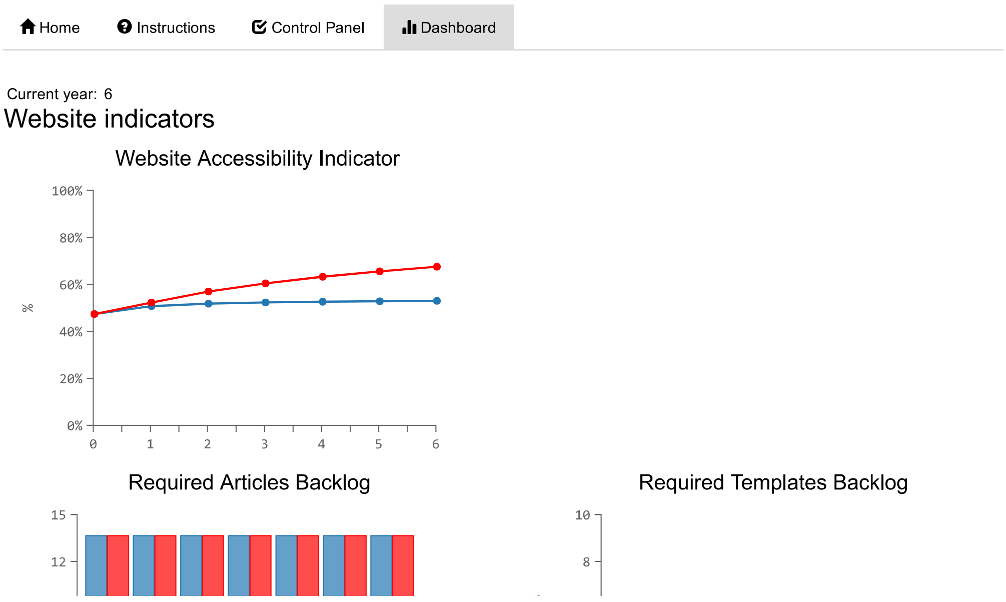
Main Editors Training on Templates Duration in Hours 0

Consult your Vendor

Consultancy Duration in Hours 0

Vendor Developers Time Dedicated to:

Figure 4: ILE Control Panel tab


Home
Instructions
Control Panel
Dashboard

Current year: 6

Website indicators

Website Accessibility Indicator
100%
80%
60%
40%
20%
0%
%
0
1
2
3
4
5
6
Required Articles Backlog
15
12
Required Templates Backlog
10
8

Figure 5: ILE Dashboard tab

TABLE II: **γ** CHANGE DETECTION

	$r_{\text{Post Then}}$	$r_{\text{Pre Post}}$	$r_{\text{Pre Then}}$	$t_{(r_{\text{Post Then}})(r_{\text{Pre Post}})}$	$t_{(r_{\text{Post Then}})(r_{\text{Pre Then}})}$	**γ**	s_{Post}	s_{Then}	s_{Pre}	$t_{(s_{\text{Post}})(s_{\text{Then}})}$	$t_{(s_{\text{Pre}})(s_{\text{Post}})}$	$t_{(s_{\text{Pre}})(s_{\text{Then}})}$	**γ**
P1	0.62^{*}	0.46	0.56^{*}	0.55	0.20	--	1.29	0.99	1.06	0.94	0.62	0.22	--
P2	-0.59^{*}	0.23	0.08	-2.10^{**}	-1.83^{**}	Yes	1.07	0.53	1.25	2.71^{**}	-0.44	2.77^{**}	--
P3	0.80^{***}	0.60^{*}	0.27	0.82	2.88^{***}	--	1.05	1.05	1.58	0.00	-1.47	1.22	--
P4	1^{***}	0.90^{***}	0.90^{***}	13.50^{***}	13.50^{***}	Yes	0.85	0.85	0.95	--	-0.70	0.70	--
P5	-0.36	0.15	0.43	-1.48^{*}	-2.05^{**}	Yes	0.84	0.88	1.20	-0.11	-1.02	1.00	--
P6^{X}	--	0.41	--	--	--	--	1.03	0.00	1.16	--	-0.36	--	--
P7	0.12	0.21	0.21	-0.19	-0.19	--	1.45	0.82	1.03	1.70	1.00	0.66	--
P8	0.73^{**}	0.85^{***}	0.65^{**}	-0.75	0.54	--	1.20	0.92	1.23	1.10	-0.14	1.10	--
P9	0.89^{***}	0.36	0.62^{**}	3.96^{***}	1.50^{*}	Yes	0.99	1.37	1.35	-2.01^{*}	-0.94	-0.05	--
P10	0.84^{***}	0.64^{**}	0.66^{**}	1.23	1.05	--	1.37	0.88	0.92	2.43^{**}	1.50	0.18	--
P11^{X}	--	--	-0.32	--	--	--	0.00	1.03	1.35	--	--	0.82	--
P12	0.70^{**}	0.18	0.08	1.37^{*}	1.75^{*}	Yes	1.62	0.92	0.92	2.37^{**}	1.72	0.00	--

* P < 0.10 ** P < 0.05 *** P < 0.01

^{X} Participant removed because of showing no variance in Pre, Post, or Then

TABLE III: **α** AND **β** CHANGE DETECTION

	$\overline{x}_{\text{Post}}$	$\overline{x}_{\text{Then}}$	$\overline{x}_{\text{Pre}}$	$t_{(\text{Then})(\text{Post})}$	$t_{(\text{Then})(\text{Pre})}$	**β**	**α**
P1	2.90	3.10	2.70	-0.61	-1.31	Yes	--
P2^{G}	3.40	2.50	3.30	1.96^{*}	1.92^{*}	--	--
P3	3.00	3.00	2.60	0.00	-0.77	Yes	--
P4^{G}	3.50	3.50	3.30	n/a	-1.50	--	--
P5^{G}	2.60	2.90	3.10	-0.67	0.56	--	--
P6^{X}	2.80	3.00	2.70	-0.61	-0.82	--	--
P7	3.10	2.70	3.20	0.80	1.34	Yes	--
P8	2.90	3.20	3.20	-1.15	0.00	--	Yes
P9^{G}	3.10	3.10	2.60	0.00	-1.34	--	--
P10	3.10	2.90	2.80	0.80	-0.43	--	Yes
P11^{X}	3.00	2.80	2.50	0.61	-0.49	--	--
P12^{G}	3.20	2.80	2.80	1.08	0.00	--	--

* P < 0.10 ** P < 0.05 *** P < 0.01 – ^{G}γ change detected

^{X} Participant removed because of showing no variance in Pre, Post, or Then

TABLE IV: **α**, **β**, AND **γ** CHANGE RESULTS VS PARTICIPANTS' PROPERTIES

	Knowledge of Math Modelling	**Knowledge of System Thinking/Dynamics**	**γ**	**β**	**α**
P1	Yes	Yes	--	Yes	--
P2	--	--	Yes	--	--
P3	--	--	--	Yes	--
P4	--	--	Yes	--	--
P5	Yes	--	Yes	--	--
P6^{X}	--	--	--	--	--
P7	--	--	--	Yes	--
P8	--	--	--	--	Yes
P9	Yes	Yes	Yes	--	--
P10	--	--	--	--	Yes
P11^{X}	--	--	--	--	--
P12	--	--	Yes	--	--

^{X} Participant removed because of showing no variance in Pre, Post, or Then

TABLE V: SYMBOLS USED IN THE PAPER AND THEIR EXPLANATIONS

Symbol	**Explanation**
α	**α** change, refers to an absolute quantitative change
β	**β** change, refers to a measurement scale intervals recalibration, i.e. a redefinition in the measurement standards
γ	**γ** change, refers to a conceptual change, i.e. a redefinition of the measurement construct
$r_{\text{Pre Then}}$	Correlation between a participant's Pre & Then answers

Symbol	Explanation
$r_{\text{Post Pre}}$	Correlation between a participant's Post & Pre answers
$r_{\text{Post Then}}$	Correlation between a participant's Post & Then answers
$t_{(r_{\text{Post Then}})(r_{\text{Pre Then}})}$	t-statistic of Williams's test for comparing correlations of two paired/dependant samples, namely $r_{\text{Post Then}}$ & $r_{\text{Pre Then}}$
$t_{(r_{\text{Post Then}})(r_{\text{Pre Post}})}$	t-statistic of Williams's test for comparing correlations of two paired/dependant samples, namely $r_{\text{Post Then}}$ & $r_{\text{Post Pre}}$
s_{Pre}	Standard deviations for participant's Pre answers
s_{Post}	Standard deviations for participant's Post answers
s_{Then}	Standard deviations for participant's Then answers
$t_{(s_{\text{Post}})(s_{\text{Then}})}$	t-statistic of Morgan Pitman test for comparing variances of two paired/dependant samples, namely a participant's Post & Then answers
$t_{(s_{\text{Pre}})(s_{\text{Then}})}$	t-statistic of Morgan Pitman test for comparing variances of two paired/dependant samples, namely a participant's Pre & Then answers
$t_{(s_{\text{Pre}})(s_{\text{Post}})}$	t-statistic of Morgan Pitman test for comparing variances of two paired/dependant samples, namely a participant's Pre & Post answers
$\overline{x}_{\text{Pre}}$	Mean values of a participant's Pre answers
$\overline{x}_{\text{Post}}$	Mean values of a participant's Post answers
$\overline{x}_{\text{Then}}$	Mean values of a participant's Then answers
$t_{(\text{Then})(\text{Pre})}$	t-statistic of Student's t-test to compare a participant's Then & Pre answers
$t_{(\text{Then})(\text{Post})}$	t-statistic of Student's t-test to compare a participant's Then & Post answers