Evaluation of BehaviorMap: *a User-Centered Behavior Language*

Fernando Wanderley
Antonio Silva
Joao Araujo

ROIS 2015

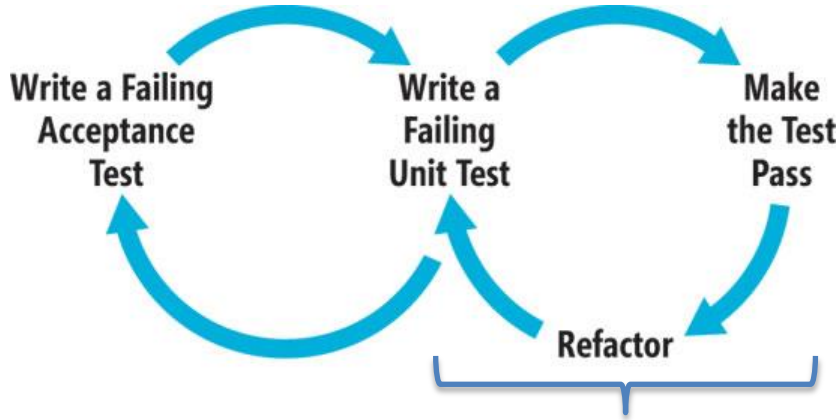
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA


I'M OUT OF HERE GUYS
CONTEXT

Context

End-User

Behavior Driven Development



Test Driven Development Cycle

jbehave

RSpec

phpspec

CppSpec

N unit

approaches are textual based and do not help domain experts understanding

And the test coverage is not easy.

Behavior Scenario Example

Feature: Map View
Scenario: Show Vessel inside Map Area

Behavior Scenario Example

```
1  Feature: Map View
2  Narrative:
3  In order to asses the situation in my area
4  As a coast guard
5  I want to see the location of each vessel marked on a map
6
7  Scenario: show vessel inside map area
8  Given vessel "Seal" at position "52.01N, 3.99E"
9  When I view the map area between "52.10N, 3.90E" and "51.90N, 4.10E"
10 Then I should see vessel "Seal" at position "52.01N, 3.99E"
```

JBehave

Behavior Scenario Template

Entities

Entities States *(before, after)*

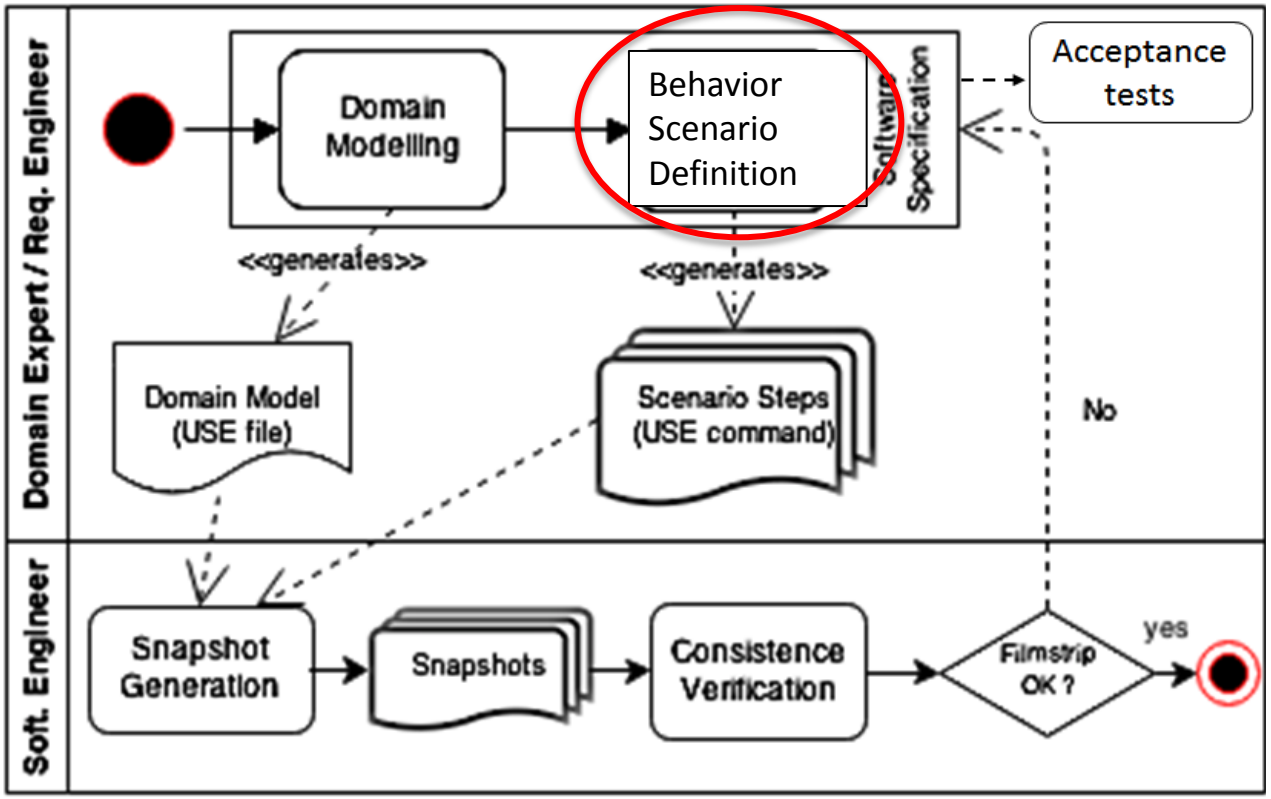
Motivation

- The use of natural language to specify requirements can convey ambiguities and loss of information when the development team reads the behavior specifications provided;

- To address these issues we designed the diagrammatic language BehaviorMap to improve cognitive aspects of BDD, through the cognitive properties of a Mind Map;

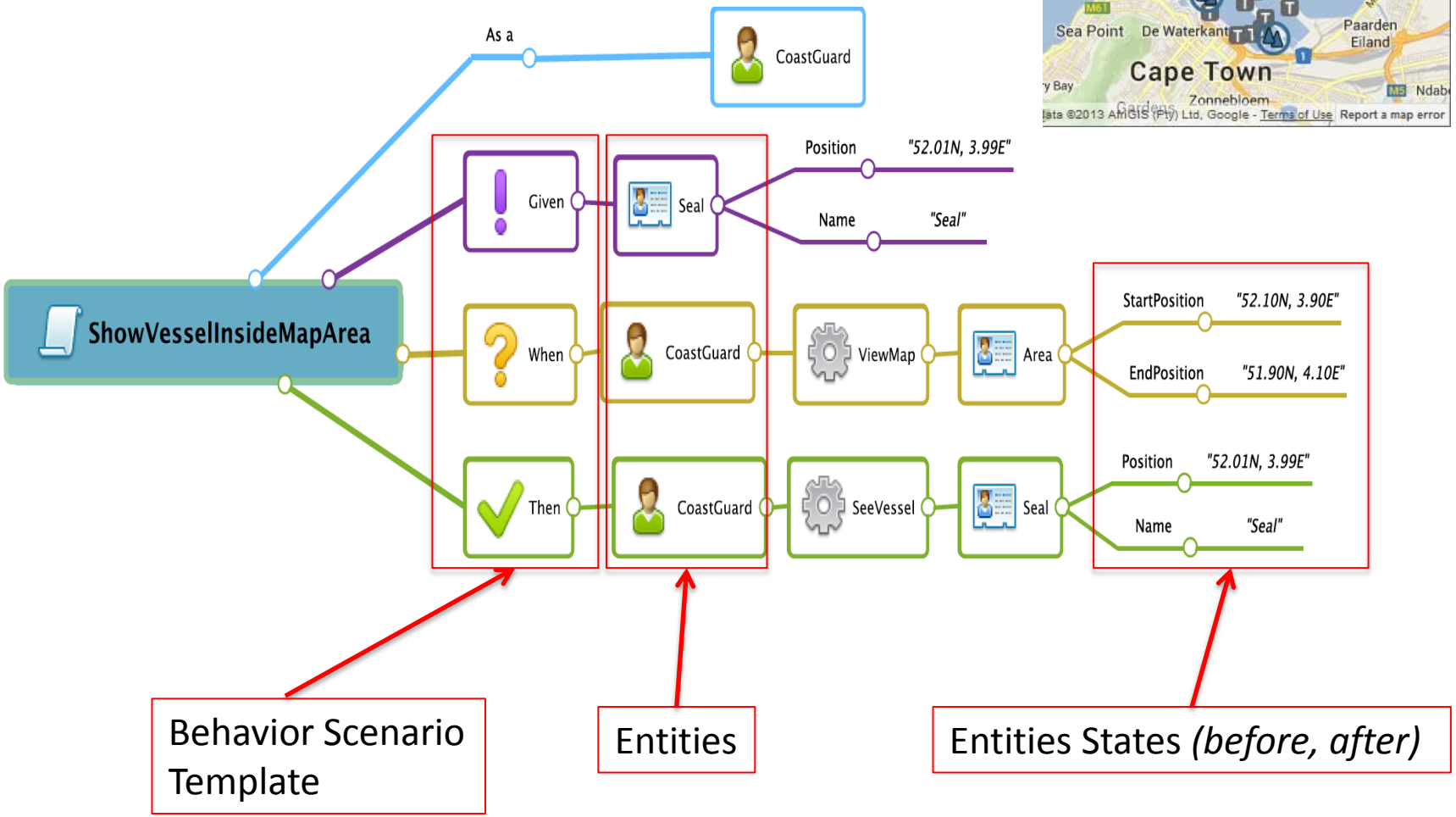
- The language belongs to a framework called SnapMind, that is composed of tools to specify both domain and behavioral models.

SnapMind Framework



[MoDre @RE'14]

BehaviorMap Example



Goals

- Our hypothesis is that by using mind map in requirements models, since it is a user-centered diagram, stakeholders will understand requirements more easily and consequently more engaged;

- To test our hypothesis, we produced an initial experimental evaluation to assess the cognitive effort of understanding BehaviorMap's and textual scenarios and;

- We used questionnaires with questions about the scenarios to measure the cognitive effort. The time effort also measured and after this tasks test coverage was performed automatically.

Experiment Design

Training Task → Practical Tasks → Comprehension Tasks

15 naïve-users
(10 of them non IT)

Training Task

- The training task lasting 30 minutes maximum, aimed to explain to the participants the elements of the experiment;

- Initially, it was explained to the participant that the scenarios were used to represent behaviour and address two different types: textual and graphical;

- The tools, JBehave (the current tool used by industry) and BehaviorMap, were explained.

Comprehension Task

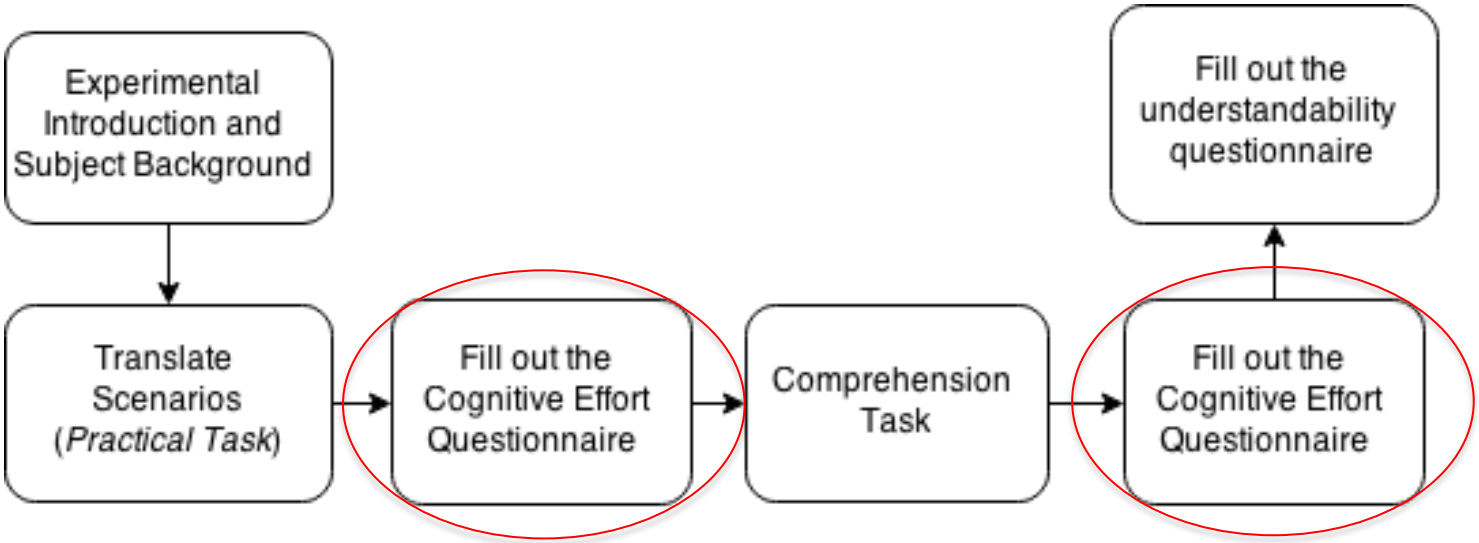
- The comprehension tasks served to assess how participants understood the graphical and textual models by answering questions about them;
  - What are the initial conditions expected?
  - What are the actions to be specified?
  - What is the expected result?

Selected Scenarios

- The metrics conceived were:
  - (i) *Scenario Size* to count the leafs in scenario branches (Given, When, Then)
  - (ii) number of Actions in *When* branch (*ActionsWhen*)
  - (iii) number Actions in *Then* branch (*ActionsThen*) and
  - (iv) Distinct Entities (*Entities*)

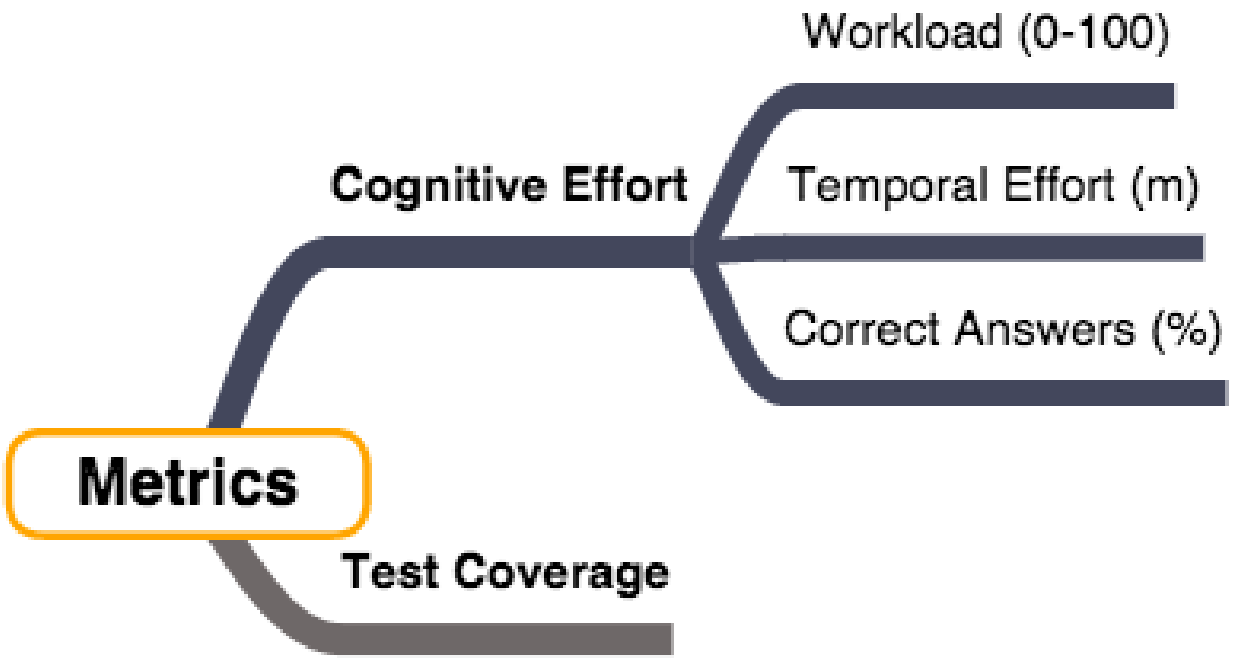
Scenario Description	*Entities*	*Scenario Size*	*Actions When*	*Actions Then*
Textual High	8	28	1	1
Graphical High	6	24	6	0
Textual Medium	2	6	1	1
Graphical Medium	2	6	1	1
Textual Low	1	3	1	0
Graphical Low	1	3	1	0
Training BM	1	4	1	0
Training Textual	2	5	1	2

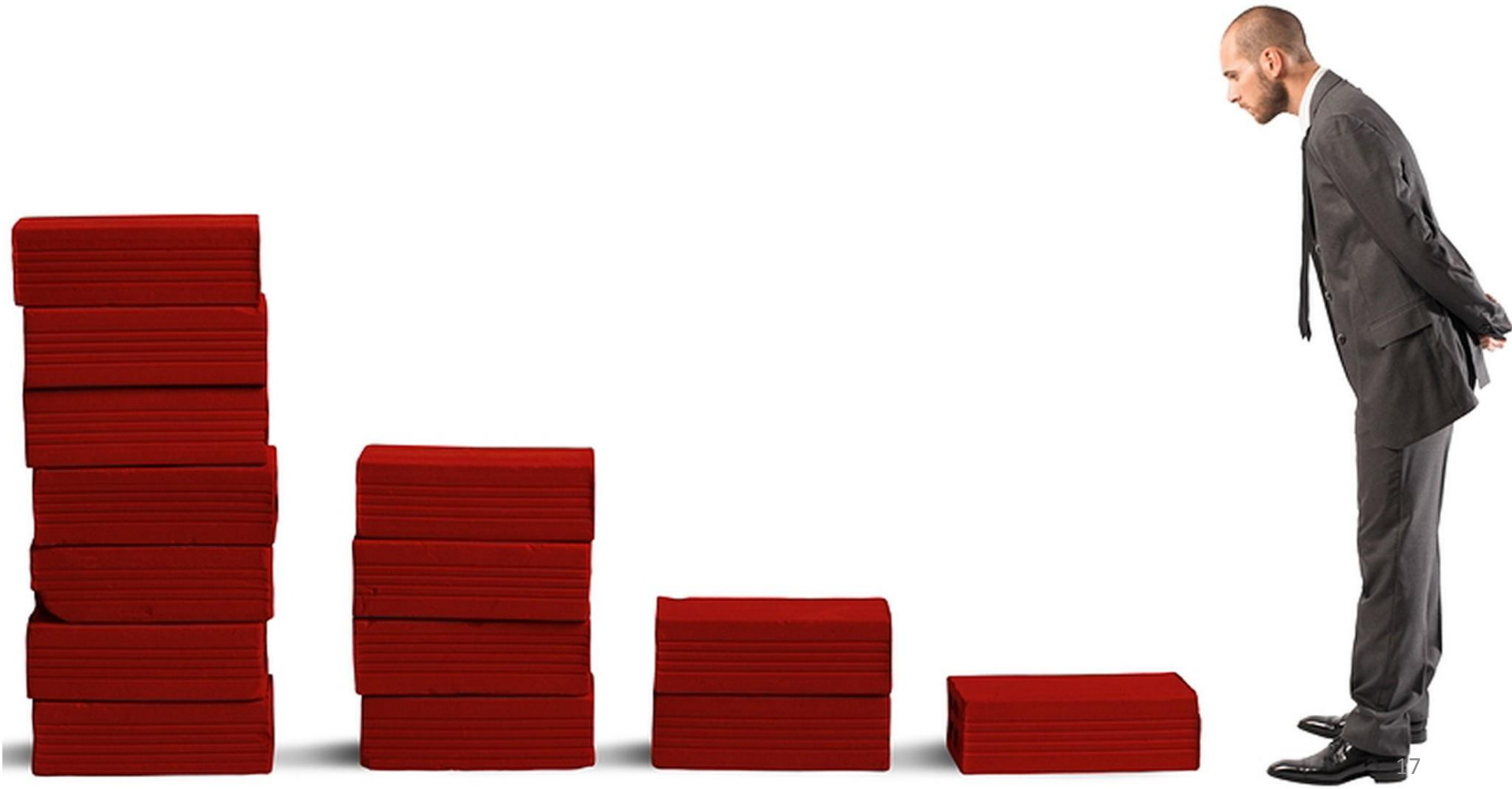
Experimental Process



- The questionnaire used was the NASA-TLX, a questionnaire quite used to assess the cognitive effort of a person performing tasks;
- He made a self-evaluation concerning the attributes of performance, mental effort, temporal effort, physical effort, level of frustration, etc.

Measurement





Results

- The data were compared using a nonparametric analysis of variance using the Kruskal-Wallis and Mann-Whitney methods ;

- The Anderson-Darling → did not have a normal distribution *(with a confidence level of 99%);*

- Analyses of variance were performed to consider two factors:
  - (i) the way the BDD scene was written (textually or graphically) and
  - (ii) its complexity level to see if there are significant differences with changing complexity;

- All results were obtained with a confidence level of 95%.

Graphical to Textual: *Better results*

- In practical tasks, the goal was to make a translation from textual to chart and vice versa;

Measurement	*Map → Text*		*Text → Map*	
	Mean	**Stdv**	**Mean**	**Stdv**
Workload (0-100)	26.26	18.17	50.70	25.00
Temporal Effort (m)	4.29	1.31	6.11	1.13
Correct Answers (%)	90.77	17.00	50.77	21.77

- An analysis of variance (Mann-Whitney) was applied → differences between means;

- The result (U = 174.0, P = 0.01) showed that the average difference in workload between tasks was not caused by random events, but rather the difference between the types of scenarios;

Comprehension Results

- In the comprehension tasks, the participants had to answer three questions for each model, in a time slot of five minutes;

Measurement	*Graphic*		*Textual*	
	Mean	**Stdv**	**Mean**	**Stdv**
Workload (0-100)	24.68	21.75	34.54	28.17
Time (m)	1.64	1.18	2.09	1.29
Correct Answers (%)	96.77	16.30	84.36	31.53

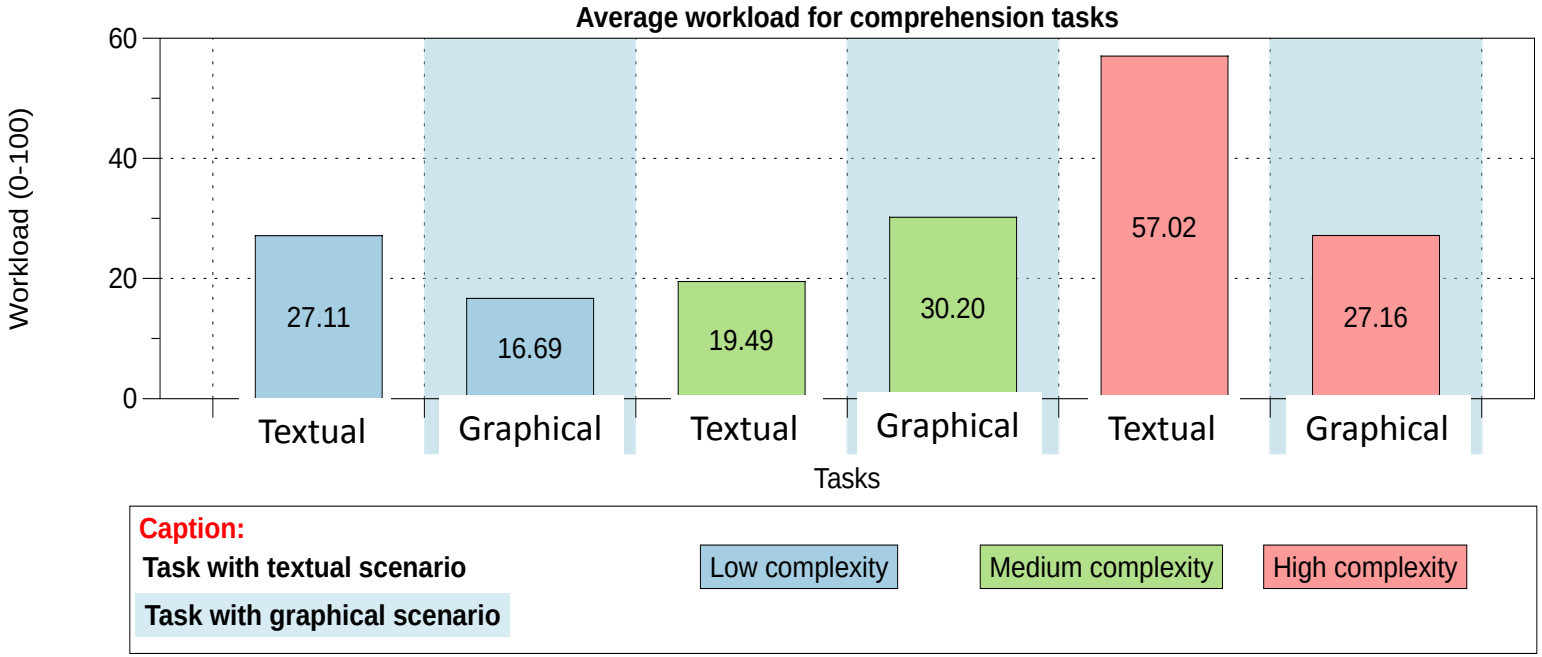
- Tasks using graphics behavioral models had lower workload and time effort and got more correct answers.

- For all the measurements, two-variance analyses were performed;

→ **One analysis fixing the scenario type and varying the complexity class (AV-I), and other fixing the complexity class and varying the scenario type (AV-II)**

Cognitive Effort

- The AV-I analysis showed that the graphical scenarios (H = 12.48, p = 0.0019) had no significant difference between them according to the changing of complexities (confidence of 95%).



Cognitive Effort

- Regarding textual scenarios (H = 3.25, *p* = 0.1969), the same did not happen.

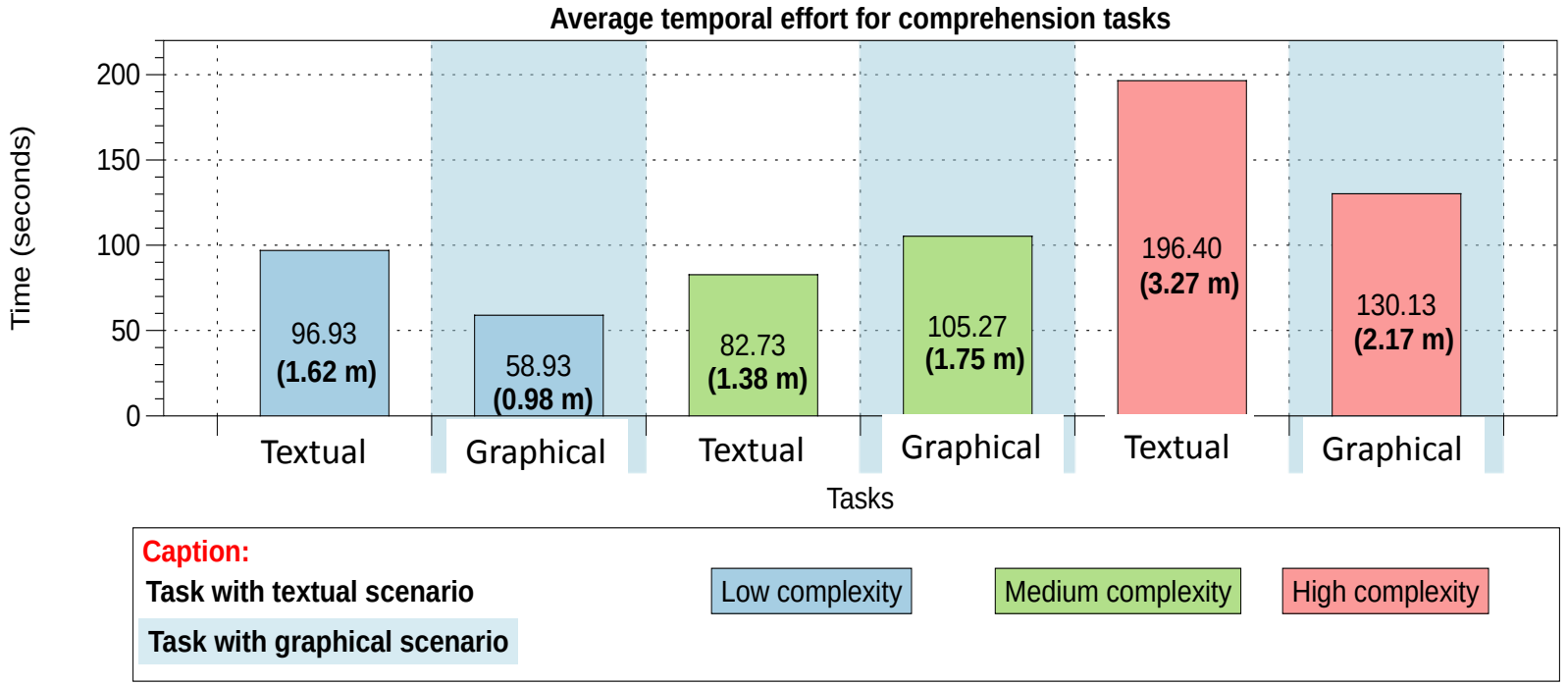
Comparing	U, *p*	Conclusion ($\alpha = 0.05$)
Low vs. Medium	133.0, *0.40*	No significant diff.
Medium vs. High	192.0, *0.00*	**Significant difference**
Low vs. High	49.0, *0.01*	**Significant difference**

- In the AV-II analysis it was found that for high complexity scenarios, the scenario type impacts the workload.

Comparing (scenario type)	U, *p*	Conclusion ($\alpha = 0.05$)
Low	78.0, *0.15*	No significant difference
Medium	141.5, *0.23*	No significant difference
High	50.0, *0.01*	**The scenario type impacts the workload**

Time Effort

- Regarding the AV-I analysis, the results for textual scenarios (H = 18.88, p < 0.0001) and for graphical scenarios (H = 11.67, p = 0.0029) showed that the complexity influenced both of them.



Time Effort

- We can conclude that the tasks with high complexity affected the time effort compared to the other two levels of complexity as expected

Comparing	U, *p*	Conclusion ($\alpha = 0.05$)
Low vs. Medium	105.5, *0.77*	No significant diff.
Medium vs. High	18.0, *0.00*	**Significant difference**
Low vs. High	198.0, *0.00*	**Significant difference**

- The results show that low complexity level has significant lower values than the middle and high levels.

Comparing	U, *p*	Conclusion ($\alpha = 0.05$)
Low vs. Medium	162.5, *0.04*	**Significant difference**
Medium vs. High	145.5, *0.17*	No significant diff.
Low vs. High	193.0, *0.00*	**Significant difference**

Time Effort

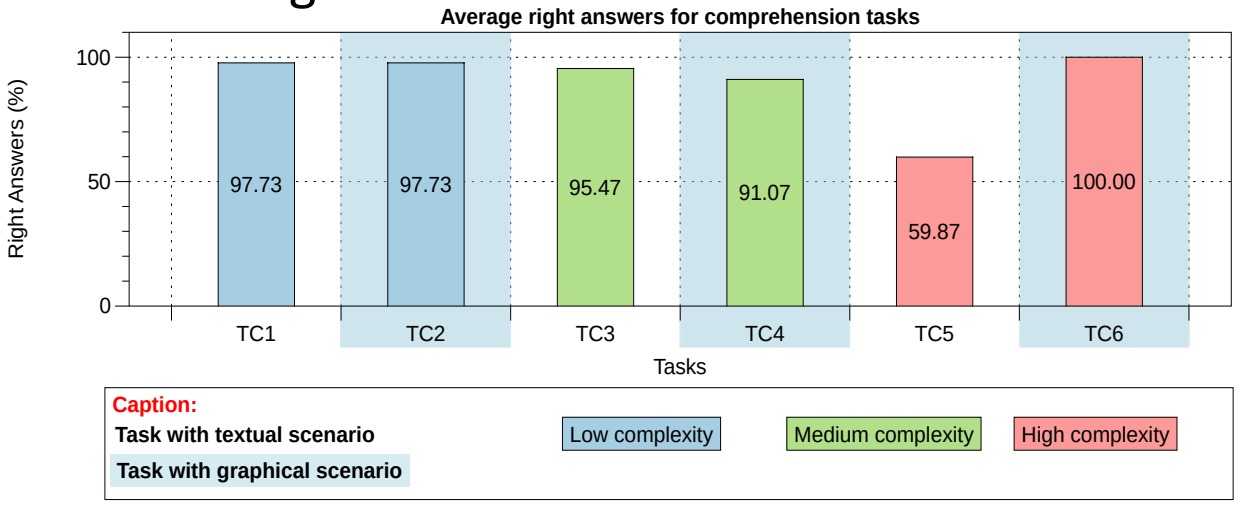
- In the AV-II analysis for the time effort, the results showed that the differences of means in the lower and higher complexity levels were not caused by random factors, but by the difference in the type of scenarios.

Comparing (scenario type)	U, *p*	Conclusion ($\alpha = 0.05$)
Low	65.0, *0.05*	**The scenario type impacts the temporal effort**
Medium	129.0, *0.49*	No significant difference
High	56.5, *0.02*	**The scenario type impacts the temporal effort**

Correct Answers

- The AV-I analysis showed that the graphical scenarios (H = 6.65, p = 0.036) had no significant difference between them according to the changing of complexities.

- Regarding textual scenarios (H = 0.4, p = 0.8187), the same did not happen. We concluded that the high complexity level affected the differences in averages

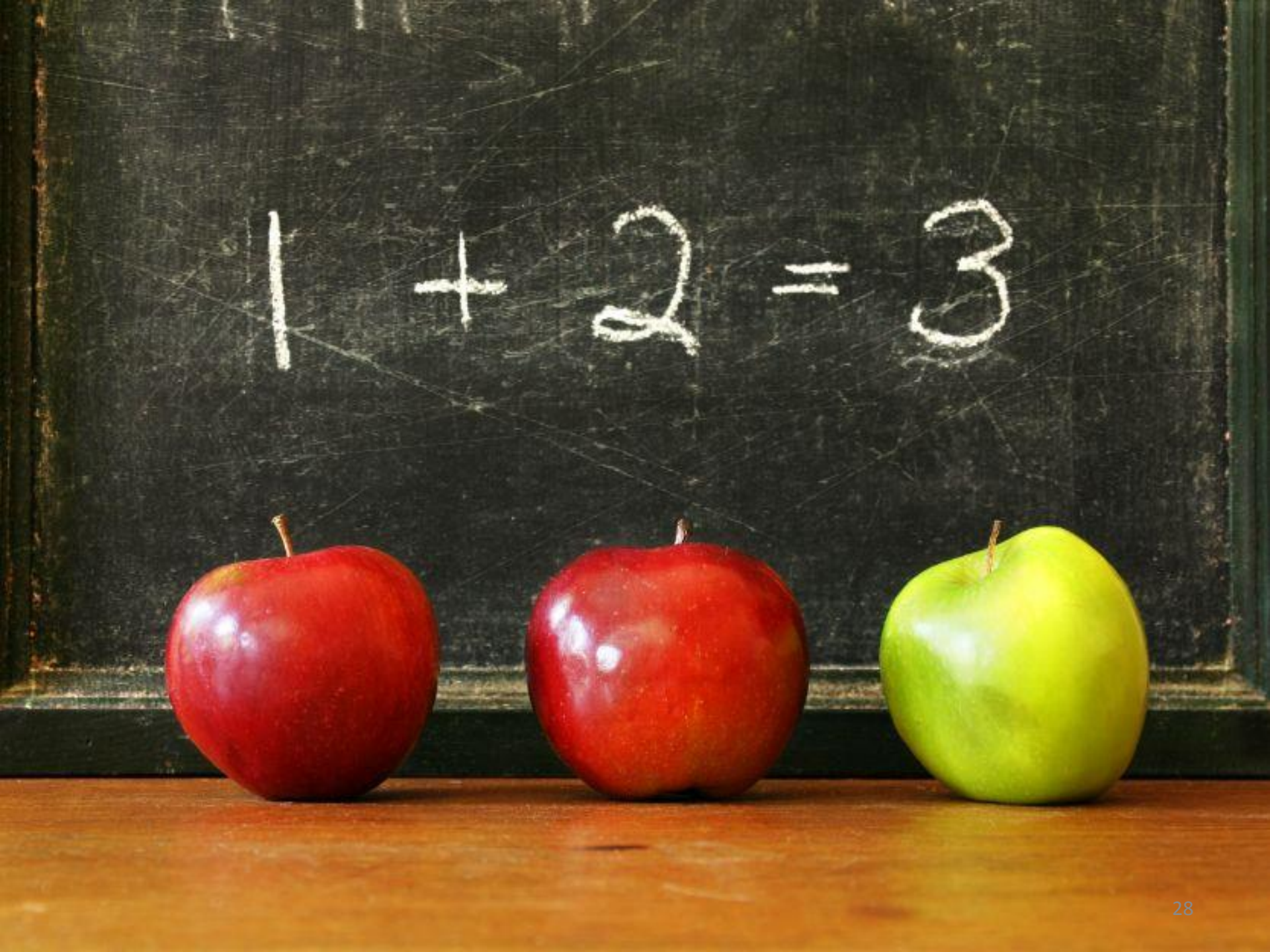


Correct Answers

- The AV-II analysis showed the high complexity level there were significant differences to affirm that type of scenario influenced the recorded responses.

Comparing	U, *p*	Conclusion ($\alpha = 0.05$)
Low vs. Medium	105.0, *0.76*	No significant diff.
Medium vs. High	163.5, 0.03	**Significant difference**
Low vs. High	168.0, 0.02	**Significant difference**

Comparing (scenario type)	U, *p*	Conclusion ($\alpha = 0.05$)
Low	112.5, 1.00	No significant difference
Medium	111.5, 0.97	No significant difference
High	172.5, 0.01	**The scenario type impacts the answers**


1 + 2 = 3

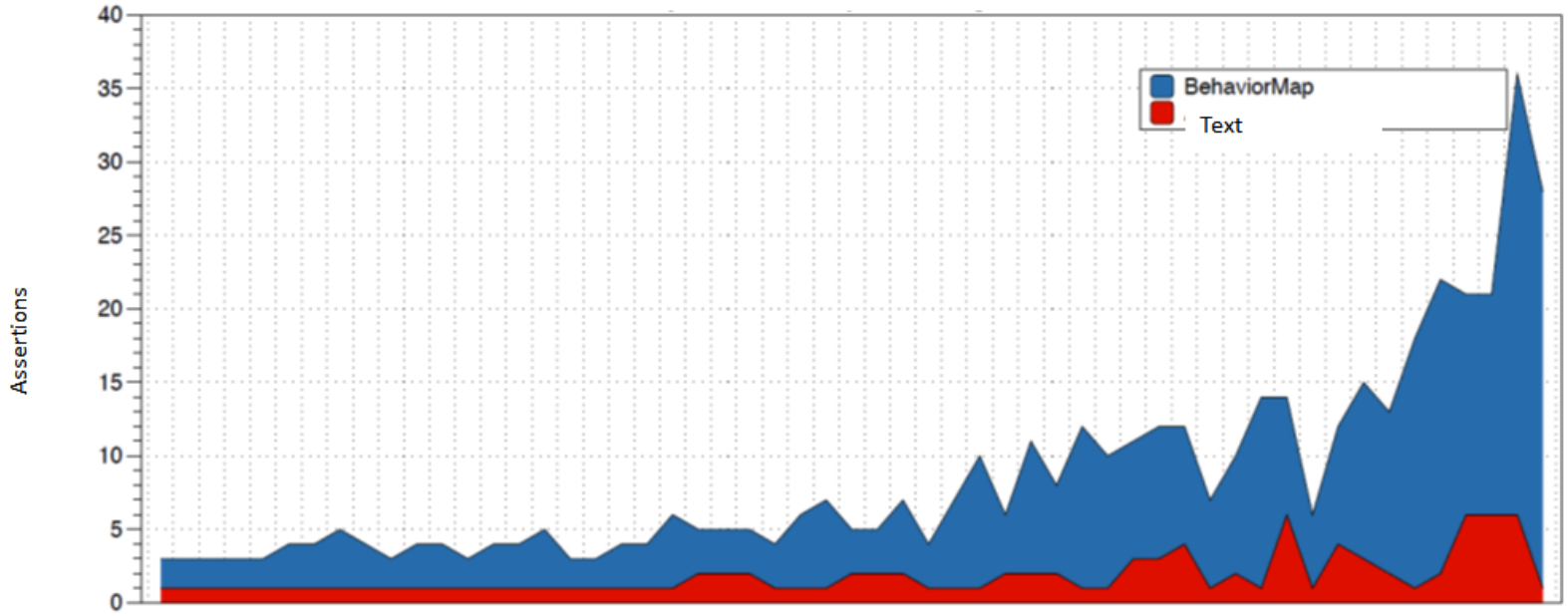
Summary for Scenarios

Measurement	*Scenario*	*Result*
Workload	Textual	More effort required in high complexity class than other classes
	Graphical	**No significant differences were recorded**
Temporal Effort	Textual	More effort required in high complexity class than other classes
	Graphical	**Less effort in low complexity class than other classes**
Correct Answers	Textual	More effort required in high complexity class than others
	Graphical	**No significant differences were recorded**

Summary for Complexities

Measurement	*Complexity*	*Result*
Workload	Low	No significant differences
	Medium	No significant differences
	High	**With significant differences (Better for BehaviorMap)**
Temporal Effort	Low	**With significant differences (Better for BehaviorMap)**
	Medium	No significant differences
	High	**With significant differences (Better for BehaviorMap)**
Correct Answers	Low	No significant differences
	Medium	No significant differences
	High	**With significant differences (Better for BehaviorMap)**

Test Coverage



- In order to verify this premise were collected a sample of 53 textual scenarios from multiple sources (academic and industry) and translate them to BehaviorMap scenarios;

- It shows that the BehaviorMap approach provides more test cases without increasing the effort time of the users, as the tests are automatically created

Conclusions

- This first experiment showed some evidence that BehaviorMap scenarios are easier to understand in relation to textual scenarios, especially when considering scenarios with higher complexity;

- Namely, in practical tasks, the results showed it was clearer and easier to translate correctly a graphical scenario to text than a textual translation to graphical;

- Regarding the results, the BehaviorMap had better results with the increasing of complexities of the scenarios, however, the textual scenarios had good performances in low and medium complexity levels;

- The BehaviorMap is a special cognitive support to Behavior Driven Development in the scenarios specification addressing the end-user understanding

Future Work

- As a clear future work is repeat this evaluation exploring user-centred usability strategies;

- One point of improvement in the experiment will be the use of biometric sensors to enhance the cognitive effort measurement precision;

- Nevertheless, replication of the experiment to substantiate this assessment is needed, with a larger number of participants and other scenarios and;

Evaluation of BehaviorMap: *a User-Centered Behavior Language*

Fernando Wanderley
Antonio Silva
Joao Araujo

ROIS 2015

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA