Experimental Evaluation of a Tool for Change Impact Prediction in Requirements Models: Design, Results, and Lessons Learned

Arda Goknil
University of Luxembourg
Luxembourg

Roderick van Domburg, Ivan Kurtev, Klaas van den Berg, Fons Wijnhoven
University of Twente
The Netherlands

***Abstract*—There are commercial tools like IBM Rational RequisitePro and DOORS that support semi-automatic change impact analysis for requirements. These tools capture the requirements relations and allow tracing the paths they form. In most of these tools, relation types do not say anything about the meaning of the relations except the direction. When a change is introduced to a requirement, the requirements engineer analyzes the impact of the change in related requirements. In case semantic information is missing to determine precisely how requirements are related to each other, the requirements engineer generally has to assume the worst case dependencies based on the available syntactic information only. We developed a tool that uses formal semantics of requirements relations to support change impact analysis and prediction in requirements models. The tool TRIC (Tool for Requirements Inferencing and Consistency checking) works on models that explicitly represent requirements and the relations among them with their formal semantics. In this paper we report on the evaluation of how TRIC improves the quality of change impact predictions. A quasi-experiment is systematically designed and executed to empirically validate the impact of TRIC. We conduct the quasi-experiment with 21 master's degree students predicting change impact for five change scenarios in a real software requirements specification. The participants are assigned with Microsoft Excel, IBM RequisitePro or TRIC to perform change impact prediction for the change scenarios. It is hypothesized that using TRIC would positively impact the quality of change impact predictions. Two formal hypotheses are developed. As a result of the experiment, we are not able to reject the null hypotheses, and thus we are not able to show experimentally the effectiveness of our tool. In the paper we discuss reasons for the failure to reject the null hypotheses in the experiment.**

***Index Terms*—Requirements management tools, change impact analysis, requirements models.**

I. Introduction

Today's software systems usually operate in a dynamic business context where business goals often change. As a result, the requirements of software systems change continuously and new requirements emerge frequently. A single requirement hardly exists in isolation: it is related to other requirements and to the software development artifacts that implement it. Thus, even a simple change in a single requirement may have a significant total effect on the whole system. Determining such an effect is usually referred to as *change impact analysis*. *Change impact prediction* is one of the results of the change impact analysis. It enumerates the set of elements expected to be impacted by a change.

Commercial tools such as *IBM Rational RequisitePro* and *DOORS* support semi-automatic change impact analysis for requirements. These tools capture the requirements relations and allow tracing the paths they form. For example, when a requirement is changed in RequisitePro, relations of the changed requirement are marked as *suspect.* RequisitePro recognizes two relation types based on the direction of the relation: *traceFrom* and *traceTo*. All requirements directly or indirectly traced from the changed requirement (with relations marked as *suspect*) are candidate for the impact. This analysis only considers the presence of a relation and does not take into account the meaning of the relation. Several requirements relation types have been identified in literature, for example, *refinement*, *part-of*, *influence*, *conflict*. The actual impact of the change depends on the semantics of the relations and for some relation types the related requirements are not impacted. Therefore, a change impact analysis technique that uses only the transitive property of the requirements relations may suggest elements that are not impacted, i.e. elements that are *false positives*. Bohner [2] calls the problem of producing a high number of false positive impacted elements as *impact explosion problem*. The reason for impact explosion is that the semantic information about relations is either missing or not utilized during change impact analysis.

We developed a tool that aims at limiting the impact explosion during change impact analysis and prediction in requirements models. TRIC (**T**ool for **R**equirements **I**nferencing and **C**onsistency checking) works on models that explicitly represent requirements and the relations among them. Five requirements relation types (*requires*, *refines*, *partially refines*, *contains*, and *conflicts*) are supported and formally defined in First-Order Logic [10] [11]. The formal semantics of the relations is used to determine if a change in a requirement has an impact on related requirements. In this way, the number of the candidate requirements is smaller than the number of the candidates given by an analysis based only on syntactical information about the relations. The technique is still semi-

automatic since the requirements engineer has to choose among several alternative changes on an impacted requirement.

Several tests performed on example models showed that TRIC eliminates a number of false positive impacted elements produced by other tools. However, this is not sufficient evidence that the tool improves change impact analysis results compared to the commercial tools. Several factors have to be considered in addition. First, the semi-automatic nature of the techniques requires input from the requirements engineers. It is possible that an experienced engineer produces excellent results on small models even if the analysis is completely manual. Second, the times taken for performing change impact analysis with different tools need to be compared.

In this paper we report on the evaluation of how TRIC improves the quality of change impact predictions by explicitly using the semantics of requirements relations. A quasi-experiment is systematically designed and executed to empirically validate TRIC. The experiment is conducted with 21 master's degree students. The students have to predict the change impact for five change scenarios in a real-life software requirements specification. The quality of change impact predictions is measured by F-score and the time for completing the predictions is measured in seconds. The independent variable is the level of tool support. The participants are assigned with Microsoft Excel, IBM Rational RequisitePro or TRIC to perform change impact prediction.

It is hypothesized that using TRIC would positively impact the quality of change impact predictions. Two formal hypotheses are developed. Null hypothesis 1 states that the F-scores of change impact predictions using TRIC will be equal to or less than those not using TRIC. Null hypothesis 2 states that the time taken to complete change impact predictions using TRIC will be equal or longer than those not using TRIC. The data are analyzed using ANOVA and $\chi 2$ statistical analyses.

Although the experiment has been designed and conducted carefully, we were not able to reject both null hypotheses. No significant difference in F-scores between TRIC and the other groups is detected. TRIC is found to be significantly slower for four out of five change impact predictions. These inferences are made at $\alpha = 0{,}05$ with a mean statistical power of 54%. We observed that using TRIC on a software requirements specification of low complexity does not yield better quality predictions but does take a longer time.

This paper is organized as follows. Section II gives details of the change impacts explosion problem in requirements. Section III gives a brief description of our tool TRIC. In Section IV, we present the planning for the experiment which serves as a blueprint for the execution of the experiment depicted in Section V, analysis and interpretation of its results given in Section VI and Section VII respectively. Section VIII concludes the paper with discussion on reasons for the failure to reject the null hypotheses in the experiment.

II. Change Impact Prediction

Change impact prediction enumerates the set of elements estimated to be impacted in change impact analysis. Table I explains the sets of elements in change impact prediction.

TABLE I. Change Impact Prediction Sets [1]

Set	Abbreviation	Description
System	-	Set of all objects under consideration.
Estimated Impact Set	EIS	Set of all objects that are estimated to be affected by the change.
Actual Impact Set	AIS	Set of all objects that were actually modified as a result of performing the change.
False Positive Impact Set	FPIS	Set of objects that were estimated to be affected during performing the change.
Discovered Impact Set	DIS	Set of objects that were not estimated to be affected, but were affected during performing the change.

The *Estimated Impact Set* may not be equal to the *Actual Impact Set*. Thus, there is a need for measuring the quality of the change impact predictions. This may be captured using a binary classifier (see the confusion matrix in Table II).

TABLE II. Confusion Matrix [7]

		Actual Impact	
		Changed	**Not changed**
Estimated Impact	**Changed**	True Positive	False Positive
	Not changed	False Negative	True Negative

Binary classifiers are also used in the domain of information retrieval. Metrics from this domain can be used to measure the quality of change impact predictions [1]. Table III shows the change impact prediction quality metrics.

TABLE III. Change Impact Prediction Quality Metrics [1]

Metric	Equation	Also known as
Recall	$\frac{\lvert EIS \cap AIS \rvert}{\lvert AIS \rvert}$	Hit rate, sensitivity, true positive rate
Precision	$\frac{\lvert EIS \cap AIS \rvert}{\lvert EIS \rvert}$	Positive predictive value
Fallout	$\frac{\lvert FPIS \rvert}{\lvert System \rvert - \lvert AIS \rvert}$	False alarm rate, false positive rate

A popular measure that combines precision and recall is the weighted harmonic mean of precision and recall, also known as the F_1-measure (see Eq. 1) because recall and precision are evenly weighted [1].

$$F_1 = (2 * \text{precision} * \text{recall}) / (\text{precision} + \text{recall}) \quad (1)$$

Measures such as $F_{0,5}$ and F_2 weigh either the precision or recall double and can be used if either precision or recall is more important than the other in a certain situation. F_1-measure is the most used one and henceforth referred to as the *F*-measure. Results on the *F*-measure are referred to as *F*-scores.

Another quality attribute of change impact predictions is the effort that it takes. While *F*-measure can be regarded as a quality measure of change impact prediction products, the measurement of change impact prediction process effort is left to human judgment [2]. Time is one plausible metric to measure effort [19].

III. TRIC: Tool for Requirements Inferencing and Consistency Checking

TRIC [11] [23] works on models that explicitly represent requirements and the relations among them. Five requirements relation types (*requires*, *refines*, *partially refines*, *contains*, and *conflicts*) are supported and formally defined in First-Order Logic. The semantics of the relations helps requirements engineers in deciding if a related requirement is really impacted by a change. TRIC provides two main features used in the experiment: (i) *managing requirements and relations,* and (ii) *reasoning on requirements relations*.

It supports two activities for reasoning on relations. First, new relations among requirements can be inferred from the initial set of given relations. Second, requirements models can be automatically checked for consistency of the relations. Both the inferred and given relations are used to propagate a change from one requirement to another requirement. The semantics of the relations can guide the requirements engineer to rule out some false positive impacted requirements.

IV. Methodology: Experimental Evaluation of TRIC for Change Impact Prediction

We present the planning for the experiment which serves as a blueprint for the execution of the experiment and interpretation of its results. The design is based on the research goal and hypotheses that support it.

A. Goal

The goal of this experiment is to analyze the real-world impact of using a software tool with formal requirements relationship types for the purpose of the evaluation of the effectiveness of tools with respect to the quality of change impact predictions.

B. Hypothesis

It is hypothesized that using TRIC has a positive impact on the quality of change impact predictions. The rationale for the hypothesis is that the available explicit requirements relation types with formal semantics in TRIC facilitate the decision for the impact. In contrast to TRIC, other tools just indicate relations without giving information about their semantics.

Hypothesis 1. The *F*-scores of change impact predictions of requirements engineers using TRIC will be equal to or less than those from requirements engineers not using TRIC.

$$H_{0,1} : \mu_1 \leq \mu_2$$
$$H_{1,1} : \mu_1 > \mu_2$$

μ is the mean F-score of change impact predictions. Population 0 consists of requirements engineers using TRIC. Population 1 consists of requirements engineers not using TRIC.

Hypothesis 2. The time taken to complete change impact predictions of requirements engineers using TRIC will be equal to or greater than those from requirements engineers not using TRIC.

$$H_{0,2} : \mu_1 \geq \mu_2$$
$$H_{1,2} : \mu_1 < \mu_2$$

μ is the mean time of change impact predictions as measured in seconds. Population 0 consists of requirements engineers using TRIC. Population 1 consists of requirements engineers not using TRIC.

The statistical significance level for testing the null hypotheses is 5% ($\alpha = 0{,}05$). A lower level would be feasible given a large enough sample size, which is not the case here due to limited time and availability of participants. From previous experiences it is known that most of the students do not volunteer for a full day. Likewise, experts from industry are too busy to participate a full day even if they are linked to the our research project as partner. Ample monetary compensation is not within the budget of this experiment and is conducive to the threat of compensatory inequality [21].

C. Design

Different groups are assigned to perform change impact analysis using a different software tool. This research setup involves control over behavioral events during change impact analysis with administrator selection, for which experimental research is the most appropriate [24].

We follow a synthetic design with three treatments to control the level of tool support within a limited amount of time. The treatment is the administration of Excel, RequisitePro, and Excel. The observation is the change impact prediction quality as measured by *F*-score and time.

D. Parameters

A single real-world software requirements specification is selected as a research object. Predetermined groups of participants perform change impact analysis on the requirements in the specification.

E. Variables

Dependent Variables. The dependent variables measured in the experiment are those required to compute the *F*-score, which is a measure of change impact prediction quality: (i) size of the *Estimated Impact Set*, (ii) size of the *False Positive Impact Set*, and (iii) size of the *Discovery Impact Set*.

Independent Variables. One independent variable in the experiment is the supplied software tool during change impact analysis. This is measured on a nominal scale: *Microsoft Excel*, *IBM Rational RequisitePro* or *TRIC*.

The nominal scale is preferred over the ordinal scale of software tool intelligence because our research is interested in the impact of TRIC on the quality of change impact predictions as a new technique versus classic techniques.

It would be a threat to internal validity if we only study the impact of using Microsoft Excel and TRIC. Such an experimental design would be biased in favor of TRIC. When assuming that requirements relationships play an important role in the results of change impact prediction, it would be logical that a software tool with dedicated support would score higher than a software tool without such support. By also studying an industrial tool such as IBM Rational RequisitePro, concerns to validity regarding the bias in tool support are addressed.

Covariate Variables. The following covariate variables are expected to influence the *F*-scores of change impact predictions

and time taken to complete them [17] [22]: *level of formal education*, *nationality*, *gender*, *current educational program*, *completion of a basic requirements engineering course*, *completion of an advanced requirements engineering course*, and *previous requirements management experience*.

F. Planning

The participants register for the experiment in advance and provide responses to the covariables. Groups are created by first assigning participants at random. The groups are equalized on covariates by manually moving participants from one group to another one.

During the experiment, the participants receive an equal and general instruction about change management for 15 minutes. Then, they receive a lecture specific to their tool for 30 minutes. Following that, they receive an equal kick-off lecture with the experimental procedure and prizes to be won for 5 minutes. Participants are then granted 60 minutes to review the software requirements specification. Following a 15-minute break, they are granted 60 minutes to perform change impact analysis for five change scenarios. Change scenarios are distributed to the participants at random to compensate learning effects. The instructions are provided by the team of researchers.

G. Participants

Participants are master students following the Software Management master course at the University of Twente. The experiment is not strictly part of the course and students are encouraged to participate on a voluntary basis. For each software tool group, there is a first prize of € 50 and a second prize of € 30. Everyone is presented with a USB memory stick.

H. Objects

Requirements Specification. The research object is a requirements specification for the WASP (Web Architectures for Services Platforms) project by the Telematica Instituut [6]. This is a public, real-world requirements specification in the context of context-aware mobile telecommunication services, with three scenarios, 16 use cases and 71 requirements (see the thesis [5] for the requirements specification).

Change Scenarios. Scenarios were created to cover a range of change scenario cases. Five cases (see Table IV) can be discerned in the theory on formal requirements relations [11].

TABLE IV. CHANGE SCENARIO CASES AND TASKS

Case	Task
Add Property to Requirement	1
Delete Property from Requirement	2, 4
Add Constraint to Property of Requirement	3
Add Requirement	-
Delete Requirement	5

Table IV shows the five change scenario cases and matching tasks. For each case, a requirement was selected at random and an appropriate change scenario was created. The change scenarios are described in the appendix of the thesis [5].

I. Instrumentation

All participants are handed out a printout of all slides that were shown to them, a copy of the software requirements specification, and a USB memory stick. The memory stick contains the requirements specification in PDF format and a digital requirements document that can be opened with their software tool. It is pre-filled with all requirements but contains no relations. The participants are told to treat the introduction, scenario and requirements chapters as leading and the use case chapter as informative.

J. Data Collection

A web application is created to support the registration of participants, distribution of experiment tasks and collection of data. The Actual Impact Set is to be determined as a golden standard from experts.

K. Analysis Procedure

The web application has built-in support to calculate the *F*-scores. For each participant, it outputs the participant number, group number, covariate scores and F-scores and times per task to a file that can be imported in SPSS 16.0. SPSS [20] is used to perform an analysis of variance using planned comparisons to test if participants in the TRIC group had significantly different *F*-scores and times than those in the Microsoft Excel or IBM RequisitePro groups. A similar test is performed for analysis of covariance. Finally, a multiple analysis of variance is used to test if there are interaction effects between the *F*-scores and times.

V. EXECUTION OF THE EXPERIMENT

We describe the steps taken to execute the experiment.

A. Sample

The experiment was conducted with 21 participants who completed the online registration before the start of the experiment to score the covariates and facilitate group matching. All registered participants showed up. The participants were distributed over three groups. 6 participants were in the Microsoft Excel group, 7 in the IBM Rational RequisitePro group and 8 in the TRIC group.

B. Preparation

Five slideshows were created: one for the general lecture, three for the specific lecture (one per group) and one for the general kick-off.

C. Data Collection Performed

All 21 participants submitted estimated impact sets for six change scenarios. The estimated impact sets of the first scenario were the result of the warm-up and not used in statistical analysis.

D. Procedure

There were some deviations from the planning with regard to the experiment location and participant distribution. The supervisors noted the following deviations:

Ambiguous Rationales for Change Scenarios: The change scenarios are not entirely unambiguous. Some students raised questions about the rationale of changes. As with the lectures, the supervisors withheld themselves from providing further explanation. This may be a reliability problem because it can induce guessing with individuals [25].

Lack of Time: Many students were not finished with adding relations before the break. After the break, some of them tried catching up by adding more relations. Others started change impact analysis with the unfinished set of relations. When this was noticed, the supervisors jointly decided to provide an extra 15 minutes. The extra time was not enough for many students. This situational factor may be a reliability problem.

Ineffective Use of Tools: Not all students used the tools with its all features and some did not use them at all. This may be a reliability problem due to differences in skills and ability if not corrected by covariates.

Lack of Precision. Some participants did not check the initially changed requirement as part of their estimated impact set. The data set was corrected to include the initially changed requirement for all participants. The underlying assumption is that this has been an oversight by the participants; however, it may just as well be a reliability problem due to a lack of motivation or reading ability.

VI. Data Analysis

A number of analyses were made regarding the representativeness of the change scenarios, the inter-rater reliability of the golden standard, the quality of participants' change impact predictions, and the time taken for the scenarios.

A. Change Scenario Representativeness

One of the authors of the WASP specification was asked to rate the representativeness of the change scenarios on an ordinal scale of low, medium or high (see Table V).

TABLE V. Representativeness of Change Scenarios

Scenario	Representativeness
1	Medium
2	Low
3	High
4	Medium
5	Low

B. Golden Standard Reliability

Four people formed their own golden standard individually; one expert (one of the authors of the WASP specification) and three academics with the software engineering department: a postdoctoral fellow, a PhD candidate and a master student.

The golden standards contain dichotomous data: a requirement is rated to be either impacted or not impacted. To form the final golden standard, it was decided to use the mode of the individual golden standards. When this was not possible initially because of a split, then the academics debated until one of them was willing to revise his prediction.

In an experimental setting, it is important to calculate the level of agreement between expert ratings [15] such as the golden standards, which is called the inter-rater reliability [15]. The inter-rater reliability was calculated as a measure of the level of agreement between the golden standards (see Table VI). The interpretation of the results of inter-rater reliability analysis is given in Section VII.B.

TABLE VI. Inter Rater Reliability Analysis

		Raw Agreement		Significance (a)		Intraclass correlation
Task	Impacted set size	Mean	Standard error	Asymptotic	Exact	Two-way Random (b)
1	3	58,1%	9,1%	0,343	0,519	0,832
2	9	78,6%	4,2%	0,438	0,544	0,936
3	1	100,0%	0,0%	-	-	1,000
4	1	100,0%	0,0%	-	-	1,000
5	6	44,9%	9,7%	0,000 (c)	0,000 (c)	0,712

a. Friedman Test

b. Using an absolute agreement definition between four raters

c. $p < 0{,}0005$

Significance levels equal to or less than 0,0005 indicate that there were significant differences between the golden standards. Exact significance levels provide more precise values than asymptotic significance levels. Asymptotic significance levels are provided for comparison with other experiments that do not list exact significance levels. The intraclass correlation score indicates the level of agreement. Higher scores are better, with a score of "0" indicating no agreement and a score of "1" indicating full agreement.

C. One-way between-groups ANOVA

One-way between-groups analysis of variance is used when there is one independent variable with three or more levels and one dependent continuous variable [20]. It tests if there are significant differences in the mean scores on the dependent variables, across the three groups.

Following *Hypothesis 1* and *Hypothesis 2*, the analysis should test if the TRIC group performed superior to both the Microsoft Excel and IBM RequisitePro groups. Planned comparisons lend themselves better to this goal than post-hoc tests because of power issues. Post-hoc tests set more stringent significance levels to reduce the risk of false positives given larger number of performed tests. Therefore, planned comparisons are more sensitive in detecting differences.

In this experiment, the independent variable is the experiment group. This experiment features two dependent variables: the F-score and the elapsed time for a task. An analysis of variance can be performed separately on both F-score and elapsed time.

A number of assumptions underlie analyses of variance. These assumptions are tested for the actual analyses to be carried out. There were some deviations while testing for normality and homogeneity of variance. A Kolmogorov-Smirnov test for normality revealed non-normality for several

results of tasks 2, 4 and 5. It was decided to analyze these tasks using a non-parametric test.

Table VII presents the results of a one-way between-groups analysis of variance to explore the impact of using three different software tools on the quality of change impact predictions, as measured by the *F*-score. Using a planned comparison for the TRIC group, there were no statistically significant differences at the $p < 0,05$ level in the *F*-scores of the three groups in either task 1 [$F(1, 18) = 0,030$; $p = 0,866$] or task 3 [$F(1, 18) = 0,242$; $p = 0,629$].

TABLE VII. ONE-WAY BETWEEN-GROUPS ANOVA ON F-SCORE

	***F*-score (higher is better)**			**ANOVA (a)**		
Task	**Group**	**Mean**	**Standard deviation**	**Significance**	**F**	**η2**
1	Excel	0,498	0,232	0,866	0,030	0,002
	ReqPro	0,658	0,187			
	TRIC	0,593	0,176			
	Total	0,588	0,198			
3	Excel	0,407	0,321	0,629	0,242	0,013
	ReqPro	0,468	0,290			
	TRIC	0,507	0,325			
	Total	0,465	0,300			

(a) Using a planned comparison with TRIC

Table VIII presents the results of a one-way between-groups analysis of variance to explore the impact of using three different software tools on the time taken to complete predicting change impact, as measured in seconds. There was a statistically significant difference at the $p < 0,05$ level in the times of the three groups for task 1 [$F(1, 18) = 24,04$; $p = 0,000$]. The effect size, calculated using η2, was 0,572. In Cohen's terms [3], the difference in mean scores between the groups is large. The TRIC group performs change impact analysis 48% slower than the Microsoft Excel group and 63% slower than the IBM Rational RequisitePro group.

TABLE VIII. ONE-WAY BETWEEN-GROUPS ON TIME

	Time (lower is better)			**ANOVA**		
Task	**Group**	**Mean**	**Standard deviation**	**Significance**	**F**	**η2**
1	Excel	193	89	0,000	24,04	0,572
	ReqPro	137	53			
	TRIC	368	117			
	Total	241	136			
3	Excel	172	70	0,219	1,753	0,088
	ReqPro	239	121			
	TRIC	314	219			
	Total	249	161			

There was no statistically significant difference at the $p < 0,05$ level in the times of the three groups for task 3 [$F(1, 18) = 1,753$; $p = 0,219$].

The attained statistical power is 56% for detecting effects with a large size, $p < 0,05$; sample size 21 and 18 degrees of freedom. The critical value for the F-test statistic to attain a significant result is 4,41. To attain a statistical power of 80%, a sample size of 34 would be required. The critical value for the F-test statistic to attain a significant result would be 4,15. This is calculated by using the G*Power 3 tool [13] because SPSS 16.0 [20] lacks the necessary support.

D. Non-parametric Testing

As a non-parametric test, χ2 test for goodness of fit can test if there are significant differences between dependent variables across multiple groups without requiring a normal data distribution [20]. It does require a sufficiently large sample size; values of 20 through 50 have been reported although there is no generally agreed threshold [9].

Table IX and Table X display the results of χ2 test for tasks 2, 4 and 5, which did not meet the requirements for analyzing them using a more sensitive analysis of variance.

Tasks 2, 4 and 5 did not meet the preconditions for performing the preferred analysis of variance; tasks 1 and 3 are tested using an analysis of variance in Section VI.C.

Table IX presents the results of a χ2 test to explore the impact of using three different software tools on the quality of change impact predictions, as measured by the *F*-score.

TABLE IX. X2 TEST FOR GOODNESS OF FIT ON F-SCORE

	***F*-score (higher is better)**				
Task	**Group**	**Mean**	**Standard deviation**	**Significance**	**χ2**
2	Excel	0,499	0,319	0,584	1,077
	ReqPro	0,517	0,129		
	TRIC	0,424	0,275		
	Total	0,476	0,242		
4	Excel	0,407	0,182	0,717	0,667
	ReqPro	0,524	0,230		
	TRIC	0,461	0,161		
	Total	0,467	0,188		
5	Excel	0,423	0,160	0,444	1,625
	ReqPro	0,528	0,100		
	TRIC	0,573	0,151		
	Total	0,515	0,146		

In Table IX, significance levels equal to or less than 0,005 indicate a significant difference in F-scores between the TRIC group and the other two groups. The χ2 value describes the test statistic for a χ2 test. It is used to describe the shape of the distribution of the χ2 test. It is reported for comparison with other experiments.

There were no statistically significant differences at the $p < 0,05$ level in the F-scores of the three groups in task 2 [$\chi2 = 1,077$; $df = 2$; $p = 0,584$], task 4 [$\chi2 = 0,667$; $df = 2$; $p = 0,717$] or task 5 [$\chi2 = 1,625$; $df = 2$; $p = 0,444$].

Table X presents the results of a χ2 test to explore the impact of using three different software tools on the time to complete change impact predictions, as measured in seconds.

There were statistically significant differences at the $p < 0,05$ level in the times of the three groups in task 2 [$\chi2 = 414$; $df = 2$; $p = 0,000$], task 4 [$\chi2 = 102$; $df = 2$; $p = 0,000$] or task 5 [$\chi2 = 612$; $df = 2$; $p = 0,000$].

Because χ2 tests do not support planned comparisons, a post-hoc comparison is required to discover how groups differ from each other. Post-hoc comparisons explore the differences for each group and can be performed using a Mann-Whitney U test, which tests for differences between two independent groups on a continuous measure [20].

A post-hoc comparison using a Mann-Whitney U test revealed that the time taken to complete task 4 was significantly different between the Microsoft Excel and TRIC

groups, p=0,020. The TRIC group performs change impact analysis 54% slower than the Microsoft Excel group.

TABLE X. X2 TEST FOR GOODNESS OF FIT ON TIME

	Time (lower is better)				
Task	Group	Mean	Standard deviation	Significance	$\chi 2$
2	Excel	133	83	0,000	414
	ReqPro	154	76		
	TRIC	222	137		
	Total	174	107		
4	Excel	213	111	0,000	102
	ReqPro	300	81		
	TRIC	467	248		
	Total	339	196		
5	Excel	324	274	0,000	612
	ReqPro	170	64		
	TRIC	342	133		
	Total	280	181		

A similar post-hoc comparison revealed that the time taken to complete task 5 were significantly different for the IBM Rational RequisitePro and TRIC groups, p=0,011. The TRIC group performs change impact analysis 50% slower than the IBM Rational RequisitePro group.

No other combination of groups yielded a significant difference in times results in the post-hoc test, including task 2.

The attained statistical power for the $\chi 2$ tests is 52% for detecting effects with a large size, $p < 0,05$, sample size 21 and two degrees of freedom. The critical $\chi 2$ value to attain a significant result is 5,99. To attain a statistical power of 80% a sample size of 39 would be required. The critical $\chi 2$ value to attain a significant result would remain 5,99. This is calculated by using G*Power 3 tool.

E. Analysis of Covariance

Analysis of covariance is an extension of analysis of variance that explores differences between groups while statistically controlling for covariates [20]. As an extension of analysis of variance, it can only be used for tasks 1 and 3 for which the initial assumptions are met.

The set of covariates should be sufficiently reliable to perform an analysis of covariance. Cronbach's alpha is an indicator of internal consistency and can be used to measure this reliability. A sufficient level of reliability as measured by Cronbach's alpha is 0,7 or above [20]. However, Cronbach's alpha for the covariates in this experiment is only 0,310 which indicates poor reliability. Attempts to eliminate one or more weak covariables resulted in a Cronbach's alpha of 0,585, which is too low to warrant an analysis of covariance and was therefore not executed.

F. Multivariate Analysis of Variance

Multivariate analysis of variance is an extension of analysis of variance when there is more than one dependent variable such as is the case with the F-score and time. The advantage of performing multivariate analyses of variance over performing separate one-way analyses of variance is that the risk of false positives is reduced [20].

An assessment of the linearity of *F*-scores and times using a Pearson product-moment correlation calculation revealed no linearity. Transformation strategies in an attempt to attain linearity over a skewed data set did not yield linearity. A multivariate analysis of variance was therefore not warranted or executed.

VII. INTERPRETATION OF RESULTS AND THREATS TO VALIDITY

In this section we interpret the findings from the analysis presented in Section VI.

A. Change Scenario Representativeness

Not all change scenarios were judged to be representative. This is both a reliability problem and a threat to internal validity. This research attempts to reflect the real world yet does not fully have real-world change scenarios.

As we depict in Section VII.B, the golden standards are very reliable. This can only be true if the change scenarios have a low level of ambiguity. This partly offsets the low representativeness. Although the change scenarios may not reflect the real world, they can still be well understood and applied to the WASP specification.

B. Golden Standard Reliability

Statistical testing for tasks 1 up to and including 4 did not reveal any significant difference between the golden standards and suggested excellent inter-rater reliability.

Statistical testing for task 5 indicates a statistically significant difference between the golden standards. However, the more precise intraclass correlation score does suggest good inter-rater reliability. The high inter-rater reliability means that the design of the tasks is feasible. Had they been too ambiguous, then it would have been likely that the inter-rater reliability would have been much lower.

C. One-way between-groups ANOVA

The quality of change impact predictions is not impacted by the tool that is being used for tasks 1 or 3. A similar conclusion can be drawn about the time taken to complete task 3. The time taken to complete task 1 is significantly different for the group that used TRIC. They performed change impact prediction of scenario 1 slower than the other groups.

D. Non-parametric Testing

The quality of change impact predictions is not impacted by the software tool that is being used for tasks 2, 4 or 5.

The time taken to complete tasks 4 and 5, who respectively remove a property and remove a requirement, are significantly different for the group using TRIC. For task 4, the TRIC group was slower than the Microsoft Excel group. For task 5, the TRIC group was slower than the IBM RequisitePro group.

The time taken to complete task 2 was indicated to be significantly different for the group using TRIC by the $\chi 2$ test, but an ensuing post-hoc comparison using a Mann-Whitney U test indicated that this result is a false positive, likely caused by a small sample size [4].

E. Analysis of Covariance

The reliability of the covariates was too low to conduct an analysis of variance. Of the strongest covariates, the first three

somehow measure the same construct. The completion of a basic requirements engineering course, completion of an advanced requirements engineering course, and months of experience, are in fact all a measure of experience with requirements management. Statistical testing detects correlations amongst these variables of medium effect size.

F. Multivariate Analysis of Variance

The assumption of linearity between the *F*-score of change impact predictions and time taken to complete them was violated, because of which a multivariate analysis of variance could not be executed. One hypothesis to explain the longer time taken yet equal *F*-score of the TRIC group is that TRIC is a more complex tool. It offers more visualization opportunities and is not as mature as the other software tools. If the benefits of TRIC are to better cope with complexity, then those may only be reaped with an appropriately complex software requirements specification.

G. Validity Evaluation

Statistical Conclusion Validity. Our research features a limited sample set. A larger sample of research objects is required for statistically valid conclusions. The observed power, required sample size for proper power and estimated error are calculated as part of the analysis.

Internal Validity. The setup of the lectures is not any fairer by assigning equal slots of time. While an equal amount of time is given to all groups for the lectures, the complexity of the tools is different. As an example, TRIC and the relation types will take more time to learn than Microsoft Excel (which is probably already known). By compressing more required knowledge into a shorter timeframe, the intensity of the lecture decreases and participants cannot be expected to understand the software tools equally well.

Construct Validity. The number of constructs and methods that are used to measure the quality of change impact prediction is monogamous; only the *F*-score is truly a measure of "product" quality, with the time taken being more of a measure of "process" quality. This may under-represent the construct of interest, complicate inferences and mix measurements of the construct with measurement of the method [21]. This experiment is subject to Hawthorne effects [21] because of participants responding differently to experimental conditions.

External Validity. Inferences are valid only as they pertain to the WASP requirements specification and the specific participants. Participants may not represent real-world requirements engineers. Finally, the instructors are three different people that may not have equal instructing aptitude.

VIII. Conclusions

The background for this research was to evaluate the impact of TRIC, a software tool that supports formal requirements relation types, on the quality of change impact predictions. It was hypothesized that using TRIC would positively impact that quality. A quasi-experiment was systematically designed and executed to empirically validate this impact.

A. Results

The results of this specific experiment do not provide a positive solution validation of TRIC. The following conclusions can be drawn with respect to the combination of participants, change scenarios and software requirements specification that were used in this experiment:

- Null hypothesis 1 stated that the F-scores of change impact predictions of requirements engineers using TRIC will be equal to or less than those from requirements engineers. Null hypothesis 1 cannot be rejected.
- Null hypothesis 2 stated that the time taken to complete change impact predictions of requirements engineers using TRIC will be equal to or longer than those from requirements engineers not using TRIC. Null hypothesis 2 cannot be rejected.

No differences in the quality of change impact predictions between using Microsoft Excel, IBM Rational RequisitePro and TRIC were detected. TRIC was detected to lead to slower change impact prediction. The mean statistical power of the tests underlying these conclusions is 54%.

Covariate reliability testing further suggested that experience with requirements management is the most covariate of all covariates, although the way it was constructed in this experiment is not reliable enough to explain any variance in F-scores or time taken to complete change impact predictions.

B. Limitations

The results of this research are subject to the following limitations:

- **Lack of control over lecture effect**. Participants require training to work with the software tools and play the role of expert. This is difficult to do reliably. First, the setup of the lecture is not fair because the same time is allotted for all three software tools, although RequisitePro and TRIC require more tutoring than Excel. Second, a reliable pre-test and post-test to measure software tool aptitude and the learning effect of the lecture is not available. The same problem is known in marketing, where there are no existing consumers of a new product. In Kotler's eight-step process of new product development [18], it is suggested that concept testing is performed with focus groups. A focus group is defined to be a small sample of typical consumers under the direction of a group leader who elicits their reaction to a stimulus such as an ad or product concept. They are one form of exploratory research that seeks to uncover new or hidden features of markets and can help solve problems. However, focus groups usually suffer from small sample sizes, limited generalizability and Hawthorne effects [18]. The problem-solving and exploratory approaches match that of action research, which seems a more plausible way of validating new software tools, though that is subject to the same challenges as focus group research [24].

- **Low participant representativeness**. There is no strong evidence to assume that master students are representative for actual requirements engineers. Although an argument can be made that a sampling of 21 master students in Computer Science and Business Information Technology can be representative for their larger population, the data set contained a sizable number of outliers for which there were no grounds for data set reduction. The experiment should be repeated with different participants to assert external validity.
- **Lack of control over change scenarios**. This research instructs participants to perform change impact prediction on a set of change scenarios. It is likely that change scenarios have influence on the results of change impact predictions, but the lack of theory surrounding change scenarios is a cause of reliability problems. Second, some students raised questions about the rationales in the change scenarios, which may have induced guessing. This limitation is partially offset by the high inter-reliability scores of the golden standards, which indicate that a group of experts interpret the change scenarios reliably and proves the usability of the experimental design if enough experts were available.
- **Small sample size**. The sample size of the research is too small to attain the generally accepted statistical power of 80%. Instead, the statistical power is 56% for the analyses of variance and 52% for the non-parametric tests. If the statistical power increases, then inferences can be made with greater confidence and smaller effects could be detected.
- **Limited comparability of software tools**. No statistical adjustments have been made for the functionality, maturity and usability of Microsoft Excel, IBM Rational RequisitePro and TRIC. Even though they all feature a traceability matrix, other tools may produce different results. Inferences can only be made with regard to these three tools.
- **Monogamous metrics**. By only using the F-score, it is possible that the quality of predictions is not measured fully and that the measurement of quality is mixed with measurement of the metric. Having more measures of quality would improve the reliability of the results.
- **Low participant reliability**. First, not all participants were as focused on the task as was expected. Second, many were under pressure to complete the experiment. Third, some participants did not check the initially changed requirement as part of their Estimated Impact Set, even though they were instructed to do so both during the lecture and by the web application. This may have led to suboptimal change impact predictions. Using experts instead of master students is not certain to produce more reliable results, because interviews have indicated that the effort of experts also depends on their stake in the project. However, shorter experiments will produce more reliable results [8].
- **Limited research object representativeness**. Specifications other than the WASP specification used can have different complexity in terms of length, structure, ambiguity, completeness and possibly other metrics which were not discussed here. This can influence the impact of using different software tools on the quality of change impact predictions. For example, an intelligent tool such as TRIC is likely to only show its benefits when tasked with a complex software requirements specification. The experiment should be repeated with a diverse set of specifications to evaluate the influence of these attributes.

C. Lessons Learned and Future Work

The experiment indicates some of the challenges in the validation of academic tools. A newly developed tool does not have an initial community of users. Usually the researchers that developed the tool are the only users and experts. Building a community of trained and experienced users take efforts and time. In some cases this may span a period of 2-3 years and is the usual duration for a PhD study. In such cases certain forms of empirical validation of the tool may be infeasible since not enough well trained participants are available.

In this experiment we did not ask the participants about their personal experience during the experiment. This information could be useful and may give insight about how the participants perceived the work process. A simple questionnaire filled-in after the experiment is a suitable instrument for collecting this information.

We hypothesized that the lack of a positive solution validation in this research can be attributed to the fact that TRIC is a more intelligent software tool and its benefits will only materialize for a sufficiently complex software requirements specification.

We expected that the presence of an explicit type and meaning of requirements relations facilitates the change impact prediction. However, this may not always be the case. Regardless which tool is used, the participants have to identify and interpret the relation. It is the proper understanding of the relation that would ultimately improve the change prediction. The identification of the relation type in TRIC may not bring immediate benefits since there were no guidelines on how to determine the change propagation on the basis of the type of the change and the relation type.

The following can be recommended to further pursue the solution validation:

- Study the state-of-the-art in change scenario theory, so that it is clear how a certain change scenario can impact change impact prediction. Much theory exists on change impact prediction, but not on the elements of change scenarios themselves. The research should be focused on real-world practice, admitting that most real-world changes will not comply to a yet to be determined academic standard. This is required to complete the necessary body of knowledge to setup a controlled experiment.
- Create multiple change scenarios of the same class. This research used an improvised classification according to the type of requirements change in terms of its parts.

The effect of this classification could not be tested because only one class of change scenarios was represented twice.

- Find a number of real-world software requirements specifications of high complexity. As with change scenario theory, there is no generally accepted criterion for what constitutes complexity, although raw indices such as page count, requirements count and tree impurity will provide a strong argument. If these specifications cannot be collected from the QuadREAD Project partners, then it is worthwhile asking governmental institutions to participate in academic research, possibly under non-disclosure agreement.
- Consider organizing an online experiment, where experts can participate from behind their own computer. This allows more time for experimentation, because the experiment can be split up into several time slots which can stretch multiple days. It also lowers the barrier to entry to participate. Given a large enough sample size, the lack of environmental control will be corrected for by randomization.
- Consider organizing multiple action research projects, where researchers can apply techniques in practical cases currently running with clients. As a precondition, it should be accepted that action research is cyclical and that TRIC must evolve as part of the cases. Give a large enough amount of action research iterations, a strong argument for generalizability may be found.
- Extend TRIC with the ability to suggest possible changes based on the semantics of the requirements relations and changes. Such an extension fully utilizes the formal semantics of the relations. This is a recently implemented feature of the tool [12] and leads to a new experiment with new hypothesis and design.

A recommendation for future work is to research the impact of classes of software tools with the same intelligence on the quality of change impact predictions. This research requires the creation of a classification scheme for levels of software tool intelligence.

ACKNOWLEDGMENT

The work has been supported by NWO (www.nwo.nl) in the Jacquard Programme and by the National Research Fund, Luxembourg (FNR/P10/03).

REFERENCES

[1] B. J. M. Abma, "Evaluation of requirements management tools with support for traceability-based change impact analysis," Master's thesis, University of Twente, Enschede, 2009.

[2] R. S. Arnold, S. A. Bohner, "Impact analysis - towards a framework for comparison," in ICSM, pp. 292-301, 1993.

[3] J. Cohen, "Statistical power analysis for the behavioral sciences," Hillsdale, New Jersey: Erlbaum, 1988.

[4] B. Dawson, R. G. Trapp, "Basic & clinical biostatistics," McGraw-Hill, 2004.

[5] R. S. A. van Domburg, "Empirical evaluation of change impact predictions using a requirements management tool with formal relation types: a quasi-experiment," Master Thesis, University of Twente, Enschede, the Netherlands, 2009.

[6] P. Ebben, "Requirements for the WASP application platform," in WASP/D2.1., Telematica Instituut, 2002.

[7] T. Fawcett, "ROC graphs: notes and practical considerations for researchers," Technical report, HP Laboratories, 2004.

[8] C. T. Fitz-Gibbon, L. Vincent, "Difficulties regarding subject difficulties: developing reasonable explanations for observable data," Oxford Review of Education, 23(3): 291-298, 1997.

[9] D. Garson, "Quantitative Research in Public Administration," from: http://faculty.chass.ncsu.edu/garson/PA765/index.htm

[10] A. Goknil, "Traceability of requirements and software architecture for change management," PhD Thesis, University of Twente, Enschede, the Netherlands, 2011.

[11] A. Goknil, I. Kurtev, K. van den Berg, and J. W. Veldhuis, "Semantics of trace relations in requirements models for consistency checking and inferencing," Software and Systems Modeling, 10(1), 31-54, 2011.

[12] A. Goknil, I. Kurtev, K. van den Berg, and W. Spijkerman, "Change impact analysis for requirements: A metamodeling approach," Information and Software Technology, 56(8): 950-972, 2014.

[13] G*Power 3. http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

[14] IBM RequisitePro. http://www-01.ibm.com/software/awdtools/reqpro/

[15] A. Jedlitschka, D. Pfahl, "Reporting guidelines for controlled experiments in software engineering," in ISESE 2005, pp. 95-104, Australia, 2005.

[16] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. K. Sjøberg, "A systematic review of quasi-experiments in software engineering," Information and Software Technology. 51: 71-82, 2009.

[17] S. Katz, J. Aronis, D. Allbritton, C. Wilson, and M. L. Soffa, "Gender and race in predicting achievement in computer science," IEEE Technology and Society Magazine, 22(3), 2003.

[18] P. Kotler, G. Armstrong, "Principles of marketing," Pearson, 2008.

[19] A. Mockus, S. G. Eick, T. L. Graves, and A. F. Karr, "On measurement and analysis of software changes," Technical report, National Institute of Statistical Sciences, 1999.

[20] J. Pallant, "SPSS survival manual: a step by step guide to data analysis using SPSS for windows," McGraw-Hill, 2001.

[21] W. R. Shadish, T. D. Cook, and D. T. Campbell, "Experimental and quasi-experimental designs for generalized causal inference," Houghton Mifflin Company, 2002.

[22] D. I. K. Sjøberg, B. Anda, E. Arisholm, T. Dyba, M. Jorgensen, A. Karahasanovic, E. F. Koren, and M. Vokac, "Conducting realistic experiments in software engineering," in ISESE, pp.17-26, Nara, Japan, 2002.

[23] TRIC. http://trese.cs.utwente.nl/tric/

[24] R. K. Yin, "Case study research: design and methods," Applied Social Research Methods, SAGE Publications, 2009.

[25] R. M. Wolf, "The validity and reliability of outcome measure," in Monitoring the Standards of Education, 1994.