Visual Notation Design 2.0: Towards User Comprehensible Requirements Engineering Notations

Patrice Caire
University of Luxembourg
Luxembourg
patrice.caire@uni.lu

Nicolas Genon, Patrick Heymans
PReCISE Research Centre
Namur, Belgium
nge, phe @fundp.ac.be

Daniel L. Moody
Ozemantics Pty Ltd
Sydney, Australia
daniel@ozemantics.com.au

Abstract—**The success of requirements engineering depends critically on effective communication between business analysts and end users, yet empirical studies show that business stakeholders understand RE notations very poorly. This paper proposes a novel approach to designing RE visual notations that actively involves naïve users in the process. We use i*, one of the most influential RE notations, to demonstrate the approach, but the same approach could be applied to any RE notation. We present the results of 5 related empirical studies that show that novices consistently outperform experts in designing symbols that are comprehensible to novices: the differences are both statistically significant and practically meaningful. Symbols designed by novices increased semantic transparency (their ability to be spontaneously interpreted by other novices) by almost 300% compared to the existing i* notation. The results challenge the conventional wisdom about visual notation design: that it should be conducted by a small group of experts; our research suggests that instead it should be conducted by large numbers of novices. This approach is consistent with principles of Web 2.0, in that it harnesses the collective intelligence of end users and actively involves them in the notation design process as "prosumers" rather than as passive consumers. We believe this approach has the potential to radically change the way visual notations are designed in the future.**

***Index Terms*—visual languages, empirical research, modeling, analysis, end user communication, requirements analysis**

I. THE PROBLEM ADDRESSED

Requirements engineering is, to a large extent, a communication problem: its success depends critically on effective communication between business analysts and end users (customers). For this reason, the research question addressed by this paper – how to design user-comprehensible visual notations – is a key issue in RE research and practice. Yet empirical studies show that we have been spectacularly unsuccessful in doing this: both field and laboratory studies show that end users understand RE models very poorly and that most analysts do not even show models to their customers [14,15,33,43]. One of the reasons why this is so difficult to do is that it is hard for experts to think like novices, a phenomenon called **the curse of knowledge** [13]. There are well-known differences in how experts and novices process diagrams [5,34,46] that are rarely taken into account in designing RE visual notations.

This paper argues that perhaps we have been going about this task the wrong way and that the solution may have been under our noses the whole time: to design notations that are understandable to end users, why not involve them in the design process? If this works in developing software systems (e.g. **participatory design**, **user-centred design**), why shouldn't it also work in developing visual notations?

A. *Semantic Transparency: "visual onomatopoeia"*

The key to designing visual notations that are understandable to naïve users is a property called **semantic transparency** [29]. Literally, this means that the meaning (*semantics*) of a symbol is clear (*transparent*) from its appearance alone. This is the visual equivalent of onomatopoeia in spoken language. **Onomatopoeia** is a literary device in which words are used whose sound suggests their meaning. Semantic transparency is the visual analogue of this, where symbols are used whose appearance suggests their meaning.

Semantic transparency is one of the most powerful tools in the visual notation designer's bag for improving understanding by novices. Semantically transparent symbols reduce cognitive load because they have built-in **mnemonics**: as a result, their meaning can be either perceived directly or easily learnt [37]. Such representations speed up recognition and improve intelligibility to naïve users [3,27]. Semantic transparency is not a binary state but a sliding scale (Fig 1):

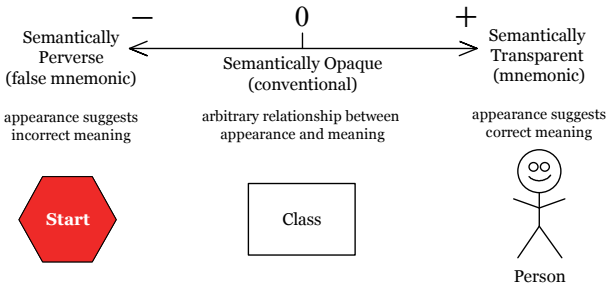

Fig 1. Semantic Transparency is a Continuum

- At the positive end of the scale, **semantic transparency** means that a novice reader could accurately infer the meaning of a symbol from its appearance (e.g. a stick figure to represent a person).
- At the zero point of the scale, **semantically opacity** means there is a purely arbitrary association between a symbol and its meaning (e.g. a rectangle to represent a UML class). Such symbols require conscious effort to remember and must be learnt by rote. Most symbols in RE notations fit into this category, as they are abstract shapes.
- At the negative end, **semantic perversity** means a novice reader would be likely to infer an incorrect meaning from the symbol's appearance (e.g. a red hexagon to indicate

"start"). Such symbols require the most effort to remember, as they require *unlearning* the familiar meaning.

Semantic transparency formalises subjective notions like "intuitiveness" or "naturalness" that are often used informally when discussing visual notations.

B. Operationalising semantic transparency.

Semantic transparency is defined as "the extent to which a novice reader can infer the meaning of symbol from its appearance alone" [29]. However, semantic transparency is typically evaluated subjectively: experts (researchers, notation designers) try to estimate the likelihood that novices will be able to infer the meaning of particular symbols [e.g. 30,31]. Experts are poorly qualified to do this because it is difficult for them to think like novices (cf. the curse of knowledge). This paper defines a way of empirically measuring (**operationalising** [9]) semantic transparency, which provides a way of objectively resolving such issues.

C. Visual Notation Design 2.0

Until now, the design of RE visual notations has been the exclusive domain of technical experts (e.g. researchers, members of OMG technical committees). Even when notations are specifically designed for communicating with business stakeholders, members of the target audience are rarely involved. For example, BPMN 2.0 is a notation designed for communicating with business stakeholders, yet no business representatives were involved in the notation design process and no testing was conducted with them prior to its release [38,44]. In the light of this, it is perhaps no surprise that RE notations are understood so poorly by business stakeholders: this is analogous to building a software system without involving end users or conducting **user acceptance testing** prior to its release, which would be a recipe for disaster.

Web 2.0 involves a radical change in the dynamics of content creation on the web, where end users can contribute their own content rather than being passive consumers [35]. For example, Threadless is a T-shirt company that does not have its own designers but allows customers to submit their own designs, which are voted on by other customers: the most popular designs are then put into production. In this paper, we apply the Web 2.0 philosophy to designing RE visual notations. We define a process for actively involving naïve users in the notation design process as *co-developers* (**prosumers**) rather than as passive consumers.

D. Research objectives

The broad research questions addressed by this paper are:

RQ1.	How can we objectively measure the semantic transparency of visual notations?
RQ2.	How can we improve the semantic transparency of visual notations?
RQ3.	Can novices design more semantically transparent symbols than experts?
RQ4.	How can we actively (and productively) involve end users in the visual notation design process?
RQ5.	How can we evaluate the user comprehensibility of visual notations prior to their release (analogous to user acceptance testing for software systems)?

II. Previous Research

A. The Physics of Notations

Traditional approaches to visual notation design are characterised by:

- An **unselfconscious** design approach [1]: there is a lack of principles for designing visual notations, meaning that designers must rely on instinct, imitation and tradition.
- Lack of **design rationale** [24]: symbols are defined without explaining why they were chosen (a common characteristic of unselfconscious design cultures [1]).

The **Physics of Notations** [29] defines a theory for designing **cognitively effective** visual notations: notations that are optimised for processing by the human mind. It consists of 9 principles based on theory and empirical evidence from a wide range of fields: semantic transparency is one of the principles. This paper extends the theory by operationalising semantic transparency and defining a way of building this into notations. It also empirically tests some of the predictions of the theory:

RQ6.	Does improving semantic transparency improve understanding by novices?
RQ7.	Does the use of explicit design principles (selfconscious design [1]) improve semantic transparency?
RQ8.	Does including explicit design rationale improve understanding by novices?

B. Goal-oriented modelling

Goal-oriented modelling is one of the most important developments in the RE field, which changes the focus from *what* and *how* (data and processes) as in traditional analysis to *who* and *why* (the actors and the goals they wish to achieve). ***i**** is one of the most widely used goal modelling languages and one of the most influential notations in the RE field [12,47,48]. Like most RE notations, it is specifically designed for communicating with business stakeholders, yet makes little or no use of semantic transparency [31]. All symbols are abstract shapes (Fig. 2) so are **semantically opaque**. Also, like most RE notations, i* lacks design rationale: symbols are defined without any explanation of why they were chosen.

Fig. 2. Standard *i** symbol set [48]

C. Applying the Physics of Notations to i*

A previous paper [31] conducted an evaluation of the i* visual notation using the Physics of Notations and proposed a revised symbol set (Fig. 3). These revisions were based on a number of principles, including Semiotic Clarity, Perceptual Discriminability, Semantic Transparency, Visual Expressiveness and Graphic Economy. Explicit design rationale was provided for each symbol. We refer to this symbol set as **PoN i*** for the remainder of the paper.

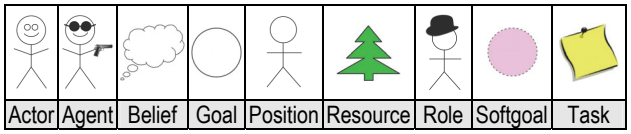

Fig. 3. Revised i* symbol set (PoN i*) [31]

The paper concluded that the standard i* visual notation was semantically opaque (all symbols were judged to be opaque except for the Belief symbol) and that the revised notation was semantically transparent. These claims were made based on expert judgement, so need to be empirically validated. This leads to two additional research questions:

RQ9. Is the i* visual notation semantically opaque?
RQ10. Is the PoN symbol set semantically transparent?

Testing these claims requires a way of empirically measuring semantic transparency (RQ1).

III. Research Design

The research design consists of 5 related empirical studies (4 experiments and 1 non-reactive study), summarised in Fig. 4. As shown in the diagram, the results of earlier studies provide inputs to later studies.

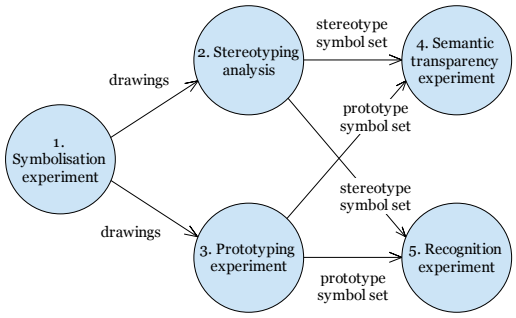

Fig. 4. Research design

1. **Symbolisation experiment**: naïve participants generated symbols for i* concepts, a task normally reserved for experts.
2. **Stereotyping analysis** (nonreactive study): we identified the most common symbols produced for each i* concept. This defined the **stereotype symbol set**.
3. **Prototyping experiment**: naïve participants identified the "best" representations for each i* concept. This defined the **prototype symbol set**.
4. **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*).
5. **Recognition experiment**: we evaluated the ability of naïve participants to learn and remember symbols from the 4 symbol sets.

The research design combines quantitative and qualitative research methods, thus providing **triangulation of method** [21]: Studies 1, 2 and 3 use qualitative methods, while Studies 4 and 5 use quantitative methods.

IV. Study 1: Symbolisation Experiment

In this experiment, we asked naïve participants to generate symbols for i* concepts. To do this, we used the **sign production technique**, developed by Howell and Fuchs [17] to design military intelligence symbols. This involves asking members of the target audience (those who will be interpreting diagrams) to generate symbols to represent concepts. The rationale behind this approach is that symbols produced by members of the target audience are more likely to be understood and recognised by other members of the target audience, due to their common cognitive profile. The results of sign production studies consistently show that symbols produced in this way are more accurately interpreted than symbols produced by experts. This approach has been used to design office equipment symbols [16], public information symbols [49] and icons for graphical user interfaces [22] but has so far not been used to design RE visual notations.

A. Participants

There were 104 participants (53 females and 51 males), all undergraduate students in Economics and Management from the University of Namur. They had no previous knowledge of goal modelling in general or i* in particular: this was a requirement for participation in the study (to ensure they were naïve). We chose business students as proxies for naïve users, as they present a similar cognitive profile: they have a business rather than technical orientation and no prior notational knowledge. IT students would not have been suitable participants in this experiment, due to their technical knowledge and orientation (i.e. the curse of knowledge).

B. Materials

Each participant was provided with a 10-page booklet, a pencil and eraser. The first page was used to ask the screening question (prior knowledge of goal modelling or i*) and to collect demographic data. The remaining pages were used to elicit symbols for the 9 i* constructs. Each construct and its definition was printed at the top of each page and participants were asked to draw the construct in the space below. To control for the size of drawings, a frame measuring 7.5cm x 7.5cm was drawn in the middle of the page.

C. Procedure

We followed the same procedures as used in previous sign production studies [e.g. 16,17,22,39]. Participants were instructed to draw the constructs in the order in which they appeared in the booklet and to produce drawings that they felt most effectively conveyed the meaning of the construct. No time limit was set but, on average, subjects took 15-25 minutes to complete the task.

D. Results

The participants produced a total of 897 drawings (corresponding to a response rate of 95.8%), which was quite a high response rate given the abstract nature of the concepts. Softgoal (9.62%) and Belief (8.65%) received the highest number of non-responses, with Actor, Position and Goal receiving less than 1% (only 1 null response out of 104).

V. Study 2: Stereotyping Analysis (nonreactive)

In this study, we analysed the drawings produced in Study 1 and identified the drawings most commonly produced for each i* concept: this defines the **population stereotype** or **median drawing**. The rationale for doing this is that the representation most commonly produced should also be the most frequently recognised as representing that concept by members of the target audience [17,22].

A. Participants

This analysis was conducted by 3 independent raters (2 of the authors of this paper plus an external rater, not involved in the research). Naïve participants were not required for this task, as stereotype identification can be done relatively objectively: it is a perceptual (pattern-matching) task rather than a cognitive task so less subject to expertise bias.

B. Procedure

We used the **judges' ranking method** [22] to identify stereotypes, which is an approach often used to achieve convergence on a common set of categories. In the first round, each judge independently categorised the drawings based on their visual and conceptual similarity. They then compared their classifications, agreed on a common set of categories and how each drawing should be classifed. Finally, they selected the most representative drawing from the category with the highest number of drawings for each concept (the **stereotypical category**), resulting in 9 stereotypical drawings.

C. Results

The primary outcome of this study was a set of 9 stereotypical drawings, one per i* construct (Fig. 5). The **degree of stereotypy** [17] or **stereotype weight** [22] represents an index of the strength of the stereotype: the level of agreement among participants about how the concept should be visually represented. All exceeded 30% but none achieved an absolute majority, which probably reflects the inherent difficulty in "concretizing" such abstract concepts [22].

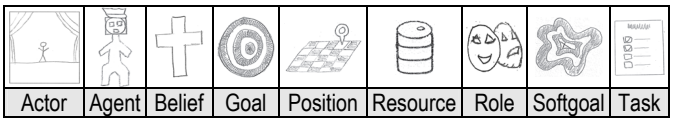

Fig. 5. Stereotype Symbol Set

VI. Study 3: Prototyping Experiment

Stereotyping has been criticised on the grounds that drawings produced most frequently may not necessarily be the ones that convey concepts most effectively. For example, Jones [22] found that in around 20% of cases the best representation of a concept as judged by members of the target audience was produced by a single person out of more than 100. In this experiment, we asked naïve participants to analyse the drawings produced in Study 1 and choose which best represents each i* construct. The drawing that received the highest rating across all participants defines the **population prototype** [22]. This represents a consensus judgement by members of the target audience about semantic transparency.

A. Participants

There were 30 naïve participants in this experiment, all students in Economics and Management from the University of Namur. We used different participants than in Study 1 but drawn from the same underlying population. It would not have been appropriate for the authors to perform this analysis, as unlike prototyping, it is not possible to do this objectively and we needed judgements by novices rather than experts. It would also not have been appropriate to use the same participants as in Study 1, as their judgements may have been biased by the drawings they produced.

B. Procedure

We conducted this experiment electronically. On the opening screen, participants were asked to answer the selection question and enter their demographic data. They then 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 below: radio buttons were provided to select the best representation. To make the task manageable, we selected a representative drawing from each category identified in Study 2 (93 in total) rather than using all 897 drawings from Study 1. Participants were asked to identify the drawing that most effectively conveyed each concept, irrespective of their artistic quality. Both the order of the screens (concepts) and the position of the drawings on each screen were randomized to counteract sequence effects. No time limit was set but subjects took 5–15 minutes to complete the task.

C. Results

The primary outcome of this experiment was a set of 9 prototypical drawings, one per i* concept (Fig. 6).

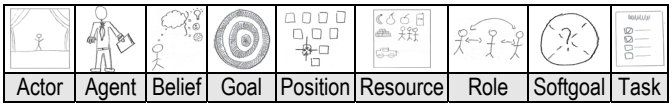

Fig. 6. Prototype Symbol Set

For 3 of the concepts, the prototypical drawings were the same as the stereotypical drawings, showing that in only a third of cases, the most common idea was the best. This supports concerns expressed in the literature about the validity of stereotyping as a basis for selecting the best symbols. For all concepts, a clear prototype emerged: there was a relatively high level of consensus among judgements of prototypicality (67% overall), with most (7/9) prototypes achieving an absolute majority.

VII. Study 4: Semantic Transparency Experiment

This experiment evaluated the ability of naïve participants to infer the meanings of symbols. To do this, we used a **blind interpretation study** (also called **comprehension test** [11,49] or **recognition test** [17]). This is the method most commonly used to measure comprehensibility of graphic symbols and is used for testing ISO standard symbols prior to their release [19]. The essence of this type of test is that participants are shown a symbol and asked to guess or infer its meaning. This corresponds very closely to the definition of semantic transparency ("the extent to which a novice reader can infer the meaning of symbol from its appearance alone"). The comprehensibility of the symbol is typically measured by the percentage of correct responses (hit rate).

A. Participants

In previous research, semantic transparency has almost always been evaluated by experts, who are poorly qualified to do this: the definition of semantic transparency [29] refers to "novice readers", which they most certainly are not. For this reason, we used naïve participants in this experiment. There were 65 participants, undergraduate students in Interpretation and Translation from the Haute Ecole Marie HAPS-Bruxelles or Accountancy from the Haute Ecole Robert Schuman-

Libramont. As in Studies 1 and 3, the participants had no previous knowledge of goal modelling or i*.

B. Experimental design

A 4 group, post-test only experimental design was used, with 1 active between-groups factor (symbol set). There were four experimental groups, corresponding to different levels of the **independent variable** (**symbol set**):

1. **Standard i*** (Fig. 2): produced by experts using intuition (unselfconscious design [1]).
2. **PoN i*** (Fig. 3): produced by experts using explicit principles (selfconscious design [1]).
3. **Stereotype i*** (Fig. 5): the most common symbols produced by novices.
4. **Prototype i*** (Fig. 6): the best symbols produced by novices (as judged by other novices).

The **dependent variables** were **hit rate** and the **semantic transparency coefficient**. The levels of the independent variable enable comparisons between unselfconscious and selfconscious design (group 1 vs 2), experts and novices (1+2 vs 3+4) and stereotyping and prototyping (3 vs 4), so provides the basis for answering RQ3 and RQ6 plus one additional question:

> RQ11. Does prototyping result in more semantically transparent symbols than stereotyping?

C. Materials

4 sets of materials were prepared, one for each symbol set. As in all previous experiments, the first page was used to ask the screening question and collect demographic data. The remaining 9 pages were used to evaluate the semantic transparency of the symbols. Each symbol was displayed at the top of the page (the **stimulus**) and the complete set of i* constructs and definitions displayed in a table below (the candidate **responses**). Participants were asked to indicate which construct they thought most likely corresponded to the symbol. Both the order in which the symbols were presented and the order in which the concepts were listed on each page were randomised to counteract sequence effects.

D. Procedure

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.

E. Hypotheses

Given that sign production studies consistently show that symbols produced by novices are more accurately interpreted than those produced by experts, we predicted that the stereotype and prototype symbol sets would outperform the standard i* and PoN symbol sets. We also predicted that the prototype symbol set would outperform the stereotype set as it represents the best drawings rather than just the most common ones. Finally, we predicted that the PoN symbol set would outperform the standard i* symbol set as it was designed based on explicit principles rather than intuition. This results in a total ordering of the symbol sets:

Prototype > *Stereotype* > *PoN* > *Standard i**

This corresponds to 12 separate hypotheses: all possible comparisons between groups on both dependent variables.

F. Results

1) Statistical significance vs practical meaningfulness

In interpreting empirical results, it is important to distinguish between **statistical significance** and **practical meaningfulness** [7]. Statistical significance is measured by the **p-value**: the probability that the result could have occurred by chance. However, significance testing only provides a binary (yes/no) response as to whether there is a difference (and whether hypotheses should be accepted or rejected), without providing any information about how *large* the difference is [6,7]. Using large enough sample sizes, it is possible to achieve statistical significance for differences that have little or no practical relevance. **Effect size (ES)** provides a way of measuring the size of differences and has been suggested as an index of practical meaningfulness [40]. Statistical significance is most important in theoretical work, while effect size is most important for applied research that addresses practical problems [6].

2) Hit rate

The traditional way of measuring comprehensibility of graphical symbols [18,20] is by measuring hit rates (percentage of correct responses). Only 6 out of the 36 symbols meet the ISO threshold for comprehensibility of 67% [11], with 5 of these from the stereotype symbol set (TABLE I).

TABLE I. HIT RATE ANALYSIS
(GREEN = ABOVE ISO THRESHOLD; UNDERLINE = BEST)

	Standard	PoN	Stereotype	Prototype
Actor	11.1%	37.5%	62.5%	43.8%
Agent	11.1%	37.5%	50.0%	37.5%
Belief	33.3%	43.8%	93.8%	31.3%
Goal	11.8%	31.3%	56.3%	31.3%
Position	5.6%	12.5%	43.8%	50.0%
Resource	11.1%	50.0%	75.0%	37.5%
Role	11.1%	43.8%	75.0%	43.8%
Softgoal	50.0%	12.5%	75.0%	50.0%
Task	11.1%	81.3%	75.0%	50.0%
Mean hit rate	**17.4%**	**38.9%**	**67.4%**	**41.7%**
Std dev	**14.5%**	**20.7%**	**15.6%**	**7.7%**
Group size (n)	**18**	**16**	**16**	**16**

In terms of overall comprehensibility, the stereotype symbol set also met the ISO threshold, which is remarkable given the abstract nature of the concepts. However none of the other symbol sets even came close (all were under 50%). The stereotype symbol set achieved a mean hit rate of almost 4 times that of standard i*, showing just how far from their potential for user comprehensibility current RE notations are.

3) Semantic transparency coefficient

The problem with conventional measures of symbol comprehension such as hit rate is that they cannot have negative values. Semantic transparency is defined as a scale from –1 to +1: it can be negative for symbols whose appearance implies an incorrect meaning (semantically perverse). In this paper, we propose a new measure of semantic transparency called the semantic transparency coefficient, based on the concept of expected and actual frequencies. Like a correlation coefficient, it varies from –1 to +1, so is consistent with the theoretical definition of semantic transparency. It is intended to measure the "correlation" between a symbol's appearance and its meaning:

positive values correspond to semantic transparency and negative values to semantic perversity. A symbol's semantic transparency coefficient is calculated using the following formula, based on the concept of Chi-square analysis:

$$\frac{\textit{maximum frequency} - \textit{expected frequency}}{\textit{total responses} - \textit{expected frequency}}$$

The **expected frequency** (number of responses expected by chance) = n/s, where n is the number of participants in the experimental group and s is the number of symbols. If the target concept receives the maximum number of responses, the coefficient will have a positive sign (transparent), while if a distractor concept is the maximum, the value will have a negative sign (perverse). The semantic transparency coefficients for all symbols are shown in TABLE II. The user generated symbols were *all* semantically transparent but only 2 of the standard i* symbols and 7 of the PoN symbols were.

TABLE II. SEMANTIC TRANSPARENCY COEFFICIENT RESULTS (GREEN = TRANSPARENT; UNDERLINE = BEST)

	Standard	PoN	Stereotype	Prototype
Actor	-0.31	0.30	0.58	0.37
Agent	-0.19	0.30	0.39	0.30
Belief	0.25	0.37	0.83	0.23
Goal	-0.07	0.23	0.45	0.23
Position	-0.19	-0.30	0.33	0.44
Resource	-0.25	0.44	0.64	0.30
Role	-0.19	0.37	0.64	0.37
Softgoal	0.44	-0.16	0.64	0.44
Task	-0.13	0.79	0.64	0.44
Mean	**-0.07**	**0.26**	**0.57**	**0.34**
Std dev	**0.25**	**0.32**	**0.15**	**0.09**
≠ 0 (one sample t-test)	**Opaque (p = .419)**	**Transparent (p = 042*)**	**Transparent (p = .000***)**	**Transparent (p = .000***)**

Overall, the standard i* symbol was found to be slightly semantically perverse, but a **one sample t-test** showed that the mean was not significantly different to zero, meaning that it is semantically opaque: this confirms RQ9. All the other symbol sets were found to be significantly semantically transparent, which confirms RQ10.

4) Hypothesis testing (differences between groups)

A **one-way analysis of variance (ANOVA)** was used to analyse differences between symbol sets on hit rate and semantic transparency. Hypothesis testing was conducted using predefined contrasts as part of the ANOVA procedure. **Cohen's d** was used to analyse the practical meaningfulness of the results (effect size) [7]. The results for the semantic transparency coefficient are summarised in TABLE III.

TABLE III. RESULTS OF HYPOTHESIS TESTING FOR SEMANTIC TRANSPARENCY COEFFICIENT (GREEN = CONFIRMED, WHITE = REJECTED, RED = CONVERSE)

Hypothesis	Statistical significance (p)	Practical meaningfulness (d)
H1: PoN > Standard	.005**	1.22+++
H2: Stereotype> Standard	.000***	3.32+++
H3: Prototype > Standard	.002***	2.19+++
H4: Stereotype > PoN	.000***	1.57+++
H5: Prototype > PoN	.703	–
H6: Prototype > Stereotype	.001***	–2.21+++

Statistical significance: * significant with $\alpha = .05$, ** $\alpha = .01$, *** $\alpha = .005$
Practical meaningfulness: + small effect ($|d| \geq .2$), ++ medium effect ($|d| \geq .5$), +++ = large effect ($|d| \geq .8$)

Only one comparison was non-significant: no difference was found between the prototype and PoN symbol sets (H5). However, contrary to our predictions, the *converse result* was found for H6: the symbols most commonly produced (stereotypes) were more semantically transparent than those judged by members of the target audience to be the best (prototypes). This results in the following ranking of the symbol sets (a partial ordering):

$$\textit{Stereotype} > \textit{Prototype} = \textit{PoN} > \textit{Standard}$$

All effect sizes are large, meaning that the differences found are also practically meaningful. Note that effect size is only reported if the difference is statistically significant; also, the effect size is negative if the effect is in the reverse direction to that predicted.

The results of hypothesis testing for hit rate are not reported here but the same differences between groups were found as for the semantic transparency coefficient: H7-H10 confirmed, H11 rejected, converse for H12. This is consistent with these variables being alternative measures of the same theoretical construct (semantic transparency).

G. Discussion

In terms of our original research questions, the conclusions from this experiment are:

RQ3: Novice generated symbols are more semantically transparent than those generated by experts: this is supported by H2, H3 and H4 but not H5. When we pooled the results of the two expert groups and the two novice groups, we found a significant difference in favour of the novice groups ($p = .000$) with a large effect size ($d = 1.47$). Remarkably, the average semantic transparency of novice-generated symbols was *more than 5 times* that of expert-generated symbols (.09 vs .46). This is consistent with the results of previous sign production studies but a very surprising result in an RE context, where the implicit assumption has always been that experts are best qualified to design symbols.

RQ7: Using explicit design principles significantly improves semantic transparency (H1). The mean hit rate for the PoN symbol set was more than twice that of the standard i* notation, meaning that symbols were more than twice as likely to be correctly interpreted without explanation.

RQ11: The superiority of the stereotype over prototype symbol set (H6) was a surprising result and challenges the standard assumption in sign production studies that drawings rated as being the best really *are* the best. This may be an example of the **preference performance paradox**: what people prefer is not necessarily what is most effective [25]. Judgements about which representation is "best" may be influenced by factors such as familiarity and aesthetics, which have nothing to do with effectiveness. It also shows the dangers of relying on subjective judgements about semantic transparency even by members of the target audience: objective, performance-based evaluations (as conducted in this experiment) provide a much more reliable basis for identifying appropriate symbols.

VIII. STUDY 5: RECOGNITION EXPERIMENT

This experiment evaluates participants' ability to learn and remember symbols from the different symbol sets. Participants were given one of the symbol sets to learn and later had to re-

call their meanings: this represents a **recognition task**. This is closer to what end users have to do in reality than guessing what symbols mean (as in Study 4), so has greater **ecological validity**. This experiment also allows us to evaluate the effect of semantic transparency on cognitive effectiveness, as recognition performance provides an early measure of cognitive effectiveness (as accurate interpretation of symbols is a prerequisite for accurate interpretation of diagrams).

A. Participants

There were 83 participants in this experiment, all undergraduate students in Accountancy from Haute Ecole Robert Schuman-Libramont or Interpretation and Translation from the Haute Ecole Marie HAPS-Bruxelles. The participants had no previous knowledge of i* or goal modelling (this was the selection criterion as in all of the previous experiments).

B. Experimental design

A 5 group, post-test only experimental design was used, with 2 active between-groups factors (symbol set and design rationale). The groups were the same as in Study 4 with one additional group: PoN with design rationale (PoN DR).

C. Materials

5 sets of materials were prepared, one for each group:

- Training materials: these defined all symbols and associated meanings for one of the symbol sets. The PoN DR symbol set included explicit design rationale for each symbol, taken from [31]. Design rationale could not be included for any of the other symbol sets because it did not exist: i* lacks design rationale for its symbols and the symbolisation experiment did not ask participants to provide design rationale for their drawings.
- Testing materials: these were used to evaluate participants' ability to accurately recognise symbols (recall their meanings). The same test materials were used as in Study 4, though in this case it was an "unblinded" interpretation test, as participants learnt the meanings of the symbols in advance.

D. Procedure

Participants were instructed 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 per page and participants had to identify the corresponding concept. Participants were not allowed to take notes during the learning phase or to refer back to the training materials during the testing phase. No time limit was set but subjects took 10-15 minutes to complete the task.

E. Hypotheses

6 of the hypotheses for this study (H13-H18) took the same form as those in the previous experiment, in that we predicted the following ranking of symbol sets:

*Prototype > Stereotype > PoN > Standard i**

We also predicted that design rationale would improve recognition performance (H19): this involves a comparison between the PoN and PoN DR groups. Finally, we predicted that semantic transparency would have a positive effect on recognition performance (H20). These last two hypotheses were based on predictions of the Physics of Notations.

F. Results

The results for the 5 symbol sets are summarised in the box and whisker plot in Fig. 7. Error rates (% incorrect responses) are used instead of accuracy scores (% correct responses) to more clearly highlight the differences between groups. Again, the stereotype symbol set performed the best (closely followed by PoN DR) and reduced the incidence of interpretation errors *by more than 5 times* (15.97% vs 3.09%) compared to standard i*.

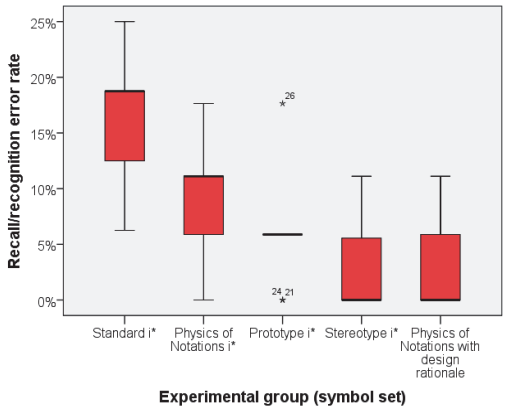

Fig. 7. Comparison of recognition error rates

1) Differences between groups

As in the previous experiment, a one-way analysis of variance (ANOVA) was used to analyse differences between groups. The results are summarised in TABLE IV.

TABLE IV. RESULTS OF HYPOTHESIS TESTING FOR RECOGNITION (GREEN = CONFIRMED, WHITE = REJECTED)

Hypothesis	Statistical significance (p)	Practical meaningfulness (d)
H13: PoN > Standard	.006**	1.16 +++
H14: Stereotype> Standard	.000***	2.32+++
H15: Prototype > Standard	.000***	1.65+++
H16: Stereotype > PoN	.022***	1.02+++
H17: Prototype > PoN	.052	–
H18: Prototype > Stereotype	.708	–
H19: PoN DR > PoN	.041*	1.05+++

Only two comparisons were not significant: no difference was found between the prototype and PoN symbol sets (as in Study 4) or between the prototype and stereotype symbol sets. All statistically significant differences were also practically meaningful, with large effect sizes in all cases.

2) Effect of semantic transparency on recall/ recognition

To evaluate RQ6, we conducted a **linear regression analysis** across all symbols and symbol sets using the semantic transparency coefficient as the **independent (predictor) variable** and recognition accuracy as the **dependent (outcome) variable** (Fig. 8). The results show that semantic transparency explains 43% of the variance in recognition performance **(r^2 = .43)** [41]. The effect is both statistically significant (p = .000) and practically meaningful ($r^2 \geq .25$ = large effect size), which confirms RQ6. The **standardised regression coefficient (β)** is

.66, meaning that for a 1% increase in semantic transparency, there will be a corresponding .66% increase in recognition accuracy. The resulting regression equation is:

Recognition accuracy (%) = 15 * *semantic transparency coefficient* + 88 (1)

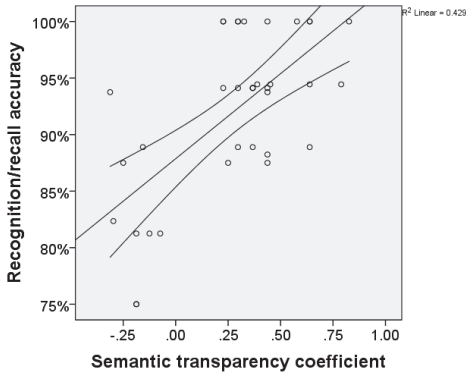

Fig. 8. Scatterplot of semantic transparency vs recognition accuracy showing line of best fit and 95% confidence intervals

3) Effect of design rationale on recognition

The difference between the PoN and PoN DR experimental groups (H13) shows that design rationale improves recognition performance over and above the effects of semantic transparency, thus confirming RQ8. Including design rationale seems to have a similar effect to more than doubling semantic transparency: PoN DR achieved a recognition accuracy of 96.3%, which based on the regression equation (Equation 1), should require a semantic transparency coefficient of .55 (rather than .26). Design rationale appears to act as an adjunct to semantic transparency in recognition processes: when it is difficult to infer the meaning of symbols from their appearance alone, design rationale helps people remember what symbols mean by creating additional semantic cues in long term memory. This emphasises the importance of explanations to the human mind, which is a well-known causal processor: even from an early age, children need to constantly know "why" [28,42].

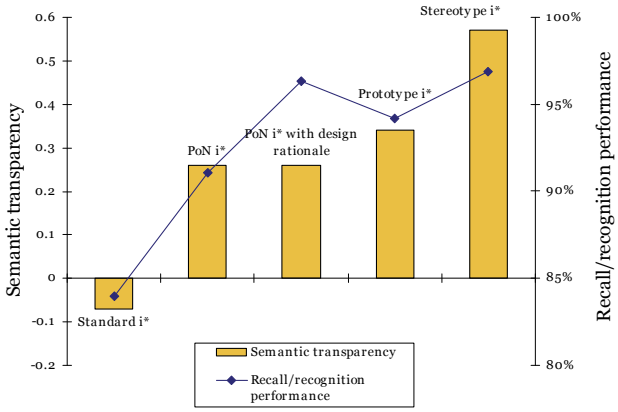

Fig. 9. Relationship between semantic transparency and recognition performance: note the "spike" introduced by design rationale

IX. CONCLUSION

This paper offers a possible solution to the conundrum of how to design visual notations that novices can understand, by getting "inside their heads" to see how they visualise the world. It is difficult, if not impossible, for notation designers to think like novices, but this paper provides a practical way of overcoming the "curse of knowledge" and actively involving end users in the notation design process. The empirical results were statistically significant and practically meaningful, so have implications for both RE theory and practice.

A. Summary of results

We summarise the findings of the paper by answering the research questions raised in Sections 1 and 2 of the paper:

RQ1. How can we objectively measure the semantic transparency of visual notations?

A: Through blind interpretation experiments (e.g. Study 4). In this paper, we have defined a new metric (semantic transparency coefficient), which has theoretical and practical advantages over traditional measures such as hit rate.

RQ2. How can we improve the semantic transparency of visual notations?

A: Through symbolisation experiments (e.g. Study 1) and stereotyping analyses (e.g. Study 2). Using this approach, we were able to increase comprehensibility of symbols (as measured by "hit rates") by naïve subjects by a factor of almost 4 (from 17% to 67%: TABLE I) and reduce interpretation errors by more than 80% (from 16% to 3.1%: Fig. 7) over notations designed in the traditional way (the existing i* symbol set). Even more surprisingly, it resulted in symbols that meet the ISO threshold for public information and safety symbols (67% hit rate): in other words, *self-explanatory symbols*.

RQ3. Can novices design more semantically transparent symbols than experts?

A: The answer to this question is emphatically "yes". 8 out of the 9 *best* (most semantically transparent) symbols for each i* concept were designed by novices and 8 out of the 9 *worst* (least semantically transparent) symbols were designed by experts (TABLE II). The average semantic transparency of novice-generated symbols was more than 5 times that of expert-generated symbols, which challenges the longstanding assumption in the RE field that experts are best qualified to design visual notations.

RQ4. How can we actively (and productively) involve end users in the visual notation design process?

A: Through symbolisation experiments (e.g. Study 1), blind interpretation experiments (e.g. Study 4) and recognition experiments (e.g. Study 5).

RQ5. How can we evaluate user comprehensibility of visual notations prior to their release?

A: Through blind interpretation studies (e.g. Study 4) and recognition studies (e.g. Study 5) using members of the target audience (or proxies). The results of such studies will help identify potential interpretation problems, which can be addressed prior to releasing them on the public. Such testing is routinely carried out for public information and safety symbols and is a requirement for their acceptance as international standards [19] but rarely, if ever, for visual notations (even international standards like UML or BPMN).

RQ6. Does improving semantic transparency improve understanding by novices?

A: Semantic transparency significantly increases recognition accuracy and reduces interpretation errors (Study 5). A 10% increase in semantic transparency leads to a 6.6% reduction in interpretation errors. Given that requirements errors are the source of more than half the errors in software development [10,23,26] and are also the most costly type of error (it is more than 100 times more costly to correct a defect post-implementation than to correct it during the requirements phase [2]), reducing requirements interpretation errors could lead to major improvements in development productivity.

RQ7. Does the use of explicit design principles improve semantic transparency?

A: Use of explicit design principles significantly improved both semantic transparency (Study 4) and cognitive effectiveness (Study 5: recognition accuracy is an early measure of cognitive effectiveness). It more than doubled the average hit rate for symbols (TABLE I) and reduced interpretation errors by almost 50% (Fig. 7).

RQ8. Does including design rationale improve understanding by novices?

A: Design rationale reduces interpretation errors by novices by more than 50% (Fig. 7).

B. Comparison to current notation design approaches

The visual notation design process described in this paper differs from traditional approaches in two important ways:

- Number of participants: in traditional notation design, typically only a single person (e.g. i* [47], ER [4]) or a small group of people (e.g. the "3 amigos" for UML [36]) is involved. In our approach, over 100 people were involved, which represents true "people power" in the style of Web 2.0.
- Expertise of the participants: notation design is normally a task reserved exclusively for technical experts, with end users not involved at all. In this paper, we used novices to generate symbols (Study 1), to choose between them (Study 3) and to evaluate their comprehensibility (Studies 4 & 5) which turns traditional notation design on its head.

We have described a novel approach to developing user comprehensible visual notations that is generalised and repeatable (the same approach could be applied to any RE notation) and is practical to apply (all studies required between 5–25 minutes to complete and used easily accessible populations). We believe this approach has the potential to change the way visual notations are designed in the future.

C. Strengths of the research

Internal validity. The following variables were controlled as part of all experiments:

- Participant characteristics: participants were randomly assigned to experimental groups to control for individual differences (**selection bias**).
- Instrumentation: the same measurement procedures were used across all experimental groups **(measurement bias**).
- Experimental setting: all groups were conducted at the same time and in the same location to eliminate **environmental effects**.
- Sequence: in all experiments, symbols were presented in random order to counteract **sequence effects**.

External validity. We used naïve participants in all experiments to increase generalisability to the target population (naïve users) and to control for **expertise bias**.

Statistical validity. We verified that the assumptions of the statistical techniques used were satisfied.

D. Limitations of the research

Most of the limitations of the research relate to the generalisability of the findings (external validity):

- A possible threat to external validity was the use of students as experimental participants. However, because the selection criteria was that they had no previous knowledge of goal modelling or i* and were business rather than IT students, they can be considered reasonable proxies for naïve users.
- The research tested only a single notation (i*): the results need to be replicated using different notations to confirm the generalisability of findings.
- All studies used participants from a single cultural background (French-Belgian): the results need to be replicated using participants from different cultures as semantic transparency of representations is often culture-dependent.
- All experiments evaluated comprehension of individual symbols rather than complete diagrams. To use a software engineering metaphor, this represents **unit testing** (testing of individual components) rather than **integration testing** (how components work together as a whole). We argue that the results are generalisable to complete diagrams as accurate interpretation of symbols is a prerequisite for accurate interpretation of diagrams [45]: research is in progress to confirm this.
- We investigated only one aspect of visual notation design (semantic transparency), which is only one of 9 principles defined in the Physics of Notations. Optimising one principle at the expense of others can have an adverse effect on overall cognitive effectiveness [29]. In particular, we did not consider ease of drawing of symbols (addressed by the Principle of Cognitive Fit), which is an important consideration in RE practice. This is one of the inherent limitations of experimental research: it is only practicable to investigate a small number of variables at a time.

E. The future of visual notation design: crowdsourcing?

This paper only represents the "tip of the iceberg" in terms of what is possible in visual notation design. Instead of getting participants to generate symbols using pencil and paper, we could do this over the web. In the same way Threadless uses its customers to design its T-shirts, we could use our customers to design our visual notations (i.e. **crowdsourcing** [8]). Throughout the history of the RE field, it has always been the case that naïve users have to learn *our* languages to communicate with us: by getting them to design the languages themselves, we may be able to overcome many of the communication problems that currently beset RE practice.

REFERENCES

[1] Alexander, C.W. (1970): *Notes On The Synthesis Of Form*. Boston, USA: Harvard University Press.

[2] Boehm, B.W. (1981): *Software Engineering Economics*. Englewood Cliffs, USA: Prentice-Hall.

[3] Britton, C. and Jones, S. (1999): "The Untrained Eye: How Languages for Software Specification Support Understanding by Untrained Users." *Human Computer Interaction* **14**: 191-244.

[4] Chen, P.P. (1976): "The Entity Relationship Model: Towards An Integrated View Of Data." *ACM Transactions On Database Systems* **1**(1): 9-36.

[5] Cheng, P.C.-H., Lowe, R.K. and Scaife, M. (2001): "Cognitive Science Approaches To Understanding Diagrammatic Representations." *Artificial Intelligence Review* **15**(1/2): 79-94.

[6] Chow, S.L. (1988): "Significance Test or Effect Size." *Psychological Bulletin* **103**(1): 105-110.

[7] Cohen, J. (1988): *Statistical Power Analysis for the Behavioural Sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Earlbaum and Associates.

[8] Doan, A., Ramakrishnan, R. and Halevy, A.Y. (2011): "Crowdsourcing Systems on the World-Wide Web." *Communications of the ACM* **54**(4): 86-96.

[9] Dubin, R. (1978): *Theory Building (revised edition)*. New York: The Free Press.

[10] Enders, A. and Rombach, H.D. (2003): *A Handbook of Software and Systems Engineering: Empirical Observations, Laws and Theories*. Reading, Massachusetts, USA: Addison-Wesley.

[11] Foster, J.J. (2001): "Graphical Symbols: Test Methods for Judged Comprehensibility and for Comprehension." *ISO Bulletin*: 11-13.

[12] Grau, G., Horkoff, J., Yu, E. and Abdulhadi, S. (2007): "i* Guide 3.0." Retrieved February 10, 2009, from http://istar.rwth-aachen.de/tiki-index.php?page_ref_id=67.

[13] Heath, C. and Heath, D. (2008): *Made to Stick: Why Some Ideas Take Hold and Others Come Unstuck*. London, England: Arrow Books.

[14] Hitchman, S. (1995): "Practitioner Perceptions on the use of Some Semantic Concepts in the Entity Relationship Model." *European Journal of Information Systems* **4**(1): 31-40.

[15] Hitchman, S. (2002): "The Details of Conceptual Modelling Notations are Important - A Comparison of Relationship Normative Language." *Communications of the AIS* **9**(10).

[16] Howard, C., O'Boyle, M.W., Eastman, V., Andre, T. and Motoyama, T. (1991): "The relative effectiveness of symbols and words to convey photocopier functions." *Applied Ergonomics* **22**(4): 218-224.

[17] Howell, W.C. and Fuchs, A.H. (1968): "Population Stereotypy in Code Design." *Organizational Behavior and Human Performance* **3**: 310-339.

[18] ISO (2003): *ISO Standard Graphical Symbols: Safety Colours and Safety Signs – Registered Safety Signs (ISO 7010:2003)*. Geneva, Switzerland, International Standards Organisation (ISO).

[19] ISO (2007): *Graphical symbols - Test methods - Part 1: Methods for testing comprehensibility (ISO 9186-1:2007)*. Geneva, Switzerland, International Standards Organisation (ISO).

[20] ISO (2007): *ISO Standard Graphical Symbols: Public Information Symbols (ISO 7001:2007)*. Geneva, Switzerland, International Standards Organisation (ISO).

[21] Jick, T.D. (1979): "Mixing Qualitative and Quantitative Methods: Triangulation in Action." *Administrative Science Quarterly* **24**: 602-611.

[22] Jones, S. (1983): "Stereotypy in pictograms of abstract concepts." *Ergonomics* **26**(6): 605-611.

[23] Lauesen, S. and Vinter, O. (2000): Preventing Requirement Defects. *Proceedings of the Sixth International Workshop on Requirements Engineering: Foundation for Software Quality (REFSQ'2000)*, Stockholm, Sweden.

[24] Lee, J. (1997): "Design Rationale Systems: Understanding the Issues." *IEEE Expert* **12**(3): 78–85.

[25] Lidwell, W., Holden, K. and Butler, J. (2003): *Universal principles of design: a cross-disciplinary reference*. Gloucester, Massachusetts: Rockport Publishers.

[26] Martin, J. (1989): *Information Engineering*. Englewood Cliffs, New Jersey, USA: Prentice Hall.

[27] Masri, K., Parker, D. and Gemino, A. (2008): "Using Iconic Graphics in Entity Relationship Diagrams: The Impact on Understanding." *Journal of Database Management* **19**(3): 22-41.

[28] Medina, J.J. (2008): *Brain Rules: 12 Principles for Surviving and Thriving at Work, Home, and School* Seattle, Washington, USA: Pear Press.

[29] Moody, D.L. (2009): "The "Physics" of Notations: Towards a Scientific Basis for Constructing Visual Notations in Software Engineering." *IEEE Transactions on Software Engineering* **35**(5): 756-777.

[30] Moody, D.L., Heymans, P. and Matulevicius, R. (2009): *Improving the Effectiveness of Visual Representations in Requirements Engineering: An Evaluation of the i* Visual Notation*. 17th IEEE International Conference on Requirements Engineering (RE09). Atlanta, Georgia, IEEE Computer Society, August 31 - September 4.

[31] Moody, D.L., Heymans, P. and Matulevicius, R. (2010): "Visual Syntax Does Matter: Improving the Cognitive Effectiveness of the i* Visual Notation." *Requirements Engineering Journal* **15**(2): 141-175.

[32] Muller, M.J. and Kuhn, S. (1993): "Participatory Design (Special Issue)." *Communications of the ACM* **36** (6): 24-28.

[33] Nordbotten, J.C. and Crosby, M.E. (1999): "The Effect of Graphic Style on Data Model Interpretation." *Information Systems Journal* **9**(2): 139-156.

[34] Novick, L.P. (2006): The Importance of Both Diagrammatic Conventions and Domain-Specific Knowledge for Diagram Literacy in Science: The Hierarchy as an Illustrative Case *Diagrammatic Representation and Inference*. D. Barker-Plummer, R. Cox and N. Swoboda. Berlin: Springer.

[35] O'Reilly, T. (2007): "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software." *Communications & Strategies* **65**(1): 17-37.

[36] OMG (2003): *Unified Modeling Language (UML) Specification, Version 1.5*: Object Management Group (OMG).

[37] Petre, M. (1995): "Why Looking Isn't Always Seeing: Readership Skills and Graphical Programming." *Communications of the ACM* **38**(6): 33-44.

[38] Recker, J. (2010): "Opportunities and Constraints: The Current Struggle with BPMN." *Business Process Management Journal* **16**(1): 181-201.

[39] Rogers, Y. and Oborne, D.J. (1987): "Pictorial communication of abstract verbs in relation to human-computer interaction." *British Journal of Psychology* **78**: 99-112.

[40] Rosenthal, R. (1991): *Meta-Analytic Procedures for Social Research*. Newbury Park, CA: Sage Publications.

[41] Rosenthal, R. and Rubin, D.B. (1979): "A Note on Percent Variance Explained as a Measure of Importance of Effects." *Journal of Educational Psychology* **74**: 395-396.

[42] Sagan, C. (1997): *The Demon-Haunted World: Science as a Candle in the Dark*. New York: Random House.

[43] Shanks, G.G. (1997): "The Challenges Of Strategic Data Planning In Practice: An Interpretive Case Study." *Journal of Strategic Information Systems* **6**(1): 69-90.

[44] Silver, B. (2009): *BPMN Method and Style: A levels-based methodology for BPM process modeling and improvement using BPMN 2.0*: Cody-Cassidy Press.

[45] Winn, W.D. (1990): "Encoding and Retrieval of Information in Maps and Diagrams." *IEEE Transactions on Professional Communication* **33**(3): 103-107.

[46] Winn, W.D. (1993): "An Account of How Readers Search for Information in Diagrams." *Contemporary Educational Psychology* **18**: 162-185.

[47] Yu, E. (1995): *Modelling Strategic Relationships for Process Reengineering (PhD thesis)*: Department of Computer Science, University of Toronto.

[48] Yu, E. (1997): *Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering*. Proceedings of the 3rd IEEE International Conference on Requirements Engineering (RE'97). Washington D.C., USA**:** 226-235, January 6-8.

[49] Zwaga, H.J. and Boersema, T. (1983): "Evaluation of a set of graphic symbols." *Applied Ergonomics* **14**(1): 43-54.