MASTER'S THESIS

Intuitive Understanding of Domain-Specific Model Languages: Proposition and application of an Evaluation Technique between two grammars

van der Kooij, M.

Award date:
2021

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us at:

pure-support@ou.nl

providing details and we will investigate your claim.

Downloaded from https://research.ou.nl/ on date: 16. May. 2025

Open Universiteit
www.ou.nl

Intuitive Understanding of Domain-Specific Model Languages: Proposition and application of an Evaluation Technique between two grammars.

Degree program:	Open University of the Netherlands, Faculty of Science Master of Science Business Process Management & IT
Course:	IM9806 Business Process Management and IT Graduation Assignment
Student:	Martin van der Kooij
Identification number:	
Date:	21 June 2021
Thesis supervisor	Ben Roelens
Second reader	Lianne Cuijpers
Third assessor	not applicable
Version number:	2.0
Status:	AF Final

Abstract

The strategic goals of an organization need to be (1) aligned with supporting business processes, and (2) communicated to their stakeholders intuitively. Enterprise Architecture (EA) focuses on these aspects. Conceptual modelling languages are used to describe the EA of organizations. One stream that can be distinguished is the domain specific modelling language (DSML). The designer of a DSML should aim for computational offloading by replacing cognitive tasks by perceptual ones. An intuitive notation can account for computational offloading, which ultimately leads to an intuitive understanding of the new DSML.
This research aims at developing an empirical evaluation method which can help makers of conceptual models to test how intuitive a new version of a modelling language is in comparison to an older version or a different modelling language which aims at delivering the same message.
In this research a new empirical evaluation method was developed and tested in a governmental setting evaluating the newly proposed and initial version of the Domain specific modelling language Process-Goal Alignment. It provides an answer whether the newly proposed version is more intuitive to understand than the initial version on the proxy indicators for intuitiveness, namely, interpretational effectiveness, interpretational efficiency, and perceived ease of use. The newly proposed version significantly performs better for the overall effectiveness and overall ease of use.

Key terms

conceptual model, domain specific modelling language (DSML), intragrammar, intuitiveness, process goal alignment (PGA), strategic fit

Summary

Conceptual modelling languages evolve over time when used. This leads to new versions and new grammars. Conceptual modelling languages are used to describe the EA of organisations. Conceptual modelling languages can be divided into two. General Purpose Modelling Languages (GPML) and Domain Specific Modelling Languages (DSML). This research focuses on the latter and in particular the Process Goal Alignment notation. The Process-Goal Alignment (PGA) modelling method is a specific DSML which targets business-oriented users as its stakeholder.

The way in which a conceptual modelling language is successful is the way in which it can be used to transform information from the creator of a model to the receiver of a model without having a sustainable cognitive load. While creation often comes with following courses to learn the ins and outs, the reading of models does not. An intuitive notation can account for computational offloading, which ultimately leads to an intuitive understanding of the new DSML. A conceptual model that originated from human creativity from a DSML therefore needs to be intuitive for a receiver to understand. The original version was created with the help of 139 students. For this research, the practical context of a Dutch governmental organization was chosen, and the following research question was formulated:

How can the intuitiveness of the PGA notation be evaluated within a practical context of a governmental organization?

For the literature review relevant papers were searched and found about evaluating the intuitiveness of conceptual modelling languages. This led towards the development of ten hypotheses and an experimental design of my own based on three proxy variables for intuitiveness: interpretational effectiveness, interpretational efficiency and perceived ease of use. Those variables were hierarchically measured on the level of individual icons, on the level of tasks and on an aggregated level of the indicator. This was done to be able to dive deeper into the results.

This experiment was started with 55 possible respondents of which 21 fully finished the experiment. All the (possible) respondents are employees of a Dutch governmental organization, with several

backgrounds (i.e. business, policy, IT, and management). The chosen experimental design is the between-subjects' design of two groups. Strategic sampling helped dividing the participating functions equally between both groups. Group A was given the experiment with questions about the initial version. Group B was given the experiment with questions about the newly proposed version. The conducted experiment was divided into two parts; a demographic part on which the strategic sampling could be performed, and the actual experiment on the PGA notation.

Respondents were presented four tasks in the actual experiment. Notation association without provided terms, notation association with provided terms, notation association with provided terms and case study, and a case study association. The correctly answered questions were used to find significant results for the interpretational effectiveness. Also, the answering time was recorded for the task and used to find significant results for the interpretational efficiency. After each task, the respondents were also asked to give their opinion about the perceived ease of use of the task. In the end of the experiment, four overall perceived ease of use questions were presented.

Task 1 is about giving textual association by means of presented icons. Task 2 is about linking correct textual associations to presented icons. Task three is the same assignment as task 2 but with a PGA notation hierarchy map presented as well. In task 4 a case model is presented, and ten comprehension questions are asked. For the first 3 tasks we investigate the possibility for a learning effect between the tasks as well. This is because only one context element had been added, and the questions were all the same.

For two of the ten formulated hypotheses namely, effectiveness and overall ease of use, significant results were found, indicating that the newly proposed version is indeed more intuitive than the initial version. However, on the level of individual icons interesting significant insights were found.

Contents

1.Introduction 1
1.1. Background 1
1.2. Exploration of the topic 1
1.3. Problem statement 2
1.4. Research objective and questions 2
1.5. Motivation/relevance 3
1.6. Main lines of approach 3
2. Theoretical framework 4
2.1. Research approach 4
2.2. Implementation 4
2.2.1. Building blocks method 5
2.2.2. Forward snowballing 6
2.2.3. Relevant articles 7
2.3. Results and conclusions 8
2.3.1. Purpose 8
2.3.2. Hypotheses 8
2.3.3. Measurement 8
2.3.4. Experimental design 12
2.3.5. Instrumentation and experimental tasks 12
2.3.6. Selection of participants 12
2.3.7. Operational procedures 12
2.3.8. Conclusion 13
2.4. Objective of the follow-up research 13
3. Methodology 14
3.1. Conceptual design: select the research method(s) 14
3.2. Technical design: elaboration of the method 14
3.2.1. Variables and measures 14
3.2.2. Hypotheses 15
3.2.3. Experimental groups 16
3.2.4. Instrumentation and experimental tasks 17
3.2.5. Selection of participants 19
3.2.6. Operational procedures 19
3.3. Data analysis 20
3.4. Reflection on validity, reliability, and ethical aspects 21
3.4.1. Construct validity 21

3.4.2. Internal validity 21
3.4.3. External validity 22
3.4.4. Reliability 22
3.4.5. Triangulation 22
3.4.6. Ethical aspects 22
4. Results 23
4.1. Analysing overall effectiveness 24
4.2. Analysing overall efficiency 26
4.3. Analysing overall ease of use 27
4.4. Analysing learning effect effectiveness 28
4.5. Analysing learning effect efficiency 30
4.6. Analysing learning effect ease of use 30
4.7. Confounding effects 31
5. Discussion, conclusions, and recommendations 32
5.1. Discussion – Reflection 32
5.1.1. Construct validity 32
5.1.2. Internal validity 32
5.1.3. External validity 32
5.1.4. Reliability 32
5.1.5. Triangulation 33
5.1.6. Ethical aspects 33
5.2. Conclusions 33
5.3. Recommendations for practice 34
5.4. Recommendations for further research 35
References 37
Appendix 1: Search strategy (added in Excel format) I
Appendix 2: Sent e-mails regarding survey part 1 (Dutch) II
Appendix 3: Survey part 1 both versions III
Appendix 4: Strategic sampling (added in Excel format) V
Appendix 5: Sent e-mails regarding survey part 2 VI
Appendix 6: Survey part 2 both versions VII
Appendix 7: Results survey part 1 and 2 (added in Excel and SPSS format) XX
Appendix 8: Shapiro Wilk test on all variables XXI
Appendix 9: Full results of T and Mann Whitney U test of all dependent variables XXIII
Appendix 10: Results of One-way ANOVA and Kruskal Wallis test demographics XXVIII

1. Introduction

1.1. Background

The strategic goals of an organization need to be (1) aligned with supporting business processes, and (2) communicated to their stakeholders intuitively. Enterprise Architecture (EA) focuses on aspects such as : “A coherent whole of principles, methods, and models that are used in the design and realisation of an enterprise’s organisational structure, business processes, information systems, and infrastructure” (Lankhorst, 2017, p. 3). Conceptual modelling languages are used to describe the EA of organisations. Conceptual modelling concerns the application of abstraction to reduce the complexity of a certain domain for a specific stakeholder purpose (Roelens & Bork, 2020).
Two main streams of conceptual modelling languages are categorised, namely (1) general purpose, and (2) domain specific (Frank, 2013; Roelens & Bork, 2020). A general purpose modelling language (GPML) is a modelling language which could describe multiple domains and their mutual coherence (The Open Group, 2019), e.g. ArchiMate , UML and BPMN. A domain specific modelling language (DSML) is a modelling language that raises the level of abstraction by specifying elements directly using domain concepts and that is specifically designed for a specific purpose (e.g. to analyse and communicate about a problem) in a particular domain often in only one company (Frank, 2013; Luoma, Kelly, & Tolvanen, 2004).
DSMLs reduce the complexity of the modelling in comparison to GPMLs. This is done by using concepts that are familiar to the intended end-users and by hiding complex model constraints in the meta-model (Frank, 2013). DSMLs are gaining popularity in the conceptual modelling field (Roelens & Bork, 2020). The designer of a DSML should aim for computational offloading by replacing cognitive tasks by perceptual ones (Bork, Schrüffer, & Karagiannis, 2019). An intuitive notation can account for computational offloading (Moody, 2009), which ultimately leads to an intuitive understanding (Michael & Mayr, 2017) of the new DSML.

1.2. Exploration of the topic

“…ensuring that PGA models can be intuitively understood by business-oriented end-users is of paramount importance to reduce the cognitive load for them.” (Roelens & Bork, 2020, p. 2)
The Process-Goal Alignment (PGA) modelling method is a specific DSML which targets business-oriented users as its stakeholder (Roelens & Bork, 2020). PGA aims at the development of a business architecture heat map following a modelling procedure that consists of three activities: (1) developing a prioritized business architecture hierarchy, (2) executing the performance measurement, and (3) performing the strategic fit improvement analysis. Strategic fit aims to align the business strategy with the internal infrastructure and processes (Henderson & Venkatraman, 1999), which enables organizations to adequately react to opportunities and threats in its external environment (Roelens & Bork, 2020).
In 2019, Roelens et al. published the first version of the PGA modelling method (Roelens, Steenacker, & Poels, 2019). The elements and their icons in the PGA modelling method are presented in a – so called – prioritized business architecture heat map. This hierarchy represents the strategic fit between (1) eight elements: Activity, Process, Competence, Value Proposition, Financial Structure, Internal Goal, Customer Goal and Financial Goal, (2) value stream relations: to link elements and show the hierarchical value structure, (3) prioritizing value stream relations in relation to their strategic importance and (4) the performance of each element.
Roelens and Bork (2020) published the second version of the PGA modelling method. The new set of icons of the PGA modelling notation is the result of evaluating the intuitiveness of the first PGA modelling method notation by an experiment among 139 master students at Gent University with an economic background. These students have indicated (1) which graphical representations they most intuitively associate with the given PGA terms, (2) which terms they associate with the given PGA

notations, and (3) which answers they could provide for comprehension questions about an example business architecture heat map. The results of this evaluation are six adjustments to the initial PGA notation. These adjustments are all icon changes for the elements: Competence, Value Proposition, Internal Goal, Customer Goal and for the Value Stream and Importance (see Figure 1).

PGA elements	Activity	Competence	Customer goal	Financial goal	Financial structure	Importance	Internal goal	Performance	Process	Value proposition	Value stream
Initial version											
Newly proposed version											

Figure 1: Suggested improvements of the PGA modelling notation (Roelens & Bork, 2020)

1.3. Problem statement

The original notation of the PGA modelling method originated from human creativity. This creativity led to the development of an initial variant of the notation. The initial variant has been evaluated via the evaluation technique of Bork et al. (Bork et al., 2019). This evaluation technique aims at how intuitive a DSML can be understood. Performing this technique on the PGA modelling method resulted in six improvements to the initial notation. These six changes should improve the intuitiveness of the PGA modelling notation. However, at this point it cannot be determined if the suggested improvements have a significant impact on the intuitiveness of the PGA modelling notation. Therefore, the main research question can be formulated as follows:

How can the intuitiveness of the PGA notation be evaluated within a practical context of a governmental organization?

The problem statement of this research consists of applying, evaluating, and benchmarking two PGA modelling notation versions in the 'expense management' case study process at the governmental organization 'Ministry of Education, Culture and Science'. The governmental-oriented end-users in this experiment are Architects, Managers/Directors, Policy officers and Finance and Control Employees. The problem statement includes:

***To apply two PGA modelling notation versions to a* practical context *of an organization*[1]**

To evaluate two applied PGA modelling notation versions on intuitiveness among governmental-oriented end-users

To select the most intuitive PGA modelling notation variant in a governmental organization

1.4. Research objective and questions

The problem statement in the previous paragraph led to a main research question and several sub questions, both theoretical (TSQ) and empirical (ESQ).

The aim of sub research question 1 is to conduct a literature review about what is known about the theoretical concepts DSML, intuitive design and experimental research design. Also, the aim is to find input on how to shape the experimental part of this research in a reliable manner.

[1] Following the three PGA modelling procedures (Roelens et al., 2019)

Sub question 1	What is known in the academic literature about evaluating the intuitiveness of PGA in the context of DSML in governmental organizations (TSQ)

The aim of sub research question 2 is to select a process which PGA can be applied to, so this can lead to a case model for the experiment.

Sub question 2	How are two PGA modelling notation versions (initial and newly proposed) applied to a process in a governmental organization? (ESQ)

The aim of sub research question 3 is to set up an experiment based on the theoretical insights of sub question 1 which will be able to measure intuitiveness of both PGA notation versions.

Sub question 3	How are two applied PGA modelling notation versions evaluated on intuitiveness among governmental-oriented end-users? (ESQ)

The aim of sub research question 4 is to analyse the results of the experiment which sub question 3 will have led to, to come up with an answer if the newly proposed version is more intuitive than the initial version of the PGA notation.

Sub question 4	How can initial and proposed PGA modelling notation be selected on intuitiveness? (ESQ)

The sub research questions 2, 3 and 4 represent the empirical part of this research – respectively the preparation phase (question 2), experiment/evaluation phase (question 3) and selection phase (question 3).

1.5. Motivation/relevance

The motivation of this research aims both at practical and academic relevance. The practical relevance of this research is related to the benefits of selecting the most intuitive PGA modelling notation for the governmental organization. The governmental organization in this research is the Ministry of Education, Culture and Science and the process in this research is expense management. An intuitive PGA modelling notation could lead to improved communication about the strategic fit between governmental process and its goals among governmental-oriented end-users. The academic relevance of this research is related to filling a scientific knowledge gap by setting up an experiment to compare two versions of the PGA modelling notation (Roelens & Bork, 2020). Prior research did not comparatively evaluate PGA and other DSMLs notation versions based on intuitiveness.

1.6. Main lines of approach

Chapter 2 provides answers for this research from the literature by answering TSQ 1. From these answers follow-up research is conducted in a practical organizational context. Chapter 3 presents the research strategy and chosen method for the research performed by answering ESQ 2 and 3. Chapter 4 discusses the quantitative results analysis – answering ESQ 4. Chapter 5 then follows with a discussion, a conclusion, the recommendations, and a reflection on the performed research.

2. Theoretical framework

The goal of the theoretical framework in this research is to conduct a literature review and summarise its main findings. Below, the following theoretical sub research question will be answered:

What is known in the academic literature about evaluating the intuitiveness of PGA in the context of DSML in governmental organizations?

2.1. Research approach

The aim of theoretical research is finding relevant literature about a scoped subject. These are detailed papers written and sorted by experts (Saunders, Lewis, & Thornhill, 2019). A literature study is an iterative process with multiple cycles (Saunders et al., 2019). For setting up a critical theoretical framework, the 'literature review process' was conducted (Saunders et al., 2019). This process led to further define the scope of the research question.

The conducted searches were performed via the OU Library Portal. This portal provides OU students with full access to available literature that meets specified search criteria. The following search strategies were followed:

1. The building blocks method (Westerkamp & Veen, 2009)
2. Forward snowballing on 'relevant' labelled articles with >50 citations in Web of Science

The search was limited by several criteria (see Table 1).

Table 1: Search criteria

Criteria	**Rationale: to ensure**
Papers must be available in the English language	That the content is readable in an understandable language.
Papers must be peer reviewed	The scientific quality of the papers
Newspapers are excluded	The scientific quality of the papers
Book reviews are excluded	The scientific quality of the papers
Papers must be fully available online	Found papers are accessible
Online citations will be ignored as article, but will be counted as search results	To limit search time without finding relevant papers
Results are sorted by relevance	Relevant papers are found in a reasonable time frame

To limit search time without finding relevant papers, the following stop criteria were used:

- When the number of found items is zero or all search results are labelled, the next query will be performed
- When five papers in a row of a query are labelled 'not relevant' the next query will be performed

2.2. Implementation

To determine if papers are relevant to this research, a specific selection strategy is developed. Papers will be scored using the title, meta data, abstract, introduction, discussion and conclusion (Jansen, 2013).

2.2.1. Building blocks method

The starting point is the theoretical sub question. Analysing this question led to the following search terms: PGA, DSML, intuitiveness and governmental organizations (see Table 2). To be able to search with these concepts, synonyms have been found using common sense and using Google and OU tutor insights. To find relevant and enough results, the search queries first focus on PGA (primary aim of this research) and expanded later to higher tier queries in DMSL and consequently conceptual modelling. Abbreviations often have multiple meanings in several scientific fields therefore the choice has been made to search for the full meaning of PGA and DSML alone. Additionally, the search queries focus on 'intuitiveness' and 'empirical evaluation'. According to the tutor 'intuitive' is not a common word in scientific literature and thus 'empirical evaluation' was suggested and used next to intuitive in queries 2.1 and 2.2. Also, we used both American English and British English spelling in the synonyms so no paper would be excluded up front.

Table 2: Search question translated to usable synonyms.

Theoretical sub question: ***What is known in the academic literature about PGA, intuitiveness, in the context of DSML in governmental organizations?***		
1	Concept: PGA/DSML	
	1.1	Process Goal Alignment
	1.2	Domain Specific Modeling Language OR Domain Specific Modelling Language
	1.3	Conceptual Modeling OR Conceptual Modelling
2	Concept: Evaluating intuitiveness	
		Evaluate Intuitive OR Intuitive Evaluation OR Evaluate Empirical OR Empirical Evaluation
3	Concept: Government organizations	
		Government Organizations OR Government Organisations OR Government

The synonyms of the selected concepts from the theoretical sub question are the building blocks for the search queries. The search is performed in a hierarchical way. The searches started at the lowest detail level 1.1, then 1.2 and then 1.3. Using at least two building blocks led to six search queries (see Table 3).

Table 3: Identified search queries based on building blocks method.

Nr	**Building blocks**	**Search query**
1	1.1 2 3	("process goal alignment") AND ((“evaluate intuitive”) OR (“intuitive evaluation”) OR (“evaluate empirical”) OR (“empirical evaluation”)) AND ((“government organizations”) OR (“government organisations”) OR (“government”))
2	1.2 2 3	((“domain specific modeling language”) OR (“domain specific modelling language”)) AND ((“evaluate intuitive”) OR (“intuitive evaluation”) OR (“evaluate empirical”) OR (“empirical evaluation”)) AND ((“government organizations”) OR (“government organisations”) OR (“government”))
3	1.3 2 3	((“conceptual modeling”) OR (“conceptual modelling”)) AND ((“evaluate intuitive”) OR (“intuitive evaluation”) OR (“evaluate empirical”) OR (“empirical evaluation”)) AND ((“government organizations”) OR (“government organisations”) OR (“government”))
4	1.1 2	("process goal alignment") AND ((“evaluate intuitive”) OR (“intuitive evaluation”) OR (“evaluate empirical”) OR (“empirical evaluation”))

Nr	Building blocks	Search query
5	1.2 2	((“domain specific modeling language”) OR (“domain specific modelling language”)) AND ((“evaluate intuitive”) OR (“intuitive evaluation”) OR (“evaluate empirical”) OR (“empirical evaluation”))
6	1.3 2	((“conceptual modeling”) OR (“conceptual modelling”)) AND ((“evaluate intuitive”) OR (“intuitive evaluation”) OR (“evaluate empirical”) OR (“empirical evaluation”))

Based on their title and abstract the found papers were labelled as ‘not relevant’ or ‘possibly relevant’. The conducted queries resulted in some double found papers, which were excluded. The outcome of the search strategy described in paragraph 2.1 can be found in Table 4 (see also Appendix 1).

Table 4: Search query outcome

Search query	Found papers	Papers assessed	Labelled possibly relevant
1	0	0	0
2	1	1	0
3	18	6	1. Gailly, Alkhaldi, Casteleyn, and Verbeke (2017)
4	0	0	0
5	16	5	0
6	121	10	2. Shanks, Tansley, Nuredini, Tobin, and Weber (2008) 3. Gemino and Wand (2004) 4. Allen and March (2012) 5. Bera, Burton-Jones, and Wand (2017)

2.2.2. Forward snowballing

Forward snowballing is starting from a possibly relevant paper and finding newer papers that have cited that paper. Three possibly relevant labelled papers had >50 citations according to the search strategy (Table 5).

Table 5: Relevant papers for forward snowballing

Papers identified for forward snowballing strategy	Web of Science citations
Gemino and Wand (2004)	111
Burton-Jones and Meso (2006)[2]	92
Shanks et al. (2008)	58

The same strategy was performed for the forward snowballing as for the search queries to label papers as ‘not relevant’ or ‘possibly relevant’ (see Table 6).

[2] The paper of Burton-Jones and Meso (2006) was not found during the search query strategy, however during the forward snowballing strategy of the paper of Gemino and Wand (2004).

Table 6: Forward snowballing outcome

Forward snowball search	Found papers	Papers assessed	Labelled possibly relevant
1	17	7	6. Mendling, Recker, Reijers, and Leopold (2018) 7. Burton-Jones and Meso (2006)
2	17	10	8. Haisjackl et al. (2016) 9. Zugal et al. (2013) 10. Gorla, Chiravuri, and Meso (2012)
3	10	5	0

2.2.3. Relevant articles

When the search strategy was finished, ten papers (five using the query strategy and five using the forward snowballing strategy) were labelled 'possibly relevant'. To determine the relevance, the next step was fully reading the introduction, discussion and conclusion and other relevant paragraphs (Jansen, 2013). After this step, two more papers Gailly et al. (2017) and Gorla et al. (2012) were excluded because of lack of relevance. The rest of the papers were marked 'relevant', and the following quality criteria were recorded (Table 7) per paper.

Table 7: Quality criteria for possibly relevant papers

Nr	Quality criteria	Rationale: when necessary
1	Reference in Web of Science (number)	Papers with more citations will be used over papers with less citations
2	Year of publication (year)	Newer papers will be used over older papers.
3	Does the paper describe a methodology or individual experiment? (M or E)	Papers that apply a methodology in a performed experiment will be used over papers that only describe a methodology.
4	Is there a link with intuitiveness? (Y or N)	Papers that describe indicators of intuitiveness like problem-solving performance, time taken, and ease of use will be used over papers that does not have these indicators.
5	Which type of questions are used? Recall, Comprehension, Problem-solving, cloze (R,Co,P,Cl)	Papers that describe usage of comprehension and/or problem-solving questions will be used over papers that does not use this type of questions.
6	Which participants: Business Experts vs Students? (B or S)	Papers that describe an experiment using business participants will be used over the papers that only using students as participants.

The following eight papers (presented in found order) are included in this literature review. The whole search strategy can be found in Appendix 1.

1. Representing Part-Whole Relations in Conceptual Modeling: An Empirical Evaluation (Shanks et al., 2008)
2. A framework for empirical evaluation of conceptual modeling techniques (Gemino & Wand, 2004)

3. A Research Note on Representing Part-Whole Relations in Conceptual Modeling (Allen & March, 2012)
4. Improving the representation of roles in conceptual modeling: theory, method, and evidence (Bera et al., 2017)
5. An Empirical Review of the Connection Between Model Viewer Characteristics and the Comprehension of Conceptual Process Models (Mendling et al., 2018)
6. Conceptualizing Systems for Understanding: An Empirical Test of Decomposition Principles in Object-Oriented Analysis (Burton-Jones & Meso, 2006)
7. Understanding Declare models: strategies, pitfalls, empirical results (Haisjackl et al., 2016)
8. Investigating expressiveness and understandability of hierarchy in declarative business process models (Zugal et al., 2013)

2.3. Results and conclusions

A conceptual model is a way to communicate information about a domain from a writer of a model to a viewer of a model. One could focus if one model is performing in an acceptable manner, or one could focus which of two or more models performs best when delivering the same message. In this research we focus on the latter. To be able to obtain definitive information about a conceptual model performance, only empirical methods can be used (Gemino & Wand, 2004). The summary of the literature review can be found in Table 8.

2.3.1. Purpose

In the reviewed literature we see seven experimental works and one methodological work of Gemino and Wand (2004). Where the purpose of the methodological work is providing a framework to perform an empirical evaluation, the experimental research has the purpose of evaluating two of more models empirically. One could evaluate a model intergrammar or intragrammar. Intergrammar means comparing one model with another model, both having their own grammar. Intragrammar means comparing one model using grammar n^1 and grammar n^x of the same modelling technique. (Gemino & Wand, 2004). In the reviewed literature (see Table 8), all experiments conducted are intragrammar, using different versions of the modelling grammar of a certain technique.

2.3.2. Hypotheses

Focusing on the used hypotheses, one can conclude that all experiments use a variant of the same hypotheses, namely: Subjects receiving conceptual model A will achieve a better performance for domain tasks then subjects receiving conceptual model B. The mostly used dependent variables in the reviewed literature are (1) problem-solving performance, (2) time taken and (3) ease of use (Allen & March, 2012; Bera et al., 2017; Burton-Jones & Meso, 2006; Mendling et al., 2018; Shanks et al., 2008).

2.3.3. Measurement

The categorization in affecting and affected variables is important to determine research questions and hypotheses (Gemino & Wand, 2004). To design empirical evaluations, one needs to determine the levels or values that affecting variables can have, and how to measure the affected variables.

Affected variables

Within a model comparison, one could focus on understanding the diagram and understanding the domain (Gemino & Wand, 2004). There are only two possible functions in a modelling comparison exercise: (1) script interpretation (i.e., reading the model) and (2) script creation (i.e., writing the model). When measuring affected variables, one could focus on the model that is created or read (i.e., product) or the process of making/reading the model itself (i.e., process) (Gemino & Wand, 2004). Within these functions one could measure two aspects; (1) Is the writer able to let the model

represent his/her view of the domain accurately or do discrepancies exist between the conception of the domain from a model creator's viewpoint and the model of the domain, and (2) can the reader of the model get the correct idea or is there any difference between what the writer intended to communicate and what the viewer understands (Gemino & Wand, 2004). However, when conducting an experiment, one must focus on both. A researcher cannot conclude if a writer's ability to develop a model is good or bad, without being able to read a model. This is also true when a researcher wants to conclude if the reader can understand the model properly, hence a created model in the research needs to be in place.
Both script interpretation and creation can be measured on efficiency and effectiveness (Gemino & Wand, 2004). All experiments conducted in the reviewed literature are using script interpretation (Allen & March, 2012; Bera et al., 2017; Burton-Jones & Meso, 2006; Haisjackl et al., 2016; Mendling et al., 2018; Shanks et al., 2008; Zugal et al., 2013). Script interpretation efficiency is measured by the ease of use/ease of learning and elapsed time (Bera et al., 2017; Burton-Jones & Meso, 2006; Gemino & Wand, 2004; Mendling et al., 2018; Shanks et al., 2008). Hereby ease of use/ease of learning is measured by the participant's perception of a task filling in a 7-point Likert scale (Burton-Jones & Meso, 2006; Shanks et al., 2008). Elapsed time is measured by the time taken to complete a task (Bera et al., 2017; Burton-Jones & Meso, 2006; Mendling et al., 2018; Shanks et al., 2008). Script interpretation effectiveness is measured by scoring answered questions of a participant with a 1 for a good answer and a 0 for a wrong answer (Allen & March, 2012; Bera et al., 2017; Burton-Jones & Meso, 2006; Haisjackl et al., 2016; Mendling et al., 2018; Shanks et al., 2008; Zugal et al., 2013)..

Affecting variables

The importance of the cases that are used (i.e., content) in the procedures cannot be overstated, as different cases will lead to different outcomes (Gemino & Wand, 2004). This is illustrated in the studies of Shanks et al. (2008) and Allen and March (2012). Both studies investigate the same question, but the outcome differs. According to Allen and March (2012), this difference is due to two flaws in the research of Shanks et al. (2008). The first one is not excluding the binary and ternary relationship as an independent variable. The second reason is that Shanks et al. (2008) do not apply the grammar correctly in the used UML models. The use of one standard case in the experiment is recommended (Gemino & Wand, 2004). The same is true when two models represent the same case, however a particular aspect of the grammar differs in both models (i.e. grammar construct). In the PGA modelling method, the grammar construct is the different set of icons used in the initial version and the newly suggested version. Also, the layout of models plays a significant role for both performance as well as completion time (Mendling et al., 2018). To validate this prior to the experiment, a model expert should be consulted to review the experimental model (Allen & March, 2012; Gemino & Wand, 2004; Haisjackl et al., 2016; Zugal et al., 2013).

When either an intra- or intergrammar comparison (i.e., nature of comparison) is chosen, it is of the utmost importance that there is no more information in the model because of using a different grammar (i.e., informational equivalence), else it can bias the outcome of the results (Gemino & Wand, 2004). In addition to this, any conclusion that can be drawn easily and quickly from the information given explicitly in a model should also be drawn easily and quickly from the information given explicitly in another model, and vice versa (i.e. computational equivalence) (Gemino & Wand, 2004).

Grammars can be evaluated by comparing them to a grammatical benchmark that contains a set of basic constructs expressing what should be modelled. If the benchmark is a set of generic constructs, it is referred to as a metamodel. If the benchmark is based on a set of beliefs on what might exist and happen in the modelled domain, it is called an ontology (Gemino & Wand, 2004, p. 255).

Table 8: Reviewed literature

Nr	Authors	Models Used (independent variables)	Testing	Subjects measured	Dependent variables	Experimental design	Domain or model understanding	Operational Procedure	Prerequisites	Questions used
1	Shanks et al. (2008)	UML Class Diagram	Should composites be represented as a relation or as an association	Business (57)	Problem solving performance Time taken Ease of use perception	Between	Domain	Room, paper	Explanation of UML symbols	Problem solving
2	Allen and March (2012)	UML Class Diagram	Are binary or ternary relationships better for understandable domain semantics	Students (31)	Problem solving performance	Within Subject	Domain	Room, paper	Several months of training	Problem solving
		UML Class Diagram	Should composites be represented as a relation or as an association	Students (82)	Problem solving performance	Within Subject	Domain	Room, paper	Several months of training	Problem solving
3	Bera et al. (2017)	EER Diagram	Testing domain understanding using guided and unguided scripts (product understanding)	Students (36)	Problem solving performance	Between	Domain	In a lab, computer	12H of training in models	Problem solving
		EER Diagram	Testing domain understanding using guided and unguided scripts (process understanding)	Students (36)	Problem solving performance Time taken	Between	Domain	In a lab, computer	12H of training in models	Eye tracking
4	Mendling et al. (2018)	BPMN models	Testing the significance of Model Viewer Characteristics for predicting model comprehension performance	Students (333) and business (197)	Model comprehension performance Time taken	Within Subject (MVC) Between (good and bad layout)	Model	Online	Need to have some modelling experience	Comprehension

Nr	Authors	Models Used (independent variables)	Testing	Subjects measured	Dependent variables	Experimental design	Domain or model understanding	Operational Procedure	Prerequisites	Questions used
5	Burton-Jones and Meso (2006)	Good, moderate, and bad versions of UML analysis diagram Use case, Class and State chart diagrams	Good decomposition models will increase analysts' understanding of a domain	Students (57)	Problem solving performance Cloze test performance Ease of use perception	Between	Domain	In a lab, paper	Object-oriented analysis course with UML	Problem solving Cloze test
6	Haisjackl et al. (2016)	Declarative Process Model	How does system analysts make sense of models (basic usage)	Students (9)	Reading the model Single building blocks Combination of constraints	Between	Domain	Room, paper	Moderate understanding of the model used obtained via mandatory training	Unknown
		Declarative Process Model	How does system analysts make sense of models (advanced usage)	Students (18)	Traces Paired constraints Hidden dependencies Existence constraints	Between	Domain	Room, paper	Moderate understanding of the model used obtained via mandatory training	Comprehension
7	Zugal et al. (2013)	Declarative Process Model	Do analysts understand the semantics of sub-processes	Students (9)	Pattern recognition Information hiding Fragmentation	Between	Domain	Room, paper	Moderate understanding of the model used obtained via mandatory training	Unknown

2.3.4. Experimental design

The commonly used experimental design is a between-subjects design (Bera et al., 2017; Burton-Jones & Meso, 2006; Haisjackl et al., 2016; Mendling et al., 2018; Shanks et al., 2008; Zugal et al., 2013). A between-subjects design ensures that no bias occurs between the first and second model, as participants only receive one. Normally, a between-subjects design has preference over a within-subjects design when the sample size is of certain amount. However, in the reviewed literature Haisjackl et al. (2016) and Zugal et al. (2013) use a between-subjects design having only nine participants, where Allen and March (2012) and Mendling et al. (2018) uses within-subjects design having 31, 82 and 530 participants.

2.3.5. Instrumentation and experimental tasks

The experimental tasks that need to be performed, should focus on either reading the model or writing the model (Gemino & Wand, 2004). For the effectiveness of model comprehension, comprehension questions should be used (Burton-Jones & Meso, 2006; Gemino & Wand, 2004; Mendling et al., 2018). Problem-solving questions are more effective measuring the understanding of a domain (Allen & March, 2012; Bera et al., 2017; Gemino & Wand, 2004; Haisjackl et al., 2016; Shanks et al., 2008; Zugal et al., 2013). Shanks et al. (2008) came to the same conclusion via a different reasoning. He states that relative to recall and comprehension performance, problem-solving performance provides a better understanding of a domain (Shanks et al., 2008). Another way of testing domain knowledge is via a Cloze test (Burton-Jones & Meso, 2006; Gemino & Wand, 2004). A cloze test is a task where the participant must fill in a blank in each text. Recall questions are questions that need to be answered from memory. (i.e. the model is first represented, and after the model is taken away, the questions are asked. These types of questions do not seem to be the best way in measuring domain understanding (Burton-Jones & Meso, 2006; Shanks et al., 2008).

2.3.6. Selection of participants

Perhaps the most important affecting variable are user characteristics. Prior work has recognised that both domain knowledge and modelling knowledge are important considerations in choosing participants (Gemino & Wand, 2004). In all papers except one (see Table 8), students are the subjects of the experiment (Allen & March, 2012; Bera et al., 2017; Burton-Jones & Meso, 2006; Haisjackl et al., 2016; Mendling et al., 2018; Zugal et al., 2013). Some argue that senior students could proxy for junior professionals (Bera et al., 2017). Others claims that students distinctively differ from practitioners (Mendling et al., 2018). In all experiments reviewed the subject needs to have at least moderate modelling experience with the modelling language used in the experiment.

2.3.7. Operational procedures

Concerning media used within the experiment, it is important that the same media is used for all models that are compared. If one test in an experiment is presented on paper and the other test is on a computer, it can bias the results. When looking at the reviewed literature, a variety of operational procedures are used concerning the location or the medium via which questions are asked (e.g. paper (Allen & March, 2012; Burton-Jones & Meso, 2006; Haisjackl et al., 2016; Shanks et al., 2008; Zugal et al., 2013), computer (Bera et al., 2017; Mendling et al., 2018), room (Allen & March, 2012; Haisjackl et al., 2016; Shanks et al., 2008; Zugal et al., 2013) or laboratory (Bera et al., 2017; Burton-Jones & Meso, 2006)) and answered (e.g. paper (Allen & March, 2012; Burton-Jones & Meso, 2006; Mendling et al., 2018), computer (Bera et al., 2017), verbally (Haisjackl et al., 2016; Shanks et al., 2008; Zugal et al., 2013)). However, there are also similarities between them. In all cases the participants could look at the model during the answering of the questions, all experiments gave an introduction to the experiment and asked demographic and experiential information about the participants and in almost all experiments (except of Mendling et al. (2018)), the model performance was tested via domain knowledge.

2.3.8. Conclusion

The literature review was conducted to be able to answer the theoretical sub question. As stated earlier in this section, there are different detail levels to this question. At the lowest level of PGA and DSMLs, the literature does not provide any concrete answers. Also, no literature has been found when focussing solely on research in a governmental context. However, searching on the higher level of detail and without the government context, the literature does give relevant insights to review.
Having reviewed the relevant literature an answer can be given to the theoretical sub question: *"What is known in the academic literature about evaluating PGA, intuitiveness, in the context of DSML in governmental organizations".*
There is a framework (Gemino & Wand, 2004) known in literature that gives principles to consider when creating a method to evaluate a conceptual model empirically. This framework is supported by the outcome and findings of the reviewed experiments. Both the methodological paper as the experimental papers give enough principles to develop a method in the upcoming chapter. These principles are selecting the research method, the technical design, the variables and measures, formulating the hypotheses, delivering experimental tasks to gather the data to test the hypotheses. These principles help to evaluate conceptual models empirically. No specific answer could be found in how to do this, when regarding intuitiveness or governmental organizations.

2.4. Objective of the follow-up research

The findings in this chapter will be used to answer the three empirical sub questions in the next chapter. The first step is to develop a PGA model using the PGA modelling method (for the initial and newly proposed version) to the chosen business context of an expense management process to be able to answer the first empirical sub question:

How are two PGA modelling notation versions (initial and newly proposed) applied to a process in a governmental organization?

The second step is developing an experiment using the initial and newly proposed PGA icons and models created so the second empirical sub question can be answered:

How can two applied PGA modelling notation versions be evaluated on intuitiveness among governmental-oriented end-users?

The third step is to conduct the experiment accordingly with ten business representatives within a practical business context of an expense management process. The outcome of these experiments will be analysed to answer the third empirical research question:

How can initial and proposed PGA modelling notation be selected on intuitiveness?

The last step is to conclude whether the performed experiment can be used for other empirical evaluations regarding intuitiveness or whether adjustments are needed.

3. Methodology

Based on the research of previous chapters the objective in this chapter is providing a methodology of comparing evaluation of the intuitive understanding of the initial and newly version of the PGA modelling notation. The methodology will be used in practice to provide answers to the empirical sub question provided in paragraph 1.4.

3.1. Conceptual design: select the research method(s)

Bork et al. (2019) gave an experimental method to provide a more intuitive notation of a conceptual model. Roelens and Bork (2020) applied this experiment to the PGA modelling notation. However, future research is needed for the evaluation of the proposed improvements (Roelens & Bork, 2020). One cannot conclude whether the newly proposed version is more intuitive than the initial version of the PGA modelling notation, as the analyses and design of both versions originated from human creativity. In this respect, it is important to delineate when a notation can be labelled as 'intuitive understandable'. The definition used in this paper is: *intuitive understandability is the ease with which a conceptual model can be understood by users immediately without any prior knowledge or training* (Jošt, Huber, Heričko, & Polančič, 2016).

The information required to achieve this objective is (1) the knowledge gathered from chapter 2 where several experiments and a methodology to evaluate conceptual models were studied, (2) the methodology which will be provided in this chapter and (3) the results of the comparative evaluation results of the intuitiveness of the initial and newly proposed version of the PGA modelling method in a governmental environment. This information can be found in literature and in a practical business context.

Prior research has evaluated the intuitiveness of PGA notation in a business context – tested on business students. This research wants to evaluate the intuitiveness of PGA notation in a governmental context – tested on professionals. By evaluating PGA in other contexts, one could expand its validity PGA beyond business contexts.

The method to provide the needed information can be obtained via a deductive approach (Saunders et al., 2019). This approach starts form theory which leads to the development of hypotheses. When the hypotheses are put forward, data collection can begin. With the collected data the hypotheses can be tested.

An experiment will be used as research strategy. *"An experiment is a research strategy whose purpose is to study the probability of a change in an independent variable causing a change in another, dependent variable. This involves the definition of null and alternative hypotheses; random allocation of participants to either an experimental or a control group; manipulation of the independent variable; measurement of changes in the dependent variable; and control of other variables." (Saunders et al., 2019, p. 803).*

3.2. Technical design: elaboration of the method

As described in the previous paragraph in this research an experimental design will be developed among business practitioners of a governmental organization.

3.2.1. Variables and measures

In all reviewed literature (see Table 8), DSML notations are the independent variables, evaluated by its desirable outcomes as dependent variables.

This research focusses on the PGA notation. Therefore the two independent variables are:

- The initial PGA modelling notation (Roelens et al., 2019)
- The newly proposed PGA modelling notation (Roelens & Bork, 2020)

The selection of the expense management process as a case model ensured no detailed domain knowledge which could have a bias to the results (Gemino & Wand, 2004). Based on literature review, it can be concluded that intuitiveness is not a commonly used word in de academic world. Three proxy words have emerged from the review, namely, effectiveness, efficiency, and ease of use.

Consequently, the dependent variables and measures are:

- Interpretational effectiveness (Allen & March, 2012; Bera et al., 2017; Burton-Jones & Meso, 2006; Gemino & Wand, 2004; Haisjackl et al., 2016; Mendling et al., 2018; Shanks et al., 2008; Zugal et al., 2013): # of correct answers to different question types
- Interpretational efficiency (Allen & March, 2012; Bera et al., 2017; Burton-Jones & Meso, 2006; Shanks et al., 2008): time needed per section of the experiment
- Perceived ease of use (Burton-Jones & Meso, 2006; Shanks et al., 2008): questionnaire items measured on a seven-point Likert scale

These independent and dependent variables are measured during the following tasks (Roelens & Bork, 2020; Roelens et al., 2019):

- Task 1: Notation association without provided terms
- Task 2: Notation association with provided terms
- Task 3: Notation association with provided terms and case study
- Task 4: Case study association

With every task, an extra context is added for the participant. First a respondent will only be given an icon. In the second task, icons, and the corresponding terms. In the third task the icons, the corresponding terms, and a case model. In the fourth task, the context of the questions changes toward the case model itself. The dependent variables will be measured at three hierarchy levels. The lowest level measures the question (icon) itself. Questions will be asked in the context of a task, which is the second level. Each task will have at least one question about the icons that differs between both versions used in this experiment. At the highest level, three indicators will be used as proxies for intuitiveness (effectiveness, efficiency, and perceived ease of use). Where applicable these three levels will be measured to give a deeper understanding and insight in the collected data. To enable the implementation in the survey tool, tasks 2 and 3 will only be measured at task level, so no results can be expected on the individual icon level.

3.2.2. Hypotheses

Based on the theoretical insights in chapter 2, ten hypotheses have been formulated:

- H01: The interpretational effectiveness of the newly proposed PGA modelling notation is higher than the initial PGA modelling notation
- H02: The interpretational efficiency of the newly proposed PGA modelling notation is higher than the initial PGA modelling notation
- H03: The perceived ease of use all tasks of the newly proposed PGA modelling notation is higher than the initial PGA modelling notation
- H04: The overall perceived ease of use of the newly proposed PGA modelling notation is higher than the initial PGA modelling notation

- H05: The learning effect of the newly proposed version is higher than the initial version for interpretational effectiveness of task 2 versus task 1
- H06: The learning effect of the newly proposed version is higher than the initial version for interpretational effectiveness of task 3 versus task 2
- H07: The learning effect of the newly proposed version is higher than the initial version for interpretational efficiency of task 2 versus task 1
- H08: The learning effect of the newly proposed version is higher than the initial version for interpretational efficiency of task 3 versus task 2
- H09: The learning effect of the newly proposed version is higher than the initial version for perceived ease of use of task 2 versus task 1
- H10: The learning effect of the newly proposed version is higher than the initial version for perceived ease of use of task 3 versus task 2

The first four hypotheses focus on the indicators themselves. H1 benchmarks both PGA notations based on interpretational effectiveness of each task. H2 benchmarks both PGA notations based on interpretational efficiency on each task. H3 benchmarks both PGA notations based on perceived ease of use within each task (measured after each task). H4 benchmarks both PGA notations based on the overall perceived ease of use (measured at the end of the experiment). The expectation is that the newly proposed version will perform significantly better than the initial version for all three indicators.

Task 1 is about giving textual association by presented icons. Task 2 is about linking correct textual associations to presented icons. Task three is the same assignment as task 2 but with a PGA notation hierarchy map presented as well. In task 4 a case model is presented, and ten comprehension questions are asked. Task 1-3 are different versions of the notation association task, while task 4 focuses on comprehension questions for a given PGA case model.
Based on these similarities of task 1-3 we expect a learning curve will be seen for both the newly proposed version as for the initial version for tasks 2 to 1 and 3 to 2. H05 and H06 focusses on the learning effect for the indicator effectiveness between tasks 2 and 1 and tasks 3 and 2. In line with H01 we expect that the learning curve of the newly proposed version will be better. This should be visible in the results with increasing relative scores after each task. All these hypotheses will be tested at the level of individual icons. H07 and H08 focuses on the learning effect for the indicator efficiency between tasks 2 and 1 and tasks 3 and 2. Also here it is expected that the newly proposed version will be performing better than the initial version. Finally, H09 and H10 focuses on the learning effect for the indicator ease of use between tasks 2 and 1 and between tasks 3 and 2. We also expect that learning effect for the perceived ease of use will be beneficial towards the newly proposed version. Not only for the individual tasks, but also for the overall perceived ease of use questions.

3.2.3. Experimental groups

To test the hypotheses a between group experiment will be held among business practitioners of a governmental organization. The choice for a between-subjects' design experiment is based on the specific characteristics of this experiment. In contrast to a within-subjects' design, between-subjects' design separates a sample group into two subgroups (or more), presenting mutual exclusive set of questions. This research does not apply a within-subjects' design because the initial and newly proposed PGA notations have strong overlap which could lead to (1) an undesirable cross-learning effect by the participants, and (2) confusion among participants what its inherent differences are. The between-subjects' design is applied to prevent this cross-learning and confusion effects – among two comparable subgroups. The sample group size of this research is numerous enough to create two subgroups consisting of professionals, representing a governmental organization.

3.2.4. Instrumentation and experimental tasks

The following instrumentation and experimental tasks are identified:

- Introduction to the experiment
- Demographic questions
- Notation association without provided terms
- Notation association with provided terms
- Notation association with provided terms and case study
- Case study association
- Perception questionnaire items

In the reviewed literature, introduction and demographic questions are the start of the experiment (Allen & March, 2012; Bera et al., 2017; Burton-Jones & Meso, 2006; Haisjackl et al., 2016; Mendling et al., 2018; Shanks et al., 2008; Zugal et al., 2013). Also, the cases used in six out of seven reviewed literature experiments are representing an actual business case (Allen & March, 2012; Bera et al., 2017; Haisjackl et al., 2016; Mendling et al., 2018; Shanks et al., 2008; Zugal et al., 2013). The difference between the starting point of the reviewed experiments and the starting point of this method is that this experiment keeps the definition on intuitiveness of Jošt et al. (2016) in a leading position. This means that the participants of this experiment will have no prior knowledge of the modelling notation used in the experiment. The three notation association tasks have emerged from the evaluation technique of Bork et al. (2019). The first notation association task in this experiment is the same as Bork et al. (2019) notation association phase. The second and third notation association task are nearly the same task, but give more context to the participant, respectively terms and case information, to be able to investigate if providing more context will benefit the intuitiveness of a modelling notation.
The first step (term association) of Bork et al. (2019) will not be adopted in this experiment, because the aim of this research is not to come up with a new version of the PGA modelling notation, but to compare them.

Introduction to the experiment

Participants to the experiment will be introduced to the experiment and given an explanation to the domain of the PGA modelling notation. This will ensure the right mindset when conducting the experiment.

Demographic questions

Demographic questions will be asked to obtain information about the profile of the participants. These control variables and measures are:

- Demographics: age, gender
- Years of experience in domain
    - Current job function: in years of experience
    - Relevant previous job functions: in years of experience
- Prior knowledge in conceptual modelling
    - Experience in reading models: in years
    - Experience in making models: in years
    - Experience in PGA modelling notation: yes/no

Notation association without provided terms

Participants will be asked to provide associated terms given a particular icon of the PGA modelling method. In this experimental task we will focus on all PGA modelling notation elements of both versions of the PGA modelling notation (one version per experimental group). In this way, not only the elements that differ between both PGA modelling notation versions can be measured, but the

whole notation could be compared with each other. The participants will be asked to give up to three possible terms for the given icon. To be able to measure the outcome of this step, the correct term will be breaking down into several 'good' terms using commonly known synonyms. Each correct icon-term association will be given a point. Also, time taken to complete the task will be measured.

Notation association with provided terms

Participants will be asked to link a given set of icons (same as previous step) to given terms, however the terms and icons are scrambled. Each correct icon-term association will be given a point. Also, time taken to complete the task will be measured.

Notation association with provided terms and case study

Participants will be given the same question presented in tasks 2. However, also the case study will be presented. Participants will be asked if they want to correct the given answers of task 2 after having seen the full case study model. Each correct icon-term association will be given a point. Also, time taken to complete the task will be measured.

Case study association

The practical context used for this experiment is the process 'expense management'. The choice for this process is that most employees will have a slight idea of the context of this process, without being an expert on it. Therefore the group of participants that can be approached is bigger. Also, it is a process that not only will be applicable in this domain. The practical context will be modelled for both versions of the PGA modelling notation. The model will be presented to an expert of the PGA modelling notation, to make sure that no flaws will emerge in the model, which can influence the results. In this experimental task, comprehension questions will be given to participants. Participants will be given multiple (four) possible answers, of which only one is correct. To counteract gambling for a correct answer also a 'No idea' option will be provided as a fifth possibility. In the introduction of the experiment, participants will be reminded not to gamble if they cannot come up with a reasoned answer, but to use the 'No idea' option instead. Comprehension question should be used for measuring model understanding/ model intuitiveness (Burton-Jones & Meso, 2006; Gemino & Wand, 2004; Mendling et al., 2018). Also Roelens and Bork (2020) state to use comprehension questions in follow up research. The correct answer will be awarded with a point. Also, time taken to complete the task will be measured.

Perception questionnaire items

After each experimental task, the participant will be presented with three questions to measure participants' perceptions about the ease of use of the experimental task performed. The questions were based upon Moore and Benbasat (1991) short-form version of the instrument developed by Davis (1989). The original questions were modified to suit the experimental tasks.
After each notation association tasks the following question will be presented:

- Overall, I believe it was easy for me to associate the notation with (the provided) terms

After the case study association task, the following question will be presented:

- Overall, I believe it was easy for me to understand the case model
- Overall, I believe that the PGA model was easy to use

The answers to these questions will be provided via a 7-point Likert scale, ranging from "Strongly agree" (7) over "Neutral" (4) to "Strongly disagree" (1).

In this way a task will not only provide an answer to whether a participant understands the icons/models provided, but also whether the task was easy to perform according to their own perception.

3.2.5. Selection of participants

The following job functions are a representation of the functions one can find in a policy-oriented ministry. From governmental workers, towards ICT, from strategical management to operations, from policy to data and finance. All will be approached to participate in the experiment.

Table 9: Complete base of possible participants

Job functions	**Count**
Policy officer/Account manager	30
Finance & Control employee	14
Director/Manager	6
Architect (domain/information)	2
Information policy worker	4
Total possible participants	55

Although the assignment of participants to the experimental groups takes place at random, it will be ensured that each function (see Table 9) will be distributed evenly between both groups. To ensure this equal distribution of functions in the two groups, strategic sampling will be used.

3.2.6. Operational procedures

Due to Covid-19 pandemic, a computer-based experiment will be conducted. The experiment will be divided into two parts. At first all 55 possible participants will be asked via an e-mail to participate in both parts 1 and 2 of the survey and if they are willing to directly complete part 1 of the survey via the link provided in the e-mail. This part gives a brief introduction to the research objective and presents the demographic questions. In this part of the survey, the PGA notation will not be disclosed to avoid any bias by searching on the Internet for PGA between part 1 and part 2 of the survey. Part 1 of the survey will be open for one week.
Based on the answers given in Part 1 of the survey, strategic sampling will be performed according to the described procedure in Table 10.

Table 10: Strategic sampling procedure

Step number	**Description**	**Order**
1.	Sort on response number	Ascending
2.	Sort on function role	Ascending
3.	Create set of responses per function/role	Not applicable
4.1 Architect	Sort on number of models read	Ascending
4.2 Director/Manager	Sort on number of models read	Descending
4.3 Finance/ Control Employee	Sort on number of models read	Ascending
4.4 Policy officer/Account manager	Sort on number of models read	Descending
4.5 Information Policy officer	Sort on number of models read	Ascending

Step number	Description	Order
5	Merge output together	Not applicable
6	Divide respondents alternately	respondent 1 to group A, respondent 2 to group B, respondent 3 to group A, etc.

The goal of this sampling strategy is to equally divide both functions/roles and the reading model knowledge to both groups. After the respondents have been assigned to the groups, both groups will receive an e-mail with an introduction to survey part 2. The text in both emails is the same, however the link to the survey will differ. Group A will be given the experiment in line with the initial version of the PGA modelling notation, group B will be given the experiment in line with the newly proposed PGA modelling notation (see Figure 2) The participants will not know if there are other participants and who they are.

Before the actual experiment will be conducted, a pre-test will be performed once the experimental design is finished. This is to ensure no technical flaws or incorrect/ non-understandable question will be in the actual version. None of the participants of the test will be a participant in the actual experiment.

The data will be collected per task (for task 1 to 3) or per question (for task 4). Once a task or question is completed, it will not be possible to change answers of previous tasks/ questions. To ensure participants will not accidently go to a next experimental task/ question a control question will be asked if the participant is indeed willing to end the task.

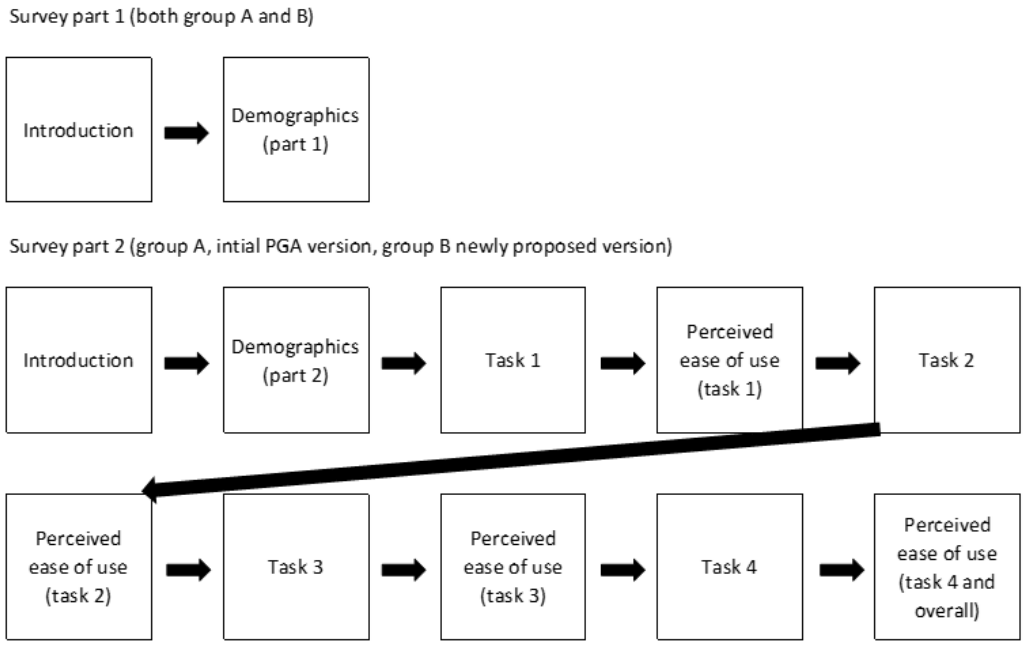
3.3. Data analysis

The data related to the dependent variables will be collected in four experimental tasks. Analyses will be performed on the collected data based on three types of total scores. The first total score measures the effectiveness per task by the number of correct answers. An answer of a participant is correct if it is comparable to the norm. The norm will be set by the researcher. Also, an expert will be asked to perform a check on overall consistency. The answers that significantly differ from the norm, are incorrect (coded: 0). The answers that somewhat are comparable to the norm, are partly correct (coded: 1). The answers that are comparable to the norm, are correct (coded: 2). For tasks 2-4 code 1 will not be provided, because the answers in these tasks can only be incorrect or correct.
The coded answers per question are discrete data. These numbers can be added up or subtracted, multiplied or divided, therefore the collected data per task are ratio data (see formulas per total score)(Saunders et al., 2019). The efficiency per task is measured by the number of seconds. Between starting and finishing a task or question, seconds are counted on a ratio scale. The perceived ease of use per task is measured by a 7-point Likert scale, which can be considered as ordinal data. Therefore medians will be used in the comparisons. Because of the between-subjects' design approach the collected data will be unpaired. Therefore, the first step is to test the normality of the dependent variables via a Shapiro-Wilk test. Based on the outcome a t-test will be performed when data is normally distributed or a Mann-Whitney U test when the data is not normally distributed.

Formulas per total score:

- Total score (Effectiveness): # correct answers/total norms * 100%
- Total score (Efficiency): # time taken in seconds
- Median score (Perceived Ease of use: all tasks): Likert scale 0 (fully agree) to 6 (fully disagree)
- Median score (Perceived Ease of use: overall): Likert scale 0 (fully agree) to 6 (fully disagree)

Figure 2: Between-subjects' experimental design



3.4. Reflection on validity, reliability, and ethical aspects

Measures will be considered to obtain good validity, reliability, and ethical aspects. The checklist of (Gelderman, 2013) will be followed.

3.4.1. Construct validity

To provide a solid construct validity, multiple sources have been investigated during the literature review. The same concepts of the studied literature have been used in the method of this research. Therefore we are confident that the operational measures are correct for the concept being studied in this research. By dividing this experiment in multiple tasks, a chain of evidence may be found in the results. Also, the questions used for measuring the perceived ease of use emerged from literature.

3.4.2. Internal validity

By using a between-subjects' design approach, participants will only conduct the experiment with one version of the PGA modelling notation. Therefore no bias or learning effect of the other version can occur. The experimental method follows a four-step design which measures the intuitiveness of the PGA modelling methods versions. The prospective participants are all having multiple years of involvement in a governmental organization. All are professional's business governmental users. The experiment will be tested by seven participants before the experiment will start, so flaws can be found before the experiment is started among the participants.
The experiment is developed by the researcher but will be tested by several participants up front. The used model will be developed by the researcher but will be tested by a PGA expert.

3.4.3. External validity

This study will be held in a nationwide governmental setting. However, generalization could be obtained for province, local level government, semi-governmental organization (like foundations) and private foundations. Although there are differences between the organizations mentioned, the similarities predominate. The selected roles that are used in this experiment are also available at these other governmental organizations.

3.4.4. Reliability

Although the experiment will be held in an online manner, all participants will attend via the same tool with the same questions to create a similar environment. Therefore all participants will have the same introduction, and all participants can ask questions and will hear all answers before the experiment will begin. After the introduction, no one will be able to talk to the researcher or other participants. The time the participant needs to finish the experimental tasks (time taken) will be automatically measured. Using a between-subjects' design will ensure that a respondent needs to invest less time in the experiment, because only one version will have to be presented. Also, this will increase the rate of respondents that will finish the experiment. A higher response rate will be beneficial for the research.

3.4.5. Triangulation

The possible participants are all known by the researcher, so the domain knowledge is subjectively determined. Nevertheless, the demographic question will also give an objective measurement. Example of triangulation in this research is that ease of use is measured by (1) participant's perception and (2) participant's time taken - per task.

3.4.6. Ethical aspects

The ethical aspects that should be considered are political opinions of the respondents that can influence the results. The researcher has a working relation with all participants which can influence the results. The privacy of the participants will be guaranteed by not asking any information that can link a person to the results of outcome in this research. All participants can participate on voluntary basis and are informed about the reasoning of this experiment. Participants will be asked upfront if they want to participate in the experiment. Even when a participant starts the experiment, he can stop the experiment at any given time.

4. Results

The experiment was started by sending e-mails (see Appendix 2) to all 55 selected possible respondents as described in paragraph 2.3.6. In this e-mail a link to LimeSurvey part 1 (see Appendix 3) was enclosed. Respondents were asked to fill in part 1, only if they had the time and the idea to also participate in part 2.
To ensure a high number of respondents, two reminder e-mails were sent as well. After a week, part 1 of the survey was closed and 25 respondents filled in a full survey. The results were extracted from LimeSurvey to be able to conduct the strategic sampling like described in paragraph 3.2.6, resulting in respondents assigned to groups A and B (see Appendix 4).
Respondents assigned to group A and B received an email with a link (see Appendix 5) for part 2 of the survey (see Appendix 6), which was open for a week in LimeSurvey. To ensure a high number of respondents to this part, two reminder e-mails were sent as well. When the survey was closed, a total number of 21 respondents filled in both parts 1 and 2 of the survey.

The results of parts 1 and part 2 were put together in one file (see Appendix 7). This was possible because of the full name question in both surveys. In line with the declaration of consent all data was anonymized directly in LimeSurvey as in the export files.

After this step, the content of the results file was reordered for analyses purposes (Demographic questions, Task 1, 2, 3 and 4, Perceptions questions and time measurement). Besides this, columns with double or irrelevant data were removed. Other cells of columns were recoded according to paragraph 3.2.6 to be able to start the analysis in SPSS. Finally, two columns were added to calculate totals of years of work experience of the respondents and the total time of task 4: case study association. This data file was imported in SPSS (version 27) and all variables were configured with the correct measure.

The analyses of the data are presented in following tables. All tables use the same colour scheme for readability purposes. Effectiveness dependent variables are presented in a green colour, efficiency dependent variables are presented in an orange colour, ease of use dependent variables are presented in a grey colour and learning effect variables are presented in a blue colour. To be able to compare the outcome of both groups the efficiency data (time in seconds) is brought back to the average time a respondent took to answer the questions.

To be able to select the correct test to test the hypotheses, the first step was to determine if the dependent variables were distributed normally via a Shapiro-Wilk test in SPSS (see Appendix 8). A p-Value higher than 0,05 means the data is following a normal distribution (H_0 accepted) If the P-Value is lower than 0,05 the data is not following a normal distribution (H_a accepted).

Based on the results of the normality test we can use non-parametric Mann-Whitney U test for non-normally distributed dependent variables (i.e., H_a accepted) or a parametric T-Test for normally distributed dependent variables (i.e., H_0 accepted). In Appendix 9 all results are presented of these tests. For readability purposes, only the tests where the newly proposed version performs significantly better than the initial version will be presented in the next paragraphs.

4.1. Analysing overall effectiveness

Table 11: Descriptive statistics effectiveness variables

Dependent Variable (effectiveness)	Score initial version	Score newly proposed version	Percentage score of the initial version	Percentage score of the newly proposed version	Difference newly proposed and initial version
Total Effectiveness: all tasks	**109**	**238**	**19%**	**39%**	**20%**
Subtotal Effectiveness: task 1	1	14	1%	11%	10%
Effectiveness: Competence, task 1	0	0	0%	0%	0%
Effectiveness: Customer goal, task 1	1	0	5%	0%	-5%
Effectiveness: Importance, task 1	0	14	0%	64%	64%
Effectiveness: Internal goal, task 1	0	0	0%	0%	0%
Effectiveness: Value proposition, task 1	0	0	0%	0%	0%
Effectiveness: Value stream, task 1	0	0	0%	0%	0%
Subtotal Effectiveness: task 2	36	68	30%	52%	22%
Effectiveness: Competence, task 2	6	6	30%	27%	-3%
Effectiveness: Customer goal, task 2	12	18	60%	82%	22%
Effectiveness: Importance, task 2	2	10	10%	45%	35%
Effectiveness: Internal goal, task 2	6	6	30%	27%	-3%
Effectiveness: Value proposition, task 2	8	14	40%	64%	24%
Effectiveness: Value stream, task 2	2	14	10%	64%	54%
Subtotal Effectiveness: task 3	28	74	23%	56%	33%
Effectiveness: Competence, task 3	2	10	10%	45%	35%
Effectiveness: Customer goal, task 3	8	14	40%	64%	24%
Effectiveness: Importance, task 3	0	14	0%	64%	64%
Effectiveness: Internal goal, task 3	10	12	50%	55%	5%
Effectiveness: Value proposition, task 3	4	8	20%	36%	16%
Effectiveness: Value stream, task 3	4	16	20%	73%	53%
Subtotal Effectiveness: task 4	44	82	22%	37%	15%
Effectiveness: Customer goal, task 4	6	8	30%	36%	6%
Effectiveness: Value proposition, task 4	0	2	0%	9%	9%
Effectiveness: Internal goal, task 4	6	2	30%	9%	-21%
Effectiveness: Importance, task 4	6	16	30%	73%	43%
Effectiveness: Value stream, task 4	2	18	10%	82%	72%
Effectiveness: Competence, Importance, Value stream, task 4	8	10	40%	45%	5%
Effectiveness: Customer goal, Internal goal, Value proposition, Value stream, task 4	2	4	10%	18%	8%
Effectiveness: Value proposition, Value stream, task 4	6	6	30%	27%	-3%
Effectiveness: Importance, Value stream, task 4	4	2	20%	9%	-11%
Effectiveness: Competence, task 4	4	14	20%	64%	44%

There are thirty-three variables for effectiveness for which the overall trend for both versions is that respondents cannot find a correct textual association out of the blue. Twenty-eight of them are at icon level, four at the task level - and on the highest level one overall indicator.

From the descriptive statistics (see Table 11) it strikes that most icon questions of task 1: notation association without provided terms, were answered incorrectly. Nevertheless, the difference between both versions is mostly beneficial towards the newly proposed version. Fourteen variables show the significant outcome that the newly proposed version is performing better than the initial versions. The newly proposed version seems to be better understood by the respondents than the initial version. The statistical test (see Table 12) of task 1 shows that the newly proposed version is significantly performing better than the initial version (P-value 0,003). However, this is the result of only one of six questions of task 1. In this respect, Importance is the only icon that has a significant result (P-value 0,002). In other words, the significance of task 1 can be explained by only the icon of Importance.

The statistical test of task 2: notation association with provided terms (see Table 12), shows that the newly proposed version is also performing significantly better than the initial version (P-value 0,026). However, this is based on only two of six questions of task 2. In this respect Importance (P-value 0,040) and Value stream (P-Value 0,007). In other words, the significance of task 2 can be explained by two icons, the Importance and Value stream.

The statistical test of task 3: notation association with provided terms and case study (see Table 12), show that the newly proposed version is performing better than the initial version (P-value 0,007). This is based on three of six questions of task 3. In this respect Competence (P-value 0,040), Importance (P-value 0,002) and Value stream (P-Value 0,009). In other words, the significance of task 3 can be explained by three icons, the Competence, Importance and Value stream.

The statistical test of task 4: case study association, (see Table 12) shows that the newly proposed version is performing better than the initial version (P-value 0,026). This is based on three questions of task 4. In this respect Competence (P-value 0,025), Importance (P-value 0,028) and Value stream (P-Value 0,001). In other words, the significance of task 4 can be explained by three icons, the Competence, Importance and Value stream. In task 4 also questions were asked that required a combination of icons. Question 6 was a question, in which all three significantly better performing icons need to be combined. However, for this question no significant (P-value 0,403) result was found that the newly proposed version was performing better than the initial version.

In Table 12 we see a P-value of 0,001 for the total effectiveness of all tasks, meaning H01: The interpretational effectiveness of the newly proposed PGA modelling notation is higher than the initial PGA modelling notation, can be accepted. This is also true for all four tasks (P-values of; 0,003, 0,026, 0,007, 0,026). However, looking at individual icon level a significant result was only found for the three icons, ‘Competence’ (task 3 and 4), ‘Importance’ (task 1-4) and ‘Value stream’ (task 2-4).

Table 12: Significant results of effectiveness variables

Dependent Variable	Performed test	Mean (rank) initial version	Mean (rank) newly proposed version	two-tailed P-Value	one-tailed P-Value	Accepted Hypothesis
Total Effectiveness: all tasks	**T-Test**	**10,900**	**21,640**	**0,001**	**0,001**	**H_a**
Subtotal Effectiveness: task 1	**Mann-Whitney U**	**7,7**	**14**	**0,006**	**0,003**	**H_a**
Effectiveness: Importance, task 1	Mann-Whitney U	7,5	14,18	0,003	0,002	H_a
Subtotal Effectiveness: task 2	**T-Test**	**3,600**	**6,180**	**0,052**	**0,026**	**H_a**
Effectiveness: Importance, task 2	Mann-Whitney U	9,05	12,77	0,080	0,040	H_a
Effectiveness: Value stream, task 2	Mann-Whitney U	8,05	13,68	0,014	0,007	H_a
Subtotal Effectiveness: task 3	**T-Test**	**2,800**	**6,730**	**0,013**	**0,007**	**H_a**
Effectiveness: Competence task 3	Mann-Whitney U	9,05	12,77	0,080	0,040	H_a
Effectiveness: Importance, task 3	Mann-Whitney U	7,5	14,18	0,003	0,002	H_a
Effectiveness: Value stream, task 3	Mann-Whitney U	8,1	13,64	0,018	0,009	H_a
Subtotal Effectiveness: task 4	**T-Test**	**4,400**	**7,450**	**0,052**	**0,026**	**H_a**
Effectiveness: Importance, task 4	Mann-Whitney U	8,65	13,14	0,056	0,028	H_a
Effectiveness: Value stream, task 4	Mann-Whitney U	7,05	14,59	0,001	0,001	H_a
Effectiveness: Competence, task 4	Mann-Whitney U	8,6	13,18	0,049	0,025	H_a

4.2. Analysing overall efficiency

For this indicator applies that the higher the score (in seconds), the lower the ease of use is perceived by the respondent

Table 13: Descriptive statistics efficiency variables

Dependent Variable (effectiveness)	Time taken initial version	Time taken newly proposed version
Total Efficiency: all tasks	**729,93**	**1152,32**
Efficiency: task 1	141,13	289,64
Efficiency: task 2	81,22	122,46
Efficiency: task 3	150,32	196,84
Subtotal Efficiency: task 4	357,26	543,38
Efficiency: Customer goal, task 4	61,15	89,29

Dependent Variable (effectiveness)	Time taken initial version	Time taken newly proposed version
Efficiency: Value proposition, task 4	29,33	49,83
Efficiency: Internal goal, task 4	29,28	45,22
Efficiency: Importance, task 4	38,47	37,30
Efficiency: Value stream, task 4	35,45	48,04
Efficiency: Competence, Importance, Value stream, task 4	35,40	53,43
Efficiency: Customer goal, Internal goal, Value Proposition, Value stream, task 4	20,75	34,82
Efficiency: Value proposition, Value stream, task 4	19,72	48,58
Efficiency: Importance, Value stream, task 4	54,96	97,42
Efficiency: Competence, task 4	32,74	39,44

The statistical test for efficiency on tasks 1, 2 and 3 did not show significant results between both versions (P-value 0,972, 0,948 and 0,898). Given the average time figures (see Table 13), it seems that the initial version is more efficient than the newly proposed version (except for Importance). To answer all questions a respondent with the initial version took 12,2 minutes (729.93 seconds) to finish the tasks (waiting times before and between tasks already removed), while a respondent of the newly proposed version took 19,2 minutes (1152.32 seconds) to finish the tasks. That is 58% less efficient.

In the collected data no evidence can be found to link these significances to individual icons because it was not measured (task 1, 2 and 3). The hypotheses H02: The interpretational efficiency of the newly proposed PGA modelling notation is higher than the initial PGA modelling notation, could therefore not be accepted.

4.3. Analysing overall ease of use

For this indicator applies that the higher the score (0 to 6), the lower the ease of use is perceived by the respondent

Table 14: Descriptive statistics ease of use variables

Dependent Variable (effectiveness)	Median initial version	Median newly proposed version
Total Ease of use: all tasks	**4,5**	**4**
Ease of use: task 1	1	1
Ease of use: task 2	5	4
Ease of use: task 3	5	4
Ease of use: task 4	5	4
Total Overall Ease of use	**4,75**	**4**
Ease of use: overall understanding	4,5	4
Ease of use: overall frustration	4,5	4
Ease of use: overall using	5	4
Ease of use: overall learning	5	4

For the task-based ease of use questions a significant difference could not be indicated between both versions. What is striking is the difference between task 1 and the rest. Only in this task, we see a median perceived ease of use of 1 for both versions. With other words, respondents found it easy to come up with textual association, regardless of if they were correct in the given context.

The perceived ease of use is slightly better in task 2, 3 and 4 for the newly proposed version (mean 4) versus the initial version (mean 5 for all tasks). However, the statistical tests show no significant difference. Because no significant results have been found H03: The overall perceived ease of use all tasks of the newly proposed PGA modelling notation is higher than the initial PGA modelling notation, could not be accepted.

For the overall ease of use questions, the newly proposed version (median 4) also outperforms the initial version (median 4,75)(see Table 14). Overall, the perceived ease of both versions scores on the less positive side of the Likert-scale for understandability, frustration, easiness of use in practise and easiness of the learning curve. Despite the negative scores, the newly proposed version performs significantly (see Table 15) better than the initial version (P-value 0,010). This is based on the overall usage (P-Value 0,016) and the overall learning question (P-value 0,011), meaning hypotheses H04: The overall perceived ease of use of the newly proposed PGA modelling notation is higher than the initial PGA modelling notation, could be accepted.

Table 15: Significant results of ease of use variables

Dependent Variable	Performed test	Mean (rank) initial version	Mean (rank) newly proposed version	two-tailed P-Value	one-tailed P-Value	Accepted Hypothesis
Total Overall Ease of use (median)	**T-Test**	**4,900**	**3,818**	**0,02**	**0,010**	**H_a**
Ease of use: overall usage	Mann-Whitney U	13,9	8,36	0,031	0,016	H_a
Ease of use: overall learning	T-Test	4,800	3,450	0,021	0,011	H_a

4.4. Analysing learning effect effectiveness

For effectiveness, learning effect analyses will be performed both for (1) the tasks and (2) the icons. When analysing the data, a positive pattern (two or more increasing scores) can be found for the newly proposed version on task level (see Table 16).

Table 16: Descriptive statistics learning effect effectiveness (task level)

Dependent Variable (effectiveness)	Percentage scores correct answers initial version	Percentage scores correct answers newly proposed version	Learning effect initial version (with previous task)	Learning effect newly proposed version (with previous task)
Subtotal Effectiveness: task 1	1%	11%	-	-
Subtotal Effectiveness: task 2	30%	52%	29%	41%
Subtotal Effectiveness: task 3	23%	56%	-7%	4%

The data show that the increase of correct answers is the highest from tasks 1 to 2 for both versions. However, the highest increase (+41%) is related to the newly proposed version. Above analyses are indicating that the newly proposed version performs better in learning effect. However, no significant results have been found in statistical analyses on task level.

Table 17: Descriptive statistics learning effect effectiveness (icon level)

Dependent Variable (effectiveness)	Percentage scores correct answers initial version	Percentage scores correct answers newly proposed version	Conversion initial version (with previous task)	Conversion newly proposed version (with previous task)
Effectiveness: Competence, task 1	0%	0%	-	-
Effectiveness: Competence, task 2	30%	27%	30%	27%
Effectiveness: Competence, task 3	10%	45%	-20%	18%
Effectiveness: Customer goal, task 1	5%	0%	-	-
Effectiveness: Customer goal, task 2	60%	82%	55%	82%
Effectiveness: Customer goal, task 3	40%	64%	-20%	-18%
Effectiveness: Importance, task 1	0%	64%	-	-
Effectiveness: Importance, task 2	10%	45%	10%	-19%
Effectiveness: Importance, task 3	0%	64%	-10%	19%
Effectiveness: Internal goal, task 1	0%	0%	-	-
Effectiveness: Internal goal, task 2	30%	27%	30%	27%
Effectiveness: Internal goal, task 3	50%	55%	20%	28%
Effectiveness: Value proposition, task 1	0%	0%	-	-
Effectiveness: Value proposition, task 2	40%	64%	40%	64%
Effectiveness: Value proposition, task 3	20%	36%	-20%	-28%
Effectiveness: Value stream, task 1	0%	0%	-	-
Effectiveness: Value stream, task 2	10%	64%	10%	64%
Effectiveness: Value stream, task 3	20%	73%	10%	9%

When analysing the data, the following positive trends can be found:

- New notation: Competence, task 1 -> 2 +27%, task 2 -> 3 +18% (total 45%)
- Initial notation: Internal goal, task 1 -> 2 +30%, task 2-> 3 +20% (total 50%)
- New notation: Internal goal, task 1 -> 2 +27%, task 2-> 3 +28% (total 55%)
- Initial notation: Value stream, task 1 -> 2 +10%, task 2-> 3 +10% (total 20%)
- New notation: Value stream, task 1 -> 2 +64%, task 2 -> 3 +9% (total 73%)

The positive trend for competence can only be found for the newly proposed version (see Table 17). However, only one significant result has been found for the icon Value stream (see Table 18).

Therefore both hypotheses, H05: The learning effect of the newly proposed version is higher than the initial version for interpretational effectiveness of task 2 versus task 1, and H06: The learning effect of the newly proposed version is higher than the initial version for interpretational effectiveness of task 3 versus task 2, cannot be accepted.

Table 18: Significant results of learning effect variables effectiveness

Dependent Variable	Performed test	Mean (rank) initial version	Mean (rank) newly proposed version	two-tailed P-Value	one-tailed P-Value	Accepted Hypothesis
Learning effect: Effectiveness Value stream, task 2 vs 1	Mann-Whitney	8,05	13,68	0,014	0,007	H_a

4.5. Analysing learning effect efficiency

Table 19: Descriptive statistics learning effect efficiency (task level)

Dependent Variable (effectiveness)	Score initial version	Score newly proposed version	Learning effect initial version	Learning effect newly proposed version
Efficiency: task 1	141,13	289,64	-	-
Efficiency: task 2	81,22	122,46	-59,91	-167,18
Efficiency: task 3	150,32	196,84	+69,1	+74,38

Table 19 shows a decrease in time effort taken for task 1 to task 2 for both versions, however an increase in time is shown for the task 2 to task 3 conversion. Also, no significant results have been found. Therefore hypotheses H07: The learning effect of the newly proposed version is higher than the initial version for interpretational efficiency of task 2 versus task 1, and H08: The learning effect of the newly proposed version is higher than the initial version for interpretational efficiency of task 3 versus task 2, cannot be accepted.

4.6. Analysing learning effect ease of use

Table 20: Descriptive statistics learning effect ease of use (task level)

Dependent Variable (effectiveness)	Score initial version	Score newly proposed version	Conversion initial version (with previous task)	Conversion newly proposed version (with previous task)
Ease of use: task 1	1	1	-	-
Ease of use: task 2	5	4	+4	+3
Ease of use: task 3	5	4	0	0

Table 20 shows a negative step for the ease of use perception for task 1 to task 2 for both versions, however a neutral step is shown for the task 2 to task 3 conversion. Also, no significant results have been found. Therefore hypotheses H09: The learning effect of the newly proposed version is higher than the initial version for perceived ease of use of task 2 versus task 1, and H10: The learning effect of the newly proposed version is higher than the initial version for perceived ease of use of task 3 versus task 2, cannot be accepted.

4.7. Confounding effects

Table 21: Significant results of demographic variables on dependent variables

Dependent Variable	Performed test	Function P-value	Total years of experience P-value	Models read P-Value	Models made P-Value	Hypothesis accepted
Total Efficiency: all tasks	**One-way ANOVA**	**0,540**	**0,005**	**0,918**	**0,775**	**H_a**

Additional testing (see Appendix 10) was performed to analyse whether we can find any confounding effects of the demographic variables in case of significant differences between the newly proposed and the initial version. For the normally distributed variables with multiple groups, the One-way ANOVA test was performed, for the non-normally distributed variables with multiple groups, the Kruskal Wallis test was performed. The analysis was executed for the following demographic variables (see Table 21):

- Function
- Total years of experience
- Models read
- Models made

For the dependent variable 'Total efficiency: all tasks' a significant influence was found for the demographic variable 'Total years of Experience' (P-value 0,005). However, this does not apply to other variables with significant differences on the icon level. Also, further investigation of the data per experience group does not show a pattern of increasing efficiency with increasing years of experience. No other significant influence was found for the other demographic variables.

5. Discussion, conclusions, and recommendations

5.1. Discussion – Reflection

For this research two survey parts were conducted to investigate if the newly proposed version of the PGA notation is performing better than the initial version. A demographic part to enable a strategic sampling strategy and a research part following a between-subjects' design to obtain data about (1) the newly proposed version and (2) the initial version. A total of 21 respondents completed both parts.

5.1.1. Construct validity

Intuitiveness is not a commonly used word in literature. Based on the literature review three proxy indicators have been found for intuitiveness in effectiveness, efficiency, and ease of use.
With these three indicators a survey had been made by the researcher. After a final concept was reached, an overall consistency check as performed by an expert. Having only access to one expert is an absolute minimum for this purpose. However, discussions about the correctness of answers were possible and led to new understandings (i.e. that the PGA notation does not provide an answer finding the largest improvement potential, or questions about the icon Value stream were too beneficial towards the newly proposed version).

5.1.2. Internal validity

The structure of adding more information to tasks 1, 2 and 3 is beneficial toward the internal validity. This, because the questions are staying the same and bias will therefore be minimized. Each task is asking the same question, only with more context information given. The critical part of this survey was the interpretation of the demographic questions. One respondent asked the researcher what a conceptual model was. Because this question was asked after part 1, the researcher was able to add some examples (besides the already given definition), however, part one already had been submitted. Also, combining multiple icons in some questions in task 4, made result comparison with tasks 1-3 not possible. For hypotheses H02, H07 and H08 one cannot claim that when the time taken in task 1 is longer than task 2, task 2 is more efficient, because in task 1 respondent had to manually give textual association (up to three) in words, while in task 2 the icons and words only had to be linked. Also, comparing tasks 2 and 3 is not correct. In task 3 extra context was given in the way of a case model. Taking this information into mind will use more time.

5.1.3. External validity

The results of this research are representative for the whole of the ministry and could be generalized. Because of the case model used (expense management), research results could also be generalized toward all ministries of the Kingdom of the Netherlands as the expense management process is the same for all ministries. The survey could be repeated without adjustments within other departments of the ministry of Education, Culture and Science. Also, this survey could be used without adjustment in all other ministries. With slight adjustments this survey would also be reusable in a business setting. The adjustments should then be made in some demographic questions (like functions) and maybe the case model.

5.1.4. Reliability

Before respondent could take part in the survey, a group of non-respondents was asked to test the survey, to test if no errors or uncertainties came up. When the test group indeed found some flaws in both surveys, these were adjusted before the survey was set.

5.1.5. Triangulation

The domain knowledge of the respondents could be subjectively determined because the researcher knew all respondents. Next to this the demographic question also gave an objective measurement. Also, three proxy indicators for intuitiveness had been measured.

5.1.6. Ethical aspects

For the ethical aspect, we needed to weigh anonymity and performing a strategic sampling to create two identical representative groups after the demographic part of the survey. With the knowledge of the researcher about LimeSurvey, a combination was not possible. Because the researcher knew all respondents and no political or other sensible questions were asked, it was decided to only request a name in both versions to be able to perform strategic sampling and to be able to link the results of both surveys. After survey 2 had been finished by all respondents, the data was linked to each other, and the respondents' data were anonymized. Possibly, this strategy was also beneficial to finding no overall influence on the demographic variables on the dependent variables.

5.2. Conclusions

In this research an answer has been found to the main research question: '*How can the intuitiveness of the PGA notation be evaluated by conducting an experiment within a practical context of a governmental organization?'.* For this a quantitive research has been conducted with the use of a between-subjects' design survey within a governmental organization, namely the ministry of Education, Culture and Science.

To be able to answer the main question four sub questions had been formulated:

Sub question 1	What is known in the academic literature about evaluating the intuitiveness of PGA in the context of DSML in governmental organizations (TSQ)

Based on the literature review, it can be concluded that intuitiveness is not a commonly used word in de academic world. Three proxy words have emerged from the review, namely, effectiveness, efficiency, and ease of use. Based on these key words, relevant papers were found concerning the evaluation of conceptual models. A governmental context was not considered in this body of literature.

Sub question 2	How are two PGA modelling notation versions (initial and newly proposed) applied to a process in a governmental organization? (ESQ)

The selection of the expense management case ensured no detailed domain knowledge which could have a bias on the results (Gemino & Wand, 2004). The possible learning effect of showing both versions to a respondent was avoided by choosing a between-subjects' design. This choice was made because only half of the icons in both versions differ. Based on my own experiences of expense management processes in both commercial as governmental environments a case study was made for the experiment.

Sub question 3	How are two applied PGA modelling notation versions evaluated on intuitiveness among governmental-oriented end-users? (ESQ)

With full respect for the anonymity of the respondents a two-part survey was conducted to perform strategic sampling, which led to perfectly balanced representative groups.

Sub question 4	How can initial and proposed PGA modelling notation be selected on intuitiveness? (ESQ)

A significant result was only found for the effectiveness indicator, which means that the newly proposed version is performing better than the initial version. Drilling down into the results, we found that this can be explained by the variables Importance, Value stream and Competence. For Importance this is true for all tasks, for Value stream for tasks 2-4 and for competence for tasks 3-4. This means that the results for Importance show the most significant difference over all tasks, followed by Value stream and Competence. If there was no usage of the proxy indicators, effectiveness, efficiency, and ease of use for intuitiveness, the significant results were probably worse.

For efficiency, the effects were beneficial towards the initial version. This is not in line with the expectations of this research. For the ease of use on task level, there are no significant results. This was not in line with the expected results.
For the overall ease of use we did find a significant result. This can be explained by two of four questions, namely overall usage, and learning.

For the first three tasks we investigated the possibility of a learning effect between the tasks as well. This because only one context element (task 1; only icon, task 2; icons and textual associations and task 3; icons, textual associations, and a case study) had been added and the questions were all the same. However, no significant results were found.

From above significant results and the maximum percentage scores for task 1-3 (56% newly proposed version, 30% initial version) and task 4 (37% newly proposed version, 22% initial version), a conclusion can be drawn that both versions are not well understood by the respondents. The PGA notation will need improvements to be able to pass the intended information from designer towards stakeholder (Norman, 1986). Also, the results from the overall ease of use question are indicators that respondents do not want to use the PGA notation in its current form.

Two of ten hypotheses led to significant results. If intuitiveness could be represented by only one proxy indicator, then there are indicators, tasks and icons that show a significant result beneficial to the newly proposed version. However, we believe that a combination of significant results for all proxy indicators, effectiveness, efficiency, and ease of use can only lead toward the answer that the newly proposed version is performing better than the initial version.

Main research question	How can the intuitiveness of the PGA notation be evaluated by conducting an experiment within a practical context of a governmental organization?

From this quantitative research the main research question can now be answered. It appears that it is possible to evaluate a conceptual model like the PGA notation on intuitiveness using the dependent variables of effectiveness, efficiency, and ease of use. This has been done by conducting an intragrammar experiment, while evaluating both versions (newly proposed and initial) of the PGA notation making use of representative stakeholders of three departments of a policy ministry of the Dutch government.

5.3. Recommendations for practice

In the opinion of the researcher, the demographic question could be obtained in a different manner. For this research simple questions were asked. However, we cannot be sure if the respondent interpreted the question in the way the researcher intended. To overcome this, we recommend to not simply ask a question about pre-knowledge, but to come up with a test such as asking about

which conceptual models they used in the last year, so the answers would be in line with the intended question asked.

The researcher intended to add extra context information between task 1-3. One could conclude that from tasks 2 to 3 not one piece of context was added (case model) but more pieces were added (descriptions of the icons like 'correctness' was given, next to the hierarchy map). Also, in task 4 questions about one icon and about multiple icons were used in questions. It is therefore recommended to add two more tasks in this structure. The new task schedule then would be:

- Task 1: Notation association without provided terms
- Task 2: Notation association with provided terms
- Task 3: Notation association with provided terms and icon description
- Task 4: Notation association with provided terms, icon description and case study
- Task 5: Case study association with questions about one icon
- Task 6: Case study association with questions about two or more icons

Before respondents were invited to take place in both survey parts, the survey parts were tested on technical flaws and unclarity by seven individuals with different backgrounds. Because still some minor questions remained, it is recommended to employ a larger group of testers, which will represent the actual respondents in a better way.

From the ethical point of view, the choice has been made to ask for the respondents' names, so that the strategic sampling could be performed and the researcher was able to link the results of survey parts 1 and 2. A technical solution should be designed to resolve this ethical choice.

5.4. Recommendations for further research

The results of this research show a rather low percentage score on correct answers for all tasks. Three to five out of six icons did not significantly perform better in all tasks.

For this we see three possible explanations that should be taken into further research.

1. Governmental usage of PGA notation
2. Cultural aspects between Belgians and Dutch need to be investigated
3. All but one icon that differ between both versions need to be altered

The third explanation is maybe the obvious one to start further research on. We recommend to first investigate if the governmental usage is biasing the results towards the PGA notation. The PGA notation has after all emerged from a business perspective (Roelens et al., 2019). In governmental perspective one speaks of civilians and organization, rather than using the term customers. Therefore maybe the icon customer goal will be less understood than the icon will score in a business setting.
While investigating possible processes for the case model, it seemed that the PGA notation is not completely usable within the governmental context. For the government legal goals are particularly important. However, the current version of the PGA notation does not provide a legal goal.
This can be done by conducting the experiment in this paper in a purely business context, by altering only the functions in the demographic part of the experiment and comparing the results with the results of this research.
We also recommend performing further research towards cultural influences on the results of this research. The icons originated from Belgian students and were applied into Dutch governmental context. This research could be done by conducting the experiment in this research for the Belgian equivalent of the Dutch ministry of Education, Culture and Science and compare the results with the results of this research.

With the results of above-described research in hand we recommend to conduct the experiment of Roelens and Bork (2020) again for the five icons that had no significant result in all tasks, with both business and governmental users to come up with new icons.

The interpretational efficiency variables were not significantly better for the newly proposed version, than for the initial version. With some variables the initial version is significantly better than the newly proposed version. This was not in line with the expectation for this research. It is also not in line with the results for the other indicator effectiveness. Therefore research is recommended to find an explanation for this. Could it be as simple that for the initial version questions were so hard to answer, respondents gave an answer just to go to the next one? An indicator for this explanation is the rated ease of use for the initial version, which is slightly worse than for the newly proposed version. However, this is not significant. But maybe other explanations can be found in research. This could be done by providing a 'do not know' option for all tasks instead of only task 4.

Also, no significant difference was found for ease of use for all tasks. One explanation could be that a 7-point Likert scale cannot measure the ease of use well enough to investigate significance. Maybe a two-step ease of use question could help finding significance. The first question could then be, 'how easy was it to answer the questions of task x', with answer possibilities such as easy, medium, hard. The second question could be, 'how easy/medium/hard was it to answer the questions of task x, with a 7-point Likert scale as answer possibility from extremely easy/medium/hard to somewhat easy/medium/hard. Another explanation could be found in what it is the respondent points out with the given Likert scale answer. Is it the easiness of the PGA notation or the easiness of the performed task besides the PGA notation? Maybe the task is easy, and the answer is easy, but the question itself is the reason a respondent finds a particular task not easy to perform.

A significant link has been found between total years of experience and the dependent variable Efficiency all tasks. The data does not provide a conclusive answer if there is a causality between the more experience you have, the faster one could answer the questions, or that another explanation is in place. From the data no pattern could been found. However, future research could dive deeper into this result.

To be able to generalize results beyond ministry or government a fictional case model was used in task 4. It should be investigated further if a fictional case model has a negative or positive influence on the results, before using this model in further practise.

Looking back at the research we have tried to create an experiment that can measure the intuitiveness of a full notation. Diving deeper into the results, this experiment gives more insight if individual icons are more intuitive than other icons and less insight if the full notation is more intuitive than another version. Future research should focus on how a full notation could be evaluated on intuitiveness, so it is able to deliver the intended message from the creator of the model towards the reader. Task 4: case association could give a good starting point for this research.

This research could be reused on the same conceptual model, intragrammarly, in a different organizational context. It could also be reused to research intragrammarly a different conceptual model in the same context, and it could also be reused to research intergrammarly two different conceptual models in the context of both governmental and business organizations.

References

Allen, G. N., & March, S. T. (2012). A Research Note on Representing Part-Whole Relations in Conceptual Modeling. *MIS quarterly, 36*(3), 945-964. doi:10.2307/41703488

Bera, P., Burton-Jones, A., & Wand, Y. (2017). Improving the representation of roles in conceptual modeling: theory, method, and evidence. *Requirements engineering, 23*(4), 465-491. doi:10.1007/s00766-017-0275-9

Bork, D., Schrüffer, C., & Karagiannis, D. (2019). *Intuitive understanding of domain-specific modeling languages: proposition and application of an evaluation technique.* Paper presented at the International Conference on Conceptual Modeling.

Burton-Jones, A., & Meso, P. N. (2006). Conceptualizing Systems for Understanding: An Empirical Test of Decomposition Principles in Object-Oriented Analysis. *Information Systems Research, 17*(1), 38-60. doi:10.1287/isre.1050.0079

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS quarterly, 13*(3), 319-340. doi:10.2307/249008

Frank, U. (2013). Domain-specific modeling languages: requirements analysis and design guidelines. In *Domain Engineering* (pp. 133-157): Springer.

Gailly, F., Alkhaldi, N., Casteleyn, S., & Verbeke, W. (2017). Recommendation-Based Conceptual Modeling and Ontology Evolution Framework (CMOE+). *Business & information systems engineering, 59*(4), 235-250. doi:10.1007/s12599-017-0488-y

Gelderman, C. J. (2013). Checklist case study-onderzoek - ontwerpvragen en verantwoording van keuzes. *Heerlen, Open Universiteit*.

Gemino, A., & Wand, Y. (2004). A framework for empirical evaluation of conceptual modeling techniques. *Requirements engineering, 9*(4), 248-260. doi:10.1007/s00766-004-0204-6

Gorla, N., Chiravuri, A., & Meso, P. (2012). Effect of personality type on structured tool comprehension performance. *Requirements engineering, 18*(3), 281-292. doi:10.1007/s00766-012-0158-z

Haisjackl, C., Barba, I., Zugal, S., Soffer, P., Hadar, I., Reichert, M., . . . Weber, B. (2016). Understanding Declare models: strategies, pitfalls, empirical results. *Software and systems modeling, 15*(2), 325-352. doi:10.1007/s10270-014-0435-z

Henderson, J. C., & Venkatraman, H. (1999). Strategic alignment: Leveraging information technology for transforming organizations. *IBM systems journal, 38*(2.3), 472-484.

Jansen, B. J. P. (2013). *Het uitvoeren van wetenschappelijk literatuuronderzoek*: Open Unversiteit.

Jošt, G., Huber, J., Heričko, M., & Polančič, G. (2016). An empirical investigation of intuitive understandability of process diagrams. *Computer standards and interfaces, 48*, 90-111. doi:10.1016/j.csi.2016.04.006

Lankhorst, M. (2017). *Enterprise Architecture at Work: Modelling, Communication and Analysis*: Springer.

Luoma, J., Kelly, S., & Tolvanen, J.-P. (2004). *Defining domain-specific modeling languages: Collected experiences.* Paper presented at the 4 th Workshop on Domain-Specific Modeling.

Mendling, J., Recker, J., Reijers, H. A., & Leopold, H. (2018). An Empirical Review of the Connection Between Model Viewer Characteristics and the Comprehension of Conceptual Process Models. *Information systems frontiers, 2018*(5), 1-25. doi:10.1007/s10796-017-9823-6

Michael, J., & Mayr, H. C. (2017). *Intuitive understanding of a modeling language.* Paper presented at the Proceedings of the Australasian Computer Science Week Multiconference.

Moody, D. (2009). The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering. *IEEE Transactions on software engineering, 35*(6), 756-779.

Moore, G. C., & Benbasat, I. (1991). Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation. *Information Systems Research, 2*(3), 192-222. doi:10.1287/isre.2.3.192

Norman, D. A. (1986). Cognitive engineering. *User centered system design, 31*, 61.

Roelens, B., & Bork, D. (2020). An Evaluation of the Intuitiveness of the PGA Modeling Language Notation. In *Enterprise, Business-Process and Information Systems Modeling* (pp. 395-410): Springer.

Roelens, B., Steenacker, W., & Poels, G. (2019). Realizing strategic fit within the business architecture: the design of a process-goal alignment modeling and analysis technique. *Software & Systems Modeling, 18*(1), 631-662.

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students 8th edition*.

Shanks, G., Tansley, E., Nuredini, J., Tobin, D., & Weber, R. (2008). Representing Part-Whole Relations in Conceptual Modeling: An Empirical Evaluation. *MIS quarterly, 32*(3), 553-573. doi:10.2307/25148856

The Open Group. (2019). *ArchiMate®3.1 Specification*.

Westerkamp, K., & Veen, M. v. (2009). *Deskresearch: Informatie selecteren, beoordelen en verwerken*.

Zugal, S., Soffer, P., Haisjackl, C., Pinggera, J., Reichert, M., & Weber, B. (2013). Investigating expressiveness and understandability of hierarchy in declarative business process models. *Software and systems modeling, 14*(3), 1081-1103. doi:10.1007/s10270-013-0356-2

Appendix 1: Search strategy (added in Excel format)

Appendix 1 Search strategy graduation

Appendix not presentable in Word. See attached Excel file.

Appendix 2: Sent e-mails regarding survey part 1 (Dutch)

Van: Kooij, Martin van der
Aan: Kooij, Martin van der
Bcc: Undisclosed
Onderwerp: Afstudeerverzoek
Datum: vrijdag 9 april 2021 15:00:10

Allen,
Via deze email wil ik vragen om 45 minuten van je tijd om mij te helpen met mijn afstudeeronderzoek naar een conceptuele modelleertaal. Van deze 45 minuten vraag ik 5 minuten komende week en 40 minuten in de week erna.
Na ruim twee jaar is het einde van mijn opleiding Business Process Management & Information Technology aan de Open Universiteit in zicht. Ik ben al ver gevorderd met mijn thesis en ben in de fase aangekomen om data te verzamelen. In mijn geval via een tweetal enquêtes. Via deze enquêtes hoop ik antwoord te krijgen op de vraag welke versie van een conceptuele modelleertaal (door mijn tutor ontwikkeld) beter te begrijpen is voor medewerkers van een overheidsorganisatie. Om bias te voorkomen wordt pas in deel twee duidelijk om welke conceptuele modelleertaal het gaat.
Enquête deel één vraag ik je in te vullen als je mee wilt doen aan beide enquêtes. In deel één verzamel ik demografische gegevens. Dit deel zal naar verwachting maximaal 5 minuten duren en geeft mij de mogelijkheid de deelnemers strategisch onder te verdelen in meerdere vervolggroepen.
Nadat deel één is afgerond en de vervolggroepen zijn samengesteld, zal ik aan elke groep een link sturen voor deel twee. Dit deel duurt naar verwachting maximaal 40 minuten.

Mocht je mij willen helpen dan kun je via deze link deelnemen aan deel één van de enquête. Deze link zal bruikbaar zijn tot en met 16 april. Deel één is prima uit te voeren op een mobiel apparaat.
Schroom niet om mij te bellen/mailen bij vragen. Voor nu, goed weekend!
NB. De enquêtes zijn volledig in het Engels. Deelnemen aan deel één of aan deel twee als je aan deel één hebt deelgenomen is GEEN verplichting. Je hebt te allen tijde de mogelijkheid om zonder opgave van redenen te stoppen, voor of tijdens het onderzoek. Tot slot wil ik meegeven dat de data die moet worden bewaard in het kader van mijn opleiding volledig geanonimiseerd zal worden.

Met vriendelijke groet,

Martin van der Kooij
Subsidieprocesmanager
.....
Erfgoed en Kunsten
Ministerie van Onderwijs, Cultuur en Wetenschap
Rijnstraat 50 | 2515 XP Den Haag | 7e verdieping
Postbus 16375| 2500 BJ Den Haag
.....
[redacted]
http://www.minocw.nl/
.....

Appendix 3: Survey part 1 both versions

Part 1

Welcome to this survey, which aims to perform a comparative evaluation of two versions of a conceptual modelling language. Conceptual modelling languages are used to describe the enterprise architecture of organisations. Conceptual modelling concerns the application of abstraction to reduce the complexity of a certain domain for a specific stakeholder purpose. Some examples of a conceptual model are policy theory models, process models and organizational charts. If you are asked the question if you ever had attended a course to read or write those models, the answer is probably, no. One of the reasons for this is that the elements used are intuitive to understand. In previous research an evaluation on the intuitiveness of an original conceptual modelling language has been performed. Based on the input of the respondents, flaws in the existing conceptual modelling language were identified and an improved notation was developed. For bias reasons we cannot disclose at this point which conceptual modelling language this research will be used.

The purpose of this research is to obtain a validation of the proposed improvements. Therefore, we set up an experiment that will be held in two parts.

Part one (this part) aims to obtain some demographic information about you as respondent on which strategic sampling can be performed.

After the samples have been made, you will be invited to the second part. This part will contain four tasks, in which the intuitiveness of the notation alternatives is compared. To avoid any bias, one sampling group will receive questions and cases based on the original version of this conceptual modelling language, the other sampling group will receive questions about the newly proposed version of this conceptual modelling language. The respondent will not know in which group they are participating.

If you have any questions about this survey, you can contact the responsible researcher:

- Martin van der Kooij, [redacted], [redacted]

This part of the survey will approximately take 5 minutes.

There are 8 questions in this survey.

Declaration of consent for participation in scientific research

- *I have been informed about part one of the research and I have read the given context information.*
- *I know how to ask questions about the research.*
- *I have been able to think about my participation to part one of the study.*
- *I understand that I can exit the investigation at any time, and I do not have to give a reason for it.*
- *I consent to the use of the data collected during part one in the research for this scientific research.*
- *I understand that all information I provide regarding this study will be collected will be anonymized after part two of this research and then will not lead back to me. When you do participate in part one, but for whatever reason will not be participating part two, also all collected data will be anonymized and will not lead back to you.*
- *I understand that the collected data of part one is kept securely for ten years.*

If you have read the above points and agree to participate in the study, please click Yes AND proceed to the Next page (button at the bottom right).

*

Please choose **only one** of the following:

- Yes
- No

What is your full name? *
Please write your answer here:
Please fill in your year of birth: *
Please write your answer here:

-

What is your current function/role?

ATTENTION: When you have a double function/roll, please select the function/roll you spend most of your time on.
*
Please choose **only one** of the following:

- Policy officer & Account manager
- Policy officer of team policy information
- Finance/Control employee
- Director/Manager
- Architect (Domain/Information)

How many years of experience do you have in your current function? *
Please write your answer here:

-

How many years of experience do you have had in the past (experience equal to your current role)? *
Please write your answer here:

-

How many conceptual models have you **read** by average during the last year (e.g. policy theory models, process models and organizational charts)? *
Please choose **only one** of the following:

- 0
- 1-5
- 6-15
- >15

How many conceptual models have you **made** by average during the last year (e.g. policy theory models, process models and organizational charts)? *
Please choose **only one** of the following:

- 0
- 1-5
- 6-15
- >15

Thank you very much for your participation in this part of the survey. If you want more information about this study, you can contact the responsible researcher:

- Martin van der Kooij, [redacted], [redacted]

16.04.2021 – 20:32 Submit your survey. Thank you for completing this survey.

Appendix 4: Strategic sampling (added in Excel format)

Appendix%204%20Results%20Part%201%

Appendix not presentable in Word. See attached Excel file.

Appendix 5: Sent e-mails regarding survey part 2

Van: Kooij, Martin van der
Aan: Kooij, Martin van der
Bcc: Undisclosed
Onderwerp: Afstudeeronderzoek deel 2
Datum: vrijdag 16 april 2021 22:29:27

Beste collega,

Wat fijn dat je deel één van mijn enquête hebt willen invullen.
Mocht je mij ook willen helpen met het invullen van deel twee van de enquête kun je via deze link deelnemen. Deze link zal bruikbaar zijn tot en met 23 april. Deel twee is alleen goed uit te voeren op een desktop/laptop.
Schroom niet om mij te bellen/mailen bij vragen. Voor nu, goed weekend!
NB. De enquêtes zijn volledig in het Engels. Deelnemen aan deel twee is GEEN verplichting. Je hebt te allen tijde de mogelijkheid om zonder opgave van redenen te stoppen, voor of tijdens het onderzoek. Tot slot wil ik meegeven dat de data die moet worden bewaard in het kader van mijn opleiding volledig geanonimiseerd zal worden.

Met vriendelijke groet,

Martin van der Kooij
Subsidieprocesmanager
.....
Erfgoed en Kunsten
Ministerie van Onderwijs, Cultuur en Wetenschap
Rijnstraat 50 | 2515 XP Den Haag | 7e verdieping
Postbus 16375| 2500 BJ Den Haag
.....
[redacted]
http://www.minocw.nl/
.....

Appendix 6: Survey part 2 both versions

Part 2: PGA Notation

Welcome to this survey, which aims to perform a comparative evaluation of two versions of the Process-Goal Alignment (PGA) notation. PGA is a conceptual modelling language oriented towards the achievement of strategic fit, which is the translation of the business strategy into concrete activities and results.

In previous research an evaluation of the intuitiveness of the original PGA notation has been performed. Based on the input of the respondents, flaws in the existing notation were identified and an improved notation was developed.

The purpose of this research is to obtain a validation of the proposed improvements. Therefore, we set up an experiment that will be held in two parts.

Part one has already been performed by you.

Part two (this part) will contain four tasks, in which the intuitiveness of the notation alternatives is compared. To avoid any bias, one sampling group will receive questions and cases based on the original version of the PGA notation, the other sampling group will receive questions about the newly proposed version of the PGA notation.

You will not know in which group you are participating.

If you have any questions about this survey, you can contact the responsible researcher:

- Martin van der Kooij, [redacted], [redacted]

According to the test phase of this survey, the average time to complete this survey is 25 minutes, with a maximum of 40 minutes.

There are 33 questions in this survey.

Declaration of consent for participation in scientific research

- *I have been informed about part one of the research and I have read the given context information.*
- *I know how to ask questions about the research.*
- *I have been able to think about my participation to part one of the study.*
- *I understand that I can exit the investigation at any time, and I do not have to give a reason for it.*
- *I consent to the use of the data collected during part two in the research for this scientific research.*
- *I understand that all information I provide regarding this study will be collected will be anonymized after part two of this research and then will not lead back to me.*
- *I understand that the collected data of part one is kept securely for ten years.*

If you have read the above points and agree to participate in the study, please click Yes AND proceed to the Next page (button at the bottom right). *

Please choose **only one** of the following:

- Yes
- No

What is your full name?

ATTENTION! Will only be used to link answers survey part 1 and part 2 together. Will be anonymized before analysing phase of data (directly after submitting part 2).

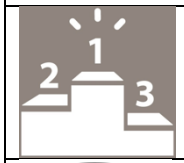
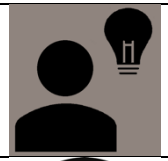
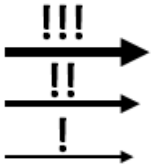
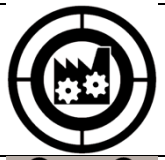
*
Please write your answer here:

Did you had prior knowledge of the PGA notation before participating to part one or two of this experiment? *
Please choose **only one** of the following:

- Yes
- No

In this first experimental task, we would like you to fill in which textual association you have to a given icon from the PGA notation. Please fill in up to three possible associations that comes to mind. At least one textual association is mandatory.

ATTENTION! Please be aware that the time you take to answer the question is recorded. If you need a break during this experiment, please use the task introductions.
Click next to proceed to the first experimental task if you are ready. After clicking, you cannot return to this information.

Initial version	Newly proposed version		Textual association 1	Textual association 2	Textual association 3
		Please fill in at least one answer			
		Please fill in at least one answer			
		Please fill in at least one answer			
		Please fill in at least one answer			
		Please fill in at least one answer			
		Please fill in at least one answer			

Overall, it was easy for me to come up with one, two or three textual associations for the given icons.
*

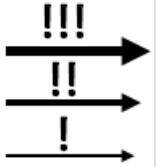
Please choose the appropriate response for each item:

	Fully agree	Agree	Somewhat agree	Neutral	Somewhat disagree	Disagree	Fully disagree

In this second experimental task, you will be presented the same icons as in the first task. This time all icons and all correct textual associations (according to PGA notation) will be given. The order is randomized. You will be asked to link an icon to its correct textual association.

ATTENTION! Please be aware that the time you take to answer the question is recorded. If you need a break during this experiment, please use the task introductions.
Click next to proceed to the first experimental task if you are ready. After clicking, you cannot return to this information.

Please link the icon to the textual association.
*
Please choose the appropriate response for each item:

Initial version	Newly proposed version	**Importance**	**Value Proposition**	**Customer Goal**	**Competence**	**Value Stream**	**Internal Goal**
							

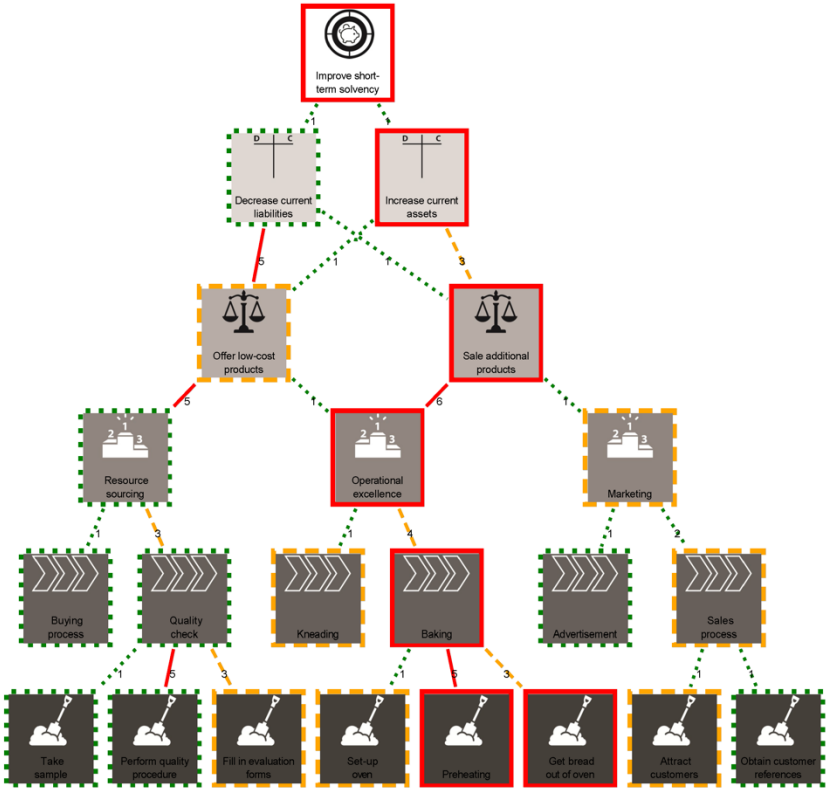
Overall, it was easy for me to come up with a link between the given icons and textual associations.
*

Please choose the appropriate response for each item:

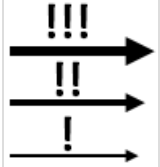
	Fully agree	Agree	Somewhat agree	Neutral	Somewhat disagree	Disagree	Fully disagree

In this third experimental task, you will be presented the same assignment as in the previous task. Except this time also a random selected hierarchy map will be shown.

ATTENTION! Please be aware that the time you take to answer the question is recorded. If you need a break during this experiment, please use the task introductions.
Click next to proceed to the first experimental task if you are ready. After clicking, you cannot return to this information.
Please link the icon to the textual association given the random hierarchy map.



Please link the icon to the textual association.
*
Please choose the appropriate response for each item:

Initial version	Newly proposed version	**Importance**	**Value Proposition**	**Customer Goal**	**Competence**	**Value Stream**	**Internal Goal**
							

Overall, it was easier for me to come up with a link between the given icons and textual associations, because of the given random hierarchy map.
*

Please choose the appropriate response for each item:

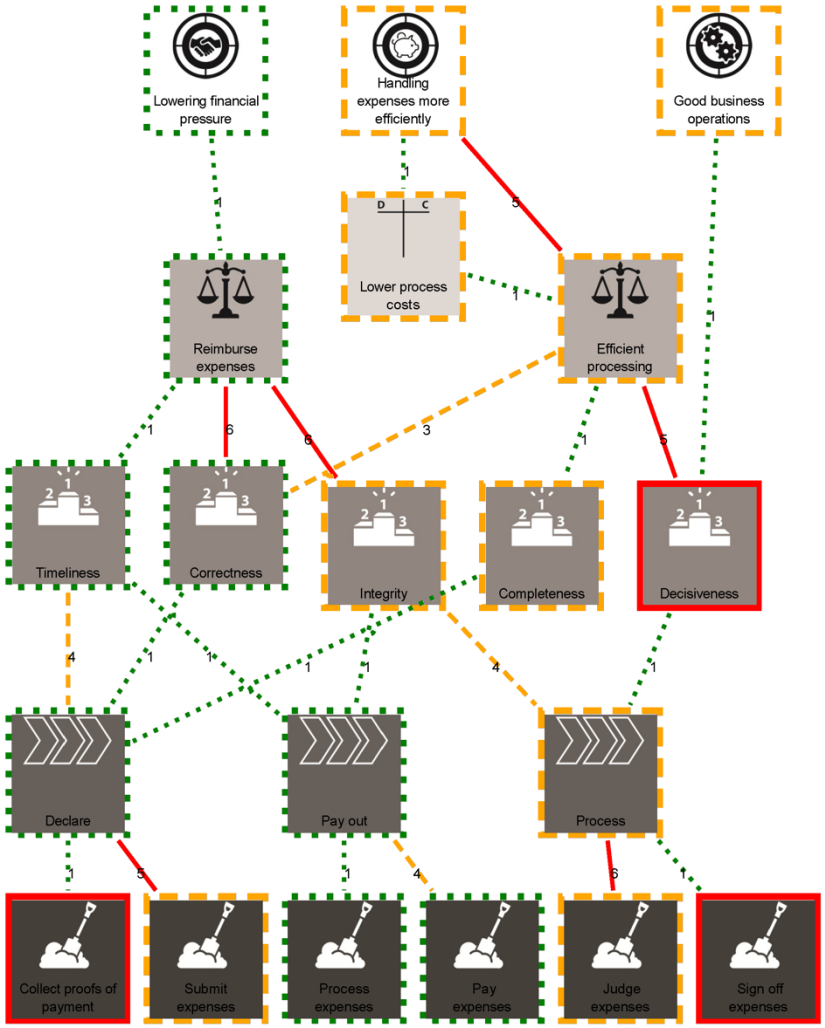
	Fully agree	Agree	Somewhat agree	Neutral	Somewhat disagree	Disagree	Fully disagree

In this fourth and final experimental task, you will be presented with a hierarchy map in a business context. Be aware that that the context is fictional. You will be given ten multiple choice questions about this model. Every question has a "No idea" option. If you cannot come up with an argued answer and are about to gamble for an answer, please use this option.

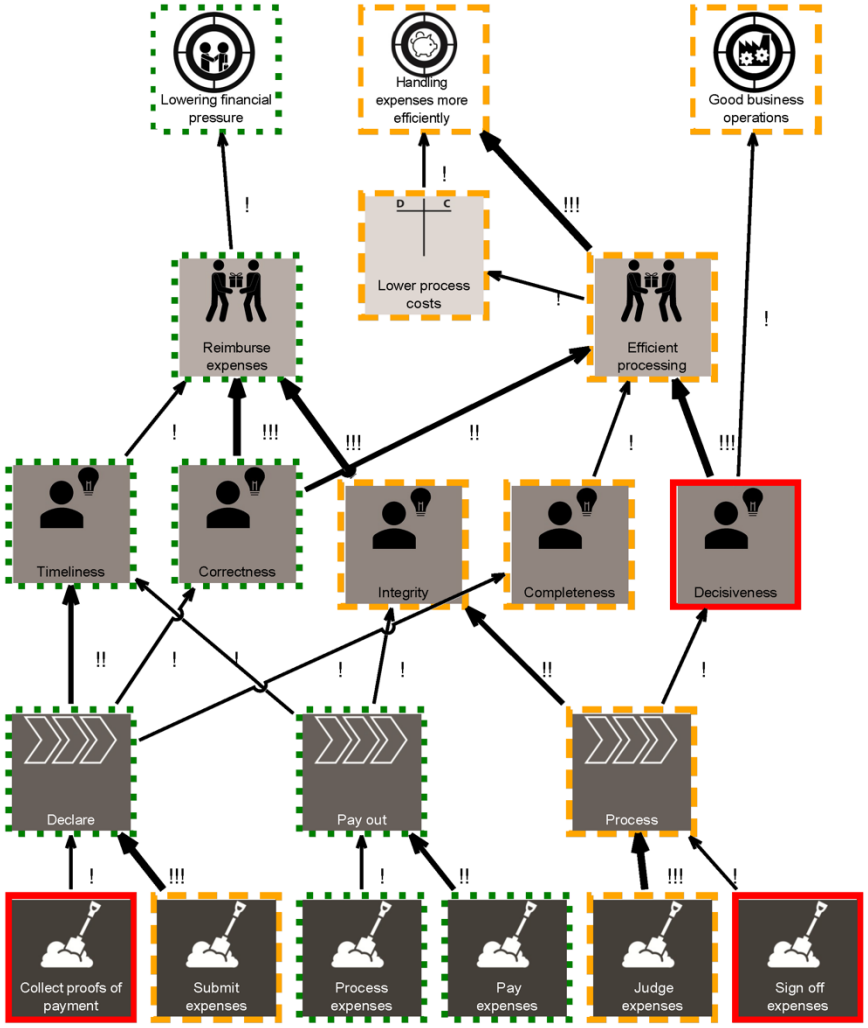
In the given hierarchy map some icons are used that you have not seen so far. Those icons do not differ between the original version and the proposed version of PGA notation. Be aware that these icons will not be needed in the answers you have to give. These icons are needed to give you a fully readable hierarchy map.

ATTENTION! Please be aware that the time you take to answer the question is recorded. If you need a break during this experiment, please use the task introductions.
Click next to proceed to the first experimental task if you are ready. After clicking, you cannot return to this information.

Initial version



Newly proposed version



Question 1:
Which customer goal has been set?
*
Please choose **only one** of the following:

- A: No Idea
- B: Lowering financial pressure
- C: Handling expenses more efficiently
- D: Good business operations
- E: Lower process costs

Question 2:
How many value propositions are shown in this case model?
*
Please choose **only one** of the following:

- A: No idea
- B: 2
- C: 3
- D: 5
- E: 6

Question 3:
Which internal goal has been set?
*
Please choose **only one** of the following:

- A: No idea
- B: Lowering financial pressure
- C: Handling expenses more efficient
- D: Good business operations
- E: Lower process costs

Question 4:
*For which element is the support of "**Efficient processing**" most important?*
*
Please choose **only one** of the following:

- A: No idea
- B: Handling expenses more efficiently
- C: Lower process costs

- D: Completeness
- E: Decisiveness

Question 5:
How many value streams does this case model have?
*
Please choose **only one** of the following:
- A: No idea
- B: 1
- C: 2
- D: 20
- E: 24

Question 6:
How many value streams with low importance on a higher hierarchy layer do the competence(s) support?
*
Please choose **only one** of the following:
- A: No idea
- B: 2
- C: 3
- D: 4
- E: 5

Question 7:
How many goals have a direct connection with value propositions?
*
Please choose **only one** of the following:
- A: No idea
- B: 1
- C: 2
- D: 3
- E: 5

Question 8:
How many value streams on a higher hierarchy level does the value proposition(s) support?
*
Please choose **only one** of the following:

- A: No idea
- B: 3
- C: 6
- D: 7
- E: 9

Question 9:
Where in this case model can you find the largest improvement potential?
*
Please choose **only one** of the following:

- A: No idea
- B: The model does not provide an answer to this question
- C: The path from "**Judge expenses**" directly towards "**Handling expenses more efficiently**"
- D: The path from "**Judge expenses**" via "**Lower process costs**" towards "**Handling expenses more efficiently**"
- E: The path from "**Process expenses**" toward "**Lowering financial pressure**"

Question 10:
*Which type of element is "**Correctness**" in this case model?*
*
Please choose **only one** of the following:

- A: No idea
- B: Competence
- C: Value Proposition
- D: Internal Goal
- E: Value Stream

Overall, it was easy for me to understand and answer the questions about the case model.
*
Please choose the appropriate response for each item:

	Fully agree	**Agree**	**Somewhat agree**	**Neutral**	**Somewhat disagree**	**Disagree**	**Fully disagree**

Overall, it was easy for me to understand which information the PGA model was giving to me.
*
Please choose the appropriate response for each item:

	Fully agree	Agree	Somewhat agree	Neutral	Somewhat disagree	Disagree	Fully disagree

Overall, the usage of PGA frustrated me.
*
Please choose the appropriate response for each item:

	Fully agree	Agree	Somewhat agree	Neutral	Somewhat disagree	Disagree	Fully disagree

Overall, it was easy for me to use the PGA model.
*
Please choose the appropriate response for each item:

	Fully agree	Agree	Somewhat agree	Neutral	Somewhat disagree	Disagree	Fully disagree

Overall, it was easy for me to learn to use the PGA model.
*
Please choose the appropriate response for each item:

	Fully agree	Agree	Somewhat agree	Neutral	Somewhat disagree	Disagree	Fully disagree

Thank you very much for your participation in this part of the survey. If you want more information about this study, you can contact the responsible researcher:

- Martin van der Kooij, [redacted], [redacted]

24.04.2021 – 05:30 Submit your survey. Thank you for completing this survey.

Appendix 7: Results survey part 1 and 2 (added in Excel and SPSS format)

Appendix%207%20Results%20Survey%20F

Appendix 7 Results in SPSS.sav

Appendix not presentable in Word. See attached Excel and SAV file.

Appendix 8: Shapiro Wilk test on all variables

In these tests the following hypotheses are tested

- **H_0 Dependent variable is distributed normally**
- **H_a Dependent variable is not distributed normally**

Dependent Variable	P Value (Shapiro -Wilk)	Accepted Hypothesis
Total Effectiveness: all tasks	**0,922**	**H_0**
Subtotal Effectiveness: task 1	0,000	H_a
Effectiveness: Competence, task 1	0,000	H_a
Effectiveness: Customer goal, task 1	0,000	H_a
Effectiveness: Importance, task 1	0,000	H_a
Effectiveness: Internal goal, task 1	0,000	H_a
Effectiveness: Value proposition, task 1	0,000	H_a
Effectiveness: Value stream, task 1	0,000	H_a
Subtotal Effectiveness: task 2	0,257	H_0
Effectiveness: Competence, task 2	0,000	H_a
Effectiveness: Customer goal, task 2	0,000	H_a
Effectiveness: Importance, task 2	0,000	H_a
Effectiveness: Internal goal, task 2	0,000	H_a
Effectiveness: Value proposition, task 2	0,000	H_a
Effectiveness: Value stream, task 2	0,000	H_a
Subtotal Effectiveness: task 3	0,052	H_0
Effectiveness: Competence, task 3	0,000	H_a
Effectiveness: Customer goal, task 3	0,000	H_a
Effectiveness: Importance, task 3	0,000	H_a
Effectiveness: Internal goal, task 3	0,000	H_a
Effectiveness: Value proposition, task 3	0,000	H_a
Effectiveness: Value stream, task 3	0,000	H_a
Subtotal Effectiveness: task 4	0,122	H_0
Effectiveness: Customer goal, task 4	0,000	H_a
Effectiveness: Value proposition, task 4	0,000	H_a
Effectiveness: Internal goal, task 4	0,000	H_a
Effectiveness: Importance, task 4	0,000	H_a
Effectiveness: Value stream, task 4	0,000	H_a
Effectiveness: Competence, Importance, Value stream, task 4	0,000	H_a
Effectiveness: Customer goal, Internal goal, Value proposition, Value stream, task 4	0,000	H_a
Effectiveness: Value proposition, Value stream, task 4	0,000	H_a
Effectiveness: Importance, Value stream, task 4	0,000	H_a
Effectiveness: Competence, task 4	0,000	H_a
Total Efficiency: all tasks	0,426	H_0
Efficiency: task 1	0,000	H_a
Efficiency: task 2	0,020	H_a
Efficiency: task 3	0,250	H_0
Subtotal Efficiency: task 4	0,796	H_0
Efficiency: Customer goal, task 4	0,046	H_a
Efficiency: Value proposition, task 4	0,000	H_a
Efficiency: Internal goal, task 4	0,384	H_0

Dependent Variable	P Value (Shapiro -Wilk)	Accepted Hypothesis
Efficiency: Importance, task 4	0,985	H_0
Efficiency: Value stream, task 4	0,688	H_0
Efficiency: Competence, Importance, Value stream, task 4	0,000	H_a
Efficiency: Customer goal, Internal goal, Value Proposition, Value stream, task 4	0,005	H_a
Efficiency: Value proposition, Value stream, task 4	0,000	H_a
Efficiency: Importance, Value stream, task 4	0,538	H_0
Efficiency: Competence, task 4	0,381	H_0
Total Ease of use: all tasks (median)	0,369	**H_0**
Ease of use: task 1	0,000	H_a
Ease of use: task 2	0,012	H_a
Ease of use: task 3	0,035	H_a
Ease of use: task 4	0,017	H_a
Total Overall Ease of use (median)	0,055	**H_0**
Ease of use: overall understanding	0,013	H_a
Ease of use: overall frustration	0,028	H_a
Ease of use: overall using	0,013	H_a
Ease of use: overall learning	0,116	H_0
Learning effect: Total Effectiveness task 2 vs 1	**0,331**	**H_0**
Learning effect: Effectiveness Percentage Competence, task 2 vs 1	0,000	H_a
Learning effect: Effectiveness Percentage Customer goal, task 2 vs 1	0,000	H_a
Learning effect: Effectiveness Percentage Importance, task 2 vs 1	0,000	H_a
Learning effect: Effectiveness Internal goal, task 2 vs 1	0,000	H_a
Learning effect: Effectiveness Value proposition, task 2 vs 1	0,000	H_a
Learning effect: Effectiveness Value stream, task 2 vs 1	0,000	H_a
Learning effect: Total Effectiveness Percentage task 3 vs 2	**0,100**	**H_0**
Learning effect: Effectiveness Percentage Competence, task 3 vs 2	0,000	H_a
Learning effect: Effectiveness Percentage Customer goal, task 3 vs 2	0,001	H_a
Learning effect: Effectiveness Percentage Importance, task 3 vs 2	0,000	H_a
Learning effect: Effectiveness Internal goal, task 3 vs 2	0,000	H_a
Learning effect: Effectiveness Value proposition, task 3 vs 2	0,000	H_a
Learning effect: Effectiveness Value stream, task 3 vs 2	0,000	H_a
Learning effect: Total Efficiency time task 2 vs 1	0,000	H_a
Learning effect: Total Efficiency time task 3 vs 2	0,001	H_a
Learning effect: Total Ease of use median task 2 vs 1	0,293	H_0
Learning effect: Total Ease of use median task 3 vs 2	0,366	H_0

Appendix 9: Full results of T and Mann Whitney U test of all dependent variables

In the tests below, the following hypotheses are tested

- **H_0 There is no difference in outcome between the initial and newly proposed version of the used dependent variable**
- **H_a The newly proposed version performs better respectively on effectiveness, efficiency, and ease of use than the initial version of the used dependent variable.**

Dependent Variable	Performed test	Mean (rank) initial version	Mean (rank) newly proposed version	two-tailed P-Value	one-tailed P-Value	Accepted Hypothesis
Total Effectiveness: all tasks	**T-Test**	**10,900**	**21,640**	**0,001**	**0,001**	**H_a**
Subtotal Effectiveness: task 1	**Mann-Whitney U**	**7,7**	**14**	**0,006**	**0,003**	**H_a**
Effectiveness: Competence task 1	Mann-Whitney U	11	11	1,000	0,500	H_0
Effectiveness: Customer goal, task 1	Mann-Whitney U	11,55	10,5	0,294	0,853	H_0
Effectiveness: Importance, task 1	Mann-Whitney U	7,5	14,18	0,003	0,002	H_a
Effectiveness: Internal goal, task 1	Mann-Whitney U	11	11	1,000	0,500	H_0
Effectiveness: Value proposition, task 1	Mann-Whitney U	11	11	1,000	0,500	H_0
Effectiveness: Value stream, task 1	Mann-Whitney U	11	11	1,000	0,500	H_0
Subtotal Effectiveness: task 2	**T-Test**	**3,600**	**6,180**	**0,052**	**0,026**	**H_a**
Effectiveness: Competence task 2	Mann-Whitney U	11,15	10,86	0,893	0,554	H_0
Effectiveness: Customer goal, task 2	Mann-Whitney U	9,8	12,09	0,281	0,141	H_0
Effectiveness: Importance, task 2	Mann-Whitney U	9,05	12,77	0,080	0,040	H_a
Effectiveness: Internal goal, task 2	Mann-Whitney U	11,15	10,86	0,893	0,554	H_0
Effectiveness: Value proposition, task 2	Mann-Whitney U	9,7	12,18	0,290	0,145	H_0
Effectiveness: Value stream, task 2	Mann-Whitney U	8,05	13,68	0,014	0,007	H_a
Subtotal Effectiveness: task 3	**T-Test**	**2,800**	**6,730**	**0,013**	**0,007**	**H_a**
Effectiveness: Competence task 3	Mann-Whitney U	9,05	12,77	0,080	0,040	H_a
Effectiveness: Customer goal, task 3	Mann-Whitney U	9,7	12,18	0,290	0,145	H_0
Effectiveness: Importance, task 3	Mann-Whitney U	7,5	14,18	0,003	0,002	H_a

Dependent Variable	Performed test	Mean (rank) initial version	Mean (rank) newly proposed version	two-tailed P-Value	one-tailed P-Value	Accepted Hypothesis
Effectiveness: Internal goal, task 3	Mann-Whitney U	10,75	11,23	0,839	0,420	H_0
Effectiveness: Value proposition, task 3	Mann-Whitney U	10,1	11,82	0,418	0,209	H_0
Effectiveness: Value stream, task 3	Mann-Whitney U	8,1	13,64	0,018	0,009	H_a
Subtotal Effectiveness: task 4	**T-Test**	**4,400**	**7,450**	**0,052**	**0,026**	**H_a**
Effectiveness: Customer goal, task 4	Mann-Whitney U	10,65	11,32	0,763	0,382	H_0
Effectiveness: Value proposition, task 4	Mann-Whitney U	10,5	11,45	0,340	0,170	H_0
Effectiveness: Internal goal, task 4	Mann-Whitney U	12,15	9,95	0,234	0,883	H_0
Effectiveness: Importance, task 4	Mann-Whitney U	8,65	13,14	0,056	0,028	H_a
Effectiveness: Value stream, task 4	Mann-Whitney U	7,05	14,59	0,001	0,001	H_a
Effectiveness: Competence, Importance, Value stream, task 4	Mann-Whitney U	10,7	11,27	0,806	0,403	H_0
Effectiveness: Customer goal, Internal goal, Value proposition, Value stream, task 4	Mann-Whitney U	10,55	11,41	0,602	0,301	H_0
Effectiveness: Value proposition, Value stream, task 4	Mann-Whitney U	11,15	10,86	0,893	0,554	H_0
Effectiveness: Importance, Value stream, task 4	Mann-Whitney U	11,6	10,45	0,486	0,757	H_0
Effectiveness: Competence, task 4	Mann-Whitney U	8,6	13,18	0,049	0,025	H_a
Total Efficiency: all tasks	**T-Test**	**729,927**	**1152,318**	**0,007**	**0,997**	**H_0**
Subtotal Efficiency: task 1	Mann-Whitney U	8,3	13,45	0,057	0,972	H_0
Subtotal Efficiency: task 2	Mann-Whitney U	8,7	13,09	0,105	0,948	H_0
Subtotal Efficiency: task 3	T-Test	150,323	196,840	0,204	0,898	H_0
Subtotal Efficiency: task 4	T-Test	357,256	543,376	0,03	0,985	H_0
Efficiency: Customer goal, task 4	Mann-Whitney U	8,7	13,09	0,105	0,948	H_0
Efficiency: Value proposition, task 4	Mann-Whitney U	10,3	11,64	0,622	0,689	H_0
Efficiency: Internal goal, task 4	T-Test	29,284	45,223	0,107	0,947	H_0
Efficiency: Importance, task 4	T-Test	38,468	37,301	0,873	0,437	H_0

Dependent Variable	Performed test	Mean (rank) initial version	Mean (rank) newly proposed version	two-tailed P-Value	one-tailed P-Value	Accepted Hypothesis
Efficiency: Value stream, task 4	T-Test	35,446	48,045	0,125	0,938	H_0
Efficiency: Competence, Importance, Value stream, task 4	Mann-Whitney U	9,6	12,27	0,324	0,838	H_0
Efficiency: Customer goal, Internal goal, Value Proposition, Value stream, task 4	Mann-Whitney U	8,7	13,09	0,105	0,948	H_0
Efficiency: Value proposition, Value stream, task 4	Mann-Whitney U	8,6	13,18	0,091	0,955	H_0
Efficiency: Importance, Value stream, task 4	T-Test	54,961	97,416	0,029	0,986	H_0
Efficiency: Competence, task 4	T-Test	32,738	39,445	0,466	0,767	H_0
Total Ease of use: all tasks (median)	**T-Test**	**4,350**	**3,727**	**0,151**	**0,076**	**H_0**
Ease of use: task 1	Mann-Whitney U	11,95	10,14	0,476	0,238	H_0
Ease of use: task 2	Mann-Whitney U	12,65	9,5	0,230	0,115	H_0
Ease of use: task 3	Mann-Whitney U	12,1	10	0,423	0,212	H_0
Ease of use: task 4	Mann-Whitney U	12,75	9,41	0,191	0,096	H_0
Total Overall Ease of use (median)	**T-Test**	**4,900**	**3,818**	**0,02**	**0,010**	**H_a**
Ease of use: overall understanding	Mann-Whitney U	12,65	9,5	0,221	0,111	H_0
Ease of use: overall frustration	Mann-Whitney U	12,95	9,23	0,153	0,077	H_0
Ease of use: overall using	Mann-Whitney U	13,9	8,36	0,031	0,016	H_a
Ease of use: overall learning	T-Test	4,800	3,450	0,021	0,011	H_a
Learning effect: Total Effectiveness task 2 vs 1	**T-Test**	**29,300**	**40,818**	**0,233**	**0,117**	**H_0**
Learning effect: Effectiveness Percentage Competence, task 2 vs 1	Mann-Whitney	11,15	10,86	0,893	0,554	H_0
Learning effect: Effectiveness Percentage Customer goal, task 2 vs 1	Mann-Whitney	9,35	12,5	0,159	0,080	H_0
Learning effect: Effectiveness Percentage Importance, task 2 vs 1	Mann-Whitney	12,45	9,68	0,093	0,954	H_0

Dependent Variable	Performed test	Mean (rank) initial version	Mean (rank) newly proposed version	two-tailed P-Value	one-tailed P-Value	Accepted Hypothesis
Learning effect: Effectiveness Internal goal, task 2 vs 1	Mann-Whitney	11,15	10,86	0,893	0,554	H_0
Learning effect: Effectiveness Value proposition, task 2 vs 1	Mann-Whitney	9,7	12,18	0,290	0,145	H_0
Learning effect: Effectiveness Value stream, task 2 vs 1	Mann-Whitney	8,05	13,68	0,014	0,007	H_a
Learning effect: Total Effectiveness Percentage task 3 vs 2	**T-Test**	**-6,600**	**4,546**	**0,482**	**0,241**	**H_0**
Learning effect: Effectiveness Percentage Competence, task 3 vs 2	Mann-Whitney	9,2	12,64	0,111	0,056	H_0
Learning effect: Effectiveness Percentage Customer goal, task 3 vs 2	Mann-Whitney	10,9	11,09	0,940	0,470	H_0
Learning effect: Effectiveness Percentage Importance, task 3 vs 2	Mann-Whitney	9,6	12,27	0,186	0,093	H_0
Learning effect: Effectiveness Internal goal, task 3 vs 2	Mann-Whitney	10,6	11,36	0,703	0,352	H_0
Learning effect: Effectiveness Value proposition, task 3 vs 2	Mann-Whitney	11,3	10,73	0,820	0,590	H_0
Learning effect: Effectiveness Value stream, task 3 vs 2	Mann-Whitney	11	11	1,000	0,500	H_0
Learning effect: Total Efficiency time task 2 vs 1	Mann-Whitney	11,8	10,27	0,573	0,714	H_0
Learning effect: Total Efficiency time task 3 vs 2	Mann-Whitney	10,6	11,36	0,778	0,389	H_0
Learning effect: Total Ease of use median task 2 vs 1	T-Test	-2,400	-2,818	0,638	0,319	H_0
Learning effect: Total Ease of use median task 3 vs 2	T-Test	-0,400	-0,273	0,880	0,560	H_0

For all dependent variables which were tested with the T-Test, equal variances could be assumed based on the outcome of the T-Test.

The P-value was calculated with the following formula for effectiveness variables:

IF (Mean (rank) newly proposed version) > (Mean (rank) initial version) THAN (2-tailed P-value)/2 ELSE (1 - ((2-tailed P-value)/2))

For efficiency and ease of use variables the formula was slightly altered to:

IF (Mean (rank) newly proposed version) < (Mean (rank) initial version) THAN (2-tailed P-value)/2 ELSE 1 - ((2-tailed P-value)/2)

The reason for these two formulas is that for effectiveness higher numbers mean a better score. For efficiency and ease of use questions, lower numbers mean a better score. Efficiency is based on time in seconds and how less seconds are used to answer a question, how more efficient a PGA version should be. For the ease of use questions a 7-point Likert scale was used, where fully agree corresponds to 0 and fully disagree corresponds to 6. Because the questions are formulated in a positive manner, a lower score means a more ease of use perception of the respondent and therefore a PGA version that is easier to use.

Appendix 10: Results of One-way ANOVA and Kruskal Wallis test demographics

In the tests below, the following hypotheses are tested

- **H_0 There is no influence of the demographic variables on the dependent variables**
- **H_a There is influence of the demographic variables on the dependent variables**

Dependent Variable	Performed test	Function P-value	Total years of experience P-value	Models read P-Value	Models made P-Value	Hypothesis accepted
Total Effectiveness: all tasks	**One-way ANOVA**	**0,568**	**0,328**	**0,564**	**0,643**	**H_0**
Subtotal Effectiveness: task 1	**Kruskal Wallis**	**0,900**	**0,543**	**0,222**	**0,518**	**H_0**
Effectiveness: Importance, task 1	Kruskal Wallis	0,818	0,538	0,564	0,490	H_0
Subtotal Effectiveness: task 2	**One-way ANOVA**	**0,753**	**0,202**	**0,114**	**0,607**	**H_0**
Effectiveness: Importance, task 2	Kruskal Wallis	0,555	0,588	0,427	0,390	H_0
Effectiveness: Value stream, task 2	Kruskal Wallis	0,209	0,359	0,380	0,151	H_0
Subtotal Effectiveness: task 3	**One-way ANOVA**	**0,156**	**0,480**	**0,227**	**0,610**	**H_0**
Effectiveness: Competence task 3	Kruskal Wallis	0,434	0,651	0,314	0,625	H_0
Effectiveness: Importance, task 3	Kruskal Wallis	0,775	0,595	0,917	0,636	H_0
Effectiveness: Value stream, task 3	Kruskal Wallis	0,216	0,384	0,084	0,367	H_0
Subtotal Effectiveness: task 4	**One-way ANOVA**	**0,593**	**0,133**	**0,058**	**0,052**	**H_0**
Effectiveness: Importance, task 4	Kruskal Wallis	0,571	0,384	0,252	0,586	H_0
Effectiveness: Value stream, task 4	Kruskal Wallis	0,589	0,477	0,611	0,586	H_0
Effectiveness: Competence, task 4	Kruskal Wallis	0,802	0,537	0,687	0,532	H_0
Total Efficiency: all tasks	**One-way ANOVA**	**0,540**	**0,005**	**0,918**	**0,775**	**H_a**
Subtotal Efficiency: task 1	Kruskal Wallis	0,430	0,357	0,894	0,825	H_0
Subtotal Efficiency: task 4	One-way ANOVA	0,526	0,145	0,984	0,767	H_0
Efficiency: Value proposition, Value stream, task 4	Kruskal Wallis	0,688	0,303	0,718	0,750	H_0
Efficiency: Importance, Value stream, task 4	One-way ANOVA	0,172	0,071	0,536	0,305	H_0
Total Overall Ease of use (median)	**One-way ANOVA**	**0,657**	**0,882**	**0,430**	**0,632**	**H_0**
Ease of use: overall using	Kruskal Wallis	0,365	0,439	0,549	0,372	H_0
Ease of use: overall learning	One-way ANOVA	0,754	0,699	0,911	0,865	H_0