An Empirical Assessment of Qualitative, Interactive i* Evaluation

Jennifer Horkoff
Department of Computer Science, University of Toronto
jenhork@cs.utoronto.ca

Abstract

Recent research has been devoted to use of the i Modeling Framework for modeling and understanding the interactions between the technical system and its human users. The benefits provided by such modeling can be increased by utilizing a qualitative, interactive i* evaluation procedure, provoking model iteration and further learning. In this work we articulate a hypothesis concerning the value of such i* evaluation, namely: the use of the i* evaluation procedure produces beneficial changes to the physical i* model and the mental model of the evaluator. These changes result in an increase in model completeness and correctness, leading to a better understanding of the domain. In order to increase our confidence in this hypothesis, we design and apply an experiment involving the application of i* evaluation and the answering of domain related questions. The results of our experiment are promising, increasing our confidence in the validity of our hypothesis and the utility of i* evaluation in system analysis.*

1. Introduction

In software system research it is becoming increasingly clear that system success depends not only on the technical soundness of a system, but how well a system successfully integrates into a human-centered environment, matching the needs of users. This creates a need to understand the interplay between technical aspects and the highly informal world of human users in order to assess the potential success of a system design. To this end, the i* conceptual modeling Framework was created [1, 2]. i* (for distributed intentionality) captures the needs of stakeholders using the concept of goals, and depicts the potential satisfaction of such goals by decomposition into further goals, eventually decomposing to concrete operationalizable tasks. The network of goal interactions is placed in the context of a social network containing dependencies amongst actors.

Although the i* Framework in and of itself may promote understanding of the domain, i* models can provide benefits beyond their initial creation. When the model is continually reconsidered, questioning the knowledge it contains, faults and deficiencies in this knowledge can be discovered, producing potentially useful domain insights. The information contained in models can be used to help answer interesting questions in the domain, predicting the effects of certain design choices on stakeholder goals. However, continual consideration of the model in an ad hoc method can be difficult. Once the modeler feels that the model is sufficiently complete it can be hard to find the motivation to continue to review and reflect on the model's contents. There is a need for a systematic method of evaluating the model contents, provoking model iteration and an increase in insights, increasing domain knowledge. To this end, a qualitative i* evaluation procedure has been introduced [3].

Since its introduction, the i* Framework has been successfully applied in multiple contexts. For instance, the Tropos Software Development Methodology [4], which uses i* in its initial stages, has been successfully applied in various domains, including a health assessment system [5]. In another application, i* models have been applied in the selects of COTS models [6]. The RESCUE Methodology, which includes a stream of i* modeling, has been applied to produce requirements for an air traffic management system [7].

When arguing for the utility of modeling frameworks such as i*, the focus is often placed solely on the perceived practical value of i* application. For instance, "We have applied the i* Framework in the following context, and have witnessed the following positive (or negative) results…" Although these accounts of practical application are useful in demonstrating the utility of the Framework, it can be useful to go beyond anecdotal evidence by attempting

to articulate the theories which underlie the perceived usefulness of i*. That is, to determine not only that the Framework is useful, but specifically *why* it is useful. In this light, the work of Ernst et al. has attempted to articulate and test the theories underlying the i* Framework [8].

In contrast to the numerous practical applications of the i* Framework, only a handful of publications have described applications of an i* evaluation procedure, for instance [9, 10]. In order to better understand the potential usefulness of such a procedure, it is necessary to precisely articulate the theories behind the procedure's utility, and to attempt to increase confidence in these theories through empirical testing. This is the primary aim of this work. In Section 2 we describe a hypothesis underlying the utility of interactive i* evaluation. In Section 3 we briefly describe the i* Framework and the i* evaluation procedure in more detail. In Section 4 we describe the experimental design we use to validate the evaluation hypothesis, with a summary of the results in Section 5. Section 6 contains an analysis of the experimental results while Section 7 explores threats to validity. Section 8 contrasts this work to related work in goal evaluation. Finally, we provide conclusions and outline future work.

2. The i* Evaluation Procedure: Value Hypothesis

We have briefly described the utility of the qualitative i* evaluation procedure. Now we attempt to rationalize the utility of the procedure by articulating an explanatory hypothesis. In order to express the effects of i* evaluation on both the model and the evaluator, we must introduce several concepts.

In this work the term *physical model* is used to refer to the concrete physical representation of the model, i.e. the drawing. Conversely, we use the term *mental model* to refer to the conception of a domain within the mind of an individual. This includes all of the knowledge that a person may hold about a domain, gained through past experience or elicitation. When an individual creates, reads or evaluates a physical model, the semantics derived are a result of the combination of the physical model being read and the mental model of the individual who is doing the reading. We call the resulting knowledge the *Mental/Physical Model*. The implication of these concepts is that the semantics derived from a physical model may differ depending on the individual who is reading the model. This is especially true for the high-level, often abstract nature of i* models. In this light, we may claim that it is difficult to make claims about the semantics of the purely physical model, as any meaning which is extracted from a physical model must be interpreted by the mental model of some individual. These terms are further explained by the sketch in Figure 1.

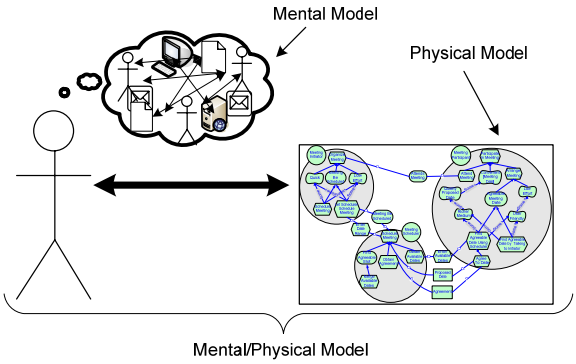

Figure 1: Depiction of a Mental, Physical and Mental/Physical Model

Now we are prepared to articulate the hypothesis underlying the value of i* evaluation:

> **Hypothesis H1**: Use of the i* evaluation procedure produces beneficial changes to the physical i* model and the mental model of the evaluator.

In order to precisely define this theory, we shall decompose it into components:

> **H1 Part 1**: Use of the i* evaluation procedure produces more changes to a) the mental model of the evaluator, b) the physical i* model and c) to the mental/physical model when compared to models of those who did not apply i* evaluation.

> **H1 Part 2**: The a) mental model b) physical model and c) physical/mental model after evaluation are: i) more complete and ii) more correct compared to the models of those who did not apply i* evaluation.

The benefits provided by making the mental, physical and mental/physical models more complete and more correct include increased knowledge on the part of the evaluator, an increase in the usefulness of the model in helping to answer domain questions, and an increase in the amount of information stored in the model, available for future reference.

The introduction of our hypotheses requires us to define a target population. Although i* modeling has been proposed as a tool in the field of software development and system analysis, it can conceptually be applied to any domain in any situation which

requires the analysis and understanding of a social system. Therefore, we shall not limit the effects of our hypothesis to a specific group of individuals with specified training. Anyone who is able to grasp the idea of capturing information via a conceptual model should experience the benefits of i* evaluation.

3. The i* Framework and Evaluation Procedure

3.1. The i* Framework

The constructs of the i* Framework include *actors*, *elements*, and *links*. ***Actors*** are divided into various types including agents, which can represent both real stakeholders and software systems, and roles, which represent the roles that agents play. ***Elements*** are intentional in that they are assigned to a particular actor, meaning that the actor has the intent to fulfill or perform the element. Such elements include *goals*, *tasks* and *resources*. In i*, there is a distinction made between *(hard) goals*, which can be satisfied by accomplishing some clear-cut criteria, and *softgoals*, whose satisfaction is not clear-cut, and are said to be sufficiently satisfied or *satisficed*. Examples of such softgoals include typical non-functional requirements such as security and performance as well as more socially motivated goals such as job satisfaction and personal success.

Links in i* represent the relationships between intentional elements. Actors depend on other actors for the satisfaction of goals, softgoals, tasks and resources, represented via *dependency* links. Internal methods to accomplish elements are represented by *decomposition*, *means-ends*, and *contribution* links. Decomposition links are used to provide more detail on how a task may be accomplished, via the accomplishment of other tasks, goals, softgoals or resources. Means-ends links show alternative tasks that could satisfy a goal.

Contribution links show the effects of elements on softgoals. These effects can be positive or negative and can be sufficient enough to *satisfice* or, conversely, *deny* a softgoal (*Make*/*Break*), or can offer weaker evidence which is not in itself sufficient to satisfy or deny a softgoal (*Help*/*Hurt*). The *Some+* and *Some-* links represent positive and negative evidence, respectively, of unknown strength, and the *Unknown* contribution represents evidence whose effect is unknown. The example model in Figure 2 includes a legend of these i* constructs.

3.2. The i* Evaluation Procedure

The i* evaluation procedure is an adaptation and expansion of an earlier qualitative procedure defined within the NFR (Non-Functional Requirement) Framework [11]. The procedure builds on the notion of element satisfaction/satisficing and denial, defining a set of six qualitative labels which represent the level of achievement or denial of an element. These labels include Satisficed (✓), Partially Satisficed (✓.), Denied (✗), Partially Denied (✗.), Conflict (≷) and Unknown (?).

The procedure involves the propagation of these labels through the links of the model. To start the procedure, initial labels are placed on the model to represent an interesting domain question to be evaluated. For instance, in the Figure 2 model, it is interesting to ask "If the Technology User does not Trust the Technology Provider, how does this affect the Technology Provider's main task of Sell Technology for Profit?" In order to represent this question, we mark the Trust softgoal of the Technology User as denied.

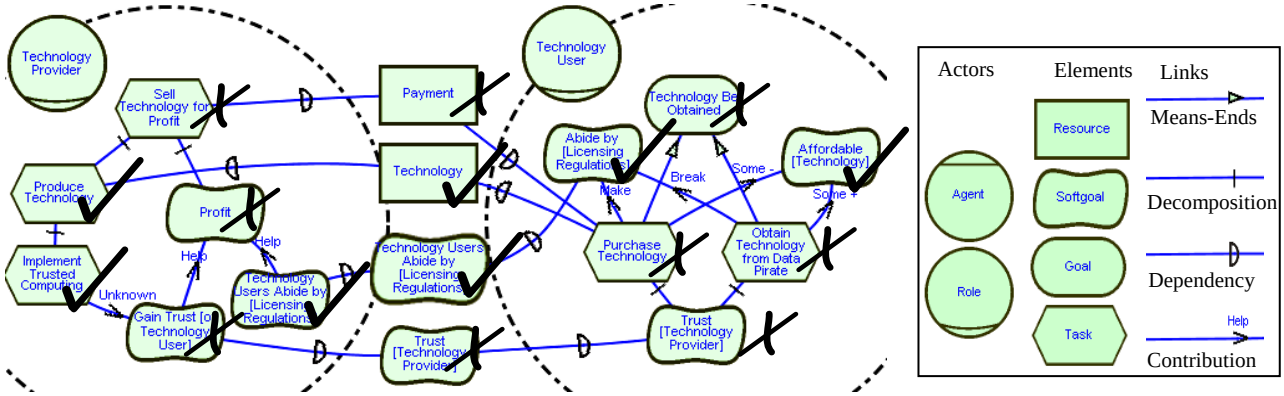

Figure 2: Example Evaluation

Once the initial labels representing an analysis question have been placed on the model, they are placed in a queue of labels to be propagated. The effects of these labels are propagated throughout the graph by repeating two steps. In step 1, rules are used to propagate the labels in the queue across links to recipient elements. Refer to [3] for a detailed description of these rules. For non-softgoal elements these rules determine the evaluation label given to the recipient elements, placed in the queue for subsequent propagation. As softgoals may be the recipient of multiple contribution links, the propagated evaluation labels for each softgoal are stored in a bag of evaluation labels. When all elements in the queue have been propagated, step 2 begins.

In step 2 the label bags of softgoals are resolved to produce a resulting label for the softgoal. This resulting label is then added to the queue for propagation in the next iteration of step 1. The bag of labels for each softgoal can be resolved in one of two ways, by qualifying for a set of automatic cases, described in [3], or by prompting the evaluator for human judgment.

When eliciting qualitative, intangible concerns in a potentially complex social network, it is not feasible to create a model that completely captures all of the goals of all of the users. Therefore, some of domain information relevant to the context of the model inevitably remains as the tacit knowledge of the modeler. This tacit knowledge comes in to play when determining the level of satisficing or denial of softgoals given the satisfaction level of contributing elements. When a softgoal has received multiple partial contributions, or contributions of different polarity, the evaluation procedure prompts the evaluator for a decision concerning the final evaluation value for a softgoal. For example, in Figure 2 the evaluator would be prompted to resolve the label bag for Gain Trust [of Technology User] in the Technology Provider.

Steps 1 and 2 are repeated until all label bags have been resolved and the queue of labels to propagate is empty. Steps to ensure algorithm convergence and termination can be found in [3].

When the evaluation labels have been propagated throughout the model, the evaluator can then perform an analysis of the results. In our example, if the Technology User does not Trust the Technology Provider, the Technology Provider cannot Sell Technology for Profit. The fundamental utility of the procedure comes in questioning the accuracy of such results, including the aspects of the model and the domain brought to light by the steps taken in propagation. For instance, is not Purchasing Technology really the same thing as Abiding by Licensing Regulations? If not, how can we adjust the model to reflect these differences? By answering questions such as these, raised by evaluation, the evaluator expands and modifies his/her mental and physical model of the domain, as expressed by our hypothesis.

4. Experimental Design

In the following section we describe an experiment designed to support or refute the hypotheses described in Section 2.

4.1. Participants.

The experiment requires a minimum of 20 participants. Ideally such participants would be selected randomly from the population to which our hypothesis applies, namely, anyone who may perform systems analysis. However, due to the difficulties in randomly selecting individuals from such a population, we will instead select participants on a volunteer basis. These participants will be members of a university community, likely students at an undergrad or graduate level. The large majority of these students will be computer science majors. It is possible that a subset of the participants may be part of an undergraduate computer science course. However, due to ethics and bias concerns, these students will not receive course credit for their participation, and their participation will be completely voluntary.

There are various independent variables which may affect the performance of participants such as i* experience, general experience in conceptual modeling, industry experience, domain experience, and level of education. A questionnaire will be developed to quiz for these factors, and will be provided to the participants in advance. In order to reduce the effects of these variables on the experimental result, we shall block for these factors when dividing the participants into two groups of equal size.

4.2. Required Materials

The creation of i* models will be based on a text document which will be provided to the participants. The domain is chosen be complicated enough to be able to produce detailed models, yet simple enough to be generally understood without experience or training. The document should be relatively high-level in order to contain many of the social, intentional aspects that i* is intended to capture. However, in order to simulate a realistic source, the document

should also contain information that is not easily captured via i* such as temporal or process-oriented information.

Work in the field of Requirements Engineering has promoted the use of model problems as a means of testing tools and procedures in comparable contexts [12]. To this end, we select the conference refereeing problem, involving the review and selection of papers for a research conference. The text document provided to the users has been composed from two sources, neither of which were explicitly intended for i* modeling. First, we have taken the general description of the conference refereeing process from the model problem description [12], and then we have taken excerpts from the work of Smith, describing the task of the referee [13]. These excerpts were chosen in such a way to provide roughly equal material on the roles of the referee, committee and author (The full text is available at www.attheendofthepaper/Text.pdf). As Smith's work focuses on both journals and conferences, there will likely be inconsistencies between the two excerpt sources. However, as these imperfections reflect the potential errors and confusion that may be found in real documents elicited from the domain, they shall be retained.

In addition to the text document, the experiment will require a set of questions concerning the document, in order to access information in the mental and mental/physical model. These questions will be designed to invoke the utility of i* models by asking "what if?", scenario-type questions, prompting for the general effects of choosing or not choosing to perform certain actions. For example:

> What are the affects of not providing sufficient justification of a paper acceptance in the referee report?

This list of 10 questions will be randomly divided into two groups of 5; these groups shall be referred to as Q_Mental and Q_MentalPhysical (abbreviated Q_MenPhy) (The full list of questions is available at: www.attheendofthepaper/Questions.pdf).

4.3. Pre-Experiment Testing

Before the actual experiment is executed, a pre-experiment shall be performed in order to create model answers to the questions and to determine reasonable time allocations for reading, modeling, evaluation and question answering. Two or three individuals who have experience in both the domain and i* will be asked to perform the experimental steps, outlined in the next section. The amount of time taken to perform each task will be measured and used to roughly determine the times given to the experiment participants, taking into account the increased experience of the pre-experiment participants. The answers to the questions produced by these individuals will be discussed and merged together to create a set of model answers.

4.4. Experimental Steps

This section outlines the concrete steps of the experiment. It should be assumed that all participants, regardless of which group they belong to, are given equal time to perform each experimental step.

1. Apply the questionnaire; analyze the results in order to block independent variables when creating groups.
2. Provide 1 to 2 hours of i* training to all participants.
3. Participants read the text document.
4. All participants are asked to create an i* model of the domain described in the document. They are able to access the document when producing the models. The models created in this step shall be referred to as M1.
5. The M1 models are taken away from the participants. The participants still have access to the document. The participants are asked to answer the Q_Mental questions. These answers will be referred to as Q_Mental_1.
6. The participants are given back their M1 models, but instructed not to make changes to them. They are asked to produce answers to the Q_MentalPhysical questions. These answers will be referred to as Q_MentalPhysical_1.
7. The two groups are separated. One group is selected at random, call this group Group E, and given 1 hour of i* evaluation procedure training. The remaining group, Group C, are given an hour of free time, with instructions not to talk to each other about the experiment or to look up further information on the domain.
8. A copy of the models produced and questions answered is made for later analysis.
9. Both groups are given their M1 models, the document, and the Q_MentalPhysical questions with their previous answers (Q_MentalPhysical_1). They are instructed to try to improve their answers, if they feel it is necessary. This time they are allowed to make changes to their models. The potentially modified versions of the Q_MentalPhysical_1 questions shall be referred to as Q_MentalPhysical_2 and the potentially

modified models will be referred to as M2. In addition, all participants are asked to produce a list of questions about the domain that they feel are not sufficiently answered by the document. The lists of questions produced will be referred to as P_Questions.

10. The i* models will again be taken away from the participants. They will still have access to the document. All participants are given the Q_Mental questions and their previous answers (Q_Mental_1). They are asked to try and improve these answers, if they think it is necessary. The potentially modified answers will be referred to as Q_Mental_2.

4.5. Post-Experiment Processing.

All resulting sets of questions shall be analyzed and given a grade reflecting their correctness. The answers will be marked by the three individuals who created the model answers. Markers are allowed to give marks for answers which do not match the model answers if they believe they are correct. The marking shall be blind in that the markers do not know whether the answers come from individuals in Group E or Group C. Each answer shall be given a number representing correctness, agreed upon by each of the three markers, ranging from 1 to 4.

In addition to marking the correctness of the questions, we shall count the number of changes in the answers for each of the two sets of questions. Changes will be classified as significant, counted as 2 points, or minor, counted as 1 point. The changes will be classified by three individuals. Disagreement on the level of changes will be discussed until the classifiers come to a consensus.

The markers shall also undertake a qualitative analysis of the modifications to the questions as well as the differences between sets of questions for the two participant groups.

The sets of models shall be analyzed to calculate the number and type of physical changes between M1 and M2. For this purpose we will consider the following model changes as one change: a change to the phrasing of an element name, the addition of an element, the removal of an element, the addition of a link, the removal of a link, changing the type of link, and changing the type of element.

Changes in model size shall also be measured. For the purposes of calculating the size of the model, we shall count each element, each actor, and each link as one unit. The number of P_Questions, questions created by the participants shall be counted; with a qualitative analysis comparing the potential question differences between groups.

5. Results (Fictional)

Due to space restrictions, we provide only the per group averages for each dependent variable in Table 1 (a table containing full (fictional) results is available at: www.attheendofthepaper/Results.pdf). In addition, box plots of the data are provided in Figure 3.

6. Analysis

Based on both the quantitative data presented and qualitative observations we shall review each section of our hypotheses in an effort to determine whether the data supports or refutes the null hypothesis.

The work of Popper tells us that we can never prove that a theory is true; instead we can only repeatedly fail to refute a hypothesis, increasing our confidence in the theory [14]. Therefore, we define our hypothesis as null hypothesis, collecting evidence to potentially refute our claims.

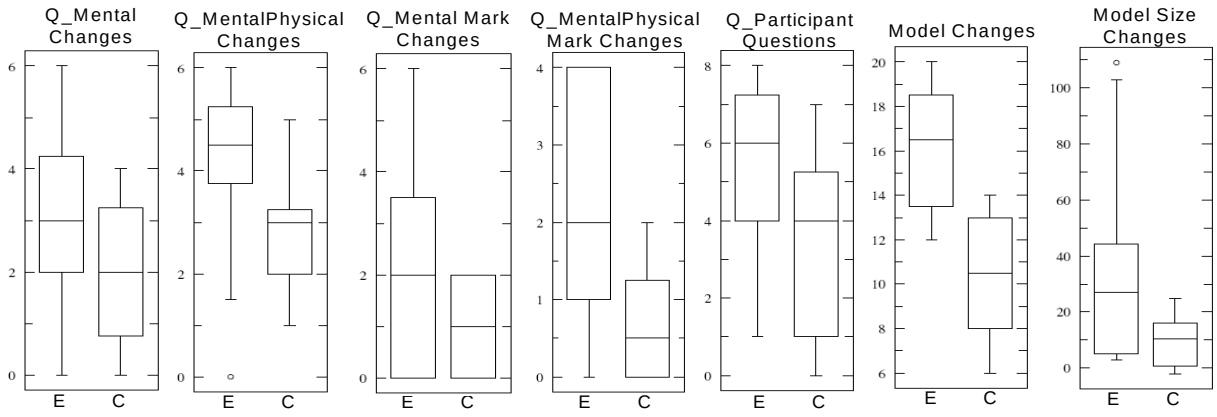

Figure 3: Box Plots of Results

Table 1: Average Results for Questions and Marks

		Group E	Group C
Average Question Changes	Q_Mental_1 to Q_Mental 2	3.10	2.10
	Q_MenPhy_1 to Q_MenPhy_2	4.20	2.80
Average Question Marks (out of 20)	Q_Mental_1	12.83	13.60
	Q_Mental_2	15.12	14.71
	Avg. Difference	2.29	1.11
	Q_MenPhy_1	14.56	13.98
	Q_MenPhy_2	16.08	14.33
	Avg. Difference	1.52	0.35
	Avg. # P_Questions	5.42	3.38
Average # Model Changes	Total M1 to M2	16.02	10.40
Average Model Size	Total M1	97.19	97.08
	Total M2	124.88	106.72
	Total Change	27.68	9.64

As the size of the sample is small, it is difficult to determine whether the results are normally distributed, a necessary condition for the application of the t test. Despite this, we have calculated the t test scores and the significance levels needed in order for the null hypotheses to be rejected for the most vital results. These scores are used in conjunction with other quantitative and qualitative data.

H0: Evaluation does not create beneficial changes: We can examine the number of questions produced by each group, P_Questions. We see that on average Group E produced 5.4 questions, while Group C produced 3.4 questions. We take these questions to represent further elicitation which would occur in a real-life application of i* evaluation. Such elicitation is likely to induce learning, produce a more complete and accurate mental model, mental/physical and physical model, helping to reject all components of H0.

Part 1) Evaluation does not create changes a) in the mental model. The results show that the average measure of changes for Group E was 1 point larger than the average measure of change for Group C, meaning that, on average, Group E made more changes to their Q_Mental answers after evaluation. The confidence level necessary in order to reject the null hypothesis (0.17) is much larger than the typical measure of 0.05; however, our box plot clearly indicates a difference between the two groups.

When performing a manual comparison of the changes in the Q_Mental questions, one can see that the E group tended to add more detail to their answers than the C group. For example, the differences between members of Group E and Group C in answering the question: "What are the affects of submitting a paper to a conference which is of very poor quality?" as shown in Table 2. One can see that in the Q_Mental_1 answers and Q_Mental_2 answer from Group C, the effects are focused mostly on the author, but in the Q_Mental_2 Group E answer the participant has expanded the focus to other actors in the domain. From such answers one can postulate that the act of propagating values across model links forced the participants to consider farther reaching effects of various possibilities, with the evaluator becoming more aware of indirect consequences.

Table 2: Example Answers for a Q_Mental Question

Group E individual: Q_Mental_1
The paper will be rejected, and there may be a long delay before the author finds out. The author may think the report is unfair and be discouraged. The referee may remember the author and think that they are not very good.
Same Group E individual: Q_Mental_2
<<same as above>> The referees and program committee will waste time that could be spent reading other papers. The standards for other papers might be lowered after the referees review the bad paper. If the referees end up reviewing lots of bad papers, they might get a negative opinion of the conference or journal.
Group C individual: Q_Mental_1
Probably the paper will be rejected, but the authors might learn a lot from the referee report and write better papers in the future.
Same Group C individual: Q_Mental_2
<<same as above>> The referee might think badly of the author, and the author might be embarrassed.

Part 1) Evaluation does not create changes b) in the physical model. Group E made, on average 6 more changes than Group C. The t score shows that we can reject the null hypothesis with a high confidence level of 0.0003. By examining the types of changes made, it is interesting to note that Group C made more changes to element types, despite making less changes overall. Experience has shown that determining the appropriate element types when creating i* models can often be problematic [15]. Perhaps the model changes prompted by evaluation caused the participants to ignore these types of changes in favor of other changes that had more effect on the evaluation results.

Part 1) Evaluation does not create changes c) in the mental/physical model. The control group made an average of 2.8 changes to the Q_MentalPhysical questions, while Group E made an average of 4.2, a difference of 1.4. The t test indicates that we can reject the null hypothesis with a confidence level of 0.048. As with the Q_Mental questions, one can see a greater increase in detail between from Q_MentalPhysical_1 to Q_MentalPhysical_2 for Group E as compared to Group C. It is interesting to note that the difference between the average changes for the two participant groups is greater for the mental/physical model than the mental model. One can interpret this result to indicate that evaluation provokes more changes in the physical model than the mental model.

Part 2) a) after evaluation the mental model i) is not more complete. As we have shown via an example in Table 2, qualitative analysis indicates that the Q_Mental questions for Group E have increased in detail after evaluation, more so than for Group C. Similarly, comparing the qualitative differences found between the answers of Q_Mental_2 for Groups E and C, the answers for Group E contain more detail compared to Group C, especially concerning indirect effects. These observations lead us to tentatively reject the corresponding components of the null hypothesis in this case.

Part 2) a) after evaluation the mental model ii) is not more accurate. We can test this hypothesis by comparing the final average marks for Q_Mental_2 for the E and C groups, noticing that the E Group has an average which is greater by 1.18. However, due to the small size of our sample, there may be differences in the inherent abilities of each group that accounts for the differences in average. Instead we focus on the average rise in marks. We can see that the average difference between the accuracy marks for Q_Mental_1 and Q_Mental_2 for Group E rises 2.29 points out of a possible 20, compared to a rise of only 1.11 for the control group. T test results show that the confidence level needed in order to reject the null hypothesis is 0.08. Examining the box plot for the Q_Mental mark changes shows a clear difference in the changes for each group. It is interesting to note in the box plot that the variance for the Group E Q_Mental changes is larger than Group C; perhaps indicating the effect of evaluation on the accuracy of the mental model varies for participants. It seems that the null hypothesis is refuted only on average results and not for every individual.

Part 2) b) after evaluation the physical model i) is not more complete. We shall attempt to measure the completeness of the physical model via its size. Our results show that the size of the physical model changed on average by 32.5 graphical components for Group E compared to 9.6 components for Group C. The confidence level needed in order to reject the null hypothesis is 0.058. One can see from the box plot for model size that the increase in size for Group E has many more outlying values than Group C, perhaps indicating that the evaluation procedure is especially effective for certain individuals. Further studies could focus on the characteristics of individual participants which encourage this effectiveness.

Part 2) b) after evaluation the physical model ii) is not more accurate. Given our claims about the semantics of the physical model, our experimental design did not allow us to test this component of our hypothesis.

Part 2) c) after evaluation the mental/physical model i) is not more complete. To evaluate this sub-hypothesis we perform a qualitative examination of the answers to the Q_MentalPhysical questions. As mentioned, the answers appeared to gain detail from Q_MentalPhysical_1 to Q_MentalPhysial_2. This gain seemed especially prevalent for Group E. When comparing the resulting Q_MentalPhysical_2 answers, we can see that, in many instances the Group E answers contain more information than the Group C answers. Similar to the analysis of the Q_Mental answers, this extra detail reflects effects which are more disconnected from the scenario described by the particular questions.

Part 2) c) after evaluation the mental/physical model ii) is not more accurate. As in 2) a), we can directly compare the Group E and C marks for Q_MentalPhysical_2, noticing that the average for Group E is 1.9 points higher. However, we again focus on comparing the marks differences from Q_MentalPhysical_1 to Q_MentalPhysical_2, with Group E showing an average difference of 2.18 compared to 0.67 for Group C. Applying the t test shows that we can reject the null hypothesis in this case with a 0.0091 degree of confidence.

We can summarize our analysis as follows:

> Use of the i* evaluation procedure makes the physical i* model more complete and the physical/mental model more accurate, and possibly more complete than if evaluation was not used, and may have the same effects on the mental model of the evaluator.

7.0. Threats to Validity

One of the foremost threats to the validity of this experiment is the potential presence of bias in favor of

the benefits of i* evaluation, as the experiment was designed by one of the proponents of the procedure. However, due to the difficulty in finding an objective party who has the resources and willingness to design and administrate such an experiment, this effect is unavoidable. As an alternative, all reasonable efforts have been made to mitigate the bias of the author.

7.1. Construct Validity

Questions as Measurement. As there is no way to directly view the mental model of an individual, or the combination of this model with a physical diagram, we have attempted to measure and compare aspects of this model via questions. How do we know whether or not the questions reflect the areas where changes in the models have occurred? By allowing the participants access to the model for the Q_MentalPhysical questions, we encourage them to expand their knowledge in the areas addressed by these questions. Therefore these questions measure the knowledge of the mental/physical model gained for questions that have been evaluated.

We have assumed that asking the participants questions without allowing them to access the model provides an accurate measure of their mental model. However, there may be difficulties in isolating the mental model from the physical representation. Even if the individual is not in contact with the physical model, the act of having interacted with the model will leave an unknown amount of residual information concerning the physical model in the mind of the modeler. By denying the participants access to the models when answering the Q_Mental questions, we test their knowledge of areas not addressed by questions that were evaluated with the model. This may increase the likelihood of testing the mental model of the individuals as opposed to the individual's memory of the physical model.

One may note that the individuals have already seen the Q_Mental questions before evaluating their model, leading to the possibility that they are evaluating these questions as well. In order to try and eliminate this effect, we ensured that the individuals do not have access to these questions with the model, and have attempted to limit the evaluation time given to the participants, only allowing them time to evaluate the Q_MentalPhysical questions. In addition, there was an hour gap between answering the Q_Mental questions and evaluating the model.

In effect, our measure of the mental/physical model measures the knowledge gained from questions that are evaluated in the model, and our measure of the mental model measures the knowledge gained from questions that are not directly evaluated in the model. The results of the second type (H1 Part 2 a) are especially interesting, showing not only that i* evaluation helps to increase knowledge on focused questions, but that it provokes general domain learning.

We have attempted to combat the subjective nature of the categorization of question changes by having three individuals agree on the level of change.

Producing Questions. It is possible that the number of P_Questions derived by the participants does not necessarily reflect the future accuracy or completeness of the mental, mental/physical or physical models. These questions may not reflect aspects of the domain which prove later to be important. However, as the future value of answering such questions is difficult to predict, we restrict ourselves to measuring their quantity.

Measuring Correctness. Differences in stakeholder viewpoints can make it difficult to establish whether or not information concerning the domain is correct. The difficulties in determining correctness are especially prevalent with information of a predictive nature, as is often produced when using conceptual models to explore early system design. Despite these difficulties, if we abandon altogether the notion of model correctness, then we are severely limited in articulating the beneficial nature of the changes prompted by i* evaluation. Therefore, we adopt the following definition of correctness:

> Information concerning a domain is considered correct when it is agreed upon by multiple sources, especially sources with some authority in the applicable area.

We have reflected this definition by assigning marks which are agreed upon by three people who are familiar with the domain.

Applied Statistics. The correctness of question answers were measured on a scale which may be considered ordinal, if the differences between the values are not considered meaningful. These marks were then averaged to produce comparative data. Although the assigning and averaging of ordinal scale marks is common in academia, it is often considered incorrect to apply such parametric statistics to an ordinal scale [16]. However, we would argue that these distances between marks are meaningful even though they are assigned manually. Our use of an agreement between three different individuals helps to reduce the subjective nature of such measures. Therefore we retain the informative measure of averages.

7.2. Internal Validity

It is necessary to assess whether the differences between results for the two groups were due to the introduction of i* evaluation to Group E or to other factors. By blocking independent variables such as domain experience, i* experience, and model experience, we have attempted to limit their effects. In addition, our focus on comparing the changes between evidence such as question marks and model size is more likely to reflect significant results, as opposed to intrinsic ability differences between groups.

By giving Group E i* evaluation training, we have inadvertently given them additional training in i*. However, if we provide additional general i* training to the control group during this time, it is likely that their increased familiarity with the basic i* concepts might also unfairly influence their performance.

7.3. External Validity

As indicated, the hypothesis underlying i* evaluation is generally aimed at anyone who may analyze a system. The participants of the experiment, however, were graduate and undergraduate Computer Science (CS) students. Although it is likely that these individuals will have to perform some sort of software system analysis at some point, not all individuals who perform system analysis necessarily have this background. Some may not have this level of formal training, some may not be explicitly trained in CS, and many may have a greater level of industry experience. As a result, we can only, with confidence, generalize our results to CS students. Future studies should attempt to test whether these results hold for a wider population.

7.4 Reliability

The use of a particular domain, descriptive document, and questions may raise concerns over experimental reliability. We have attempted to demonstrate the reliability of our results by choosing a domain description which is not specifically aimed for i* modeling, containing extraneous and contradictory information, as may be found in real-life situations. However, we have deliberately chosen a domain and domain questions for which i* and i* evaluation is suited, involving intentional desires, interactions between multiple actors, and the tracing of qualitative effects. If a significantly different domain, without information on such aspects, was chosen both i* modeling and i* evaluation would be less useful. By articulating hypothesis for a procedure which extends the capabilities of i* modeling, we adopt the limitations of the Framework as articulated in the theories of Ernst et al [8]. We see this not as a deficit of i* or i* evaluation, but as a natural consequence of conceptual representation. Not every framework can be suitable for every application domain.

8. Related Work

The i* Framework shares concepts and notations with goal modeling Frameworks such as the NFR [11] and KAOS Frameworks [17]. Various procedures to measure the achievement of goals have been introduced for these Frameworks [18, 19]. The focus of these procedures is on answering questions regarding the satisfaction of goals, given a prescriptive design option.

The i* Framework differentiates itself from these goal modeling Frameworks in part by focusing on exploratory as opposed to prescriptive design. Similarly, the i* evaluation procedure differs itself from procedures intended for goal modeling by explicitly encouraging model iteration and learning, with the knowledge gained often going beyond the focus of the original analysis question.

9.0 Conclusions and Future Work

The presence of clear foundational theories for modeling Frameworks such as i* and procedures such as i* evaluation can assist potential users in determining the contexts and applications for which such tools may be best suited. The key is to be able to understand not only which tools can be useful, but why they are useful, in order to effectively select the best tool(s) for each application.

To this end we have articulated the hypothesis underlying the utility of the i* evaluation procedure, namely, that this procedure produces beneficial changes to both the physical i* model and the mental model of the evaluator. The domain understanding gained through these changes can lead to a more effective articulation of system requirements, better accounting for the integration of a system into a social environment. Through the design and administration of an experiment, we have increased our confidence in the validity of this hypothesis.

This work has focused on the individual effect of i* evaluation. In the future, we plan to articulate and test similar theories for the potential effects of i* evaluation on group communication and agreement.

References

[1] E. Yu, *Modelling Strategic Relationships for Process Reengineering*, Ph.D. thesis, also Tech. Report DKBS-TR-94-6, Dept. of Computer Science, University of Toronto, 1995.

[2] E. Yu, "Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering", *Proceedings of the 3rd IEEE Int. Symp. on Requirements Engineering (RE'97)*, Washington D.C., USA, Jan. 6-8, 1997, pp. 226-235.

[3] J. Horkoff, *Using i* for Evaluation*, (Master's Thesis), Department of Computer Science, University of Toronto, 2006.

[4] J. Mylopoulos and J. Castro, "Tropos: A Framework for Requirements-Driven Software Development", In J. Brinkkemper and A. Solvberg (eds.), *Information Systems Engineering: State of the Art and Research Themes*, *Lecture Notes in Computer Science*, *Springer-Verlag*, 2000, p. 261-273.

[5] H. Mouratidis, P. Giorgini, G. Manson, I. Philp, "Using Tropos methodology to Model an Integrated Health Assessment System", *Proceedings of the 4th International Bi-Conference Workshop on Agent-Oriented Information Systems (AOIS-2002)*, Toronto, Ontario, May 2002.

[6] X. Franch, N.A.M. Maiden, "Modelling Component Dependencies to Inform Their Selection", *Lecture Notes in Computer Science*, *Volume 2580*, Jan 2003, pp. 81- 91.

[7] N. A. M. Maiden, S. V. Jones, S. Manning, J. Greenwood, L. Renou, "Model-Driven Requirements Engineering: Synchronising Models in an Air Traffic Management Case Study", *16th International Conference*, *CAiSE 2004*, Riga, Latvia, June 7-11, 2004, pp. 368-383.

[8] N. Ernst, J. Aranda, and J. Horkoff, "An Empirical Framework for Model Assessment", in progress, University of Toronto, 2006.

[9] L. Liu, E. Yu, J. Mylopoulos, "Security and Privacy Requirements Analysis within a Social Setting", *International Conference on Requirements Engineering (RE'03)*, Monterey, California, 2003, pp. 151-161.

[10] L. Liu, E. Yu, "Designing Information Systems in Social Context: A Goal and Scenario Modelling Approach", *Information Systems*, *Elservier*, *29(2)*, 2004, pp. 187-203.

[11] L. Chung, B.A. Nixon, E. Yu, J. Mylopoulos, *Non-Functional Requirements in Software Engineering*, Kluwer Academic Publishers, 2000.

[12] M. Shaw, D. Garlan, R. Allen, D. Klein, J. Ockerbloom, C. Scott, M. Schumacher, *Model Problems*, Retrieved July 2006 from http://www.cs.cmu.edu/~ModProb/

[13] A.J. Smith, "The Task of the Referee", Computer Science Division, EECS Department, University of California Berkeley, California 94720, USA

[14] K. R. Popper, *The logic of scientific discovery*, New York, Basic Books, 1959.

[15] J. Horkoff, J. Aranda, S. Easterbrook, E. Yu, "Feature Prioritization and Analysis Using i* Models", Submitted to *14th IEEE International Requirements Engineering Conference (RE'06).*

[16] L. Briand, K. El Emam, and S. Morasca, "On the Application of Measurement Theory to Software Engineering", *Empirical Software Eng.--An Int'l J.*, vol. 1, no. 1, 1996.

[17] A. Dardenne, S. Fickas and A. van Lamsweerde, "Goal-Directed Concept Acquisition in Requirements Elicitation", *Pmc. IWSSD-6 - 6th Intl. Workshop on Software Specificationand Design*, 1991, pp. 14-21.

[18] P. Giorgini, J.Mylopoulos, R. Sebastiani, "Simple and Minimum-Cost Satisfiability for Goal Models", *16th Conference On Advanced Information Systems Engineering (CAiSE*04)*, Springer Verlag, 2004.

[19] E. Letier, A. van Lamsweerde, "Reasoning about Partial Goal Satisfaction for Requirements and Design Engineering", *Proceedings of FSE'04, 12th ACM International Symp. on the Foundations of Software Engineering*, Newport Beach, CA, Nov. 2004, pp. 53-62.

Note: This is the material which would be hypothetically available online. It is not officially part of the paper, but is attached here for convenience.

Domain Description: Conference Refereeing

From: http://www.cs.cmu.edu/~ModProb/CR.html

Professional conferences are held in order to announce and discuss new results. The core activity of organizing a conference centers on selecting the papers to be presented. Usually this is done by making an open invitation calling for papers to be submitted, circulating the submitted papers to a (geographically distributed) panel of reviewers, then selecting the best papers to appear on the program. A system to automate conference refereeing should do the following:

1. The program committee announces "call for papers."

2. Authors receive the call for papers and decide to will submit papers on their work. They write papers and send them to the program committee. A given paper may have several authors, but only one reply address.

3. The program committee registers the contributed papers upon receipt.

4. At a certain point in time the program committee distributes the papers among the panel of referees. Each paper is sent to three distinct referees, none of whom is an author of the paper.

5. The program committee continuously collects reports from the referees.

6. At a certain point in time the program committee selects papers for inclusion in the program and notifies the authors about the selection. This may involve obtaining additional opinions from the referees.

7. The program committee advises the authors of the selection results.

From:
The Task of the Referee*
Alan Jay Smith, Computer Science Division, EECS Department, University of California
Berkeley, California 94720, USA
http://www.idt.mdh.se/phd/Smith-TaskOfTheReferee.pdf

“The task of the referee is to evaluate in a timely manner a paper for publication in a specific journal or conference proceedings. This involves determining if the work presented is correct, if the problem studied and the results obtained are new and significant, if the quality of the presentation is satisfactory or can be made so, and what revisions and changes to the paper are necessary and/or desirable. The evaluation must be with regard to the coverage and degree of selectivity of the specific publication.

…

There is a constant stream of papers written and submitted for publication to conferences, journals, newsletters, anthologies, annuals, trade journals and newspapers, and other periodicals. Many such publications use referees as impartial, external experts to evaluate papers. This approach is often called *peer review.* Refereeing is a public service, one of the professional obligations of a computer science and engineering professional. Typically, referees learn to produce referee reports without any formal

instruction: by practice, by feedback from editors, by seeing referee reports for their own papers, and by reading referee reports written by others.

…

A paper is publishable if it makes a *sufficient contribution*. A contribution can be new and interesting research results, a new and insightful synthesis of existing results, a useful survey of or tutorial on a field, or a combination of those types. To quote a referee for this article itself: ‘‘small results which are surprising and might spark new research should be published; papers which are mostly repetitions of other papers should not; papers which have good ideas badly expressed should not be published but the authors should be encouraged to rewrite them in a better, more comprehensible fashion.’’ The role of the referee is to provide an *opinion* as to whether the paper makes such a sufficient contribution. There is seldom a single correct evaluation of a paper, and equally skilled and unbiased readers will differ.

The Task of the Referee

The two major components of a referee report are:

(a) A recommendation for or against publication in a specific publication or presentation at a specific forum. An equivocal recommendation is acceptable if adequate discussion is provided for the guidance of the editor or program committee. If rejection is recommended, and if the paper does contain some publishable research, the report can suggest another place to publish. In all cases, sufficient discussion *must* be provided to justify the recommendation.

(b) A list of necessary and recommended changes and revisions. A recommendation to reject the paper does not excuse the referee from suggesting changes that might permit the paper to be published elsewhere, or after resubmission. The extent of necessary revisions, for journal publication, is largely separate from the recommendation for (eventual) publication; for a conference, the short time available for revisions, and the difficulty of arranging for a second (or n’th) round of revisions generally means that a paper which requires substantial revision cannot be accepted.

It is very important that the referee walk the uncertain line between being overly permissive (“publish everything”) and overly restrictive (“nothing is good enough to publish”). If the referee is insufficiently critical, poor research is encouraged, recognition (of a sort) and honors (of a sort) are given to those who don’t deserve it, the naive and inexperienced reader is misled, the author is misled as to what is publishable, disrespect for the field is encouraged, commercial development is distorted, as are hiring, promotion and tenure decisions, and the paper may actually subtract from the general store of knowledge; consider the Piltdown man fraud. As has been noted in [Thom84] and elsewhere, one of the worst problems with unrestrained publication is to bury the professional under mounds of paper, only a very small fraction of which can be examined, let alone read. If the referee is overly critical of research, he blocks good research from publication, or causes it to be delayed in publication, wastes the time of authors, damages careers, and perhaps leaves journals with nothing to publish and conferences with nothing to present. It is particularly important not to reject new and significant work which runs counter to the prevailing wisdom or current fashions.

It is important for a referee who wants to be taken seriously to have a middle of the road view, to be able to distinguish good from bad work, and major from minor from negative contributions to the literature. A referee who always says “yes” or always says “no” is not helpful.

…

Conflicts of Interest

If you have a conflict of interest, you should make it known to the editor. If the conflict is severe, you should not referee the paper, but should instead return it to the editor. For example, if you have a feud with an author, or a significant personal disagreement, it would be wise to send the paper back. If you are competing with the author for funding, and this is a proposal, you should make that known to the program officer.

The opposite type of conflict also occurs - you are being asked to referee a paper written by a friend, colleague, former or current student, boss or subordinate, or former advisor. If you feel that you cannot provide an objective review, then you should return the paper to the editor.

Role of the Editor or Program Chairman

The editor has several tasks [Bish84]. Here we refer to both the editor in chief, who typically has the authority to decide whether to accept a paper, and the associate editors, who solicit the referee reports and recommend to the editor in chief whether to publish. The editor receives the paper from the author and maintains correspondence with the author. The editor selects the referees, sends them each a copy of the paper with suitable instructions, and awaits their results. The editor *should* remind tardy referees, and find new referees after a certain period if no response has been received.

The editor should select referees who are knowledgeable in the subject matter of the paper, and can be relied upon to provide a fair and objective evaluation. Unfortunately, it is not always possible to do this - there are too many papers to be reviewed, and too few people known to be sufficiently expert and responsible. There is also another problem - by definition, people in area X believe that work in area X is worthwhile. A report received from someone in area X will evaluate the paper in area X by the standards of area X, but will seldom, if ever, say that work in area X is pointless and should be discontinued. It is, however, quite possible that such a response is appropriate; if one wants to debunk alchemy, one sends the paper to a chemist, not an alchemist. If you receive a paper to referee which is outside your area, you should consider whether it has been sent to you deliberately, and for that reason. Someone has to say that the emperor has no clothes.

…

In the case of a conference, the program chairman is responsible for selecting referees and collecting and tallying their reports. Typically, the program committee, in a meeting or conference call, will decide which papers to accept by majority vote. The program chair may or may not have a vote that is larger than that of the others on the committee, but he seldom has the authority to accept or reject papers over the opposition of a majority of the program committee. Due to the large number of papers to be handled in a very short time, referees and authors are not usually given the personal attention provided by an editor who handles only one or a few papers per month. Note that program committees often use numerical scores to prepare ranked lists of papers; such scores should be assigned carefully and should be viewed skeptically by the committee.

When You Are the Author

This article has been directed at the referee, but instructions to the referee are also instructions to the author. When starting research, when writing a paper, when finishing the paper, and when deciding where to submit it, ask yourself: how will this paper do when refereed according to the criteria given here? Some specific things to think about are: Are you submitting the paper to the right place? Some journals and conferences will not consider material outside a specific scope; why waste 3-12 months to find out that your paper wasn't appropriate? Likewise, if you know that your paper is minor, why send it to a highly selective forum; send it somewhere where it has a reasonable chance of being accepted.

If you suspect that further work is needed before publication, do that work; it may turn an unpublishable paper into a publishable one, without the 3-12 month extra delay. A look at an issue of the publication to which you are considering submission will answer many of these questions; it is also helpful to look over the information provided by the journal to prospective authors; e.g. [CACM89, IEEE84].

Keep in mind that a good referee report is immensely valuable, even if it tears your paper apart. Consider - each report was prepared without charge by someone whose time you could not buy. All the errors they find, all the mistaken interpretations they make are things that you can correct before publication. Appreciate referee reports, and make use of them. Some authors feel insulted, and ignore referee reports; that is a waste of an invaluable resource. An author receiving a negative referee report often suspects that the editor, program committee, program chair, and/or referees are incompetent, biased, or otherwise unfair. While this sometimes happens, it is the exception; individual referee reports are often wrong, but a *set* of negative referee reports is an accurate indication that your paper has a problem, and needs to be either rewritten or redone before resubmission, or discarded as unpublishable or embarrassing. Note particularly that the reader of a paper forms an opinion of the author; if the quality of a paper is such as to reflect badly on the author, it should not even be submitted for publication. Authors are particularly referred to [Day77], [Levi83], [Mano81], and [Wegm86], which provide discussions of how to write technical papers. Refereeing is also a good way to learn to write better papers; evaluating the work of others gives one insight into one's own.

Domain Questions

Q_Mental

1. What are the affects of accepting a paper of poor quality for a research conference?
2. What are the affects of submitting a paper to a conference which is of very poor quality?
3. What are the affects of rejecting a paper of good quality from a research conference?
4. What are the affects of submitting a referee report late?
5. What are the affects of a referee accepting a paper to review when he/she has a conflict of interest with that paper?

Q_MentalPhysical

6. What are the affects of not providing sufficient justification of a paper rejection in the referee report?
7. What are the affects of the author not making the changes recommended to them by the referees, when the paper is accepted to the conference?
8. What are the affects if the Program Chair chooses referees who are not knowledgeable in the subject matter of the submitted paper?
9. What are the affects of submitting a paper to a conference when the paper is out of the scope of the subject matter of the conference?
10. What are the affects of the referee being very critical in the report on a submitted paper?

Detailed Results

Question Results

Participants	Question Changes: Q_Mental_1 to Q_Mental 2	Question Changes: Q_Physical_1 to Q_Physical_2	Question Marks (out of 20): Q_Mental_1	Q_Mental_2	Difference	Q_Physical_1	Q_Physical_2	Difference	New Questions: # P_Questions
E1	2	4	16	16	0	16	18	2	1
E2	5	5	10	15	6	15	17	2	8
E3	4	3	12	17	5	17	18	1	5
E4	3	0	18	18	0	15	19	4	5
E5	3	5	12	14	2	14	16	2	5
E6	6	4	10	13	3	16	16	0	1
E7	0	6	14	14	0	12	16	4	8
E8	3	5	9	11	2	11	13	1	7
E9	3	4	15	17	2	12	16	4	7
E10	2	6	13	16	3	12	14	2	7
C1	0	3	15	15	0	12	12	0	7
C2	4	4	13	14	1	14	14	0	4
C3	3	2	14	16	2	14	16	2	4
C4	2	3	18	18	0	17	17	0	0
C5	4	2	14	16	2	17	17	0	6
C6	0	5	11	11	0	15	16	1	1
C7	1	3	8	9	1	12	13	1	1
C8	2	2	16	17	1	10	13	2	4
C9	3	3	17	19	2	9	11	1	3
C10	2	1	10	11	1	15	15	0	5

Avg. E	Avg. E	Avg. E	Avg. E	E Avg. Diff	Avg. E	Avg. E	E Avg. Diff	Avg. E
3.1	4.2	12.829	15.116	2.287	14.022	16.197	2.175	5.418

Avg. C	Avg. C	Avg. C	Avg. C	C Avg. Diff	Avg. C	Avg. C	C Avg. Diff	Avg. C
2.1	2.8	13.598	14.705	1.107	13.590	14.262	0.672	3.383
Diff A - B	Diff A - B							
1	1.4							
t= 1.43	t= 2.12			t= 1.86			t= 2.92	t= 1.74
sdev= 1.56	sdev= 1.48			sdev= 1.57			sdev= 1.15	sdev= 2.44
degrees of freedom = 18 The probability of this result, assuming the null hypothesis, is 0.17	degrees of freedom = 18 The probability of this result, assuming the null hypothesis, is 0.048			degrees of freedom = 18 The probability of this result, assuming the null hypothesis, is 0.080			degrees of freedom = 18 The probability of this result, assuming the null hypothesis, is 0.0091	degrees of freedom = 18 The probability of this result, assuming the null hypothesis, is 0.098

Model Results

	# Model Changes								Model Size					
Participants	Total M1 to M2	Name	Add Element	Remove Element	Add Link	Remove Link	Link Type	Element Type	Total M1	Elements	Links	Total M2	Elements	Links	Total Change
E1	17	4	2	2	1	2	4	2	68	22	46	146	56	90	78
E2	18	3	0	1	3	4	2	4	91	34	57	97	34	63	6
E3	12	4	2	3	1	0	1	0	90	27	63	124	55	68	33
E4	17	2	2	3	2	2	4	2	64	22	42	173	58	116	109
E5	16	3	3	3	3	1	2	1	97	42	55	128	43	85	31
E6	14	3	2	1	2	3	0	3	116	35	81	119	52	67	3
E7	20	2	3	3	3	3	4	2	127	49	79	135	51	84	8
E8	14	0	3	2	4	3	2	1	81	26	55	110	37	73	29
E9	12	2	1	1	2	2	1	2	135	41	94	138	43	95	3
E10	20	2	4	2	5	3	3	1	103	33	70	128	50	78	25
C1	14	3	1	1	2	2	2	3	92	36	56	118	36	81	25
C2	13	2	3	1	3	2	1	2	88	39	49	104	42	62	16
C3	9	3	0	0	2	1	2	2	112	33	79	113	33	80	1
C4	13	3	2	2	1	1	1	3	96	38	58	99	45	54	2
C5	9	2	2	0	2	2	2	1	96	34	63	106	46	60	10
C6	12	1	3	1	1	1	1	4	98	42	56	98	44	54	0
C7	12	2	1	2	2	1	1	3	96	38	58	113	62	51	16
C8	8	0	0	1	2	2	2	0	94	45	48	92	45	47	-2
C9	8	1	1	0	3	0	0	3	94	26	68	110	23	87	15
C10	6	0	0	1	2	1	2	1	104	41	64	115	49	66	11

Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E	Avg. E
16.017	2.420	2.090	2.191	2.666	2.356	2.333	1.961	97.193	33.027	64.166	129.728	47.871	81.856	32.535
Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C	Avg. C
10.403	1.560	1.265	1.015	1.885	1.213	1.324	2.141	97.078	37.264	59.814	106.718	42.518	64.201	9.640

t= 4.42
sdev= 2.83

degrees of freedom = 18 The probability of this result, assuming the null hypothesis, is 0.0003

t= 2.02
sdev= 25.5

degrees of freedom = 18 The probability of this result, assuming the null hypothesis, is 0.058