MASTER'S THESIS

An evaluation of the intuitiveness of the PGA notation in a practical business context

Hage, R.

Award date:
2022

An evaluation of the intuitiveness of the PGA notation in a practical business context

Opleiding:	Open Universiteit, faculteit Betawetenschappen
	Masteropleiding Business Process Management & IT
Degree programme:	Open University of the Netherlands, Faculty Science
	Master of Science Business Process Management & IT
Course:	IM0602 BPMIT Graduation Assignment Preparation IM9806 Business Process Management and IT Graduation Assignment
Student:	Roger
Hage Identification number:	
Date:	16-01-2022
Thesis supervisor	Ben Roelens
Second reader	Harry Martin
Version number:	1.0
Status:	Final

Abstract

The Process-Goal Alignment (PGA) technique is a domain-specific modeling language (DSML) especially designed for realizing strategic fit within the business architecture of an organization. The PGA technique uses its own modeling notation, showing how one can get strategies defined and executed within the business architecture. In prior research, adaptations to the initial PGA notation were proposed to make it more intuitive. However, it could not ascertain that the adapted PGA notation was an improvement compared to the initial one. This research introduced a method to compare the intuitiveness of two notations in order to determine whether the adapted notation performs significantly better than the initial version.

With a significant improvement observed for the overall interpretational effectiveness, the adapted PGA notation seemed to perform better than the initial one, and based on the overall perceived ease of use and usefulness, the participants also had the feeling that the adapted PGA notation performed better. Detailed analysis revealed that from the six suggested adaptations only one element performed convincingly better, namely Importance, and two less convincing, namely Competence and Value stream. With the mixed results obtained in this research, it is not possible to unambiguously answer whether the adapted variant of the notation performs better. Further rethinking and redesign is thus needed.

Key terms

Process-Goal Alignment (PGA), domain-specific modeling language (DSML), Notation, Intuitiveness.

Summary

Organizations are using an Enterprise Architecture (EA) to plan and implement business and IT strategies and serves alignment between business and IT stakeholders. Conceptual models are an important instrument for an EA. Models reduce complexity by only considering relevant items of the real world and are used as a means of communication about a specific purpose for specific stakeholders. This can be realized by a domain-specific modeling language (DSML), in which only the required concepts and terms from a particular domain are specified. This results in a compact language that is completely tailored to specific needs within the domain. A DSML contains a specific graphical notation to visualize the underlying domain concepts and terms, which promotes the use and comprehensibility of the respective model.

The Process-Goal Alignment (PGA) technique is a DSML especially designed for realizing strategic fit within the business architecture of an organization. The PGA technique uses its own modeling notation, showing how one can get strategies defined and executed within the business architecture. The notation was originally designed by the creators of the PGA technique.

In prior research, adaptations to the initial PGA notation were proposed. However, it could not be ascertained that the adapted PGA notation was an improvement compared to the initial one. The objective of this research is to find a way to determine whether the intuitiveness (i.e. semantic transparency) of the adapted PGA notation is significantly better than the initial PGA notation. A notation that is semantically transparent will lower the cognitive load for end-users because the used graphical symbols act as mnemonics. In this way, novice end-users can visually recognize known domain concepts and terms.

This research introduced a method to compare the intuitiveness of two notations in order to determine whether the adapted notation performs significantly better than the initial version. To determine the intuitiveness of the different PGA notations three dependent variables were used, one objective (i.e. interpretational effectiveness) and two subjective (i.e. perceived ease of use and perceived usefulness). To examine the interpretational effectiveness, three experiments were designed, a semantic transparency experiment, a case study with comprehension questions and a recall/recognition experiment. To examine the perceived ease of use and perceived usefulness, questions were asked about the overall perceived ease of use, the perceived ease of use of every individual PGA concept and the overall perceived usefulness. As experimental design, a between-subjects design was used, where the participants were divided in two groups. One group is shown the initial PGA notation, the other group is shown the adapted PGA notation.

With a significant improvement observed during the case study experiment for the overall interpretational effectiveness, the adapted PGA notation seems to perform better than the initial one. Based on the overall perceived ease of use and usefulness, the participants also had the feeling that the adapted PGA notation performs better. Based on this overall result, further analysis was done to find out for which concepts of the initial PGA notation it was worth considering to replace with the adapted ones. From the six suggested adaptations, only one performed convincingly better, namely Importance, and two less convincing, namely Competence and Value stream.
With the mixed results obtained in this research, it is thus not possible to unambiguously answer whether the adapted variant of the notation performs better.

Further research is required to:

- determine whether the questions asked during the semantic transparency experiment, case study, recall/recognition experiment and perceived ease of use should be asked differently (e.g. open questions instead of multiple choice) or that other experiments should be conducted to measure the interpretational effectiveness and perceived ease of use.
- examine whether a within-subjects design yields better results, for instance with an experiment in which both the initial and adapted notations of the various concepts are shown side by side and the participant can indicate a preference.
- give participants the possibility to give reasons why they did not completely fill out the survey, in order to make necessary adjustments when needed.
- examine whether results will differ when the survey is conducted with a larger number of participants. Besides conducting a survey within one financial organization, the survey could be conducted with participants from different financial organizations and/or from organizations within other economic sectors.

Contents

1. Introduction

1.1. Background

Organizations are using an Enterprise Architecture (EA) to plan and implement business strategies. Lankhorst (2013, p. 3) defines an EA 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". The EA is applied to plan and implement business and IT strategies and serves alignment between business and IT stakeholders. To this end, an EA is usually divided into the following four architecture domains: Business Architecture, Application Architecture, Data / Information Architecture, and Technology Architecture (Kotusev, 2019).

Conceptual models are an important instrument for EA. Lankhorst (2013, p. 123) defines a conceptual model as "an unambiguous, abstract conception of some parts or aspects of the real world". Models reduce complexity by only considering relevant items of the real world and are used as a means of communication about a specific purpose for specific stakeholders. By using a modeling language, an EA can be made visible and described uniformly. There are two types of modeling languages, general purpose modeling languages (GPMLs) and domain-specific modeling languages (DSMLs). A GPML provides modellers with abstract concepts and terms, not specifically aimed at a particular domain. Due to its abstract design, a GPML can be widely deployed within organizations and be used to map out the most diverse issues. An example of a GPML, which is also one of the most used EA modeling languages at this moment, is ArchiMate. The Open Group defines ArchiMate as "a visual language with a set of default iconography for describing, analysing, and communicating many concerns of Enterprise Architectures as they change over time" (The Open Group, 2019, p. 1). By using ArchiMate, the four different architecture domains and their relations and dependencies can be described and visualized in an integral way. In addition to a GPML, such as ArchiMate, there are DSMLs. With a DSML, only the required concepts and terms from a particular domain are specified (Luoma, Kelly, & Tolvanen, 2004). This results in a compact language that is completely tailored to specific needs within the domain.

A DSML contains a specific graphical notation to visualize the underlying domain concepts and terms, which promotes the use and comprehensibility of the respective model. The graphical notation should serve both the modeller who creates the model and the user who interprets the model. Relative to an average modeller, an average user is less or even not trained to read a graphical notation. Consequently, the development of an intuitive notation is of crucial importance. A more formal term for this intuitive notation is semantic transparency (Caire, Genon, Heymans, & Moody, 2013). Semantic transparency is defined by Moody (2009, p.15) as "the extent to which a novice reader can infer the meaning of a symbol from its appearance alone". A notation that is semantically transparent will lower the cognitive load for end-users as they can visually recognize the domain concepts and terms. To achieve this, coordination with end-users is necessary with regard to their knowledge, beliefs and expectations (Bork, Schrüffer, & Karagiannis, 2019).

1.2. Exploration of the topic

The Process-Goal Alignment (PGA) technique is a DSML especially designed for realizing strategic fit within the business architecture of an organization (Roelens, Steenacker, & Poels, 2019). Roelens et al. (2019) noticed that in organizations there is a gap between the strategies that an organization defines and the actual execution of those strategies. Research revealed that 65% of organizations have an agreed upon strategy, only 14% of the employees understand this strategy and only 10% of the organizations successfully execute the strategy (Roelens et al., 2019). The goal of the PGA technique is to solve this issue by a holistic approach. By only using relevant domain concepts and terms, the technique is easily understandable for business users and serves as a means of communication by making the strategies visible from top (i.e. strategy definition) to bottom (i.e. strategy execution) via the intermediate business infrastructure. The PGA technique uses its own modeling notation, showing how one can get strategies defined and executed within the business architecture. The top-layer consists of the strategic goals. It states which financial goals an organization wants to reach, which customers should be serviced at what satisfaction level, and which internal objectives are needed to accomplish that. The infrastructural part states which structures should be in place. It consists of financial structures (i.e. cost and revenues), value propositions (i.e. products and services) and competences (i.e. internal knowledge and skills). The execution part holds the processes and underlying activities, explaining what work should be performed. The PGA model includes performance metrics in combination with a prioritization mechanism. This enables to visualize areas of interest, with their corresponding priority, by different colours in the models. This heat mapping technique notifies business users where they should pay attention to.

As mentioned, the PGA technique uses its own modeling notation, which consists of icons. The notation was originally designed by the creators of the PGA technique, who were guided by the principle of semantic transparency. Roelens and Bork (2020) did an empirical evaluation regarding the intuitiveness of the initial PGA notation among Master students of Ghent University. As it was the goal to evaluate the intuitiveness of the PGA notation, only a brief introduction of the PGA technique was given, without showing any notation. The evaluation consisted of three phases. In the term association phase, participants had to draw their own intuitive notation of the concepts used in the PGA technique. During the Notation Association phase, the participants were asked to write down three intuitive associations that came up in their minds for the existing notation of the PGA technique. During the Case Study phase, an example of a business architecture heat map was shown to the participants. With respect to this business architecture heat map, several questions were asked about the concepts shown, their relationships, purpose and meaning. After analysing the results of the different experimental phases, adaptations to the initial notation were proposed. This concerned an alternative visualization for the following six PGA concepts: Competence, Value proposition, Internal goal, Customer goal, Value stream and Importance.

1.3. Problem statement

The notation of the PGA technique was originally designed by the creators, who were guided by the principle of semantic transparency. However, the design of a semantically transparent notation is a subjective matter. It depends on people's background and experience. What is intuitive for person A may not hold for person B. The same holds for the creators and the intended end-users, as what is designed to be intuitive for the creators may not hold for the intended end-users. Roelens and Bork (2020) conducted an evaluation of the notation of the PGA technique with the previously described experimental technique. This resulted in proposed adaptations to the initial notation, which should lead to the design of an improved, more intuitive, notation. To investigate whether the intuitiveness of the adapted notation is significantly better than the initial notation, further experimental research is needed.

This leads to the central research question that can be formulated as follows:

How can be determined whether the intuitiveness of the adapted PGA notation is significantly better than the initial PGA notation?

1.4. Research objective and questions

The objective of this research is to design a method to determine whether the intuitiveness of the adapted PGA notation is significantly better than the initial PGA notation.

In order to answer the central research question this research will first elaborate on intuitiveness, and more specifically on intuitiveness of a graphical notation of conceptual models.
This results in the next sub-questions:
Sub-question 1:
What is meant by intuitiveness of a graphical notation of conceptual models?

Sub-question 2:
How to determine whether a graphical notation of a conceptual model is intuitive?

Next, this research will elaborate more deeply on the PGA technique. The research will in particular focus on the PGA notation. Roelens and Bork (2020) conducted an empirical evaluation of the initial notation of the PGA technique, which resulted in adaptations to the notation for some of the concepts and terms. This evaluation took place among students which are not really the intended end-users. The PGA technique is designed to be used within organizations and the intended end-users will be business-oriented end-users. In order to be able to compare the two different notations, the PGA technique should be empirically evaluated within a practical business context.
This results in the next sub-question:

Sub-question 3:
How can the intuitiveness of the different PGA notations be empirically evaluated within a practical business context?

In order to know which PGA notation is preferred in a practical business context, the initial PGA notation and the version with the proposed adaptations of the PGA notation should be compared with each other. Only when the adapted notation performs significantly better than the initial notation, a change of notation should be implemented.
This results in the next sub-question:

Sub-question 4:
Does the adapted PGA notation perform intuitively better in a practical business context?

1.5. Motivation/relevance

This research has theoretical relevance because it will extend previous research done by Roelens and Bork (2020), who performed an experimental technique to evaluate the intuitiveness of the initial PGA notation. This resulted in proposals for adaptations to the PGA notation. This research will extend the experimental technique with one or more steps to evaluate and compare the different visual notations in order to determine which notation is significantly better. This paves a path to objectively choose the most intuitive notation.

This research will also have practical implications. It elaborates on the PGA technique, a DSML designed for realizing strategic fit within the business architecture of an organization (Roelens et al., 2019). It will show the practical benefits of a DSML compared to a GPML. These benefits consist of (i) the use of concepts and terms from the targeted domain, defined by experts of that domain, ready

to use and tailored for that domain and (ii) a specific graphical notation, also tailored for that domain, which makes the models easier to read and understand (Frank, 2011). This research especially concentrates on this last feature, showing the practical benefits of a DSML, in this case the PGA technique, with an intuitive notation.
The PGA technique is designed to be used within organizations, by business-oriented end-users. In this research different PGA notations will be evaluated within a practical business context, which will ultimately lead to the identification of the most intuitive PGA notation that can be used within organizations.

1.6. Main lines of approach

This chapter started with an introduction of this research. It highlighted the background and explored the topic in more detail. A problem statement was formulated, which led to the central research question: "How can be determined whether the intuitiveness of the adapted PGA notation is significantly better than the initial PGA notation?". This central research question was broken down in sub-questions. The last section described the motivation and relevance of this research.
Chapter 2 will elaborate on the current literature which outlines the theoretical framework. This theoretical framework will be built by answering the first three sub-questions and will be used as guidance for answering the central research question. In chapter 3 the methodology which will be used in this research will be described in detail. This methodology will be conducted in a practical business context and the results will be analysed in chapter 4. In chapter 5 the analysis of the results will be discussed, the last sub-question will be answered and conclusions will be drawn. This will lead to recommendations for practice and further research.

2. Theoretical framework

2.1. Research approach

In this chapter a literature review is conducted. The goal of a literature review is to investigate the current state of the research field, what are the key theories, concepts and ideas (Saunders, Lewis, & Thornhill, 2019). While performing the literature review a theoretical framework is developed. This theoretical framework is used as guidance for answering the central research question: “How can be determined whether the intuitiveness of the adapted PGA notation is significantly better than the initial PGA notation?”.

The process of writing a literature review is an iterative process that consists of the following steps: (i) define search parameters for the research questions and objectives, (ii) generate search terms, (iii) conduct the search, (iv) obtain literature, (v) evaluate the literature, (vi) record the literature, (vii) start drafting the review and, if appropriate, iterate further like (i) with revised search parameters (Saunders et al., 2019). Items (i) and (ii) are described in this section, (iii) till (vii) in the consequent sections.

To answer the central research question, it was split into four sub-questions. This chapter will elaborate on the first three sub-questions to develop a theoretical framework. The first two sub-questions elaborate on intuitiveness, and more specifically on intuitiveness of a graphical notation of conceptual models. The third sub-question elaborates on how to empirically evaluate the intuitiveness of the different PGA notations and find out how to compare these different notations in order to determine which performs significantly better. This division of sub-questions resulted in a following search for the literature review.

Table 1 describes for all queries the search parameters (i.e. step i) that were used. For each parameter, a reason is given why the corresponding value is chosen.

Table 1 Generic search parameters

Parameter	**Value**	**Reason**
Used database	Open University library portal	Permission to download article
Filter	Full text online	Only interested in complete articles, not only abstracts
Year of publication	At least 2004	In 2004 Gemino and Wand published a basic work about a framework for empirical evaluation of conceptual modeling techniques (Gemino, & Wand, 2004).
Content type	All	No restrictions on content types
Discipline	All	No restrictions on disciplines
Language	English	Most scientific articles are written in English
Exclude from results	Newspaper article and Book reviews	Articles should be (scientific) papers, books, journals, not newspapers or book reviews

The next search terms (i.e. step ii) were chosen because they are used in the sub-questions itself or are a synonym.
For the first two sub-questions, combinations of the following search terms were used (see Table 2):

Table 2 Search terms first two sub-questions

Search terms	Sub-question / Synonym	Sub-questions
conceptual model	Sub-question	What is meant by intuitiveness of a graphical notation of conceptual models?
intuitiveness	Sub-question	
graphical notation	Sub-question	How to determine whether a graphical notation of a conceptual model is intuitive?
semantic transparency	Synonym of intuitiveness	

For the third sub-question combinations of the following search terms were used (see Table 3):

Table 3 Search term third sub-question

Search terms	Sub-question / Synonym	Sub-question
empirical evaluation	Sub-question	How can the intuitiveness of the different PGA notations be empirically evaluated within a practical business context?
intuitiveness	Sub-question	
semantic transparency	Synonym of intuitiveness	

Besides the database search, backward and forward snowballing was executed. With backward snowballing the reference list of an article is used to identify other relevant articles, which by definition will lie in the past. Forward snowballing implies that a given article is cited by (a) relevant article(s), which by definition was written more recently (Wohlin, 2014). In this way, based on further investigating the relevant articles, a set of new relevant articles was added to the list. Another important source used to find relevant articles was the thesis supervisor, who shared relevant literature at the start of the research and during three-weekly meetings.

All search results were categorized based on relevancy. Because of time limitations, only the top ten articles of every search were taken into account. The results were saved in separate excel sheets, with separate tabs per search term(s). Table 4 describes the evaluation criteria which were scored by reading the abstract of the article:

Table 4 Article evaluation criteria

Criterion	Description	Possible values
Web of Science	This criterion indicates whether the article is cited in other scientific articles, published in Web of Science.	Y / N
Intuitiveness / Semantic transparency	Intuitiveness (synonym Semantic transparency) is the extent to which a novice reader can infer the meaning of a symbol from its appearance alone, and one of the principles for designing effective visual notations (Moody, 2009). This criterion indicates whether the article is primarily focused on intuitiveness, a main topic of this research.	Y / N
PGA / DSML / Conceptual Model	Process-Goal Alignment (PGA) is a specific Domain-Specific Modeling Language (DSML) and more generally a Conceptual Model. This criterion indicates whether the article is primarily focused on conceptual models, DSML or PGA, which is a main topic of this research.	Y / N
Graphical Notation	A notation based on graphical elements rather than text. This criterion indicates whether the article is primarily focused on a graphical notation, which is a main topic of this research.	Y / N

Criterion	Description	Possible values
Methodology/Experiment	This criterion indicates whether the article is constructing a generic methodology (M), conducting individual experiments (E) or both (M / E).	M / E

A first selection was done by only selecting articles with positive (Y) outcomes for the criteria "Intuitiveness / Semantic transparency", "PGA / DSML / Conceptual Model" and "Graphical Notation". To reach the final set of articles, the whole text of the selected articles was read and based on the same criteria determined if the article would still be relevant or not.

2.2. Implementation

With the defined search parameters (i.e. step i) and generated search terms (i.e. step ii) the database search was conducted (i.e. step iii). Next, the literature, with a maximum of ten articles, was obtained (i.e. step iv) and evaluated (i.e. step v). First only the abstract of each article was evaluated, later eventually the whole article. All steps conducted were recorded (i.e. step vi) in excel sheets.

The "Total number of relevant articles after first selection " column shows, after reading the abstract of the first ten results, all articles with positive (Y) outcomes for the criteria "Intuitiveness / Semantic transparency", "PGA / DSML / Conceptual Model" and "Graphical Notation". The "References to final relevant articles" column shows, after reading the whole text and based on the same criteria, the APA references of the articles that still were relevant.

Table 5 not only shows an overview of the results of the database search (more details can be found in Appendix 1), but also the results of snowballing and the relevant articles mentioned by the thesis supervisor (more details can be found in Appendix 2).

Table 5 Overview of the results of the database search, snowballing and thesis supervisor

Search terms / Snowballing / Thesis supervisor	Total number of articles	Total number of relevant articles after first selection	References to final relevant articles
"conceptual model" AND "intuitiveness"	33	4	0
"conceptual model" AND "semantic transparency"	4	2	0
"graphical notation" AND "intuitiveness"	7	2	0
"graphical notation" AND "semantic transparency"	8	3	1. El Kouhen, Gherbi, Dumoulin, & Khendek, (2015)
"empirical evaluation" AND "intuitiveness"	10	3	1. Bork, Schrüffer, & Karagiannis (2019)
"empirical evaluation" AND "semantic transparency"	16	7	1. Bork, Schrüffer, & Karagiannis (2019) 2. Genon, Caire, Toussaint, Heymans, & Moody (2012) 3. Santos, Gralha, Goulão, & Araújo (2018)
Backward snowballing	6	6	1. Caire, Genon, Heymans, & Moody (2013) 2. Moody, Genon, Heymans, & Caire (2012)
Forward snowballing	1	1	0

Search terms / Snowballing / Thesis supervisor	Total number of articles	Total number of relevant articles after first selection	References to final relevant articles
Thesis supervisor	4	4	1. Roelens, & Bork (2020)

2.3. Results and conclusions

This section summarizes the results based on the final set of relevant articles. With these results, the first three sub-questions are answered in the next two sub sections. In the last sub section conclusions are drawn about the literature review.

2.3.1. Intuitiveness of a graphical notation

The first two sub-questions elaborate on intuitiveness, and more specifically on intuitiveness of a graphical notation of conceptual models.

Sub-question 1: What is meant by intuitiveness of a graphical notation of conceptual models?

A more formal term for the intuitiveness of a (graphical) notation of a conceptual model is semantic transparency (Caire et al., 2013). Semantic transparency is one of the nine principles of the Physics of Notations theory (Moody, 2009) and it means that the meaning (semantics) of a symbol is clear (transparent) from its appearance alone (Caire et al., 2013). A notation that is semantically transparent will lower the cognitive load for end-users because the used graphical symbols act as mnemonics. In this way, novice end-users can visually recognize known domain concepts and terms. Semantic transparency is a continuum, as shown in figure 1.

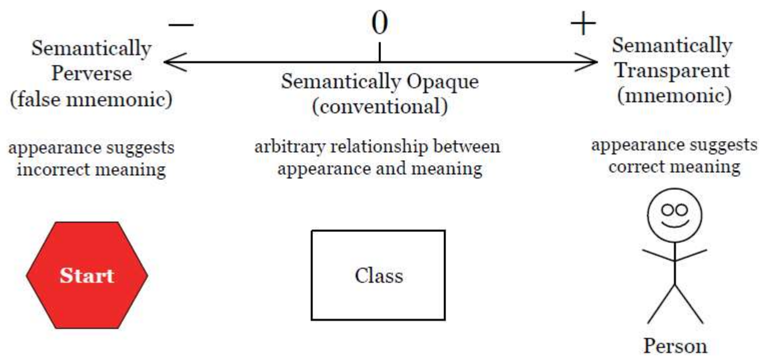


Figure 1 Semantic Transparency is a continuum (Caire et al., 2013, p. 115)

When a graphical notation is semantically transparent (+), the appearance of a symbol suggests the correct meaning and it can act as a mnemonic for a novice end-user. As an example Caire et al. (2013) drew a stick figure to represent a person.

When a graphical notation is semantically opaque (0), there is an arbitrary relationship between the appearance and the meaning of it. As an example Caire et al. (2013) drew a rectangle to represent a UML class.

When a graphical notation is semantically perverse (-), the appearance of a symbol suggests an incorrect meaning and it can act as a false mnemonic for a novice end-user. As an example Caire et al. (2013) drew a red hexagon to indicate start.

Sub-question 2: How to determine whether a graphical notation of a conceptual model is intuitive?

As Caire et al. (2013) state, semantic transparency is typically evaluated subjectively by experts (e.g. researchers, notation designers). While designing the graphical notation, experts try to think like novice end-users, but because of their inherent knowledge, that is difficult to accomplish. To objectively determine whether a graphical notation of a conceptual model is intuitive, the semantic transparency should be evaluated empirically, with novice end-users.

2.3.2. Empirical evaluation of the intuitiveness

The third sub-question elaborates on how to empirically evaluate the intuitiveness of the different PGA notations and find out how to compare these different notations in order to determine which performs better.

Sub-question 3: How can the intuitiveness of the different PGA notations be empirically evaluated within a practical business context?

To investigate how to empirically evaluate the intuitiveness of different PGA notations, a broader view is taken in the literature to analyse the empirical evaluation of conceptual models. Each relevant article is examined concerning the aspects measures, hypotheses, experimental design, instrumentation and experimental tasks, selection of participants and operational procedures (see Appendix 3). Table 6 provides a summary per aspect with references to the corresponding articles.

Table 6 Article summary per aspect with references to the corresponding articles

Aspect	Description	References
Variables and Measures	Articles with **different variants of a notation** (independent variables).	Bork et al. (2019) Caire et al. (2013) El Kouhen et al. (2015) Moody et al. (2012) Roelens & Bork (2020) Santos et al. (2018)
	Articles with measures regarding **interpretational effectiveness** (dependent variable) measuring the number of correct terms/concepts associated to a certain notation.	Bork et al. (2019) Caire et al. (2013) El Kouhen et al. (2015) Moody et al. (2012) Roelens & Bork (2020) Santos et al. (2018)
	Article with measures regarding **interpretational effectiveness** (dependent variable) measuring the number of correct drafted models.	Bork et al. (2019)
	Article with measures regarding **interpretational effectiveness** (dependent variable) measuring the number of correct interpreted concepts.	Roelens & Bork (2020)
	Article with measures regarding **interpretational effectiveness** (dependent variable) measuring the degree of proximity between a symbol and the semantic construct represented by it.	Santos et al. (2018)
	Article with measures regarding **interpretational efficiency** (dependent variable) measuring the time needed to draft a notation for a certain term.	Bork et al. (2019)

Aspect	Description	References
	Article with measures regarding **interpretational efficiency** (dependent variable) measuring the time needed to associate terms to a given notation.	Bork et al. (2019)
	Article with measures regarding **interpretational efficiency** (dependent variable) measuring the time needed to draft models.	Bork et al. (2019)
	Articles with measures regarding **perceived usefulness** (dependent variable), i.e. the usefulness of a specific notation a participant experienced. This is measured based on answers from participants to questions about usefulness on a five point Likert scale.	Genon et al. (2012) Moody et al. (2012)
	Articles with measures regarding **perceived ease of use** (dependent variable) measuring the experience of how easily a participant could draw a concept of a specific notation, based on a five point Likert scale.	Genon et al. (2012) Moody et al. (2012)
	Articles with measures regarding **perceived ease of use** (dependent variable) measuring the experience of how easily a participant could select the best drawing of the mentioned concept of a specific notation, based on a five point Likert scale.	Genon et al. (2012) Moody et al. (2012)
Hypotheses	To evaluate the intuitiveness of a conceptual model, all hypotheses compare **two different variants of a notation** (i.e. the independent variables).	
	Articles with (multiple) hypotheses regarding **interpretational effectiveness** (dependent variable).	Bork et al. (2019) Caire et al. (2013) El Kouhen et al. (2015) Moody et al. (2012) Roelens & Bork (2020) Santos et al. (2018)
	Article with a hypothesis regarding **interpretational efficiency** (dependent variable).	Bork et al. (2019)
	Articles with multiple hypotheses regarding **perceived usefulness** (dependent variable).	Genon et al. (2012) Moody et al. (2012)
	Articles with multiple hypotheses regarding **perceived ease of use** (dependent variable).	Genon et al. (2012) Moody et al. (2012)
Experimental design	Articles with a **between-subjects** experimental design**.** Saunders et al. (2019) defines a between-subjects design as "an experimental design allowing a comparison of results to be made between an experimental group and a control group" (Saunders et al., 2019, p. 797).	Caire et al. (2013) El Kouhen et al. (2015) Genon et al. (2012) Moody et al. (2012) Santos et al. (2018)
	Articles with a **within-subjects** experimental design**.** Saunders et al. (2019) defines a within-subjects design as "an experimental design using only a single group where every participant is exposed to the planned intervention or series of interventions" (Saunders et al., 2019, p. 821).	Bork et al. (2019) Roelens & Bork (2020)
Instrumentation and experimental tasks	Articles with **term association tasks,** where terms are provided to a participant and the participant has to draw one or more graphical representations that he/she deems as the most intuitive for the term. Experimental tasks with **term association** are also described as a **symbolisation experiment**.	Bork et al. (2019) Roelens & Bork (2020)
	Articles with a **symbolisation experiment** (see **term association**).	Caire et al. (2013) El Kouhen et al. (2015) Genon et al. (2012) Moody et al. (2012) Santos et al. (2018)

Aspect	Description	References
	Articles with experimental tasks with **notation association**, where samples of a notation are provided and participants register intuitive associations that popup when looking at them. Experimental tasks with **notation association** are also described as a **semantic transparency experiment**.	Bork et al. (2019) Roelens & Bork (2020)
	Articles with a **semantic transparency experiment** (see **notation association**).	Caire et al. (2013) El Kouhen et al. (2015) Moody et al. (2012) Santos et al. (2018)
	Articles with experimental tasks including a **case study**, where participants are asked comprehension questions about an example model in order to determine the participants` model understanding.	Bork et al. (2019) Roelens & Bork (2020)
	Articles with a **prototyping experiment**, where naïve participants identify the best graphical representations which were produced by other naïve participants (e.g. during term association tasks).	Caire et al. (2013) El Kouhen et al. (2015) Genon et al. (2012) Moody et al. (2012) Santos et al. (2018)
	Articles with a **recall/recognition experiment**, where the ability of naïve participants to learn and remember graphical representations is evaluated.	Caire et al. (2013) Moody et al. (2012)
	Articles with a **"best of breed" symbols identification**, where based on the results of the semantic transparency experiment and recall/recognition experiment, the most cognitively effective graphical representation for each term is identified by comparing the different symbol sets with each other.	El Kouhen et al. (2015) Moody et al. (2012)
Selection of participants	Only naïve (with respect to the modeling language investigated) undergraduate/master **students** were selected. A clear distinction is made between the students **with** or **without previous knowledge of modeling**.	
	Articles with naïve undergraduate/master **students without previous knowledge of modeling**.	Bork et al. (2019) Caire et al. (2013) El Kouhen et al. (2015) Genon et al. (2012) Moody et al. (2012) Santos et al. (2018)
	Articles with naïve undergraduate/master **students with previous knowledge of modeling**.	Santos et al. (2018)
Operational procedures	Articles with a phase to **collect participants' demographic data, modeling experience and domain understanding**.	Bork et al. (2019) Genon et al. (2012) Roelens & Bork (2020)
	Articles with an **introduction phase**, which gives a brief introduction to the topic at hand.	Bork et al. (2019) Roelens & Bork (2020)
	Articles with an **experimental phase**, where one or more experiments, as described in the "Instrumentation and experimental tasks", are performed in order to collect various measures regarding interpretational effectiveness, interpretational efficiency, perceived usefulness and/or perceived ease of use.	Bork et al. (2019) Caire et al. (2013) El Kouhen et al. (2015) Genon et al. (2012) Moody et al. (2012) Santos et al. (2018)
	Articles with a **concluding phase**, where participants are asked to fill out a feedback survey in order to provide positive and negative feedback and improvement suggestions.	Bork et al. (2019) Roelens & Bork (2020)

2.3.3. Conclusions

Based on sections 2.3.1. (Intuitiveness of a graphical notation) and 2.3.2. (Empirical evaluation of the intuitiveness), the next conclusions can be drawn.

Intuitiveness of a graphical notation of a conceptual model

Intuitiveness of a (graphical) conceptual model notation is formally known as semantic transparency. Semantic transparency is measured on a scale that runs from semantically perverse (-) to semantically opaque (0) and ends at semantically transparent (+). To objectively determine whether a graphical notation of a conceptual model is intuitive, the semantic transparency should be evaluated empirically, with novice end-users.

Variables and Measures

The most common way to measure the interpretational effectiveness is by measuring the number of correct terms/concepts associated to a certain element of the notation (Bork et al., 2019; Caire et al., 2013; El Kouhen et al., 2015; Moody et al., 2012; Roelens & Bork, 2020; Santos et al., 2018). Other, less common ways to measure the interpretational effectiveness are: measuring the number of correctly drafted models (Bork et al., 2019), measuring the number of correctly interpreted concepts (Roelens & Bork, 2020) and measuring the degree of proximity between a symbol and the semantic construct represented by it (Santos et al., 2018).
Measuring interpretational efficiency can be implemented by registering the time taken by a participant to fulfil a certain task, which is only described by Bork et al. (2019).
Perceived usefulness and perceived ease of use are both perceptions. Perceived usefulness, described twice (Genon et al., 2012; Moody et al., 2012), is measured by asking participants questions about how they experience the usefulness of a specific notation, based on a five point Likert scale.
Perceived ease of use, also described twice (Genon et al., 2012; Moody et al., 2012), is measured by asking participants questions about how easily they could draw a concept (on a five point Likert scale) and asking participants questions about easily they could interpret the meaning of a mentioned concept of a specific notation (on a five point Likert scale).

Hypotheses

The independent variable used in all hypotheses are the different variants of a notation. The most common way to evaluate the intuitiveness of these notations are hypotheses regarding interpretational effectiveness (Bork et al., 2019; Caire et al., 2013; El Kouhen et al., 2015; Moody et al., 2012; Roelens & Bork, 2020; Santos et al., 2018). Three other, less common, ways to evaluate intuitiveness are interpretational efficiency (Bork et al., 2019), perceived usefulness (Genon et al., 2012; Moody et al., 2012) and perceived ease of use (Genon et al., 2012; Moody et al., 2012).

Experimental design

Two ways of experimental design are described: between-subjects and within-subjects. The most commonly used experimental design is the between-subjects design (Caire et al., 2013; El Kouhen et al., 2015; Genon et al., 2012; Moody et al., 2012; Santos et al., 2018). The between-subjects design has several advantages. Firstly, it prevents practice or learning effects, where participants get familiar with the subject while conducting an experiment. This familiarity effect could be of large influence while comparing the intuitiveness of different notations. Secondly, participants have to

conduct less experiments, so use less time, which can prevent fatigue effects. Fatigue effects occur when participants become tired, and can influence the results of the conducted experiments.
Less common is the within-subjects design (Bork et al., 2019; Roelens & Bork, 2020).

Instrumentation and experimental tasks

All articles describe experimental tasks for evaluating the intuitiveness of the notations. Most articles concentrate on the individual symbols instead of complete diagrams/models.
The most described experimental tasks are the symbolisation experiment (also known as term association), semantic transparency experiment (also known as notation association) and prototyping experiment (Bork et al., 2019; Caire et al., 2013; El Kouhen et al., 2015; Genon et al., 2012; Moody et al., 2012; Roelens & Bork, 2020; Santos et al., 2018). In some cases, after the semantic transparency experiment, a recall/recognition experiment (Caire et al., 2013; Moody et al., 2012) and a "best of breed" symbols identification (El Kouhen et al., 2015; Moody et al., 2012) is performed.
Less described are experiments including a case study with comprehension questions (Bork et al., 2019; Roelens & Bork, 2020).

Selection of participants

All participants to the experiments are naïve undergraduate/master students.
The most commonly selected participants are students without previous knowledge of modeling (Bork et al., 2019; Caire et al., 2013; El Kouhen et al., 2015; Moody et al., 2012; Roelens & Bork, 2020; Santos et al., 2018). Only Santos et al. (2018) selected students with previous knowledge of modeling, trying to better simulate a practical business context.

Operational procedures

All operational procedures have an experimental phase, where the experimental tasks are performed (Bork et al., 2019; Caire et al., 2013; El Kouhen et al., 2015; Moody et al., 2012; Roelens & Bork, 2020; Santos et al., 2018).
Less common are procedures that include a phase to collect participants' demographic data, modeling experience and domain understanding (Bork et al.,2019; Genon et al., 2012; Roelens & Bork, 2020).
An introduction phase, which gives a brief introduction to the topic at hand, and a concluding phase, where participants are asked to provide feedback and improvement suggestions, is only described by Bork et al. (2019) and Roelens and Bork (2020).

2.4. Objective of the follow-up research

Conclusions drawn during this chapter effectively answered the first three sub-questions regarding the intuitiveness of a graphical notation and the empirical evaluation of the intuitiveness. The objective of the follow-up research is to set up an experimental method for the evaluation of the intuitiveness of the PGA notation in order to find out whether the intuitiveness of the adapted PGA notation is significantly better than the initial PGA notation. The research will be conducted by experiments within a practical business context. Given the scope of this thesis, around 10-15 business end-users will be selected as participants. Next, the intuitiveness of the PGA notation variants will be analysed based on the results of the conducted experiments. Finally, the evaluation also helps to identify necessary adjustments to the experimental setup for evaluating the intuitiveness of other modeling notations in future research.

3. Methodology

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

The objective of this research is to answer the central research question:

How can be determined whether the intuitiveness of the adapted PGA notation is significantly better than the initial PGA notation?

The notation of the PGA technique is originally designed by the creators, which depends on human creativity. The research of Bork et al. (2019) set up an experimental technique to achieve a more intuitive notation. Roelens and Bork (2020) used this technique with the goal of achieving a more intuitive PGA notation. That resulted in six suggested adaptations to the initial PGA notation, concerning the following elements: Competence, Value proposition, Internal goal, Customer goal, Value stream and Importance (see Appendix 4 for detailed information). However, the technique cannot statistically ascertain that the adapted PGA notation is an improvement compared to the initial one.

The objective of this research is to find a way to determine whether the intuitiveness of the adapted PGA notation is significantly better than the initial PGA notation, i.e. whether the improvement actually occurs.

The information required to achieve the objective can, at first, be found by examining methods described in the current literature (chapter 2). The methods that are relevant to evaluate the intuitiveness of a graphical notation will be used to set up a detailed research method, described later on in this chapter, to evaluate the intuitiveness of the different PGA notations. This research method will be applied in a practical business context. This will yield comparative evaluation results of the intuitiveness of the PGA notation (chapter 4).

The relevant research approach to deliver this information is described by Saunders et al. (2019) as deduction. With a deductive approach the research starts with building up a theory, based on current literature. This theory is then tested by designing a research strategy, which collects data and elaborates on hypothesis testing. For this research, the deductive approach is chosen. Current literature (see chapter 2) contains a solid base to build up a theory about how to evaluate the intuitiveness of a graphical notation.

In the literature found in chapter 2, a clear preference emerged about the research strategy. All examined literature conducted one or more experiments to evaluate the intuitiveness of a graphical notation. Saunders et al. (2019, p. 190) define the purpose of an experiment "to study the probability of a change in an independent variable causing a change in another dependent variable". As noted in chapter 2, the independent variables are the initial and adapted PGA notations. Four possible dependent variables are interpretational effectiveness, interpretational efficiency, perceived usefulness and perceived ease of use (see paragraph 2.3.3).

3.2. Technical design: elaboration of the method

The experiment is chosen as the relevant research strategy. This section describes how this research will be conducted in detail. The same aspects as used in the literature will be used: variables and measures, hypotheses, experimental design, instrumentation and experimental tasks, selection of participants and operational procedures.

The next paragraphs will describe each aspect in more detail and explain which choices are made.

Variables and Measures

In this research two independent variables will be evaluated and compared:

- The initial PGA notation.
  The initial notation consists of 11 PGA concepts. This research builds on the research done by Roelens and Bork (2020), resulting in the following six PGA concepts that will be taken into account: Competence, Value proposition, Internal goal, Customer goal, Value stream and Importance.
- The adapted PGA notation.
  The six suggested adaptations mentioned in the research done by Roelens and Bork (2020).

To determine the intuitiveness, the literature found in chapter 2 reveals four dependent variables. Two of them are objective variables: interpretational effectiveness, mentioned in six articles (Bork et al., 2019; Caire et al., 2013; El Kouhen et al., 2015; Moody et al., 2012; Roelens & Bork, 2020; Santos et al., 2018) and interpretational efficiency, mentioned in one article (Bork et al., 2019). Two of them are subjective variables: perceived usefulness, mentioned in two articles (Genon et al., 2012; Moody et al., 2012) and perceived ease of use, mentioned in two articles (Genon et al., 2012; Moody et al., 2012). Interpretational efficiency is only mentioned in one article and left out of this research, meaning that the next three dependent, one objective and two subjective, variables will be examined.

Objective dependent variable:

- Interpretational effectiveness
  This will be measured by the number of correct answers (hit rate) to a number of different questions.

Subjective dependent variables:

- Perceived ease of use
  This will be measured by asking participants questions about how easily they could interpret the meaning of a mentioned concept of a specific notation, based on a five point Likert scale, ranging from ‘Strongly disagree’ to ’ Strongly agree’.
- Perceived usefulness
  This will be measured by asking participants questions about how they experience the usefulness of a specific notation, based on a five point Likert scale, ranging from ‘Strongly disagree’ to ’ Strongly agree’.

Hypotheses

The hypotheses are formulated to compare the variables of intuitiveness between the two independent variables, the initial PGA notation and the adapted PGA notation. The initial PGA notation is designed by the creators of the PGA technique. The adapted PGA notation is the result of an experimental technique, conducted by Roelens and Bork (2020), with the goal of achieving a more intuitive PGA notation. Therefore, it is expected that every variable of intuitiveness of the adapted PGA notation will be higher than that of the initial PGA notation.

Based on the dependent and independent variables, the next hypotheses are formulated:

$H_{effectivenessTotal}$: The interpretational effectiveness of the adapted PGA notation (i.e. the total of all six concepts) is higher than the interpretational effectiveness of the initial PGA notation (i.e. the total of all six concepts).

$H_{effectivenessCompetence}$: The interpretational effectiveness of the adapted PGA notation for the concept Competence is higher than the interpretational effectiveness of the initial PGA notation for the concept Competence.

$H_{effectivenessValueProposition}$: The interpretational effectiveness of the adapted PGA notation for the concept Value Proposition is higher than the interpretational effectiveness of the initial PGA notation for the concept Value Proposition.

$H_{effectivenessInternalGoal}$: The interpretational effectiveness of the adapted PGA notation for the concept Internal Goal is higher than the interpretational effectiveness of the initial PGA notation for the concept Internal Goal.

$H_{effectivenessCustomerGoal}$: The interpretational effectiveness of the adapted PGA notation for the concept Customer Goal is higher than the interpretational effectiveness of the initial PGA notation for the concept Customer Goal.

$H_{effectivenessValueStream}$: The interpretational effectiveness of the adapted PGA notation for the concept Value Stream is higher than the interpretational effectiveness of the initial PGA notation for the concept Value Stream.

$H_{effectivenessImportance}$: The interpretational effectiveness of the adapted PGA notation for the concept Importance is higher than the interpretational effectiveness of the initial PGA notation for the concept Importance.

$H_{ease\ of\ useOverall}$: The overall perceived ease of use of the adapted PGA notation is higher than the overall perceived ease of use of the initial PGA notation.

$H_{ease\ of\ useCompetence}$: The perceived ease of use of the adapted PGA notation for the concept Competence is higher than the perceived ease of use of the initial PGA notation for the concept Competence.

$H_{ease\ of\ useValueProposition}$: The perceived ease of use of the adapted PGA notation for the concept Value Proposition is higher than the perceived ease of use of the initial PGA notation for the concept Value Proposition.

$H_{ease\ of\ useInternalGoal}$: The perceived ease of use of the adapted PGA notation for the concept Internal Goal is higher than the perceived ease of use of the initial PGA notation for the concept Internal Goal.

$H_{ease\ of\ useCustomerGoal}$: The perceived ease of use of the adapted PGA notation for the concept Customer Goal is higher than the perceived ease of use of the initial PGA notation for the concept Customer Goal.

$H_{ease\ of\ useValueStream}$: The perceived ease of use of the adapted PGA notation for the concept Value Stream is higher than the perceived ease of use of the initial PGA notation for the concept Value Stream.

$H_{\text{ease of useImportance}}$: The perceived ease of use of the adapted PGA notation for the concept Importance is higher than the perceived ease of use of the initial PGA notation for the concept Importance.

$H_{\text{usefulnessOverall}}$: The overall perceived usefulness of the adapted PGA notation is higher than the overall perceived usefulness of the initial PGA notation.

The perceived usefulness concerns the usefulness of the complete notation, how it can describe the entire business architecture. Because this concerns the overall perception of usefulness, no hypotheses are formulated for the perceived usefulness of individual concepts.

Experimental design

In the current literature, the most commonly used experimental design is the between-subjects design (Caire et al., 2013; El Kouhen et al., 2015; Genon et al., 2012; Moody et al., 2012; Santos et al., 2018). The studied literature made use of a high number of participants, which results in two relative large groups after dividing the participants. This research compares the intuitiveness of two different PGA notations. With a within-subjects design, which could be a better choice with a small number of participants, care must be taken for carryover effects where familiarity might affect the outcomes (Saunders et al., 2019). Because this familiarity effect could be of large influence while comparing the intuitiveness of two different PGA notations with a closely related graphical design, a between-subjects design is chosen, despite the small number of participants.

Instrumentation and experimental tasks

This research focusses on comparing two different PGA notations that are already present, the initial one and the adapted one. Three different experiments, designed to evaluate and compare notations will be conducted (Bork et al., 2019; Caire et al., 2013; El Kouhen et al., 2015; Moody et al., 2012; Roelens & Bork, 2020; Santos et al., 2018): the semantic transparency experiment, an experiment including a case study with comprehension questions and the recall/recognition experiment. Although the experiment including a case study with comprehension questions is less described in literature, it has been included in this research. During this experiment not only individual concepts are looked at, but complete diagrams/models, with several different concepts, which is more in line with the practical business context.

With the semantic transparency experiment (also known as notation association) the interpretational effectiveness will be measured by showing a graphical notation of a PGA concept and ask the participant to choose the correct meaning of this concept. During the semantic transparency experiment, the participants are shown PGA notations of one particular set (the initial or adapted one), which they have never seen before.

During the experiment including a case study with comprehension questions a case, represented in a particular PGA notation (the initial or adapted one), is shown to the participant and different comprehension questions about this case are asked. The PGA notation shown will be the same one shown during the semantic transparency experiment. With this experiment the interpretational effectiveness-will be measured.

In the recall/recognition experiment, all PGA concepts and their meaning are shown so a participant can remember them. During the recall/recognition experiment, the participants are shown PGA notations with corresponding meaning of one particular set (the initial or adapted one), the same PGA notation as the one shown during the previous two experiments. After the remembrance phase, the interpretational effectiveness will be measured by showing a graphical notation of a PGA concept and ask the participant to choose the correct meaning of this concept.

Selection of participants

In the current literature, experiments are conducted with naïve undergraduate/master students, commonly without previous knowledge of modeling. This research will be conducted in a practical business context within the financial sector, with around 10-15 business-oriented end-users. These participants are working for a large financial organization and are holding (non-)management positions, which makes it possible that their knowledge and experience show more fluctuations than in the case of naïve students. This means that it is important to know their demographics, their modeling experience and their domain understanding about strategic fit.

Operational procedures

The group of participants will receive an online survey consisting of six phases. To overcome the earlier described familiarity effect, participants are randomly split into two groups. One group is shown the initial PGA notation, the other group is shown the adapted PGA notation. The content of the experimental tasks conducted during phase 2-4 is described in the "Instrumentation and experimental tasks" section.

In summary, the operational procedure consists of the next five phases:

1. Initiation
   First, the participant is explained that taking part in this research is voluntary and that he/she can stop at any time. The results of every participant will be anonymized to guarantee their privacy. After the informed consent, based on the research done by Roelens and Bork (2020), a short introduction is given about the purpose of PGA, without showing any details about the notation. Participants are asked to fill out demographic information about their age, gender, current position and number of years of experience in their current position. Next questions about their modeling experience are asked: how they rate their modeling expertise and which modeling languages they know. At last, questions about the PGA domain are asked: what understanding do they have about strategic planning and strategic fit.
2. Semantic transparency experiment
   During this phase the semantic transparency experiment (also known as notation association) is conducted.
3. Case study
   During this phase an experiment including a case study with comprehension questions is conducted.
4. Recall/recognition experiment
   During this phase the recall/recognition experiment is conducted.
5. Perceived Ease of use and Perceived Usefulness
   During this phase the participants are asked questions about the overall perceived ease of use and overall perceived usefulness. For every PGA concept (Competence, Value proposition, Internal goal, Customer goal, Value stream and Importance) a separate question will be asked

about the perceived ease of use, analogous to Moody et al. (2012), where, during the symbolisation and prototyping experiments, questions were asked to find out how easily each concept could be drawn and interpreted.

6. Conclusion
During this phase participants are asked to give feedback about the completed procedure (Roelens & Bork, 2020).

When all participants concluded their survey, the collected data is verified and made ready for data analysis .

3.3. Data analysis

During all six phases of the operational procedure quantitative data is collected. This data will be used to answer the different hypotheses. To know which statistical tests should be conducted, the collected data needs to be classified for both the dependent and independent variables.

The first classification concerns the type of the dependent variables. During the Semantic transparency experiment, the Recall/recognition experiment and the Case study, information is collected about the interpretational effectiveness of the two different PGA notations. The interpretational effectiveness is measured with the number of correct answers (hit rate), which is numerical interval data. It cannot be determined in advance whether these data are normally distributed. This will be analysed after collecting the data. If so, interpretational effectiveness will be treated as normally distributed data.

The perceived usefulness and the perceived ease of use, both measured based on a five point Likert scale, are both ordinal data.

The second classification is to know which type the independent variables are. The independent variables consist of the different concepts of the initial and the adapted PGA notation. This means that the independent variables are categorical, consisting of two groups. These two groups of categorical data needs to be classified as paired or unpaired. Paired means that both samples consist of the same test subjects. Unpaired means that both samples consist of distinct test subjects. The experiments are conducted with a between-subjects design, the participants are always exposed to one kind of PGA notation at a time. This results in unpaired data regarding the PGA notations.

Table 7 shows the corresponding statistical tests that need to be conducted.

Table 7 Statistical tests

Dependent variable	Independent variable Two groups Unpaired
Ordinal	Wilcoxon-Mann-Whitney Test
Interval	Wilcoxon-Mann-Whitney Test
Normal	t-Test (for unpaired)

3.4. Reflection w.r.t. validity, reliability and ethical aspects

Construct validity

Saunders et al. (2019) defines construct validity as “the extent to which your measurement questions actually measure the presence of those constructs you intended them to measure” (Saunders et al., 2019, p. 799).

Measurements done in this research are based on measurements described in current literature. The questions asked in the survey are based on questions asked in the current literature, which measure the intuitiveness of graphical notations.

Internal validity

Saunders et al. (2019) defines internal validity as “the extent to which findings can be attributed to interventions rather than any flaws in your research design” (Saunders et al., 2019, p. 806).

For this research experiments are conducted. The experiments are digitally conducted, by completing a survey, by a small group of participants. The operational procedures guarantee that every participant gets the same kind of information. To exclude disruptive factors as much as possible the participants are asked to conduct the survey within a limited time period and besides that asked to not communicate about the survey with other participants.

The participants are business-oriented end users. Their profile of novice end-users is suitable to objectively determine whether a PGA notation is intuitive (see paragraph 2.3.1).

The experiments will first be pre-tested to see if they are technically implemented correctly. This will be done with a small group of respondents without much prior knowledge (e.g. family or acquaintances) who will not participate in the final research.

External validity

Saunders et al. (2019) defines external validity as “the extent to which the research results from a particular study are generalizable to all relevant contexts” (Saunders et al., 2019, p. 803).

The experiments are conducted in a practical business context within the financial sector. All end-users are working for a large financial organization. This context is more in line with the intended, business-oriented, end users of the PGA method. It also results in a mixed composition of end-users, based on their demographic information, management position, modeling experience and knowledge about the PGA domain. This results in greater generalizability.

A threat for the external validity is the limited number of participants. Because of this low number of participants, the likelihood that the results of this research are generalizable is low.

Reliability

Saunders et al. (2019) defines reliability as “the extent to which data collection technique or techniques will yield consistent findings, similar observations would be made or conclusions reached by other researchers or there is transparency in how sense was made from the raw data" (Saunders et al., 2019, p. 815).

To increase the reliability of this research, three dependent variables will be measured. One objective variable (i.e. interpretational effectiveness) and two subjective (i.e. perceived usefulness and perceived ease of use) variables. Objective variables are based on facts, e.g. number of correct answers (hit rate). Subjective variables are based on personal opinion, a perception a certain person experiences. Using both objective and subjective variables, allows for triangulation of data. Saunders et al. (2019) defines triangulation as "use of two or more independent sources of data or data-collection methods within one study in order to help ensure that the data are telling you what you think they are telling you" (Saunders et al., 2019, p. 819). The different data sources will make it possible to statistically check whether the adapted version will be more intuitive.

Besides triangulation of data, the operational procedures, including experiments and survey will be documented. To avoid influencing the participants, they are asked to fill out the survey digitally, at a place and time they can choose themselves, without intervention of any researcher.

Research ethics

Saunders et al. (2019) defines research ethics as "standards of the researcher's behaviour in relation to the rights of those who become subject of a research project, or who are affected by it" (Saunders et al., 2019, p. 815).

During the introduction phase, every participant is explained that taking part in this research is voluntary and that he/she can stop at any time. The results of every participant will be anonymized to guarantee their privacy. The organization where this research will take place will be anonymized too, and confidentiality agreements will be made. For the assessment, all research data will be submitted to the OU for verification and stored in a secure and anonymous manner.

4. Results

In chapter 3, a method is designed in order to answer sub-question 3: 'How can the intuitiveness of the different PGA notations be empirically evaluated within a practical business context?'. This method has been conducted in practice, and the results are described in this chapter.

Paragraph 4.1 describes how the various experiments were conducted. The next paragraphs will show the results of the conducted experiments. After accepting the informed consent, questions were asked about the demographics of the participants (paragraph 4.2). The interpretational effectiveness was measured by the semantic transparency experiment, the case study and the recall/recognition experiment (paragraph 4.3). Then questions were asked about the perceived ease of use (paragraph 4.4) and perceived usefulness (paragraph 4.5). At last respondents were asked to provide feedback about the survey itself and/or about the PGA notation (paragraph 4.6).

4.1. Conducting the experiments

The results of this research are obtained by conducting an online survey (see appendix 5). Before the survey was finally distributed among the intended participants, it was tested among family and colleagues and updated based on their comments. In total 35 participants within a large financial organization were asked to fill out this survey. After the deadline of two and a half weeks had passed, 25 participants filled out the survey, of which 11 partially and 14 completely. The responses were downloaded in Excel and transformed in such a way that they could be used directly for the statistical analysis in SPSS (see appendix 6).

During the analysis of the results for this research, the 14 completely filled out surveys were taken into account. As elaborated in chapter 3, a between-subjects design was chosen for this research. Based on their year of birth, the participants were divided in two groups. Six respondents , with an odd year of birth, were shown the initial PGA notation and eight respondents, with an even year of birth, were shown the adapted PGA notation.

The relevant statistical tests to obtain the results of paragraphs 4.3, 4.4 and 4.5 are described in appendix 7. The raw results of all statistical tests used in paragraphs 4.2, 4.3, 4.4 and 4.5 are included in appendix 8.

4.2. Demographics

After accepting the informed consent, questions were asked about the demographics of the participants. This resulted in the next observations:

- Six respondents hold a non-management position within the organization, seven a tactical management position and one a strategic management position.
- Two respondents rated their model experience as very low, three rated their model experience as low, eight rated their model experience as medium and one rated his/her model experience as high.
- Four respondents had no experience with modelling languages, five had experience with one kind of modelling languages, two had experience with two kinds of modelling languages and three had experience with three kinds of modelling languages.

- Three respondents rated their understanding of strategic fit as low, nine respondents as medium and two respondents as high.

4.3. Interpretational effectiveness

The interpretational effectiveness was measured by the semantic transparency experiment, the case study and the recall/recognition experiment. During these experiments the interpretational effectiveness of the six PGA concepts (Competence, Value proposition, Internal goal, Customer goal, Value stream, Importance) was measured.

Based on the dependent and independent variables, the next hypotheses were formulated:

$H_{effectivenessTotal}$: The interpretational effectiveness of the adapted PGA notation (i.e. the total of all six concepts) is higher than the interpretational effectiveness of the initial PGA notation (i.e. the total of all six concepts).

$H_{effectivenessCompetence}$: The interpretational effectiveness of the adapted PGA notation for the concept Competence is higher than the interpretational effectiveness of the initial PGA notation for the concept Competence.

$H_{effectivenessValueProposition}$: The interpretational effectiveness of the adapted PGA notation for the concept Value Proposition is higher than the interpretational effectiveness of the initial PGA notation for the concept Value Proposition.

$H_{effectivenessInternalGoal}$: The interpretational effectiveness of the adapted PGA notation for the concept Internal Goal is higher than the interpretational effectiveness of the initial PGA notation fo the concept Internal Goal.

$H_{effectivenessCustomerGoal}$: The interpretational effectiveness of the adapted PGA notation for the concept Customer Goal is higher than the interpretational effectiveness of the initial PGA notation for the concept Customer Goal.

$H_{effectivenessValueStream}$: The interpretational effectiveness of the adapted PGA notation for the concept Value Stream is higher than the interpretational effectiveness of the initial PGA notation for the concept Value Stream.

$H_{effectivenessImportance}$: The interpretational effectiveness of the adapted PGA notation for the concept Importance is higher than the interpretational effectiveness of the initial PGA notation for the concept Importance.

Semantic transparency experiment

During the semantic transparency experiment, the interpretational effectiveness was measured by showing a graphical notation of a PGA concept and ask the participant to choose the correct meaning of this concept. The six PGA concepts shown were: Competence, Value proposition, Internal goal, Customer goal, Value stream and Importance.

Table 8 shows, per PGA concept, the results of the semantic transparency experiment. The last row shows the results of the total of all six PGA concepts.

Table 8 Results semantic transparency experiment

Dependent variable	Count Initial	Mean (Rank) Initial	Count Adapted	Mean (Rank) Adapted	Mean (Rank) Adapted > Mean (Rank) Initial	Calculated one sided p-value	Significant
ST Competence	6	8,83	8	6,5	No	0,869	No
ST ValueProposition	6	6,75	8	8,06	Yes	0,269	No
ST InternalGoal	6	6,67	8	8,13	Yes	0,2455	No
ST CustomerGoal	6	7,33	8	7,63	Yes	0,416	No
ST ValueStream	6	6,5	8	8,25	Yes	0,189	No
ST Importance	6	4,75	8	9,56	Yes	0,012	Yes
ST total	6	5,5	8	7,25	Yes	0,143	No

The results of the total of all the individual PGA concepts indicate that the adapted PGA notation (mean 7,25) performs better than the initial PGA notation (mean 5,50).

To find out if the difference is significant enough, first the one-sided significance is 0.143 is calculated. The significance of 0.143 is greater than 0.05, meaning that the alternative hypothesis H1 will be rejected. Based on the total result it cannot be stated that the adapted PGA notation performs significantly better than the initial PGA notation.

When analysing the results of the six individual PGA concepts, it shows that this statement is supported by all PGA concepts, except Importance. The adapted PGA notation performs better than the initial PGA notation for PGA concept Importance.

Case study

A case, represented in a particular PGA notation (the initial or adapted one), was shown to the participant and comprehension questions about this case were asked. For each of the six PGA concepts (Competence, Value proposition, Internal goal, Customer goal, Value stream and Importance) one comprehension question was asked.

Table 9 shows, per PGA concept, the results of the case study experiment. The last row shows the results of the total of all six PGA concepts.

Table 9 Results case study experiment

Dependent variable	Count Initial	Mean (Rank) Initial	Count Adapted	Mean (Rank) Adapted	Mean (Rank) Adapted > Mean (Rank) Initial	Calculated one sided p-value	Significant
C Competence	6	6,67	8	8,13	Yes	0,2055	No
C ValueProposition	6	6,83	8	8	Yes	0,274	No
C InternalGoal	6	7,17	8	7,75	Yes	0,3585	No
C CustomerGoal	6	6,33	8	8,38	Yes	0,149	No
C ValueStream	6	6,17	8	8,5	Yes	0,1075	No
C Importance	6	5,83	8	8,75	Yes	0,0665	No
C total	6	1,5	8	3	Yes	0,02	Yes

The results of the total of all the individual PGA concepts indicate that the adapted PGA notation (mean 3) performs better than the initial PGA notation (mean 1,5).

To find out if the difference is statistically significant, first the one-sided significance is 0.02 is calculated. The significance of 0.02 is less than 0.05, meaning that the alternative hypothesis H1 will be accepted. Based on the total result it can be stated that the adapted PGA notation performs significantly better than the initial PGA notation.

When analysing the results of the six individual PGA concepts, it shows that this effect cannot be clearly explained by any of the individual PGA concepts.

Recall/recognition experiment

In the recall/recognition experiment, all PGA concepts and their meaning were shown so a participant could remember them. After the remembrance phase, the interpretational effectiveness was measured by showing a graphical notation of a PGA concept and ask the participant to choose the correct meaning of this concept. The six PGA concepts shown were: Competence, Value proposition, Internal goal, Customer goal, Value stream and Importance.

Table 10 shows, per PGA concept, the results of the recall/recognition experiment. The last row shows the results of the total of all six PGA concepts.

Table 10 Results recall/recognition experiment

Dependent variable	**Count Initial**	**Mean Rank Initial**	**Count Adapted**	**Mean Rank Adapted**	**Mean Rank Adapted > Mean Rank Initial**	**Calculated one sided p-value**	**Significant**
REC Competence	6	6,17	8	8,5	Yes	0,0445	Yes
REC ValueProposition	6	9	8	6,38	No	0,9485	No
REC InternalGoal	6	7,5	8	7,5	No	0,5	No
REC CustomerGoal	6	7,5	8	7,5	No	0,5	No
REC ValueStream	6	6,83	8	8	Yes	0,124	No
REC Importance	6	6,17	8	8,5	Yes	0,0445	Yes
REC total	6	6,75	8	8,06	Yes	0,254	No

The results of the total of all the individual PGA concepts indicate that the adapted PGA notation (mean 8,06) performs better than the initial PGA notation (mean 6,75).

To find out if the difference is statistically significant, first the one-sided significance is 0.254 is calculated. The significance of 0.254 is greater than 0.05, meaning that the alternative hypothesis H1 will be rejected. Based on the total result it cannot be stated that the adapted PGA notation performs significantly better than the initial PGA notation.

When analysing the results of the six individual PGA concepts, it shows that a significant improvement is observed for two individual PGA concepts. The adapted PGA notation performs better than the initial notation for the concepts Competence and Importance.

4.4. Perceived ease of use

After the three experiments regarding interpretational effectiveness, questions were raised about the perceived ease of use of the PGA notations. First four questions were raised about the overall perceived ease of use. After that, for each of the six shown PGA concepts (Competence, Value proposition, Internal goal, Customer goal, Value stream, Importance) a separate question was raised about the perceived ease of use of that PGA concept.

Based on the dependent and independent variables, the next hypotheses were formulated:

$H_{ease\ of\ useOverall}$: The overall perceived ease of use of the adapted PGA notation is higher than the overall perceived ease of use of the initial PGA notation.

$H_{ease\ of\ useCompetence}$: The perceived ease of use of the adapted PGA notation for the concept Competence is higher than the perceived ease of use of the initial PGA notation for the concept Competence.

$H_{ease\ of\ useValueProposition}$: The perceived ease of use of the adapted PGA notation for the concept Value Proposition is higher than the perceived ease of use of the initial PGA notation for the concept Value Proposition.

$H_{ease\ of\ useInternalGoal}$: The perceived ease of use of the adapted PGA notation for the concept Internal Goal is higher than the perceived ease of use of the initial PGA notation for the concept Internal Goal.

$H_{ease\ of\ useCustomerGoal}$: The perceived ease of use of the adapted PGA notation for the concept Customer Goal is higher than the perceived ease of use of the initial PGA notation for the concept Customer Goal.

$H_{ease\ of\ useValueStream}$: The perceived ease of use of the adapted PGA notation for the concept Value Stream is higher than the perceived ease of use of the initial PGA notation for the concept Value Stream.

$H_{ease\ of\ useImportance}$: The perceived ease of use of the adapted PGA notation for the concept Importance is higher than the perceived ease of use of the initial PGA notation for the concept Importance.

Table 11 shows, per PGA concept, the results of the perceived ease of use questions. The last row shows the results of the overall perceived ease of use of the PGA notation.

Table 11 Results perceived ease of use questions

PGA Concept	Count Initial	Mean Rank Initial	Count Adapted	Mean Rank Adapted	Mean Rank Adapted > Mean Rank Initial	Calculated one sided p-value	Significant
Competence	6	7,33	8	7,63	Yes	0,4445	No
ValueProposition	6	5,67	8	8,88	Yes	0,067	No
InternalGoal	6	6,33	8	8,38	Yes	0,1715	No
CustomerGoal	6	6,08	8	8,56	Yes	0,1105	No
ValueStream	6	5,42	8	9,06	Yes	0,0485	Yes
Importance	6	3,92	8	10,19	Yes	0,002	Yes
Overall perceived Ease of Use	6	4,17	8	10	Yes	0,0045	Yes

The results of the overall perceived Ease of use indicate that the adapted PGA notation (mean 10) performs better than the initial PGA notation (mean 4,17).

To find out if the difference is statistically significant, first the one-sided significance is 0.0045 is calculated. The significance of 0.0045 is less than 0.05, meaning that the alternative hypothesis H1 will be accepted. Based on the overall perceived Ease of use result it can be stated that the adapted PGA notation performs significantly better than the initial PGA notation.

When analysing the results of the six individual PGA concepts, it shows that this statement is supported by two individual PGA concepts. The adapted PGA notation performs significantly better than the initial notation for concepts Value stream and Importance.

4.5. Perceived usefulness

After the questions about the perceived ease of use, three questions were raised about the overall perceived usefulness of the PGA notation.

Based on the dependent and independent variables, the next hypothesis was formulated:

$H_{usefulnessOverall}$: The overall perceived usefulness of the adapted PGA notation is higher than the overall perceived usefulness of the initial PGA notation.

Table 12 shows the results of the overall perceived usefulness of the PGA notation.

Table 12 Results overall perceived usefulness

	Count Initial	Mean Rank Initial	Count Adapted	Mean Rank Adapted	Mean Rank Adapted > Mean Rank Initial	Calculated one sided p-value	Significant
Overall perceived Usefulness	6	5,33	8	9,13	Yes	0,0305	Yes

The results of the overall perceived Usefulness indicate that the adapted PGA notation (mean 9,13) performs better than the initial PGA notation (mean 5,33).

To find out if the difference is statistically significant, first the one-sided significance is 0.0305 is calculated. The significance of 0.0305 is less than 0.05, meaning that the alternative hypothesis H1 will be accepted. Based on the results, it can be stated that the overall perceived usefulness of the adapted PGA notation is significantly higher than the overall perceived usefulness of the initial PGA notation.

4.6. Feedback

At the end of the survey the participants could provide positive and/or negative feedback about the survey itself and about the PGA notation.

The participants indicated that the notation was difficult to understand at first, during the semantic transparency experiment and the case study, but a lot easier after the recall/recognition experiment. This effect is also clearly visible in the hit rates of these experimental tasks.

5. Discussion, conclusions and recommendations

This chapter contains a discussion of the results as described in chapter 4. Next, reflections on validity, reliability and research ethics are done. Thereafter, conclusions are drawn and recommendations for practice and further research will be given.

5.1. Discussion

This research builds on the research done by Roelens and Bork (2020), which aimed to achieve a more intuitive PGA notation. It resulted in six suggested adaptations to the initial PGA notation, concerning: Competence, Value proposition, Internal goal, Customer goal, Value stream and Importance. In this research, the intuitiveness of the initial and suggested adaptations of the PGA notation were evaluated and compared in a practical business context.

As described in chapter 3, three dependent variables are examined to determine the intuitiveness of the two PGA notations: one objective dependent variable (i.e. the interpretational effectiveness) and two subjective dependent variables (i.e. perceived ease of use and perceived usefulness). In chapter 3, hypotheses were also formulated regarding the intuitiveness of the independent variable(i.e. the initial PGA notation and the adapted PGA notation), with the expectation that every variable of intuitiveness of the adapted PGA notation will be higher than that of the initial PGA notation. These hypotheses were tested in chapter 4 and, in order to answer sub-question 4, 'Does the adapted PGA notation perform intuitively better in a practical business context?', the results are discussed below.

Overall

Table 13 shows the hypotheses that were formulated regarding the overall effectiveness, perceived ease of use and perceived usefulness of the two PGA notations and whether there is a significant improvement observed during the various experiments.

Table 13 Results hypotheses overall effectiveness, perceived ease of use and perceived usefulness

Hypothesis	**Significant improvement observed?**				
	Semantic transparency	**Case study**	**Recall / recognition**	**perceived ease of use**	**perceived usefulness**
$H_{effectivenessTotal}$: The interpretational effectiveness of the adapted PGA notation (i.e. the total of all six concepts) is higher than the interpretational effectiveness of the initial PGA notation (i.e. the total of all six concepts)	No	Yes	No	n/a	n/a
$H_{ease\ of\ useOverall}$: The overall perceived ease of use of the adapted PGA notation is higher than the overall perceived ease of use of the initial PGA notation	n/a	n/a	n/a	Yes	n/a
$H_{usefulnessOverall}$: The overall perceived usefulness of the adapted PGA notation is higher than the overall perceived usefulness of the initial PGA notation	n/a	n/a	n/a	n/a	Yes

Conclusion: with a significant improvement observed during the case study experiment for the overall interpretational effectiveness, the adapted PGA notation seems to perform better than the initial one, and based on the overall perceived ease of use and usefulness, the participants also have the feeling that the adapted PGA notation performs better. Based on this overall result, it seems worth considering to replace certain concepts from the initial PGA notation with the adapted ones.

Next sections will elaborate on each PGA concept separately to draw a conclusion for the individual concepts. This conclusion is based on a comparison between the concepts.

Competence

Table 14 shows the hypotheses that were formulated for the concept Competence regarding the effectiveness and perceived ease of use of the two PGA notations and whether there is a significant improvement observed during the various experiments.

Table 14 Results hypotheses concept Competence

Hypothesis	**Significant improvement observed?**			
	Semantic transparency	**Case study**	**Recall / recognition**	**perceived ease of use**
$H_{effectivenessCompetence}$: The interpretational effectiveness of the adapted PGA notation for the concept Competence is higher than the interpretational effectiveness of the initial PGA notation for the concept Competence	No	No	Yes	n/a
$H_{ease\ of\ useCompetence}$: The perceived ease of use of the adapted PGA notation for the concept Competence is higher than the perceived ease of use of the initial PGA notation for the concept Competence	n/a	n/a	n/a	No

For the concept Competence , a significant improvement is observed during the recall / recognition experiment for the interpretational effectiveness. Compared to the other five adapted PGA concepts, of which four (i.e. Value Proposition, Internal Goal, Customer Goal, Value Stream) do not show any significant improvement for the interpretational effectiveness, it seems worth considering to replace the initial notation of the concept with the adapted one. However, this significant improvement for the interpretational effectiveness (objective), cannot be supported by the perceived ease of use (subjective), which shows no significant improvement. Despite the (limited) statistically significant result, the adapted notation can provide a suitable starting point for further redesign. This results in the conclusion that the notation of this concept still requires rethinking in order to arrive at a further improved notation.

Value Proposition

Table 15 shows the hypotheses that were formulated for the concept Value Proposition regarding the effectiveness and perceived ease of use of the two PGA notations and whether there is a significant improvement observed during the various experiments.

Table 15 Results hypotheses concept Value Proposition

Hypothesis	**Significant improvement observed?**			
	Semantic transparency	**Case study**	**Recall / recognition**	**perceived ease of use**
$H_{effectivenessValueProposition}$: The interpretational effectiveness of the adapted PGA notation for the concept Value Proposition is higher than the interpretational effectiveness of the initial PGA notation for the concept Value Proposition	No	No	No	n/a
$H_{ease\ of\ useValueProposition}$: The perceived ease of use of the adapted PGA notation of concept Value Proposition is higher than the perceived ease of use of the initial PGA notation of concept Value Proposition	n/a	n/a	n/a	No

For the concept Value Proposition, no significant improvement is observed for both the interpretational effectiveness and perceived ease of use. This results in the conclusion that the notation of this concept requires further redesign in order to arrive at an improved notation.

Internal Goal

Table 16 shows the hypotheses that were formulated for the concept Internal Goal regarding the effectiveness and perceived ease of use of the two PGA notations and whether there is a significant improvement observed during the various experiments.

Table 16 Results hypotheses concept Internal Goal

Hypothesis	**Significant improvement observed?**			
	Semantic transparency	**Case study**	**Recall / recognition**	**perceived ease of use**
$H_{effectivenessInternalGoal}$: The interpretational effectiveness of the adapted PGA notation for the concept Internal Goal is higher than the interpretational effectiveness of the initial PGA notation for the concept Internal Goal	No	No	No	n/a
$H_{ease\ of\ useInternalGoal}$: The perceived ease of use of the adapted PGA notation for the concept Internal Goal is higher than the perceived ease of use of the initial PGA notation for the concept Internal Goal	n/a	n/a	n/a	No

For the concept Internal Goal, no significant improvement is observed for both the interpretational effectiveness and perceived ease of use. This results in the conclusion that the notation of this concept requires further redesign in order to arrive at an improved notation.

Customer Goal

Table 17 shows the hypotheses that were formulated for the concept Customer Goal regarding the effectiveness and perceived ease of use of the two PGA notations and whether there is a significant improvement observed during the various experiments.

Table 17 Results hypotheses concept Customer Goal

Hypothesis	**Significant improvement observed?**			
	Semantic transparency	**Case study**	**Recall / recognition**	**perceived ease of use**
$H_{effectivenessCustomerGoal}$: The interpretational effectiveness of the adapted PGA notation for the concept Customer Goal is higher than the interpretational effectiveness of the initial PGA notation for the concept Customer Goal	No	No	No	n/a
$H_{ease\ of\ useCustomerGoal}$: The perceived ease of use of the adapted PGA notation for the concept Customer Goal is higher than the perceived ease of use of the initial PGA notation for the concept Customer Goal	n/a	n/a	n/a	No

For the concept Customer Goal, no significant improvement is observed for both the interpretational effectiveness and perceived ease of use. This results in the conclusion that the notation of this concept requires further redesign in order to arrive at an improved notation.

Value Stream

Table 18 shows the hypotheses that were formulated for the concept Value Stream regarding the effectiveness and perceived ease of use of the two PGA notations and whether there is a significant improvement observed during the various experiments.

Table 18 Results hypotheses concept Value Stream

Hypothesis	**Significant improvement observed?**			
	Semantic transparency	**Case study**	**Recall / recognition**	**perceived ease of use**
$H_{effectivenessValueStream}$: The interpretational effectiveness of the adapted PGA notation for the concept Value Stream is higher than the interpretational effectiveness of the initial PGA notation for the concept Value Stream	No	No	No	n/a
$H_{ease\ of\ useValueStream}$: The perceived ease of use of the adapted PGA notation for the concept Value Stream is higher than the perceived ease of use of the initial PGA notation for the concept Value Stream	n/a	n/a	n/a	Yes

For the concept Value Stream, a significant improvement is observed for the perceived ease of use. Compared to the other five concepts, of which four (i.e. Competence, Value Proposition, Internal

Goal, Customer Goal) do not show any significant improvement for the perceived ease of use, this seems worth considering to replace the initial notation of the concept with the adapted one. However, this significant improvement for the perceived ease of use (subjective), cannot be supported by the interpretational effectiveness (objective), which shows no significant improvement. Despite the (limited) statistically significant result, the adapted notation can provide a suitable starting point for further redesign. This results in the conclusion that the notation of this concept still requires rethinking in order to arrive at a further improved notation.

Importance

Table 19 shows the hypotheses that were formulated for the concept Importance regarding the effectiveness and perceived ease of use of the two PGA notations and whether there is a significant improvement observed during the various experiments.

Table 19 Results hypotheses concept Importance

Hypothesis	**Significant improvement observed?**			
	Semantic transparency	**Case study**	**Recall / recognition**	**perceived ease of use**
$H_{effectivenessImportance}$: The interpretational effectiveness of the adapted PGA notation for the concept Importance is higher than the interpretational effectiveness of the initial PGA notation for the concept Importance	Yes	No	Yes	n/a
$H_{ease\ of\ useImportance}$: The perceived ease of use of the adapted PGA notation for the concept Importance is higher than the perceived ease of use of the initial PGA notation for the concept Importance	n/a	n/a	n/a	Yes

For the concept Importance, a significant improvement is observed for two experiments of the interpretational effectiveness (objective). This result is supported by a significant improvement for the perceived ease of use (subjective). Compared to the other concepts, of which four (i.e. Value Proposition, Internal Goal, Customer Goal, Value Stream) do not show any significant improvement for the interpretational effectiveness and four (i.e. Competence, Value Proposition, Internal Goal, Customer Goal) that do not show any significant improvement for the perceived ease of use, this results in the conclusion that it is recommended to replace the initial notation of the concept with the adapted one.

5.2. Reflection w.r.t. validity, reliability and ethical aspects

Based on the reflection of chapter 3, the following sections describe the reflection done after conducting the experiments.

Construct validity

As mentioned in chapter 3, measurements done in this research are based on measurements described in current literature. The questions asked in the survey are based on questions asked in the current literature, which measure the intuitiveness of graphical notations.

The six adapted PGA notations were the result of experiments done by Roelens and Bork (2020) with the goal of achieving a more intuitive PGA notation. It was expected that the adapted PGA notation would be more intuitive than the initial PGA notation. The conducted experiments should show significant improvements.

The results (see paragraph 5.1) show that expectations were only partially met for the measurements of the interpretational effectiveness and perceived ease of use. Therefore, further analysis is required to replicate the research and determine whether the questions asked during the semantic transparency experiment, case study, recall/recognition experiment and perceived ease of use should be asked differently or that other experiments should be conducted to measure the interpretational effectiveness and perceived ease of use in a more valid way.

Internal validity

The survey was created with LimeSurvey, which is a free, open source web application for creating online surveys (see Appendix 5).

Before distributing the survey to the intended participants, it was first pre-tested by two colleagues and one family member to see if the questions were technically implemented correctly. The pre-test participants had little to no prior knowledge with modeling notations and they did not participate in the final research. Their remarks (i.e. changing answer options for gender to Male, Female, Other and Prefer not to say and moving the answer option None from first position to last position when asking which conceptual modeling languages the participant knows) were incorporated in the final version of the survey.

The final version of the survey was distributed to all intended participants by email. In this email they were asked to fill out the survey alone, in one go and at a convenient time for themselves. Completing the survey would take about 15 to 20 minutes. The indicated deadline was two weeks later. All information they provided was collected anonymously. By conducting this survey online, it was guaranteed that every participant got the same kind of information. Intervention of the researcher to split the participants in two groups (to show the initial or adapted PGA notation) was avoided by splitting them randomly, based on their year of birth. Other disruptive factors were excluded as much as possible by asking the participants to conduct the survey within a limited time period and not to communicate about it with other participants.

Despite the above mentioned precautions it is noteworthy to mention that a high number of participants only partially completed the survey (11 out of 25). Six participants stopped after filling in the demographics section, four stopped after filling in the semantic transparency section and one stopped after filling in the recall/recognition section. Because the participants could stop the survey at any point without giving any reason and there was no possibility to give a reason for those who wanted to, it is unclear why so many participants did not completely fill out the survey.

External validity

As mentioned in chapter 2, in former research, participants to the experiments were mainly undergraduate/master students without previous knowledge of modeling (Bork et al., 2019; Caire et al., 2013; El Kouhen et al., 2015; Moody et al., 2012; Roelens & Bork, 2020; Santos et al., 2018). The experiments for this research were conducted in a practical business context within the financial sector. All end-users are working for a large financial organization. It resulted in a mixed composition of participants, based on their demographic information (seven female, seven male, ages from 31 to 57), management position (six non-management, seven tactical management, one strategic management), modeling experience (two very low, three low, eight medium and one high), experience with modeling languages (four had no experience, five with one language, two with two languages and three with three languages) and knowledge about the PGA domain (three low, nine high). This composition results in a greater generalizability of the results.

In this research, the research method, designed in chapter 3, is used to evaluate two PGA notations. The generic experimental setup makes this research method suitable for investigating other DSMLs as well, resulting in a greater generalizability.

However, the limited number of 14 participants is an inherent drawback, which negatively influences the generalizability of the results.

Reliability

To increase the reliability of this research, three dependent variables were measured. One objective variable (i.e. interpretational effectiveness) and two subjective (i.e. perceived usefulness and perceived ease of use) variables. Using both objective and subjective variables, allowed for triangulation of data. Saunders et al. (2019) defines triangulation as "use of two or more independent sources of data or data-collection methods within one study in order to help ensure that the data are telling you what you think they are telling you" (Saunders et al., 2019, p. 819). In this research, three independent data collection methods within one study are used to statistically check whether the adapted version will be more intuitive than the initial one.

As another independent source of data, positive and/or negative feedback about the survey itself and about the PGA notation was asked at the end of the survey. There were comments about the survey itself, mainly about the difficulty of the subject and the ordering of experiments, but no substantive feedback was given about the PGA notation itself. In the end, this feedback could not be used for further triangulation of the data.

Influencing of the participants was avoided by asking them to fill out the survey digitally, at a place and time they could choose themselves, without intervention of any researcher.

The operational procedures, including experimental tasks, survey, raw response information, data transformations done and results of the statistical tests, are documented (see appendices 5 to 8).

Research ethics

The survey starts with a 'Declaration of consent for participation in scientific research', stating that the participant:

- has been informed about the research and has read the given context information
- knows how to ask questions about the research
- has been able to think about his/her participation in the study
- understands that he/she can exit the investigation at any time and does not have to give a reason for it
- consents to the use of the data collected during this research for this scientific research
- understands that all information he/she provides regarding this study will be collected anonymously and will not lead back to him/her

- understands that the collected data is kept securely for 10 years

Only when the participant understands and agrees to all of these points, the survey is started. As mentioned, the participant can stop at any time without giving any reason. The results of every participant were anonymized to guarantee their privacy. The organization where this research was conducted was anonymized too, and confidentiality agreements were made. For the assessment, all research data will be submitted to the OU for verification and stored in a secure and anonymous manner.

5.3. Conclusions

To answer the main research question 'How can be determined whether the intuitiveness of the adapted PGA notation is significantly better than the initial PGA notation?', four sub-questions were formulated and answered.

Sub-question 1, 'What is meant by intuitiveness of a graphical notation of conceptual models?', introduced a more formal term for the intuitiveness of a (graphical) notation of a conceptual model: semantic transparency (Caire et al., 2013). This is one of the nine principles of the Physics of Notations theory (Moody, 2009). It means that the meaning (semantics) of a symbol is clear (transparent) from its appearance alone (Caire et al., 2013). A notation that is semantically transparent will lower the cognitive load for end-users because the used graphical symbols act as mnemonics. In this way, novice end-users can visually recognize known domain concepts and terms.

Sub-question 2, 'How to determine whether a graphical notation of a conceptual model is intuitive?', explained that semantic transparency is typically evaluated subjectively by experts (e.g. researchers, notation designers) (Caire et al.,2013). While designing the graphical notation, they try to think like novice end-users, but because of their inherent knowledge, that is difficult to accomplish. To objectively determine whether a graphical notation of a conceptual model is intuitive, the semantic transparency should be evaluated empirically, with novice end-users.

Sub-question 3, 'How can the intuitiveness of the different PGA notations be empirically evaluated within a practical business context?', elaborated on the evaluation of the PGA notation. Roelens and Bork (2020) conducted an empirical evaluation of the initial notation of the PGA technique, which resulted in adaptations to the notation for six PGA concepts, but they could not statistically ascertain that the adapted PGA notation was an improvement compared to the initial one.
This research introduced a method to compare two notations in order to determine which notation performs significantly better than the other. To determine the intuitiveness of the different PGA notations three dependent variables were used, one objective (i.e. interpretational effectiveness) and two subjective (i.e. perceived ease of use and perceived usefulness). To examine the interpretational effectiveness, three experiments were designed, a semantic transparency experiment, a case study with comprehension questions and a recall/recognition experiment. To examine the perceived ease of use and perceived usefulness questions were asked about the overall perceived ease of use, the perceived ease of use for every PGA concept and the overall perceived usefulness. As experimental design, a between-subjects design was used, where the participants were divided in two groups. One group is shown the initial PGA notation, the other group is shown the adapted PGA notation.

Sub-question 4, ‘Does the adapted PGA notation perform intuitively better in a practical business context’, is answered by putting the method, introduced for this research, in practice. When the adapted PGA notation performs significantly better than the initial PGA notation, a change of notation should be implemented.

An online survey was conducted in a practical business context within the financial sector. In total 35 participants within a large financial organization were asked to fill out this survey. After the deadline of two and a half weeks had passed, 25 participants filled out the survey, of which 11 partially and 14 completely. Besides the aforementioned experimental tasks and perception questions, participants were asked to fill out demographic information, which resulted in a mixed composition of participants, based on their age, gender, current position, number of years of experience in their current position and their modeling experience. Based on the year of birth, the participants were randomly split in two groups. For one group, which consisted of six participants, all questions asked concerned the initial PGA notation. For the other group, which consisted of eight participants, the questions concerned the adapted PGA notation.

The results of the experiments conducted during this research show that:

- the interpretational effectiveness during the semantic transparency experiment performs significantly better for one PGA concept, Importance.
- the interpretational effectiveness during the case study performs significantly better for the total result of all six PGA concepts, but this effect cannot be clearly explained by any of the individual PGA concepts.
- the interpretational effectiveness during the recall/recognition experiment performs significantly better for two PGA concepts, Competence and Importance.
- the overall perceived ease of use performs significantly better, but this is only supported by two PGA concepts, Value stream and Importance.
- the overall perceived usefulness performs significantly better.

With a significant improvement observed during the case study experiment for the overall interpretational effectiveness, the adapted PGA notation seems to perform better than the initial one, and based on the overall perceived ease of use and usefulness, the participants also had the feeling that the adapted PGA notation performs better. Based on this overall result, further analysis was done to find out for which concepts from the initial PGA notation it was worth considering to replace with the adapted ones. From the six suggested adaptations, only one performed convincingly better, namely Importance, and two less convincing, namely Competence and Value stream.

With the mixed results obtained in this research, it is thus not possible to unambiguously answer whether the adapted variant of the notation performs better.

5.4. Recommendations for practice

This research elaborated on the PGA technique, a DSML designed for realizing strategic fit within the business architecture of an organization (Roelens et al., 2019), meant to be used by business-oriented end-users. In this research two different PGA notations, the initial one, described by Roelens et al. (2019), and an adapted one, described by Roelens and Bork (2020), were evaluated within a practical business context to determine the most intuitive PGA notation to be used within organizations. This resulted in a practical recommendation to replace the initial PGA notation of the concept Importance with the adapted one and consider the adapted notations for the concepts Competence and Value stream as suitable starting points for further redesign.

5.5. Recommendations for further research

This research extends previous research, done by Roelens and Bork (2020), by evaluating and comparing two different visual notations in order to determine which notation performs significantly better. It also paves a path to objectively choose the most intuitive notation for other DSMLs.

While examining the construct validity, it was expected that the adapted PGA notation would be more intuitive than the initial PGA notation. The conducted experiments should show significant improvements, but expectations were only partially met for the measurements of the interpretational effectiveness and perceived ease of use. Further research is required to determine whether the questions asked during the semantic transparency experiment, case study, recall/recognition experiment and perceived ease of use should be asked differently (e.g. open questions instead of multiple choice) or that other experiments should be conducted to measure the interpretational effectiveness and perceived ease of use.
In this research a between-subjects design was chosen, where participants were split in two groups, and each group only was shown one kind of PGA notation. Further research could examine whether a within-subjects design yields better results, for instance with an experiment in which both the initial and adapted notations of the various concepts is shown side by side and the participant can indicate a preference.

While examining the internal validity, it was mentioned that a high number of participants only partially completed the survey (i.e. 11 out of 25). Because the participants could stop the survey at any point without giving any reason and there was no possibility to give a reason for those who wanted to, it was unclear why so many participants did not completely fill out the survey. In future research participants should be given the possibility to give reasons why they did not completely fill out the survey, in order to make the necessary adjustments when needed.

While examining the external validity, it was mentioned that the limited number of 14 participants resulted in a lower generalizability of the results of this research. Further research with a larger number of participants should be conducted to examine whether results will differ from the results in this research. Besides conducting a survey within one financial organization, the survey could be conducted with participants from different financial organizations and/or from organizations within other economic sectors. This enables to investigate whether specific demographic variables influence the results. Results can be broken down to age, management position, modeling experience and knowledge about a domain to find out whether certain groups of participants perceive the intuitiveness of a modelling notation in a different way.

While examining the reliability, it was mentioned that some feedback about the survey itself was given, but no substantive feedback about the PGA notation itself. Besides the experiments and perceived ease of use and usefulness questions, for further triangulation of the data, further research should collect more feedback from participants, for instance by asking open questions about the notation. Participants could for instance be asked to do suggestions about a certain notation of a concept themselves or what kind of experimental tasks or questions they are missing in the survey.

References

Bork, D., Schrüffer, C., & Karagiannis, D. (2019). Intuitive understanding of domain-specific modeling languages: proposition and application of an evaluation technique. In International conference on conceptual modeling (pp. 311-319). Springer, Cham.

Caire, P., Genon, N., Heymans, P., & Moody, D. L. (2013). Visual notation design 2.0: Towards user comprehensible requirements engineering notations. In 2013 21st IEEE International Requirements Engineering Conference (RE) (pp. 115-124). IEEE.

El Kouhen, A., Gherbi, A., Dumoulin, C., & Khendek, F. (2015). On the semantic transparency of visual notations: experiments with UML. In International SDL Forum (pp. 122-137). Springer, Cham.

Frank, U. (2011). Some guidelines for the conception of domain-specific modelling languages. Enterprise modelling and information systems architectures (EMISA 2011).

Gemino, A., & Wand, Y. (2004). A framework for empirical evaluation of conceptual modeling techniques. Requirements Engineering, 9(4), 248-260.

Genon, N., Caire, P., Toussaint, H., Heymans, P., & Moody, D. (2012). Towards a more semantically transparent i* visual syntax. In International Working Conference on Requirements Engineering: Foundation for Software Quality (pp. 140-146). Springer, Berlin, Heidelberg.

Kotusev, S. (2019). Enterprise architecture and enterprise architecture artifacts: Questioning the old concept in light of new findings. Journal of Information technology, 34(2), 102-128.

Lankhorst, M. (2013). Enterprise Architecture at Work. Springer, Berlin, Heidelberg.

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

Moody, D.: The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering. IEEE Trans. Softw. Eng. 35(6), 756–779 (2009)

Moody, D. L., Genon, N., Heymans, P., & Caire, P. (2012). Visual Notation Design 2.0: Designing User-Comprehensible Diagramming Notations.

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, Cham.

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.

Santos, M., Gralha, C., Goulão, M., & Araújo, J. (2018). Increasing the semantic transparency of the KAOS goal model concrete syntax. In International Conference on Conceptual Modeling (pp. 424-439). Springer, Cham.

Saunders, M., Lewis, P., & Thornhill, A. (2019). Research methods for business students. Pearson education.

The Open Group (2019). ArchiMate® 3.1 Specification. Retrieved from https://pubs.opengroup.org/architecture/archimate3-doc/chap01.html#_Toc10045266.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th international conference on evaluation and assessment in software engineering (pp. 1-10).

Appendix 1 Results database search

In the next tables the results of the search criteria are listed per sub-question and search term(s). When the total results is greater than ten results, only the first ten results are taken into account.

When a line is marked white, the article did not meet the criteria and was not added to the list of relevant articles. When a line is marked yellow, the article is added to the list of relevant articles. If the article is marked red, after reading the abstract it met the criteria and seemed to be relevant, but after reading the whole article it was discarded as not relevant enough.

Database search Sub-question 1 and 2

Search	"conceptual model" AND "intuitiveness"
Total results	33

Nr	Year of publication	Web of Science	Intuitiveness / Semantic transparency	PGA / DSML / Conceptual Model	Graphical Notation	Methodology / Experiment
1	2019	N	Y	N	N	M
2	2012	Y	Y	Y	Y	M
3	2020	N	Y	Y	Y	E
4	2015	N	N	N	Y	M
5	2004	N	N	N	N	M
6	2013	Y	N	N	N	M
7	2015	N	N	Y	Y	M
8	2010	Y	Y	Y	Y	M
9	2019	N	N	N	N	M
10	2017	Y	Y	Y	Y	M

Nr	Author(s)	Title
1	Radka Nacheva; Latinka Todoranova; Snezhana Sulova ; Bonimir Penchev	Concept Map Mining Approach Based on the Mental Models Retrieval
2	Cuadra, Dolores; Martínez, Paloma ; Castro, Elena; Al-Jumaily, Harith	Guidelines for representing complex cardinality constraints in binary and ternary relationships
3	Locoro, Angela; Cabitza, Federico; Ravarini, Aurelio ; Buono, Paolo	IGV Short Scale to Assess Implicit Value of Visualizations through Explicit Interaction
4	Celentano, Augusto; Dubois, Emmanuel	Evaluating metaphor reification in tangible interfaces
5	Laurini, Robert; Paolino, Luca; Sebillo, Monica ; Tortora, Genoveffa; Vitiello, Giuliana	Dealing with geographic continuous fields: the way to a visual GIS environment
6	Brzostowski, Jakub; Brzostowski, Jakub; Wachowicz, Tomasz ; Wachowicz, Tomasz	NegoManage: A System for Supporting Bilateral Negotiations

7	Derntl, Michael; Nicolaescu, Petru; Erdtmann, Stephan ; Klamma, Ralf; Jarke, Matthias	Near Real-Time Collaborative Conceptual Modeling on the Web
8	Moody, D.L.; Heymans, Patrick; Matulevicius, Raimundas	Visual syntax does matter: improving the cognitive effectiveness of the i* visual notation
9	Yoxall, Alaster; Gonzalez, Victor; Best, Joshua ; Rodriguez-Falcon, Elena M; Rowson, Jennifer	As you like it: Understanding the relationship between packing design and accessibility
10	Figl, Kathrin	Comprehension of Procedural Visual Business Process Models

Database search Sub-question 1 and 2

Search	"conceptual model" AND "semantic transparency"
Total results	4

Nr	Year of publication	Web of Science	Intuitiveness / Semantic transparency	PGA / DSML / Conceptual Model	Graphical Notation	Methodology / Experiment
1	2018	N	Y	Y	Y	M
2	2016	N	N	Y	Y	M
3	2013	N	Y	Y	Y	M
4	2011	N	N	N	N	M

Nr	Author(s)	Title
1	Zehnder, Eloïse; Mayer, Nicolas; Gronier, Guillaume	Evaluation of the Cognitive Effectiveness of the CORAS Modelling Language
2	Silva, Lyrene; Moreira, Ana; Araújo, João ; Gralha, Catarina; Goulão, Miguel; Amaral, Vasco	Exploring Views for Goal-Oriented Requirements Comprehension
3	Le Pallec, Xavier; Dupuy-Chessa, Sophie	Support for quality metrics in metamodelling
4	Kohlhase, Andrea; Kohlhase, Michael	Towards a flexible notion of document context

Database search Sub-question 1 and 2

Search	"graphical notation" AND "intuitiveness"
Total results	7

Nr	Year of publication	Web of Science	Intuitiveness / Semantic transparency	PGA / DSML / Conceptual Model	Graphical Notation	Methodology / Experiment
1	2014	Y	N	Y	N	M

2	2017	Y	N	N	N	M
3	2010	Y	Y	Y	Y	M
4	2012	Y	Y	Y	Y	M
5	2015	N	N	Y	Y	M
6	2014	N	N	N	N	M
7	2014	N	N	N	N	M

Nr	Author(s)	Title
1	Klimek, Radosław	A system for deduction-based formal verification of workflow-oriented software models
2	Bashroush, Rabih; Garba, Muhammad; Rabiser, Rick; Groher, Iris; Botterweck, Goetz;	CASE Tool Support for Variability Management in Software Product Lines
3	Moody, D. L.; Heymans, Patrick; Matulevicius, Raimundas	Visual syntax does matter: Improving the cognitive effectiveness of the i visual notation
4	Cuadra, Dolores; Martínez, Paloma; Castro, Elena; Al-Jumaily, Harith	Guidelines for representing complex cardinality constraints in binary and ternary relationships
5	Derntl, Michael; Nicolaescu, Petru; Erdtmann, Stephan ; Klamma, Ralf; Jarke, Matthias	Near Real-Time Collaborative Conceptual Modeling on the Web
6	Bazydlo, Grzegorz; Adamski, Marian; Wegrzyn, Marek; Munoz, Alfredo Rosado	From UML Specification into FPGA Implementation
7	Przemyslaw, Plecka; Krzysztof, Bzdyra	Usefulness of Software Valuation Methods at Initial Stages of ERP Implementation

Database search Sub-question 1 and 2

Search	"graphical notation" AND "semantic transparency"
Total results	8

Nr	Year of publication	Web of Science	Intuitiveness / Semantic transparency	PGA / DSML / Conceptual Model	Graphical Notation	Methodology / Experiment
1	2014	Y	N	N	N	M
2	2019	Y	N	Y	Y	M
3	2015	N	Y	Y	Y	E
4	2010	Y	Y	Y	Y	M
5	2017	N	N	Y	Y	M
6	2016	N	Y	Y	Y	M
7	2018	Y	N	N	N	M

8	2019	N	N	Y	Y	

Nr	Author(s)	Title
1	Cheng, Peter C.	Graphical notations for syllogisms: How alternative representations impact the accessibility of concepts
2	del-Río-Ortega, Adela; Resinas, Manuel; Durán, Amador; Bernárdez, Beatriz; Ruiz-Cortés, Antonio; Toro, Miguel	Visual ppinot: A Graphical Notation for Process Performance Indicators
3	El Kouhen, Amine; Gherbi, Abdelouahed; Dumoulin, Cédric; Khendek, Ferhat	On the Semantic Transparency of Visual Notations: Experiments with UML
4	Moody, D. L.; Heymans, Patrick; Matulevicius, Raimundas	Visual syntax does matter: Improving the cognitive effectiveness of the i visual notation
5	Erdogan, Gencer; Stølen, Ketil	Design Decisions in the Development of a Graphical Language for Risk-Driven Security Testing
6	Cánovas Izquierdo, Javier Luis; Cabot Sagrera, Jordi	Collaboro: a collaborative (meta) modeling tool
7	Ober, Ileana; Palyart, Marc; Bruel, Jean-Michel; Lugato, David	On the use of models for high-performance scientific computing applications: an experience report
8	Pablo Carvallo, Juan; Franch, Xavier	An empirical study on the use of i by non-technical stakeholders: the case of strategic dependency diagrams

Database search Sub-question 3

Search	"empirical evaluation" AND "intuitiveness"
Total results	10

Nr	Year of publication	Web of Science	Intuitiveness / Semantic transparency	PGA / DSML / Conceptual Model	Graphical Notation	Methodology / Experiment
1	2020	Y	Y	N	Y	M
2	2009	N	Y	Y	Y	E
3	2020	N	Y	N	N	E
4	2012	Y	Y	N	N	M
5	2016	N	Y	N	N	M
6	2011	N	Y	N	N	M / E
7	2011	Y	Y	N	N	M
8	2013	Y	Y	N	N	E
9	2016	Y	Y	Y	Y	E
10	2019	N	Y	Y	Y	M

Nr	Author(s)	Title
1	Blumenthal, David B.; Boria, Nicolas; Gamper, Johann; Bougleux, Sebastien; Brun, Luc	Comparing heuristics for graph edit distance computation
2	Fischer, Sandrine; Itoh, Makoto; Inagaki, Toshiyuki	A cognitive schema approach to diagnose intuitiveness: an application to onboard computers
3	Lou, Xiaolong; Li, Xiangdong; Hansen, Preben; Feng, Zhipeng	An Empirical Evaluation on Arm Fatigue in Free Hand Interaction and Guidelines for Designing Natural User Interfaces in VR
4	Wills, Andy J.; Pothos, Emmanuel M.	On the Adequacy of Current Empirical Evaluations of Formal Models of Categorization
5	Albertetti, Fabrizio; Grossrieder, Lionel; Ribaux, Olivier; Stoffel, Kilian	Change points detection in crime-related time series: An on-line fuzzy approach based on a shape space representation
6	Tian, Feng; Cao, Xiang; Lu, Fei; Dai, Guozhong; Zhang, Xiaolong; Wang, Hongan	Empirical studies of pen tilting performance in pen-based user interfaces
7	Martens, David; Vanhoutte, Christine; De Winne, Sophie; Baesens, Bart; Sels, Luc; Mues, Christophe	Identifying financially successful start-up profiles with data mining
8	Hürst, Wolfgang; van Wezel, Casper	Gesture-based interaction via finger tracking for mobile augmented reality
9	Shamim, Azra; Balakrishnan, Vimala; Tahir, Muhammad; Ahsan Qureshi, Muhammad	Age and domain specific usability analysis of opinion visualisation techniques
10	Bork, Dominik; Schrüffer, Christine; Karagiannis, Dimitris	Intuitive Understanding of Domain-Specific Modeling Languages: Proposition and Application of an Evaluation Technique

Database search Sub-question 3

Search	"empirical evaluation" AND "semantic transparency"
Total results	16

Nr	Year of publication	Web of Science	Intuitiveness / Semantic transparency	PGA / DSML / Conceptual Model	Graphical Notation	Methodology / Experiment
1	2018	N	Y	Y	Y	E
2	2012	N	Y	Y	Y	M / E
3	2015	Y	Y	N	Y	M
4	2013	Y	Y	Y	Y	E
5	2011	Y	Y	Y	Y	M
6	2018	N	Y	N	Y	E
7	2010	Y	Y	Y	Y	M / E
8	2019	N	Y	Y	Y	M
9	2018	Y	Y	N	N	M

10	2017	Y	Y	Y	Y	M

Nr	Author(s)	Title
1	Santos, Mafalda; Gralha, Catarina; Goulão, Miguel; Araújo, João	Increasing the Semantic Transparency of the KAOS Goal Model Concrete Syntax
2	Genon, Nicolas; Caire, Patrice; Toussaint, Hubert; Heymans, Patrick; Moody, Daniel	Towards a More Semantically Transparent i Visual Syntax
3	Luqman, Hamza; Karpati, Peter; Sindre, Guttorm; Opdahl, Andreas L.	Extending the UML Statecharts Notation to Model Security Aspects
4	Figl, Kathrin; Mendling, Jan; Strembeck, Mark	The Influence of Notational Deficiencies on Process Model Comprehension
5	Cheng, Peter C.	Probably Good Diagrams for Learning: Representational Epistemic Recodification of Probability Theory
6	Henriques, Henrique; Lourenço, Hugo; Amaral, Vasco; Goulão, Miguel	Improving the Developer Experience with a Low-Code Process Modelling Language
7	Moody, D. L.; Heymans, Patrick; Matulevicius, Raimundas	Visual syntax does matter: Improving the cognitive effectiveness of the i visual notation
8	Bork, Dominik; Schrüffer, Christine; Karagiannis, Dimitris	Intuitive Understanding of Domain-Specific Modeling Languages: Proposition and Application of an Evaluation Technique
9	Tenbergen, Bastian; Weyer, Thorsten; Pohl, Klaus	Hazard Relation Diagrams: a diagrammatic representation to increase validation objectivity of requirements-based hazard mitigations
10	Figl, Kathrin	Comprehension of Procedural Visual Business Process Models: A Literature Review

Appendix 2 Results snowballing and articles thesis supervisor

In the next tables the articles of the backward and forward snowballing are listed. The last table contains the articles mentioned by the thesis supervisor.

When a line is marked yellow, the article is added to the list of relevant articles. If the article is marked red, after reading the abstract it met the criteria and seemed to be relevant, but after reading the whole article it was discarded as not relevant enough.

Backward snowballing articles

Nr	Year of publication	Web of Science	Intuitiveness / Semantic transparency	PGA / DSML / Conceptual Model	Graphical Notation	Methodology / Experiment
1	2009	Y	Y	Y	Y	M
2	2016	Y	Y	Y	Y	M
3	2014	N	N	Y	Y	M
4	2013	N	Y	Y	Y	M / E
5	Draft only	N	Y	Y	Y	M / E
6	2003	Y	N	Y	Y	M

Nr	Original source author(s)	Original source title
1	Roelens, Ben; Bork, Dominik	An Evaluation of the Intuitiveness of the PGA Modeling Language Notation
2	Bork, Dominik; Schrüffer, Christine; Karagiannis, Dimitris	Intuitive Understanding of Domain-Specific Modeling Languages: Proposition and Application of an Evaluation Technique
3	Roelens, Ben; Steenacker, Wout ; Poels, Geert	Realizing strategic fit within the business architecture: the design of a Process-Goal Alignment modeling and analysis technique
4	Bork, Dominik; Schrüffer, Christine; Karagiannis, Dimitris	Intuitive Understanding of Domain-Specific Modeling Languages: Proposition and Application of an Evaluation Technique
5	Caire, Patrice; Genon, Nicolas; Heymans, Patrick ; Moody, Daniel L	Visual notation design 2.0: Towards user comprehensible requirements engineering notations
6	Gemino, Andrew; Wand, Yair	A framework for empirical evaluation of conceptual modeling techniques

Nr	Author(s)	Title
1	Moody, D	The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering
2	Gulden, Jens; van der Linden, Dirk; Aysolmaz, Banu	A Research Agenda on Visualizations in Information Systems Engineering
3	Roelens, Ben; Poels, Geert	The creation of business architecture heat maps to support strategy-aligned organizational decisions
4	Caire, Patrice; Genon, Nicolas; Heymans, Patrick ; Moody, Daniel L	Visual notation design 2.0: Towards user comprehensible requirements engineering notations

5	Moody, DL; Genon, N; Heymans, P; Caire, P	Visual Notation Design 2.0: Designing User-Comprehensible Diagramming Notations
6	Gemino, Andrew; Wand, Yair	Evaluating modeling techniques based on models of learning

Forward snowballing articles

Nr	Year of publication	Web of Science	Intuitiveness / Semantic transparency	PGA / DSML / Conceptual Model	Graphical Notation	Methodology / Experiment
1	2019	Y	Y	Y	Y	M

Nr	Original source author(s)	Original source title
1	Moody, D	The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering

Nr	Author(s)	Title
1	van der Linden, Dirk; Hadar, Irit ; Zamansky, Anna	What practitioners really want: requirements for visual notations in conceptual modeling

Thesis supervisor articles

Nr	Author(s)	Year of publication	Title
1	Roelens, Ben; Steenacker, Wout ; Poels, Geert	2019	Realizing strategic fit within the business architecture: the design of a Process-Goal Alignment modeling and analysis technique
2	Roelens, Ben; Bork, Dominik	2020	An Evaluation of the Intuitiveness of the PGA Modeling Language Notation
3	Gemino, Andrew; Wand, Yair	2004	A framework for empirical evaluation of conceptual modeling techniques
4	Lankhorst, Marc	2013	Enterprise Architecture at Work: Modelling, Communication and Analysis

Appendix 3 Relevant articles examination

In this appendix each article of the final set of relevant articles (see section 2.2) is elaborated with the purpose, hypotheses, measures, experimental design, instrumentation and experimental tasks, selection of participants and operational procedures.

Paper / Book	**Bork, D., Schrüffer, C., & Karagiannis, D. (2019). Intuitive understanding of domain-specific modeling languages: proposition and application of an evaluation technique. In International conference on conceptual modeling (pp. 311-319). Springer, Cham.**
Purpose	To propose an empirical evaluation technique for evaluating the intuitiveness of a notation.
Hypotheses	The interpretational effectiveness of the new BCM notation is higher than the interpretational effectiveness of the old BCM notation The interpretational efficiency of the new BCM notation is higher than the interpretational efficiency of the old BCM notation
Measures	Interpretational effectiveness number of correct BCM terms associated to the right notation number of correct drafted BCM models Interpretational efficiency time needed to draft a notation for each BCM term (max 10 minutes) time needed to associate BCM terms to a given notation (max 10 minutes) time needed to draft five BCM models (max 30 minutes)
Experimental design	Within-subjects
Instrumentation and experimental tasks	Three independent experiments: term association, notation association, and case study
Selection of participants	In total, 15 information science Master students participated in the evaluation. Most participants are male (87%), between 25 and 29 years old, and are in the second semester of their Masters
Operational procedures	Initiation phase where the participants are briefly introduced to the domain and the building blocks of the modeling method to be evaluated. Phase 1: term association where participants are provided terms that refer to names of modeling language concepts. Each participant then individually drafts one or more graphical representations for each term he/she deems most intuitive. Phase 2: notation association where participants are presented notations of the current modeling language. They are then asked to record their up to three intuitive associations that pop out when looking at the notations. Phase 3: case study to test whether participants are able to intuitively combine the modeling language concepts in order to solve the presented case. Conclusion phase where the conductor presents the solution of the case study before the participants are asked to fill out a feedback survey. Participants are asked to provide positive and negative feedback, and improvement suggestions.

Paper	**Caire, P., Genon, N., Heymans, P., & Moody, D. L. (2013). Visual notation design 2.0: Towards user comprehensible requirements engineering notations. In 2013 21st IEEE International Requirements Engineering Conference (RE) (pp. 115-124). IEEE.**
Purpose	To propose a novel approach to designing RE visual notations that actively involves naïve users in the process.
Hypotheses	Four different i* notations were evaluated: Standard i* notations based on the original i* model notation PoN i* notations based on a previous paper that proposed a revised i* symbol set using Physics of Notations.

	Stereotype i* notations based on the most common symbols for each i* concept that was produced by naïve participants in an earlier conducted experiment. Prototype i* notations based on the by naïve participants best recognized representations for each i* concept that was produced by naïve participants in an earlier conducted experiment. This resulted in the next hypotheses: The interpretational effectiveness of the PoN i* notation is higher than the interpretational effectiveness of the Standard i* notation The interpretational effectiveness of the Stereotype i* notation is higher than the interpretational effectiveness of the Standard i* notation The interpretational effectiveness of the Prototype i* notation is higher than the interpretational effectiveness of the Standard i* notation The interpretational effectiveness of the Stereotype i* notation is higher than the interpretational effectiveness of the PoN i* notation The interpretational effectiveness of the Prototype i* notation is higher than the interpretational effectiveness of the PoN i* notation The interpretational effectiveness of the Prototype i* notation is higher than the interpretational effectiveness of the Stereotype i* notation
Measures	Interpretational effectiveness number of correct i* concepts associated to the right notation
Experimental design	Between-subjects
Instrumentation and experimental tasks	Five experiments/studies were conducted: Symbolisation experiment: naïve participants generated symbols for i* concepts, a task normally reserved for experts. Stereotyping analysis (nonreactive study): we identified the most common symbols produced for each i* concept. This defined the stereotype symbol set. Prototyping experiment: naïve participants identified the “best” representations for each i* concept. This defined the prototype symbol set. Semantic transparency experiment: we evaluated the ability of naïve participants to infer the meanings of novice- designed symbols (stereotype and prototype symbol set) compared to expert-designed symbols (standard i* and PoN i*). Recognition experiment: we evaluated the ability of naïve participants to learn and remember symbols from the 4 symbol sets.
Selection of participants	Selected participants for all experiments were naïve undergraduate students with no previous knowledge of goal modeling or i*.
Operational procedures	Participants were randomly assigned to experimental groups and provided with a copy of the experimental materials. They were instructed to work alone and not discuss their responses with other participants. No time limit was set but subjects took 10-15 minutes to complete the task.

Paper	**El Kouhen, A., Gherbi, A., Dumoulin, C., & Khendek, F. (2015). On the semantic transparency of visual notations: experiments with UML. In International SDL Forum (pp. 122-137). Springer, Cham.**
Purpose	To report on a set of experiments that confirm the lack of semantic transparency of the Unified Modeling Language (UML) as designed by OMG and to compare this standard to alternative solutions where naive users are involved in the design of the notations to speed-up the learning of these languages to new users. The purpose is not to redefine the visual syntax of UML but to show the importance of involving end-users into the design decisions made generally by experts.
Hypotheses	Three different UML notations were evaluated: Standard UML notations based on the original UML model notation PoN UML notations based on a previous research that proposed a revised UML symbol set using Physics of Notations.

	Prototype UML notations based on the best symbols produced by naive users as judged by other naive users. This resulted in the next hypotheses: The interpretational effectiveness of the PoN UML notation is higher than the interpretational effectiveness of the Standard UML notation The interpretational effectiveness of the Prototype UML notation is higher than the interpretational effectiveness of the Standard UML notation The interpretational effectiveness of the Prototype UML notation is higher than the interpretational effectiveness of the PoN UML notation
Measures	Interpretational effectiveness number of correct UML concepts associated to the right notation
Experimental design	Between-subjects
Instrumentation and experimental tasks	The experiments are limited to a few elements of UML visual syntax. Five experiments were conducted: Symbolization experiment: naive participants (i.e. with background on Object-Oriented paradigm but without previous knowledge on UML) generated symbols for UML concepts (drawings). Stereotyping analysis: we analysed the results of Experiment 1 and identified the most common symbols produced for each UML concept (stereotype symbols sets). Prototyping experiment: other group of naive participants (different from the first one) analysed the drawings produced in Experiment 1 and identified the "best" representations for each UML concept (prototype symbols set). Semantic transparency experiment: another group of naive users were asked to infer the meaning of 3 sets of symbols from their appearance alone : Prototypes from Experiment 3 and two external inputs, which are the Standard UML notation and the UML notation based on Physics of Notation Theory (PoN). Identify "best of breed" symbols: based on the results of the semantic transparency experiment, we identified the most cognitively effective symbols for each UML concept across all symbol sets. The study evaluated the comprehension of symbols as visual unit rather than complete diagrams.
Selection of participants	Computer science students (naive users) as they must know Object-Oriented concepts to draw them, but only students with no previous knowledge of modeling languages in general or UML.
Operational procedures	Participants were randomly assigned to experimental groups and provided with a copy of the experimental materials. They were instructed to work alone and not discuss their responses with other participants.

Paper	**Genon, N., Caire, P., Toussaint, H., Heymans, P., & Moody, D. (2012). Towards a more semantically transparent i* visual syntax. In International Working Conference on Requirements Engineering: Foundation for Software Quality (pp. 140-146). Springer, Berlin, Heidelberg.**
Purpose	Investigate how to design cognitively effective requirements modeling notations, concentrating on the visual syntax definition.
Hypotheses	The perceived usefulness of one i* concept is higher than the perceived usefulness of another i* concept. The perceived ease of use of one i* concept is higher than the perceived ease of use of another i* concept.
Measures	Perceived usefulness Percentage per drawn i* concept. Perceived ease of use Ease of drawing an i* concept, queried with a 5-point scale with values "easy", "fairly easy", "neither easy nor difficult", "fairly difficult" and "difficult".

	Ease of selecting the best drawing of the mentioned i* concept, queried with a 5-point scale with values “easy”, “fairly easy”, “neither easy nor difficult”, “fairly difficult” and “difficult”.
Experimental design	Between-subjects
Instrumentation and experimental tasks	Three experiments were conducted: production of drawings. The goal of this experiment is to obtain drawings hand-sketched by participants to represent each i* concept. identifying the stereotypical drawings out of the results of the first experiment. population prototype. This experiment looks for the drawing that best represents the corresponding i* concept. The personal opinion of participants is taken to elect the drawing that is the most semantically transparent for a given referent concept.
Selection of participants	All participants were students which had no previous knowledge of i* or modeling in general.
Operational procedures	Experiment 1: Each participant was provided with a 14-page booklet, a pencil and an eraser. The first page presented a form to collect participants’ demographic data. The remaining 13 pages were respectively dedicated to the 13 i* concepts. A (3” x 3”) frame where participants were asked to sketch their drawing was printed in the middle of the page. For each i* concept, participants were asked (a) to sketch what they estimate to be the best drawing to represent the name and the definition of this concept. There was no time limit but they were asked to sketch as quickly as possible. The intent was to capture their intuition.Each time a drawing was produced, the participant had to evaluate the difficulty of the task on a 5-point scale. Participants were also told to respond one page at time and not to go back in the booklet. Experiment 2: a judges’ ranking method was applied. Concept per concept, three of the authors categorised the drawings obtained from Experiment 1 based on the similarity of ideas that they expressed. Experiment 3: Each student was provided with the name and the definition of the concept. The set of drawings was composed of 160 drawings: the 13 stereotypes along with one representative of each category of the 13 i* concepts. An online questionnaire was set up. The participants were asked to enter their demographic data on the first page and then they navigated through 13 pages, one per i* concept. Each page displayed the name and the definition of the concept at the top of the page. The middle of the page was dedicated to instructions for selecting (using radio buttons) the best drawing among the matrix of representatives. The difficulty of the selection task was evaluated on a visual analogue scale (VAS) at the bottom of the page.

Paper	**Moody, D. L., Genon, N., Heymans, P., & Caire, P. (2012). Visual Notation Design 2.0: Designing User-Comprehensible Diagramming Notations.**
Purpose	To propose a radical new approach to designing diagramming notations that actively involves end-users in the process.
Hypotheses	Four different i* notations were evaluated: Standard i* notations based on the original i* model notation PoN i* notations based on a previous paper that proposed a revised i* symbol set using Physics of Notations. Stereotype i* notations based on the most common symbols for each i* concept that was produced by naïve participants in an earlier conducted experiment. Prototype i* notations based on the by naïve participants best recognized representations for each i* concept that was produced by naïve participants in an earlier conducted experiment. This resulted in the next interpretational effectiveness hypotheses: The interpretational effectiveness of the PoN i* notation is higher than the interpretational effectiveness of the Standard i* notation

	The interpretational effectiveness of the Stereotype i* notation is higher than the interpretational effectiveness of the Standard i* notation The interpretational effectiveness of the Prototype i* notation is higher than the interpretational effectiveness of the Standard i* notation The interpretational effectiveness of the Stereotype i* notation is higher than the interpretational effectiveness of the PoN i* notation The interpretational effectiveness of the Prototype i* notation is higher than the interpretational effectiveness of the PoN i* notation The interpretational effectiveness of the Prototype i* notation is higher than the interpretational effectiveness of the Stereotype i* notation Besides the interpretational effectiveness hypotheses the next hypotheses were included: The perceived usefulness of one i* concept is higher than the perceived usefulness of another i* concept. The perceived ease of use of one i* concept is higher than the perceived ease of use of another i* concept.
Measures	Interpretational effectiveness number of correct i* concepts associated to the right notation Perceived usefulness Percentage per drawn i* concept. Perceived ease of use Ease of drawing an i* concept, queried with a 5-point scale with values “easy”, “fairly easy”, “neither easy nor difficult”, “fairly difficult” and “difficult”. Ease of selecting the best drawing of the mentioned i* concept, queried with a 5-point scale with values “easy”, “fairly easy”, “neither easy nor difficult”, “fairly difficult” and “difficult”.
Experimental design	Between-subjects
Instrumentation and experimental tasks	The research design consists of 6 related empirical studies (4 experiments and 2 non-reactive studies). The results of earlier studies provide inputs to later studies. Symbolisation experiment: naïve participants generated symbols for i* concepts, a task normally reserved for experts. Stereotyping analysis: analyses of the results of Experiment 1 to identify the most common symbols produced for each i* concept. These defined the stereotype symbol set. Prototyping experiment: naïve participants analysed the drawings produced in Experiment 1 and identified the “best” representations for each i* concept. These defined the prototype symbol set. Semantic transparency experiment: naïve users were asked to infer the meaning of symbols from their appearance alone. The symbols were from one of 4 symbol sets, two designed by experts (standard i* and PoN i*) and two designed by novices (the stereotype and prototype symbol sets from experiments 2 and 3). Recall/recognition experiment: naïve users were given the meanings of the symbols from one of the 4 symbol sets and then had to recall them. This experiment also evaluated the effect of design rationale and semantic transparency on their performance. Identify “best of breed” symbols: based on the results of steps 4 and 5, identify the most cognitively effective symbols for each i* construct across all symbol sets.
Selection of participants	All participants were undergraduate business students. They had no previous knowledge of goal modeling in general or i* in particular to ensure participants were truly naïve. Business students were chosen as proxies for end-users, as they are similar in important characteristics to the target population: they have a business rather than technical orientation and have no previous knowledge of the notation being tested, so present a similar cognitive profile.
Operational procedures	Experiment 1: a similar procedure as used in sign production studies was used. Participants were asked to draw the constructs in the order in which they appeared in

	the booklet. They were instructed to produce drawings that they felt most effectively conveyed the meaning of the construct. Experiment 2: the judges' ranking method was used, which is a common approach for achieving convergence on a set of categories. Experiment 3: an on-line questionnaire was conducted. Participants navigated through 9 screens, one for each i* concept. The name and definition of the concept was displayed at the top of the screen with the candidate drawings (representatives from each category identified in the stereotyping study) displayed below. Participants were asked to select one drawing that most effectively conveyed each concept. Experiment 4: an on-line questionnaire was conducted. Participants were asked to choose one concept for each symbol presented but that each choice was independent: they could choose the same concept in response to multiple symbols. The purpose of this was to reduce the cognitive difficulty of the task (having to remember previous choices) and to tap directly into their intuition about what symbols meant. Experiment 5: Participants were asked to study the training materials until they understood all symbols and their meanings (learning phase). They then proceeded to the testing phase, where symbols were presented one at a time and participants had to recall their meanings. Each response required choosing from the complete set of i* concepts (closed questions), so more closely approximated recognition rather than recall. Participants were not allowed to take notes during the learning phase or to refer back to the training materials during the testing phase. Experiment 6: This was a (quantitative) non-reactive study, as it involves meta-analysis of the results from Experiments 3 and 4.

Paper	**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, Cham.**
Purpose	To test the current PGA notation by applying an evaluation technique for testing the intuitiveness of domain-specific modeling languages.
Hypotheses	The interpretational effectiveness of the new proposed PGA notation is higher than the interpretational effectiveness of the current PGA notation
Measures	Interpretational effectiveness number of correct PGA terms associated to the right notation number of correct interpreted PGA concepts
Experimental design	Within-subjects
Instrumentation and experimental tasks	Three independent experiments: term association, notation association, and case study
Selection of participants	The evaluation tasks were performed by 139 Master's students of Ghent University with an elaborate economical background and basic modeling experience.
Operational procedures	Initiation phase where the participants are briefly introduced to the relevant domain (i.e., strategic fit in the business architecture) and the building blocks of the PGA modeling method without showing any visual aspects like language concepts or sample models. Phase 1: term association where participants were provided terms that refer to names of PGA modeling language concepts. Each participant then individually drafted one or more graphical representations that he/she deems as the most intuitive for the element. Phase 2: notation association where samples of the current PGA notation were presented and participants were asked to record up to three intuitive associations that pop out when looking at them. The notations were presented without any hint of e.g., the name or the semantics of this concept. The concepts forming part of the term association were different to the ones of the notation association to omit hampering intuitiveness.

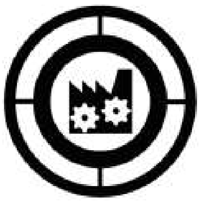
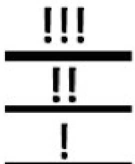
	Phase 3: case study task which included comprehension questions targeting an example of a business architecture heat map. Conclusion phase where the participants were asked to provide qualitative feedback and improvement suggestions about the current PGA notation.

Paper	**Santos, M., Gralha, C., Goulão, M., & Araújo, J. (2018). Increasing the semantic transparency of the KAOS goal model concrete syntax. In International Conference on Conceptual Modeling (pp. 424-439). Springer, Cham.**
Purpose	To propose an alternative concrete syntax for KAOS that increases its semantic transparency leading to a significantly higher correct symbol identification by novices.
Hypotheses	Three different KAOS notations were evaluated: Standard KAOS symbol set based on the original KAOS model notation Stereotype KAOS symbol set based on the most representative symbols produced by novice users. Prototype KAOS symbol set based on the best symbols produced by novice users as judged by other novice users. This resulted in the next hypotheses: The interpretational effectiveness of the Stereotype KAOS symbol set is higher than the interpretational effectiveness of the Standard KAOS symbol set The interpretational effectiveness of the Prototype KAOS symbol set is higher than the interpretational effectiveness of the Standard KAOS symbol set The interpretational effectiveness of the Prototype KAOS symbol set is higher than the interpretational effectiveness of the Stereotype KAOS symbol set
Measures	Interpretational effectiveness number of correct KAOS concepts associated to the right notation the degree of proximity between a symbol and the semantic construct represented by it
Experimental design	Between-subjects
Instrumentation and experimental tasks	The research design consists of 4 related empirical studies, where the results of the earlier studies provide inputs to the later studies. Symbolisation experiment: a group of 99 novice participants designed symbols for KAOS concepts, a task normally reserved for experts; Stereotyping analysis: we identified and organised categories with the most common symbols produced for each KAOS concept. This defined the stereotype symbol set. Prototyping experiment: a group of 88 novice-participants chose the symbols they consider to better represent each KAOS concept. The most voted symbols for each KAOS concept defined the prototype symbol set. Semantic transparency experiment: we evaluated the ability of 52 participants to infer the meanings of novice-designed symbols (stereotype and prototype symbol set) compared to expert-designed symbols (standard KAOS).
Selection of participants	All participants were students from different courses (Mechanical Engineering, Industrial Engineering and Management, Environmental Engineering, Civil Engineering, and Computer Science). This diversity was deliberately chosen because in this way the participants would act as surrogates of stakeholders from different backgrounds who will interact with requirements engineers. The participants were categorised in: With No Knowledge in Modeling Languages (WNKML) and With Knowledge in Modeling Languages (WKML).
Operational procedures	Experiment 1: Participants were instructed to produce drawings expressing the meaning of each concept. Experiment 2: The symbols produced in experiment 1 were classified into symbol categories. For both WKML and WNKML groups, we categorised the symbols based on their visual and conceptual similarity. We then combined the categories of symbols produced by both groups and counted the number of members in each category. We

	then selected the most representative category for each concept, resulting in the stereotype symbol set. Experiment 3: Participants were verbally instructed to answer a questionnaire, by choosing the symbols that, in their opinion, best expressed the meaning of the KAOS concepts. Experiment 4: Participants were instructed verbally to answer a questionnaire, by selecting a symbol from each set that better described each KAOS concept.

Appendix 4 Suggested adaptations to the initial PGA notation

In this appendix the six suggested adaptations to the initial PGA notation are listed (Roelens & Bork, 2020, Table 4).

PGA Concept	Initial notation	Suggested new notation
Competence		
Value proposition		
Internal goal		
Customer goal		
Value stream		
Importance		

Appendix 5 Survey

The online survey was created with LimeSurvey. The structure (see also paragraph 3.2 Operational procedures) consists of the next sections:

- Consent
  The participant is explained that taking part in this research is voluntary and that he/she can stop at any time. The results of every participant will be anonymized to guarantee their privacy. After the informed consent, a short introduction is given about the purpose of PGA, without showing any details about the notation.
  This section contains one question.
- Demographics
  Participants are asked to fill out demographic information about their age, gender, current position and number of years of experience in their current position. Next questions about their modeling experience are asked: how they rate their modeling expertise and which modeling languages they know. At last, questions about the PGA domain are asked: what understanding do they have about strategic planning and strategic fit.
  This section contains seven questions. Based on the year of birth questions about the initial or adapted PGA notation were asked in the following sections.
- Notation Association
  During this phase the semantic transparency experiment (also known as notation association) is conducted.
  This section contains a small intro and six questions about the initial PGA notation and six questions about the adapted PGA notation.
- Case
  During this phase an experiment including a case study with comprehension questions is conducted.
  This section contains a small intro and six questions about the initial PGA notation and six questions about the adapted PGA notation.
- Recall
  During this phase the recall/recognition experiment is conducted.
  This section contains a small intro. After the intro the six PGA notations, of the intended PGA notation, are shown for a maximum of 60 seconds. Subsequently this section contains six questions about the initial PGA notation and six questions about the adapted PGA notation..
- Perceived
  During this phase the participants are asked questions about the overall perceived ease of use and overall perceived usefulness. Per concept, separate questions are asked about the perceived ease of use.
  This section contains a small intro, and subsequently per PGA notation four questions about the overall perceived ease of use, six questions about each concept of the PGA notation and three questions about the overall perceived usefulness.
- Feedback
  During this phase participants are asked to give feedback about the completed procedure.

The next link shows the Limesurvey structure as described above.

limesurvey_survey_592294.lss

Appendix 6 Survey responses

The responses of the survey (see appendix 5) were transformed with Excel in such a way that they could be used directly for the statistical analysis in SPSS. The next table shows the transformations done.

Element	**Data**	**Code**
PGAVersion	Initial	0
	New	1
Position	Non management	0
	Operational management	1
	Tactical management	2
	Strategic management	3
ExperiencePosition	< 1 year	0
	>= 1 year and < 2 years	1
	>= 2 years and < 3 years	2
	>= 3 years and < 5 years	3
	>= 5 years	4
ModelExperience	1: Very low	0
	2: Low	1
	3: Medium	2
	4: High	3
	5: Very high	4
ModelingLanguages	None	0
	Checked 1	1
	Checked 2	2
	Checked 3	3
	Checked 4	4
StrategicUnderstand	1: Very low	0
	2: Low	1
	3: Medium	2
	4: High	3
	5: Very high	4
NotationAssociation	No association	0
	Wrong answer	1
	Good answer	2
Case	Wrong answer	0
	Good answer	1
Recall	No association	0
	Wrong answer	1
	Good answer	2
Perception questions	Strongly disagree	0
	Disagree	1
	Neutral	2
	Agree	3
	Strongly agree	4

The next link shows the Excel with all responses and translations.

Survey responses
reformatted for SPS!

Appendix 7 Relevant statistical tests

As described in chapter 3.3, to find out which statistical analysis needs to be performed, the type of the dependent variables were classified. The interpretational effectiveness, represented by Semantic transparency experiment, the Recall/recognition experiment and the Case study, is measured by the number of correct answers (hit rate), which is numerical interval data. To know which analysis to perform it needs to be determined whether the data of the three experiments is normally distributed or not.

To test the normality of a small sample (14 completely filled out surveys), a Shapiro-Wilk analysis is performed. The next hypotheses are tested:

Hypothesis 0: The data is normally distributed.

Hypothesis 1: The data is not normally distributed.

When the significance is greater than 0.05, hypothesis 0 is met, so the data are normally distributed. If not, Hypothesis 1 is met, so the data are not normally distributed.

The next table shows if the data of the Semantic transparency experiment (ST total) is normally distributed or not.

Dependent variable	**Shapiro-Wilk**		**Normally distributed**	**Type of dependent variable**	**Independent variable Two groups Unpaired**
	df	**Sig.**			
ST Competence	14	0,009	No	Interval	Wilcoxon-Mann-Whitney Test
ST ValueProposition	14	0,004	No	Interval	Wilcoxon-Mann-Whitney Test
ST InternalGoal	14	0,003	No	Interval	Wilcoxon-Mann-Whitney Test
ST CustomerGoal	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
ST ValueStream	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
ST Importance	14	0,007	No	Interval	Wilcoxon-Mann-Whitney Test
ST total	14	0,081	Yes	Normal	t-Test (for unpaired)

The next table shows if the data of the Case study (C total) is normally distributed or not.

Dependent variable	**Shapiro-Wilk**		**Normally distributed**	**Type of dependent variable**	**Independent variable Two groups Unpaired**
	df	**Sig.**			
C Competence	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
C ValueProposition	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
C InternalGoal	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
C CustomerGoal	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
C ValueStream	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
C Importance	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
C total	14	0,694	Yes	Normal	t-Test (for unpaired)

The next table shows if the data of the Recall/recognition experiment (REC total) is normally distributed or not.

Dependent variable	Shapiro-Wilk		Normally distributed	Type of dependent variable	Independent variable Two groups Unpaired
	df	Sig.			
REC Competence	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
REC ValueProposition	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
REC InternalGoal	14		No	Interval	Wilcoxon-Mann-Whitney Test
REC CustomerGoal	14		No	Interval	Wilcoxon-Mann-Whitney Test
REC ValueStream	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
REC Importance	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test
REC total	14	<,001	No	Interval	Wilcoxon-Mann-Whitney Test

The perceived usefulness and the perceived ease of use, both measured based on a five point Likert scale, are both ordinal data. All ordinal data will be statistically tested with the Wilcoxon-Mann-Whitney Test.

Appendix 8 Raw results statistical tests

The next tables show the raw results of all statistical tests, conducted in SPSS version 26,which are used to obtain the results of paragraphs 4.2, 4.3, 4.4 and 4.5 and appendices 6 and 7.

First, statistical tests for normality are conducted for all interval variables, to determine which test should be conducted.

Next, the T-test results are shown. Only for variables NA total and C total these test is conducted.

Next, for all other variables Mann-Whitney Tests are conducted.

Normality

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
NA total	,211	14	,093	,890	14	,081
C total	,173	14	,200*	,958	14	,694
REC total	,320	14	<,001	,679	14	<,001

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
NACompetence	,253	14	,015	,821	14	,009
NAValueProposition	,230	14	,043	,792	14	,004
NAInternalGoal	,268	14	,007	,786	14	,003
NACustomerGoal	,510	14	<,001	,428	14	<,001
NAValueStream	,361	14	<,001	,675	14	<,001
NAImportance	,226	14	,051	,810	14	,007

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
C01Competences	,443	14	<,001	,576	14	<,001
C02Importance	,369	14	<,001	,639	14	<,001
C03ValueStream	,407	14	<,001	,616	14	<,001

C04InternalGoal	,478	14	<,001	,516	14	<,001
C05CustomerGoal	,332	14	<,001	,646	14	<,001
C06ValueProposition	,369	14	<,001	,639	14	<,001

a. Lilliefors Significance Correction

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RECCompetence	,510	14	<,001	,428	14	<,001
RECValueProposition	,478	14	<,001	,516	14	<,001
RECInternalGoal	.	14	.	.	14	.
RECCustomerGoal	.	14	.	.	14	.
RECValueStream	,534	14	<,001	,297	14	<,001
RECImportance	,510	14	<,001	,428	14	<,001

a. Lilliefors Significance Correction

T-Test

Group Statistics

	PGAVersion	N	Mean	Std. Deviation	Std. Error Mean
NA total	0	6	5,50	2,588	1,057
	1	8	7,25	3,105	1,098
C total	0	6	1,50	1,049	,428
	1	8	3,00	1,309	,463

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
NA total	Equal variances assumed	,050	,826	-1,117	12	,143	,286	-1,750	1,567	-5,164	1,664
	Equal variances not assumed			-1,148	11,800	,137	,274	-1,750	1,524	-5,076	1,576
C total	Equal variances assumed	,214	,652	-2,300	12	,020	,040	-1,500	,652	-2,921	-,079
	Equal variances not assumed			-2,379	11,903	,017	,035	-1,500	,631	-2,875	-,125

Mann-Whitney Test

Ranks

	PGAVersion	N	Mean Rank	Sum of Ranks
REC total	0	6	6,75	40,50
	1	8	8,06	64,50
	Total	14		
EASE med	0	6	4,17	25,00
	1	8	10,00	80,00
	Total	14		
USE med	0	6	5,33	32,00
	1	8	9,13	73,00
	Total	14		

Test Statistics[a]

	REC total	EASE med	USE med
Mann-Whitney U	19,500	4,000	11,000

Wilcoxon W	40,500	25,000	32,000
Z	-,661	-2,629	-1,871
Asymp. Sig. (2-tailed)	,508	,009	,061
Exact Sig. [2*(1-tailed Sig.)]	,573[b]	,008[b]	,108[b]

Ranks

	PGAVersion	N	Mean Rank	Sum of Ranks
NACompetence	0	6	8,83	53,00
	1	8	6,50	52,00
	Total	14		
NAValueProposition	0	6	6,75	40,50
	1	8	8,06	64,50
	Total	14		
NAInternalGoal	0	6	6,67	40,00
	1	8	8,13	65,00
	Total	14		
NACustomerGoal	0	6	7,33	44,00

	1	8	7,63	61,00
	Total	14		
NAValueStream	0	6	6,50	39,00
	1	8	8,25	66,00
	Total	14		
NAImportance	0	6	4,75	28,50
	1	8	9,56	76,50
	Total	14		

Test Statistics[a]

	NACompetence	NAValueProposition	NAInternalGoal	NACustomerGoal	NAValueStream	NAImportance
Mann-Whitney U	16,000	19,500	19,000	23,000	18,000	7,500
Wilcoxon W	52,000	40,500	40,000	44,000	39,000	28,500
Z	-1,123	-,616	-,688	-,212	-,882	-2,258
Asymp. Sig. (2-tailed)	,262	,538	,491	,832	,378	,024
Exact Sig. [2*(1-tailed Sig.)]	,345[b]	,573[b]	,573[b]	,950[b]	,491[b]	,029[b]

Ranks

	PGAVersion	N	Mean Rank	Sum of Ranks
C01Competences	0	6	6,67	40,00
	1	8	8,13	65,00
	Total	14		
C02Importance	0	6	5,83	35,00
	1	8	8,75	70,00
	Total	14		
C03ValueStream	0	6	6,17	37,00
	1	8	8,50	68,00
	Total	14		
C04InternalGoal	0	6	7,17	43,00
	1	8	7,75	62,00
	Total	14		
C05CustomerGoal	0	6	6,33	38,00
	1	8	8,38	67,00
	Total	14		
C06ValueProposition	0	6	6,83	41,00
	1	8	8,00	64,00

Total		14		

Test Statistics[a]

	C01Competences	C02Importance	C03ValueStream	C04InternalGoal	C05CustomerGoal	C06ValueProposition
Mann-Whitney U	19,000	14,000	16,000	22,000	17,000	20,000
Wilcoxon W	40,000	35,000	37,000	43,000	38,000	41,000
Z	-,823	-1,502	-1,241	-,362	-1,041	-,601
Asymp. Sig. (2-tailed)	,411	,133	,215	,717	,298	,548
Exact Sig. [2*(1-tailed Sig.)]	,573[b]	,228[b]	,345[b]	,852[b]	,414[b]	,662[b]

Ranks

	PGAVersion	N	Mean Rank	Sum of Ranks
RECCompetence	0	6	6,17	37,00
	1	8	8,50	68,00
	Total	14		
RECValueProposition	0	6	9,00	54,00
	1	8	6,38	51,00
	Total	14		

RECInternalGoal	0	6	7,50	45,00
	1	8	7,50	60,00
	Total	14		
RECCustomerGoal	0	6	7,50	45,00
	1	8	7,50	60,00
	Total	14		
RECValueStream	0	6	6,83	41,00
	1	8	8,00	64,00
	Total	14		
RECImportance	0	6	6,17	37,00
	1	8	8,50	68,00
	Total	14		

Test Statistics[a]

	RECCompetence	RECValueProposition	RECInternalGoal	RECCustomerGoal	RECValueStream	RECImportance
Mann-Whitney U	16,000	15,000	24,000	24,000	20,000	16,000
Wilcoxon W	37,000	51,000	60,000	60,000	41,000	37,000

Z	-1,700	-1,631	,000	,000	-1,155	-1,700
Asymp. Sig. (2-tailed)	,089	,103	1,000	1,000	,248	,089
Exact Sig. [2*(1-tailed Sig.)]	,345[b]	,282[b]	1,000[b]	1,000[b]	,662[b]	,345[b]

Ranks

	PGAVersion	N	Mean Rank	Sum of Ranks
EASECOMP	0	6	7,33	44,00
	1	8	7,63	61,00
	Total	14		
EASECUST	0	6	6,08	36,50
	1	8	8,56	68,50
	Total	14		
EASEIMPO	0	6	3,92	23,50
	1	8	10,19	81,50
	Total	14		
EASEINTE	0	6	6,33	38,00
	1	8	8,38	67,00
	Total	14		

EASEVALU	0	6	5,67	34,00
	1	8	8,88	71,00
	Total	14		
EASESTRE	0	6	5,42	32,50
	1	8	9,06	72,50
	Total	14		

Test Statistics[a]

	EASECOMP	EASECUST	EASEIMPO	EASEINTE	EASEVALU	EASESTRE
Mann-Whitney U	23,000	15,500	2,500	17,000	13,000	11,500
Wilcoxon W	44,000	36,500	23,500	38,000	34,000	32,500
Z	-,140	-1,224	-2,910	-,947	-1,500	-1,660
Asymp. Sig. (2-tailed)	,889	,221	,004	,343	,134	,097
Exact Sig. [2*(1-tailed Sig.)]	,950[b]	,282[b]	,003[b]	,414[b]	,181[b]	,108[b]