Expert Systems With Applications 136 (2019) 145–158

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Comprehension of business process models: Insight into cognitive strategies via eye tracking

Miles Tallon[a,d,*], Michael Winter[b], Rüdiger Pryss[b], Katrin Rakoczy[c], Manfred Reichert[b], Mark W. Greenlee[a], Ulrich Frick[d]

[a] Institute for Experimental Psychology, University of Regensburg, Regensburg, Germany
[b] Institute of Databases and Information Systems, Ulm University, Ulm, Germany
[c] Leibniz Institute for Research and Information in Education, Frankfurt/Main, Germany
[d] HSD Research Centre, HSD – University of Applied Sciences, Cologne, Germany

ARTICLE INFO

Article history:
Received 3 February 2019
Revised 30 May 2019
Accepted 15 June 2019
Available online 17 June 2019

Keywords:
Visual literacy
Business process model
Eye tracking
Latent class analysis
Cognitive workload

ABSTRACT

Process Models (PM) are visual documentations of the business processes within or across enterprises. Activities (tasks) are arranged together into a model (i.e., similar to flowcharts). This study aimed at understanding the underlying structure of PM comprehension. Though standards for describing PM have been defined, the cognitive work load they evoke, their structure, and the efficacy of information transmission are only partially understood. Two studies were conducted to better differentiate the concept of *visual literacy* (VL) and *logical reasoning* in interpreting PM.

Study I: A total of 1047 students from 52 school classes were assessed. Three different process models of increasing complexity were presented on tablets. Additionally, written labels of the models' elements were randomly allocated to scholars in a 3-group between-subjects design. Comprehension of process models was assessed by a series of 3×4 (=12) dichotomous test items. Latent Class Analysis of solved items revealed 6 qualitatively differing solution patterns, suggesting that a single test score is insufficient to reflect participants' performance.

Study II: Overall, 21 experts and 15 novices with respect to visual literacy were presented the same set of PMs as in Study I, while wearing eye tracking glasses. The fixation duration on relevant parts of the PM and on questions were recorded, as well as the total time needed to solve all 12 test items. The number of gaze transitions between process model and comprehension questions was measured as well. Being an expert in visual literacy did not alter the capability of correctly understanding graphical logical PMs. Presenting PMs that are labelled by single letters had a significant influence on reducing the time spent on irrelevant model parts but did not affect the fixation duration on relevant areas of interest.

Both samples' participants required longer response times with increasing model complexity. The number of toggles (i.e., gaze transitions between model and statement area of interest) was predictive for membership in one of the latent classes. Contrary to expectations, denoting the PM events and decisions not with real-world descriptions, but with single letters, led to lower cognitive workload in responding to comprehension questions and to better results. Visual Literacy experts could neither outperform novices nor high-school students in comprehending PM.

1. Introduction

1.1. What are process models?

A process model (PM) is a textual or visual representation, which documents all steps of an entire process (Schultheiss & Heiliger, 1963). Thereby, visual process models, inter alia, allow the depiction of complex algorithms, business steps, or logistical operations in a descriptive form (Aguilar-Saven, 2004; Bharathi et al., 2008; Rojas, Munoz-Gama, Sepúlveda & Capurro, 2016).

Abbreviations: PM, Process Model; VL, Visual Literacy.

* Corresponding author at: Institute for Experimental Psychology, Universitätsstraße 31, 93053 Regensburg, Germany.
E-mail addresses: miles.tallon@stud.uni-regensburg.de, m.tallon@hs-doepfer.de (M. Tallon), michael.winter@uni-ulm.de (M. Winter), ruediger.pryss@uni-ulm.de (R. Pryss), rakoczy@dipf.de (K. Rakoczy), manfred.reichert@uni-ulm.de (M. Reichert), mark.greenlee@ur.de (M.W. Greenlee), u.frick@hs-doepfer.de (U. Frick).

https://doi.org/10.1016/j.eswa.2019.06.032
0957-4174/

PM should be designed such that practitioners can apply them for their tasks at hand (Roehm, Tiarks, Koschke & Maalej, 2012; Ungan, 2006). Moreover, PMs have to be understandable by all practitioners (Reggio, Ricca, Scanniello, Di Cerbo & Dodero, 2015; Zimoch, Pryss, Probst, et al., 2017). Existing research on process model comprehension therefore has considered two groups of factors: (1) Subjective capability (e.g., model reader expertise) should be distinguished from (2) objective characteristics of the model itself (e.g., process model complexity).

For objective factors, a framework has been proposed (Moody, Sindre, Brasethvik & Sølvberg, 2002) to evaluate the quality of process models. Notational deficiencies (e.g., semantic transparency) and their influence on the comprehension of process models have been reported by Figl, Mendling and Strembeck (2013). Regarding subjective factors, Recker and Dreiling (2007) compared two popular process modeling languages (business process model notation BPMN and event-driven process chain EPC). These studies focus on subjective aspects of PM comprehension, since they conclude that subjective factors have a greater impact than objective factors. A recent overview on studies investigating subjective as well as objective factors of PM comprehension is provided by Figl (2017).

Understanding PMs may not only be regarded as an endpoint depending on both factors described above, but also as a key competence for a multitude of cognitive tasks that share in common the classification and ordering of events and decisions into meaningful sequences (Dumas, La Rosa, Mendling & Reijers, 2013). As PMs are mostly presented as charts following specific rules of formalization in a standardized notation, it seems to be of interest to analyse the interplay between the visual inspection of charts representing PMs and their comprehension (Barthet & Hanachi, 1991; Dumas et al., 2012).

1.2. Semantic notation of PM

After a series of experiments with both subjective (i.e., cognitive load, Sweller, Ayres & Kalyuga, 2011) and objective factors (i.e., semiotic theory), Mendling, Strembeck and Recker (2012) conclude that additional semantic information impedes syntax comprehension, whereas theoretical knowledge facilitates syntax comprehension.

The study at hand tries to open up the perspective of PM comprehension from pure graphical notation to semantic notions (real-world problem descriptions versus symbolic notation) as well as to personal capacities necessary for model comprehension (psychometric measurement of competence types or levels). Recker and Dreiling (2011) also highlight the importance of understanding subjective factors to enable development of understandable PMs.

1.3. Visual literacy

Subjective factors play a key role in the understanding of PMs. It is therefore of interest to take a closer look at the ability of attentively analysing and interpreting images, an ability that is coined as Visual Literacy (VL; see Avgerinou & Petterson, 2011). From the review by Figl (2017), it becomes clear that the construct of VL has not yet been used to analyse potential interactions between subjective and objective factors with respect to model comprehension. To the best of our knowledge, with the exception of a recent study (Bačić & Fadlalla, 2016), whose authors focused more on visual *intelligence* than on *literacy*, no study has yet been published dealing with the concept of Visual Literacy and its impact on PM comprehension. This is even more astonishing considering that VL has been postulated as a basic competence underlying the precise deciphering of images (receptive component of VL), the production of such images, as well as the reflection on the constituent processes (Wagner & Schönau, 2016). Images guide our perception of the world, our preferences, and our decisions, and VL is considered a central goal of arts education (Wagner & Schönau, 2016). Whether or not a good capability of analysing, memorizing, and envisaging visual stimuli is helpful for the comprehension or production of PMs (Brumberger, 2011), has yet to be determined.

It also remains unclear whether VL can be measured like an IQ score on a continuum of homogeneous tasks representing the same, continuously distributed latent trait, best assessed by a "Rasch scale" (see (Boy, Rensink, Bertini & Fekete, 2014) for an example in the field of visualization capability). By contrast, VL might also represent a categorical model (Brill & Maribe Branch, 2007), for which different groups of people have specific gifts and talents in common, qualitatively differing from each other without the possibility of representing these differences by a single score (latent class model, see (McCutcheon, 1987)).

1.4. Eye tracking as measurement for PM comprehension

Eye tracking methods help to understand and visualize underlying cognitive processes in problem solving (Bednarik & Tukiainen, 2006). Thus, eye tracking can help to externally validate the measurement method of VL. Eye tracking has been established in the investigation of competence and competence acquisition (Jarodzka, Gruber & Holmqvist, 2017). Conclusions about strategies or procedural knowledge can be drawn by analysing the processing of visual tasks that, otherwise, could not have been verbalized or could only be partially verbalized by the subjects retrospectively (Reingold & Sheridan, 2011; Sheridan & Reingold, 2014). The underlying cognitive processes thus may be better understood (Lai et al., 2013). Eye tracking measures have provided insights into differences in experts and novices (Gegenfurtner, Lehtinen & Säljö, 2011; Vogt & Magnussen, 2007), the prediction of fluid intelligence (Laurence, Mecca, Serpa, Martin & Macedo, 2018), as well as distinguishing between strategies in spatial problem solving (Chen & Yang, 2014).

PM comprehension has been studied by means of eye tracking (Figl, 2017; Hogrebe, Gehrke & Nüttgens, 2011; Petrusel & Mendling, 2013; Zimoch, Mohring, et al., 2017, 2018), but not from the viewpoint of VL. It could be shown that subjects providing correct responses to comprehension questions after regarding a graphical model had fixated longer on relevant parts of the respective PM than on irrelevant parts (Petrusel & Mendling, 2013;Zimoch et al., 2018).

Cognitive strategies analysed by eye movements have been studied for graphically oriented intelligence tests (Hayes, Petrov & Sederberg, 2011;Vakil & Lifshitz-Zehavi, 2012). A recent study by Laurence et al. (2018) could predict from eye movement indicators approximately 45% of the variance of "Wiener Matrizen Test 2″ (Formann, Waldherr & Piswanger, 2011) test results. Toggling (gaze transition between two areas of interest) has been shown to be the most reliable measure (Laurence et al., 2018) in this context. Other typical measurements include pupillometry (Van Der Meer et al., 2010) or fixation distribution (Bucher & Schumacher, 2006; Najemnik & Geisler, 2005). Based on previous results on the analysis of matrix-based cognitive tests, the present study enhances the spectrum of visual tasks and tries to compare similar output measures for the comprehension of PMs.

To conclude, this study contributes to further analysing comprehension of PMs by using eye tracking data. Previous studies have shown that experts in their professional domain (e.g. art, medicine, chess) fixate longer on task relevant parts and shorter on task redundant parts (Gegenfurtner et al., 2011). It has yet to be determined how the comprehension of graphically presented logical models is influenced by VL.

1.5. Research goals and objectives

This study aims to apply psychometric concepts to the field of PM research. Moreover, we try to corroborate these efforts by using innovative technology (i.e., eye tracking measurements). Notably, the role of expertise in VL for solving visual tasks seems unclear, and even questionable for comprehending PMs.

Based on the previous research on process model comprehension, this paper wants to contribute empirically to the influences on process model comprehension. Methodologically, this is accomplished by means of (1) latent class analysis (LCA) and (2) eye tracking. Through LCA, we are able to determine if the answers given by students follow a homogeneous latent trait or should better be interpreted as qualitatively differing solution patterns. The use of eye tracking helps to identify potential differences in participants' understanding by analysing where and for how long subjects fixate PM aspects. Cognitive load theory (Sweller et al., 2011) interprets these measurements as indicators for cognitive workload.

In summary, three major research questions are addressed in this paper:

(1) How can the comprehension of PMs be measured in a population of students? More specifically, do answering patterns follow a homogeneous latent trait or should they be interpreted as qualitatively differing solution patterns?
(2) How do features of PMs have an impact on the general PM comprehension?
   a. Do students successfully decipher the graphical notation (e.g., logical symbols like arrows, "x" or "+")?
   b. How does the semantic notation of PMs influence the response time and the PM comprehension?
   c. What effect does the model complexity have on response time and comprehension?
(3) How does the competence level in analysing and interpreting images (VL) covary with PM comprehension?
   a. How do VL experts and novices differ in fixation duration on relevant rsp. redundant parts of the PMs?
   b. How does the expertise in VL covary with the eye movement's volatility of gaze transitions?

2. Materials and methods

2.1. Subjects

Sample I comprised 1047 high-school students from 52 classes (9th to 13th grade: 21, 28, 1, 1, 1) in 29 schools in Germany. Overall, 52.5% were female, the average age was 15.27 years (SD = 0.94). Schools were recruited in the federal states of Hessen, North-Rhine Westphalia, Schleswig-Holstein, and Rhineland Palatinate via leaflets, letters and personal recommendations. The test was conducted in regular classrooms. Up to 30 students were able to participate in the test simultaneously. In Sample I understanding PM was one segment of a longer (duration: 45 min) test on Visual Literacy. All answers were given via touchscreen input by the participants. School classes were offered a lump sum of 100€ as collective compensation.

Participants in Sample II were enrolled as experts in visual literacy ($n = 21$), if they were members of the European Network of Visual Literacy (ENViL) or working in professions requiring a high visual competence (photographer, gallerist, art educator, art designer, art students, or self-employed artists). Novices ($n = 15$) in visual literacy were adults from the clerical and academic staff of various educational settings declaring themselves as not overwhelmingly talented or familiar with arts and visual design. The age span ranged from 16 to 66 years ($M = 29.5$). All participants had normal or corrected-to-normal vision. Student participants in Sample II received 20€ each as compensation. Other participants, including the expert group, who were intrinsically interested in the topic of Visual Literacy and eye tracking, participated without further compensation.

The study was conducted according to the guidelines for human research outlined by the Declaration of Helsinki and was approved by the Ethics Committee of Research of the Leibniz Institute for Research and Information in Education (DIPF, 01JK1606A). All subjects (and their legal representatives respectively) had given written informed consent.

2.2. Materials and procedure

The assessment in both samples was conducted on Android A6 Tablets with 10.1-inch screen size. All test items were programmed specifically for the assessment tool (Andrews et al., 2018). The process models were created in BPMN 2.0 (OMG, 2011 OMG Specification, Object Management Group.). This language serves as an industry standard and constitutes the most widely used process modeling language (Allweyer, 2016).

All participants were given the identical instruction on the tablet screen: *"In the following, different processes are presented in the form of process models. A process model visualizes the sequence of events and decisions. Try to understand the process in the process model and select all correct statements (multiple statements can be correct)."*

Participants were required to inspect three subsequently presented PMs and to evaluate 4 statements based on the respective model, thereby representing a within-subject factor with three factor levels (Fig. 1). Statements were balanced for affirmation and rejection to indicate the correct response. The models were ordered in increasing complexity, where each new model included more activities (boxes) and gateways (inclusive, exclusive or parallel paths). Furthermore, in order to ensure a proper increase in process model complexity, the process models were created using the guidelines from Becker, Rosemann and Von Uthmann (2000) and the adopted cognitive complexity measure proposed in Gruhn and Laue (2006). The comprehension statements as well as the activity-labels in the respective "boxes" of each process model were randomly allocated to each subject in one of three different verbal frames, thereby representing a between-subjects factor with the following factor levels: Letters (L), Sentences (S) and Pseudo Sentences (P). This manipulation means that events in the process models as well as in the comprehension test items were either denoted with a single letter (e.g. "execute F"), a meaningful sentence describing an everyday situation (e.g. "read Facebook message"), or with a pseudo sentence (e.g. "An ecap with mistives cannot be handed over") using meaningless artificial nouns to describe the events.

For Sample II, SMI eye tracking glasses were used (SMI ETG 2w Analysis Pro). The glasses were positioned onto the subject's head, and the subjects were free to move their heads during task completion. Subjects were seated 50–80 cm away from the tablet screen. All eye tracking data were recorded at 60 Hz. Saccades and fixations (as well as blinks) were recorded binocularly and computed by the SMI event detection algorithm. Each session started with a 3-point calibration following the standard procedures for SMI iView™. The default eye movement parameters from SMI BeGaze™ version 3.7 were used. A fixation cross was displayed between each trial for 2 s. More details of the procedure and on data processing for eye tracking measurements are given in a supplementary e-appendix.

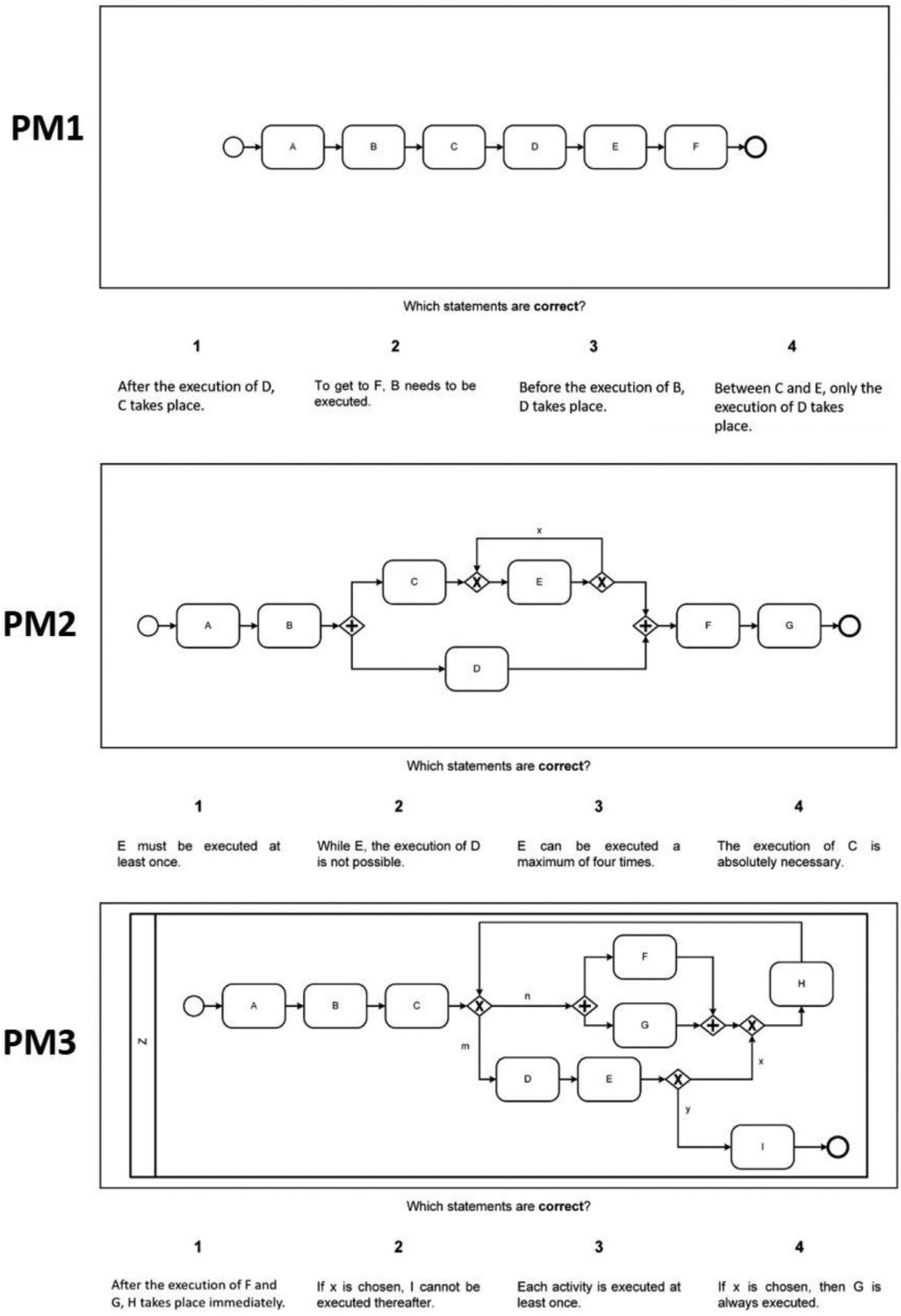

Fig. 1. Process Models (PM1, PM2, PM3) in the letter condition. PMs were presented to respondents in increasing complexity. The boxes (activities) include actions to be performed, the arrows (sequence flow) define the execution order of activities, the x (an exclusive gateway) splits the routes of the sequence flow to exactly one of the outgoing branches. The + symbolizes a parallel gateway that is used to activate all outgoing branches simultaneously.

2.3. Measurement and data analysis

The vector of 12 responses given on the tablets was transformed into 12 dichotomous items x representing each a correct judgement of the underlying verbal statement (1 = correct). The vector $\boldsymbol{x}_v$ of judgements then was analysed by latent class models (Dayton & Macready, 2006) describing typical solution patterns among the participants.

$$p(\boldsymbol{x}_v) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{k} \pi_{ixg} \quad \text{where :} \quad \sum_{g=1}^{G} \pi_g = 1 \tag{1}$$

with g:= number of latent class (*1 .. G*), x:= response chosen on item i *(1 .. k)*, $\boldsymbol{x}_v$ vector of correct judgments, π_g:= relative size

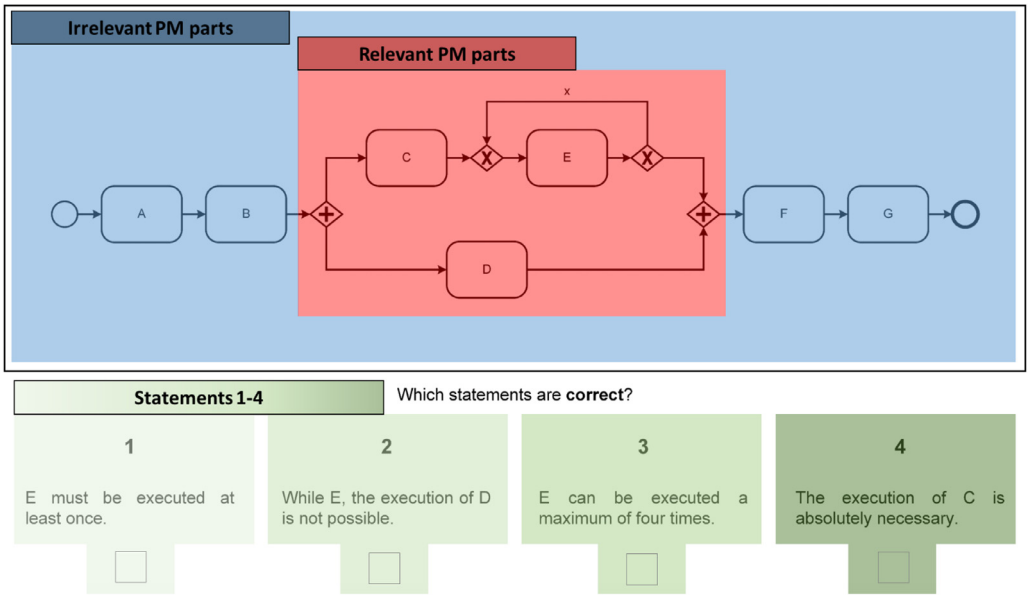

Fig. 2. AOI distribution for PM 2 (parallel paths, 1 loop) – Irrelevant PM parts (blue), relevant PM parts (red), and relevant parts of answers 1–4 (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of class g, and π_{ixg} probability of choosing response x on item i given class g.

Model parameters (π_g, $\pi_{ix|g}$) were estimated with MPLUS (6.0) software for all LCA solutions between 2 and 8 latent classes. The best number of latent classes was decided on model fit criteria (AIC, BIC) and the Vuong-Lo-Mendell-Rubin Likelihood Ratio Test, as well as the Lo-Mendell-Rubin adjusted LR test implemented in MPLUS (Asparouhov & Muthén, 2012). In order to prevent local maxima of the likelihood function of the estimated parameters, the number of initial stage random starts was set to 1000, and the number of final stage optimizations to 50 for each number of classes. The estimated model parameters (π_g, π_{ixg}) can be used to calculate membership probabilities for each participant in every latent class g in the following way (see equation 37, Rost and Langeheine (1997) p. 29).

$$p(g|x_v) = \frac{\pi_g \prod_{i=1}^{k} \pi_{ixg}}{\sum_{h=1}^{G} \pi_h \prod_{i=1}^{k} \pi_{ixh}} \quad (2)$$

Based on the modal value, each participant was classified in his/her most probable latent class. Participants from Sample II were also classified using their response patterns and the item parameters estimated from Sample I. Additional measurements in Sample II were based on the following eye tracking characteristics: a) response latency, which is the time spent on each trial in seconds, b) fixation duration on PM, which is the sum of all fixation durations on the model, c) fixation time on statements, which is the time spent on fixating the four response statements, d) number of toggles, which is the number of transitions between model and responses, and e) toggling rate, which is the number of toggles between model and responses divided by response latency. Transitions between model and responses were counted each time the subject's gaze moved from model area of interest (AOI) to any statement AOI or vice versa. Whenever the gaze would stop to fixate on regions that were not defined by any AOI ("White Space"), the transition was not counted as a toggle.

Fixations for each trial were mapped on corresponding reference images by a single rater (MT) using SMI fixation-by-fixation semantic gaze mapping. For a comparison to frame-by-frame mapping see Vansteenkiste, Cardon, Philippaerts and Lenoir (2015). Independent ratings were performed (by MW) based on complete datasets of two randomly chosen subjects. In our study we reached a high inter-rater-reliability (Cohen's Kappa > 0.94 for all PMs). Fig. 2 shows the AOIs of the second PM. Relevant parts of the graphical model (coloured in red) that were necessary for correctly accepting/rejecting a statement were a priori determined by process modeling experts from Ulm University (Zimoch, Pryss, Schobel, et al., 2017). The wording of all test items (in German) was also a result of expert discussions within the same group. All gaze data was acquired by SMI iView ETG™ software. The analyses were carried out with SMI eye tracking software "BeGaze 3.7″. Further information on the eye tracking equipment, technical settings and calibration procedure can be found in the e-appendix of this article.

Differences between PMs were analysed using repeated measurement ANOVA models for all eye movement indicators. Due to the relatively small sample size, differences between groups of respondents on the same indicators (e.g. status of expertise) were tested using univariate GLM models. In order to test significant associations between latent class membership and eye movement indicators, dummy variables for the larger groups (LC4, LC5, and LC6, see Section 3.2) were constructed. In separate models, response latency, fixation duration on redundant or relevant parts of PM2 (second model in order of appearance), fixation duration on response statements, and number of toggles between PM2 and answering statements were tested as predictors of class membership via logistic regression models. All subjects not classified into one of the three larger groups were incorporated as part of the respective reference group, against which the impact of, for example, toggles was tested to predict membership. Again, due to small sample size these calculations were performed only in univariate analyses (only one predictor) omitting multivariate relationships and interaction effects during these explorative analyses. All statistical tests beyond the experimental variation of conditions are regarded as purely explorative and therefore not subject to measures against inflation of Type-I error risk.

3. Results

3.1. Solution patterns in scholars in sample I

Both criteria (AIC and BIC) displayed substantial improvement of model fit until the introduction of a sixth latent class to be es-

Table 1
Process Model complexity and latent class parameters in Sample I.

Model complexity	Item	Test items (wording for Letter condition*)	Solution	Probability of correct solution in latent class (corresponds to π_{ixg} in formula 1)						
				LC1	LC2	LC3	LC4	LC5	LC6	Total Sample I
linear model	Q1	After the execution of D, C takes place.	reject	0.727	0.611	0.906	0.896	0.979	0.998	0.889
	Q2	To get to F, B needs to be executed.	accept	0.566	0.578	0.632	0.652	0.876	0.918	0.757
	Q3	Before the execution of B, D takes place.	reject	0.594	0.394	0.726	0.822	1	1	0.825
	Q4	Between C and E, only the execution of D takes place.	accept	0.337	0.351	0.34	0.489	0.766	0.576	0.541
parallel paths, 1 loop	Q1	E must be executed at least once.	accept	0	1	0	0	0.094	1	0.377
	Q2	While E, the execution of D is not possible.	reject	0	0.815	1	1	0	0.871	0.527
	Q3	E can be executed a maximum of four times.	reject	0.872	0.836	0	0.933	0.932	0.948	0.822
	Q4	The execution of C is absolutely necessary.	accept	0.126	0.208	0	1	0.064	0.436	0.287
linear, exclusive and inclusive gateways, 2 loops	Q1	After the execution of F and G, H takes place immediately.	accept	0.306	0.565	0.217	0.385	0.598	0.55	0.481
	Q2	If x is chosen, I cannot be executed thereafter	accept	0.244	0.323	0.396	0.341	0.597	0.548	0.456
	Q3	Each activity is executed at least once.	reject	0.448	0.497	0.604	0.644	0.78	0.608	0.630
	Q4	If x is chosen, then G is always executed.	reject	0.538	0.579	0.736	0.696	0.877	0.72	0.726

*Table 1 gives model parameters for all conditions.

timated. A seventh class resulted in deterioration of the BIC index, and no statistically significant differences could be demonstrated compared to the more parsimonious model with 6 latent classes in both the Vuong-Lo-Mendell-Rubin Likelihood Ratio Test, and the Lo-Mendell-Rubin adjusted LR test. Therefore, six latent classes were chosen as the final solution.

Table 1 gives an overview on the item parameters $\pi_{ix|g}$, which denote the probability of a correct solution in each of the six latent classes for each comprehension item.

Red-shaded cells in Table 1 depict below-average probabilities ($>$ |10%|) of solutions for the respective item in each latent class. Green-shaded cells signify above-average probabilities ($>$ 10%) of correctly solved items.

Interpretation of latent class 1 (LC1) and latent class 6 (LC6) seems straightforward: LC1 represents a group of persons with rather poor chances to solve each of the comprehension items. Members display probabilities at least 10% below the chance rates of the whole sample. This group comprised about 13% of the sample and was called “under performers”. On the contrary, LC6 consists of about 31% of the participants with excellent performance: members had no comprehension probability below sample average, but most items were solved with slightly or clearly better (green cells: $>$ 10%) probabilities than the total sample. LC6 were called “logic champions”.

LC2 (24%) closely resembles LC1 except that participants are most likely able to respond correctly to items 1 and 2 of the “parallel paths – 1 loop” model (PM2), which had zero probability in LC1. On the other hand, the group LC5 (10%) is quite similar to the largest group “logic champions” class (LC6), but it fails to recognize the correct solutions for question 1, 2 and 4 of the “parallel paths – 1 loop” model (PM2). LC2 can be labelled as “under-performers with understanding of simultaneous tasks”, and LC5 as “logically correct thinking with misinterpretation of parallel paths”.

LC3 represents a typical response pattern (12%) that is performing at an average level for all test items requiring a comparison of not more than two activities. But when 3 or more information units have to be combined for a correct solution, LC3 strongly underperforms (e.g. “After the execution of D, C takes place“ (PM1,Q1) vs. “After the execution of F and G, H takes place immediately” (PM3, Q1). Therefore they were called “binary thinking group”. Finally, the solution probabilities in LC4 (size 10%) display an excellent understanding of parallel paths (but misunderstand the “x” notation of loops), and a slightly below average comprehension of PM1 and PM3. Accordingly, this group was therefore called “multi-tasking group”.

Both the fact of numerous intersections of solution profiles in Table 1 and a formal model test of a Rasch scale (Andersen LR Test score $=$ 104.99; df $=$ 11, $p < 0.0001$) reject a homogenous latent trait as adequate psychometric model of PM comprehension, as measured by the given 12 items (see Andersen, 1973, Rost, 1988). It is therefore not meaningful to interpret the sum of correctly solved items as a simple measure to quantify a latent, continuous ability of high-school students to understand graphical models. Instead, it seems necessary to compare the interrelations of the typical comprehension patterns as qualitatively differing groups according to other variables like sociocultural background and task-relevant eye movements.

When events and decisions were presented under the “P”-condition (pseudo sentences), latent classes 3 (binary thinking group) and 4 (multi-tasking group) were more prevalent (each by 12%) than expected under the assumption of having no association between model condition and problem-solving pattern (see Table 2), while the better performing groups LC5 and LC6 were under-represented. Thus, describing processes with pseudo sentences seems to prohibit correct deciphering of more complex loop structures. When PM were presented with meaningful sentences (condition “S”), latent classes 2 (under performers with understanding of simultaneous tasks) and 5 (misinterpretation of parallel paths) were clearly over-frequented (by 15% and 26% respectively). Finally, under the condition of solely mentioning

Table 2
Number of latent class members by model condition in Sample I.

Condition		Latent class						*N* total
		1	2	3	4	5	6	
Letter (L)	Frequency	35	56	20	24	11	191	337 (32.19%)
	Row%	10.39	16.62	5.93	7.12	3.26	56.68	
	Column%	25.93	22.13	16.26	22.64	10.58	58.59	
Sentence (S)	Frequency	49	121	46	33	63	42	354 (33.81%)
	Row%	13.84	34.18	12.99	9.32	17.8	11.86	
	Column%	36.3	47.83	37.4	31.13	60.58	12.88	
Pseudo Sentence (P)	Frequency	51	76	57	49	30	93	356 (34%)
	Row%	14.33	21.35	16.01	13.76	8.43	26.12	
	Column%	37.78	30.04	46.34	46.23	28.85	28.53	
Total	**Frequency**	135	253	123	106	104	326	1047
	%	12.89	24.16	11.75	10.12	9.93	31.14	100

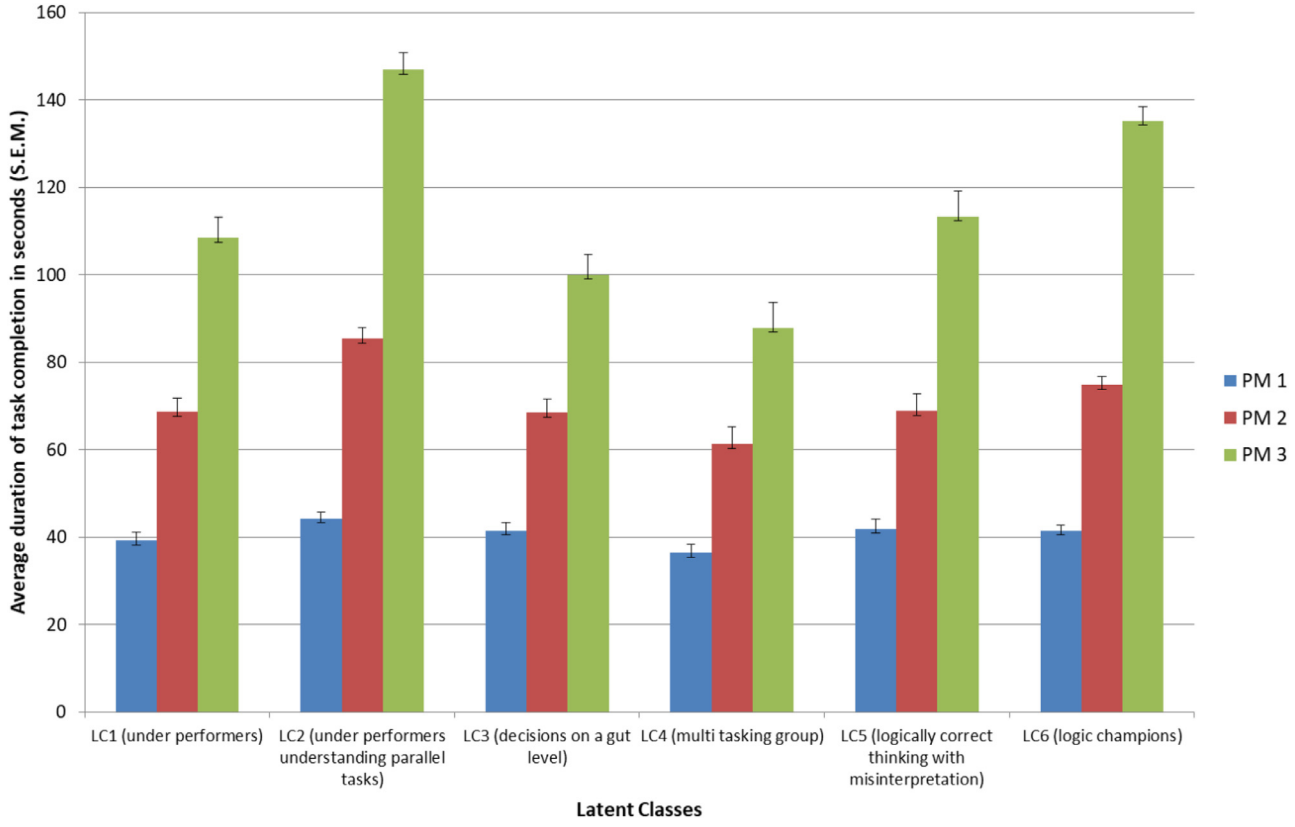

Fig. 3. Impact of increasing complexity of PMs on task completion durations (=response latencies) in Sample I.

letters for events and decisions of a PM (condition "L"), latent class 6 (logic champions) was the most prominent cognitive solution pattern, and a clear under-representation of LC3 (binary thinking) and LC5 (misinterpretation of parallel paths) was observed. Denoting PMs with only letters thus favours good task performance. These effects are statistically significant (Pearson χ^2 (d.f. 10) = 202.99; $p < 0.0001$) and can be interpreted causally, as each participant's allocation to one of the conditions was randomly chosen.

Neither age nor gender of the participants, nor parental educational background or students' self-ratings of being gifted with visual imagination could be shown to interact with class membership (results not shown here). The condition of PM presentation clearly resulted in differing durations of problem solving. Overall, task completion for the letters condition required, on average, 206.2 s (SD = 85.8) and meaningful sentences 239.2 s (SD = 82.0). In turn pseudo sentences required a mean duration of 290.7 s (SD = 149.9) before completely responding to all 12 items.

Increasing complexity of PMs required more time over all six latent classes ($F_{2,2080} = 2059.7$, $p < 0.001$) (see Fig. 3). Though differences between latent classes ($F_{5,1040} = 16.3$, $p < 0.001$) and an interaction effect of complexity*latentclass ($F_{10,2080} = 30.8$, $p < 0.001$) in the respective ANOVA model proved also significant, this is mainly due to the large sample size. Effect sizes were 0.30 (eta squared) for complexity, but only 0.06 for latent classes and 0.03 for the interaction effect.

3.2. Solution patterns and corresponding eye movement parameters in sample II

Table 3 displays descriptive statistics for the eye tracking measurements broken down by a) status of respondents' expertise, b) condition of the PM phrasing, and c) membership of the respondents in latent class.

Bonferroni-adjusted post-hoc analysis revealed a significant difference in response latency between PM1 and PM3 (−36.44 s.,

Table 3
Descriptive statistics for the eye tracking measurements in Sample II.

		Expertise status		Model condition			Membership in latent class			
	Total sample II ($N = 36$) Mean (SD)	VL Experts ($N = 21$) Mean (SD)[b]	VL Novices ($N = 15$)	Letters ($N = 14$) Mean (SD)[c]	Sentences ($N = 12$)	Pseudo ($N = 10$)	LC4 ($N = 6$) Mean (SD)	LC5 ($N = 11$)	LC6 ($N = 16$)	Other ($N = 3$)
Response latency (sec)	78.10 (33.14)	**87.07 (30.66)**	**65.55 (33.37)**	**60.92 (26.49)**	**84.34 (36.61)**	**94.66 (28.36)**	57.80 (9.25)	85.40 (37.12)	82.12 (35.62)	70.52 (29.26)
Fixation duration on models (sec)	38.15 (19.91)	41.51 (17.87)	33.44 (19.91)	27.74 (15.40)	43.82 (23.99)	45.91 (14.76)	24.10 (5.05)	41.38 (19.45)	42.30 (22.84)	32.23 (15.22)
Fixation duration on models (%)	47.54 (7.39)	46.59 (7.52)	48.87 (7.39)	44.49 (7.28)	50.29 (8.14)	48.50 (5.42)	42.00 (7.23)	47.77 (4.89)	49.75 (7.21)	46.00 (13.45)
Fixation duration on Relevant (Red)[a] (sec)	29.30 (17.86)	29.64 (13.1)	28.83 (23.48)	28.23 (15.02)	32.62 (25.65)	26.83 (9.31)	21.65 (8.67)	29.26 (12.41)	33.67 (23.38)	21.46 (10.14)
Fixation duration on Irrelevant (Blue)[a] (sec)	10.41 (8.31)	12.14 (7.64)	7.98 (8.86)	**4.01 (2.37)**	**13.70 (9.53)**	**15.42 (6.59)**	5.00 (1.87)	9.96 (8.05)	12.29 (9.57)	12.86 (7.92)
Fixation duration on statements (sec)	25.81 (9.81)	**29.71 (9.68)**	**20.34 (9.82)**	21.29 (8.55)	25.79 (9.67)	32.16 (8.84)	22.91 (5.67)	28.13 (11.14)	25.29 (10.01)	25.82 (13.32)
Fixation duration on statements (%)	33.96 (6.45)	34.98 (6.73)	32.54 (6.45)	35.22 (3.80)	32.21 (9.83)	34.31 (4.03)	39.45 (6.05)	33.47 (3.11)	31.64 (6.70)	37.20 (10.09)
PM2 fixation duration on statements (sec)	21.37 (8.93)	24.44 (8.03)	17.06 (8.55)	18.99 (8.39)	22.43 (9.86)	23.42 (8.64)	20.32 (8.13)	20.90 (8.70)	22.46 (10.16)	19.59 (7.72)
PM2 fixation duration on statements (%)	29.76 (8.89)	31.09 (8.84)	27.90 (8.93)	30.26 (7.18)	29.32 (12.74)	29.57 (5.81)	35.14 (9.19)	28.75 (6.99)	27.99 (8.39)	32.10 (16.55)
Number of toggles	19.87 (8.24)	21.76 (9.34)	17.22 (8.24)	19.83 (9.34)	18.97 (9.41)	21.00 (5.24)	13.61 (3.55)	23.15 (9.63)	20.79 (7.06)	15.44 (10.36)
Rate of toggling	0.267 (0.083)	0.253 (0.083)	0.287 (0.092)	0.333 (0.078)	0.223 (0.067)	0.228 (0.040)	0.239 (0.061)	0.295 (0.102)	0.271 (0.074)	0.20 (0.07)

As in Sample I, increasing model complexity required longer response latencies ($F_{2, 70} = 12.31$, $p < 0.001$, $\eta^2 = 0.260$). With rising complexity, the fixation duration on models rose as well ($F_{2, 70} = 31.46$, $p < 0.001$, $\eta^2 = 0.466$) and the number of toggles increased ($F_{2, 70} = 7.49$, $p = 0.001$, $\eta^2 = 0.181$).

[a] exclusively for PM2.

[b] bold font: significant $p < 0.05$ for fixation duration on statements, marginally significant $p = 0.053$ for response latency (t-test).

[c] bold font: significant $p < 0.05$ (F-test).

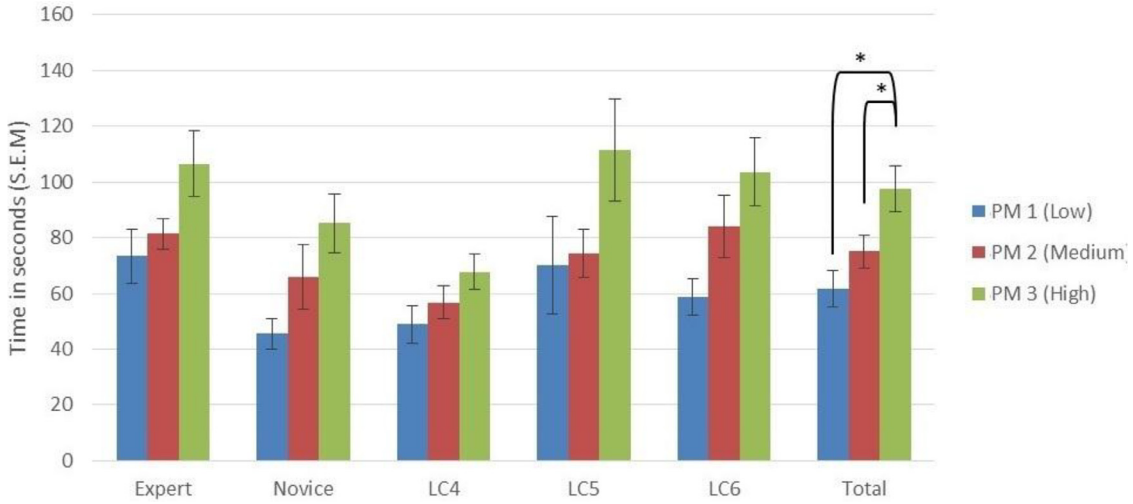

Fig. 4. Response latency in seconds (SEM) on each model by PM complexity, expertise level, and latent class membership in Sample II. [*significant on ($p < 0.05$)].

nominal $p = 0.001$) and between PM2 and PM3 (−22.11 s., nominal $p = 0.002$) (see Fig. 4). Additionally, number of toggles for PM3 was significantly higher than for PM1 (+ 7.6 toggles, nominal $p = 0.004$). Furthermore, response latency in the letter condition differed significantly from the one in the pseudo sentences condition (-33.74 s., $p < 0.05$) with an average duration being about 34 s longer in the pseudo sentences compared to the letter conditions.

No differences could be shown between VL experts and novices concerning eye movements, with the exception of fixation duration on statements, which differed significantly with VL experts spending more time on the possible responses than novices ($M_{Experts} = 29.71$ s., $M_{Novices} = 20.34$ s.; $F_{1,34} = 6.994$, $p < 0.05$, $\eta^2 = 0.171$). Also, task completion duration of VL experts tended to last longer ($p = 0.053$). VL experts tended to invest more time in arriving at any solution, but failed to outperform novices. There were no statistically significant differences between the VL experts and novices in fixation durations on relevant ($F_{1,34} = 0.017$, n.s.) or redundant model parts ($F_{1,34} = 2.274$, n.s.) of PM2,. We could also not demonstrate an association between expertise status and latent class membership (χ^2 (3, $N = 36$) = 1.870, $p = 0.600$). The number of toggles between PM2 and statements was inversely predictive for LC4 (OR = 0.785 [0.622–0.992]). Other eye tracking measurements (fixation durations on either part of the model) were not associated with membership in latent classes. Membership in latent class, model condition, and visual expertise did not interact significantly with the main effect of increasing complexity. However sta-

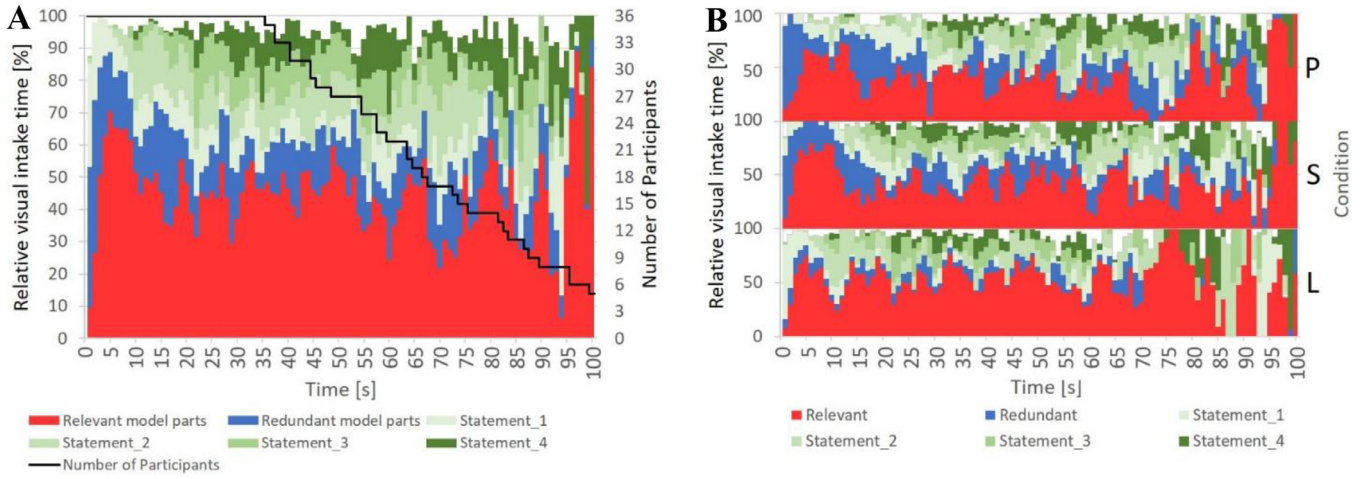

Fig. 5. A (left) and 5B (right). Histogram of AOI hit distribution for PM2 over the first 100 s (A) and by model condition (L=Letter, S=Sentences, P=Pseudo sentences) (B).

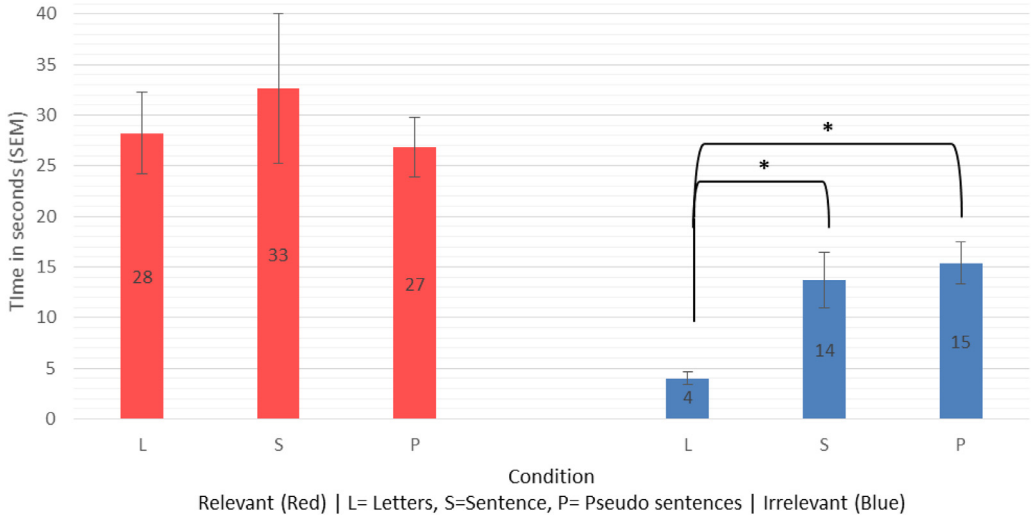

Fig. 6. Average fixation duration on relevant and irrelevant parts of PM2 by condition [*significant on ($p < 0.05$)].

tistical power is quite low for most of the variables in Table 3 (e.g. 1-β ranging from 0.069 up to 0.643 for the observed differences).

For a hypothetical "small" effect size in variable "response latencies (Cohen's $d = 0.22$), meaning that experts were on average 7 s faster than non-experts, statistical power would reach 0.16. For a medium effect size ($d = 0.40$, 14 s difference) power would reach 0.35,.and for a large effect size ($d = 0.66$, 21 s difference) power would reach 0.60.

Fig 5A and B shows the AOI hit distribution over the first 100 s of PM2. Different colours represent different AOI (see Fig. 2). As can be seen from Fig. 5A, median response latency of PM2 (right vertical axis, solid black step function) in Sample II was reached in about 66–70 s. After this time, 50% of all participants in Sample II had made their decision for PM2, only 5 participants needed longer than 100 s to respond. PM2 was chosen as an example, as it proved to differentiate between the participants' problem-solving patterns in Sample I most prominently. On average, participants directed their fixations primarily to relevant parts (red) of the model (29.3 s.; SD 17.9), which is about three times longer than the time inspecting the irrelevant parts (blue) of PM2 (10.4 s.; SD 8.3).

However, as can be seen in Fig. 5B, there were characteristic differences between the three model conditions in attention distribution as measured by fixation durations.

Further investigating the relationship between the different model conditions (L, S, P) and the time spent on fixating different (relevant/irrelevant) parts of the PMs revealed an advantage of the letter condition with respect to the redundant parts of the model: Separately analysing fixation durations by model condition (Fig. 6) indicates that the letter condition is associated with shorter fixation periods on irrelevant parts of the process model ($M = 4.01$ s., $SD = 2.37$) compared to the sentence ($M = 13.70$ s, $SD = 9.64$) and pseudo sentence ($M = 15.42$ s., $SD = 6.59$) condition ($F_{2, 33} = 10.757$ s., $p < 0.05$, $\eta^2 = 0.395$).

Fig. 7 illustrates the total time spent on the process model (= response latency, left half A) and fixation duration on each process model (right half B) as part of the total response latency.

4. Discussion

4.1. Measurement of PM comprehension: solution patterns

Six latent classes with qualitatively differing solution profiles were adequate to classify scholars in Sample I. These configurative and non-ordered profiles can be interpreted as separate solution patterns, where specific model parts are understood better than others. Beyond very good performers (LC6 "logic champions") and quite bad performers (LC1 "under performers") there exist other groups of students at intermediate "levels", which can be related to qualitatively differing errors. E.g. isolated good comprehension of simultaneous activities in process models (LC2) in front of

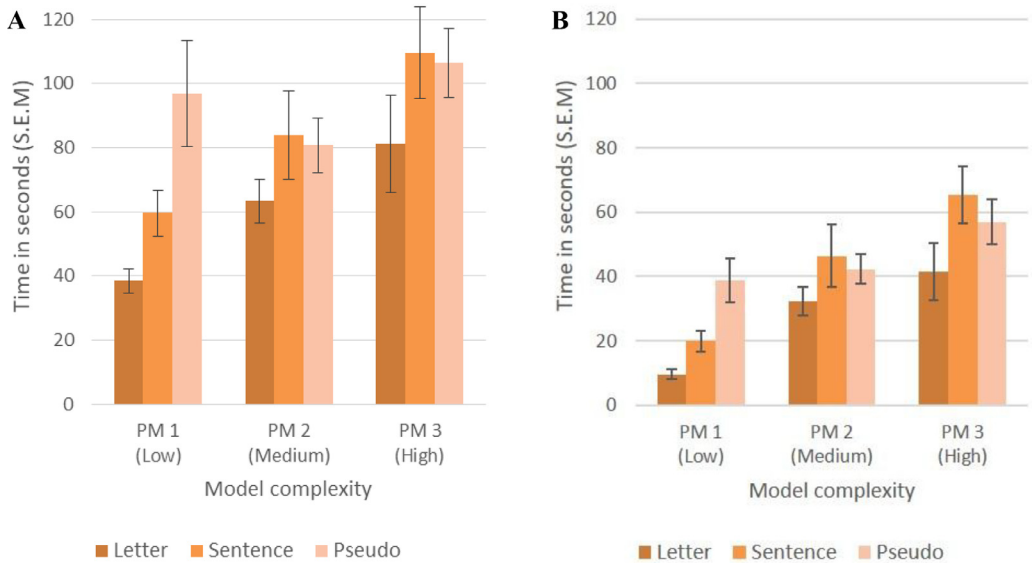

Fig. 7. A (left) and 7B (right). Bar charts of average response latencies **(A)** and average fixation duration **(B)** on each PM by model complexity and letter, sentence and pseudo sentence condition.

otherwise bad performance, or isolated lacking comprehension of parallel paths (LC5), or lacking capacity to compare more than 2 relevant facts (LC3). Participants in LC4 are best in understanding the concept of parallel pathways, but at the same time do not easily understand repeating loops.

Thus, an interpretation of the total number of correct responses would disregard important differences between different cognitive strategies mainly for "average" good participants. Given the unknown increase in cognitive workload with more complex graphical models, and given the experimentally varied wording conditions of graphs and test items, and thirdly given the differing logical problems formulated by test items, a grouping algorithm like LCA seems to be a good choice to differentiate students according to their capacity to decipher process models.

Moreover, differentiating specific comprehension errors has also a practical implication: Within educative context, it is important to know, which specific concepts and tasks are still misunderstood or are already understood in order to give meaningful feedback (Shute, 2008). Knowing which solution profile a learner applies helps to give meaningful feedback and derive adequate strategies for improvement.

In Sample II, the majority of the participants responded in a similar fashion to the profiles of LC5 ("logically correct thinking, with misinterpretation") or LC6 ("logic champions"). This better performance might partly be explained by the higher mean age and, resulting from that, the longer formal education of these participants. Nevertheless, solution patterns were only weakly connected to aspects of eye movements while working on the tasks. Only the number of toggles (gaze transitions) between the graphical model and the written statements was negatively associated with membership in LC4 ("multi-taskers"). The lower the number of toggles in PM2, the more likely the participant displayed a correct understanding of parallel pathways (even better than LC6), while failing to understand the notion of loops. In other studies a high rate of toggling was negatively correlated with intelligence scores that used visual tasks as a measurement basis (e.g. the Wiener Matrizen Test 2, see (Laurence et al., 2018)). Excessive toggling characterized a strategy to eliminate mutual contradictory responses instead of finding logical sequences within systematically ordered matrices of pictograms (Arendasy & Sommer, 2013; Bethell-Fox, Lohman & Snow, 1984). In our study, the four statements underneath each PM often addressed similar activities. In PM2 there were two statements addressing the notion of loops, which could have been weighted against each other by means of toggling (Q1: "E must be executed at least once" vs. Q3: "E can be executed a maximum of four times").

Even though LC5 and LC6 were quite different in the comprehension of PM2, other eye tracking measurements like the participants' fixation durations on either part of the model (classified into various areas of interest) were not associated with membership in latent classes. But finding no differences could be due to low statistical power.

4.2. Features impacting comprehension: PM complexity

Model complexity was handled as a within-subject factor in each condition. With increasing model complexity, the time required to respond to the comprehension questions rose. This is true for both Sample I and Sample II. Concerning eye movement indicators, the same increase could be observed for fixation duration on the models and the total number of toggles. This demonstrates that participants aspired to find the correct solutions and were not prone to click a response alternative quickly or randomly, in reaction to overly excessive demands. While we do not have comparable eye tracking data in Sample I, the participants had been asked whether they thought the test was too difficult to be solved and whether they understood the tasks. Only 25 participants (of 1047) responded in the affirmative to the former and 23 denied the latter question. Therefore, we assume a high aspiration level across both samples, which supports a preliminary interpretation of the determined latent classes as potential "cognitive styles". It should be kept in mind, that the interpretation of latent classes as "cognitive styles" is based on a purely data driven approach and should be regarded a preliminary tentative interpretation of empirical solution patterns. Further studies should focus on a convincing link between cognitive theory and solution patterns, as the latter might change with alternative operationalisations of PM complexity.

4.3. Features impacting comprehension: semantic notation

The PM conditions, i.e., whether the PM components had been labelled by letters, sentences, or pseudo sentences, were associated with a different prevalence of latent classes in Sample I

(see Table 2). They also exerted a systematic influence on some of the eye tracking variables. Contrary to our expectations, sentences representing everyday processes as naturalistic scenarios were not associated with a higher prevalence of the "logic champions" LC6, as earlier studies would have predicted (Van Merrienboer & Sweller, 2005;Sweller & Sweller, 2006). Instead, in more than half of the participants single letters as denotation generated a solution pattern of the "logic champion" type. This is in line with the finding of Mendling et al. (2012) on the impeding effects of additional semantic information on syntax comprehension.

Stimulus features nested in the PMs appear to impose a high extraneous cognitive load (Sweller, 2005) that requires working memory resources. Longer fixation duration (as measured in Sample II) can be understood as prolonged cognitive processing (Sweller et al., 2011, p. 81). Eye tracking data can indicate where and for how long the subject focuses his or her attention, implying corresponding variations on cognitive load. When splitting the model AOI into relevant and irrelevant parts, as we did with PM2 (see Fig. 2), the fixation duration on irrelevant parts was significantly shorter in the letter condition than in the sentences and pseudo sentences condition. On the other hand, fixating relevant parts of the models displayed no significant differences between conditions (see Fig. 6). The relevant parts all had about the same fixation time in all three stimulus conditions (see the percentage of red and blue in Fig. 5B).

Additional verbal workload, regardless of sentences content (pseudo or real sentences) does not increase the time needed to focus on relevant model parts; additional time is only spent on verifying irrelevant model activities. Verbal attributes seem to distract from identifying the relevant model parts, but do not increase the time needed to focus on the relevant parts of the model. One might assume that for PMs that only include letters, the fixation duration could be expected to decrease on every part of the model corresponding to less reading time. However, this is not the case here. So, what contributes to this effect?

We assume three different types of cognitive processes, which are needed to come up with a solution to the statements presented below each model. First you need to *read and understand* (A) the sentences and model activities, then *find and compare* (B) the statements with the relevant model parts, and finally *evaluate and decide* (C) whether or not the statement is correct. This follows the idea of the so-called SOI model ("Selection-Organization-Integration"), which has been elaborated for cognitive load theory in multimedia learning (Mayer, 1996; 1999). The time spent on irrelevant parts is only used for *reading and understanding* (A) as well as *finding and comparing* (B), but not for *evaluating* (C) the statements. A and B take significantly longer in the sentences and pseudo sentences condition as the structure of the sentence and the meaning of words need to be understood before it can be rejected as irrelevant. The relevant parts of PM2 include logical gateways, which were essential for answering most questions. These gateway symbols did not differ between conditions. From this point of view the fixation duration on relevant model parts should not differ between conditions, as the symbols did not change between conditions and the time spent on relevant model parts prominently included the time to *evaluate and decide* (C), whether the statement is true or false.

It might be speculated that a model, which combines letters for redundant model parts and sentences for important model parts, would be the most efficient design implementation for reducing the time spent fixating on the model as a whole. The practical implication would be that the most important information can be presented in a more natural verbal form (sentences), where other information should be presented in a short "logic-inducing" variant (e.g. letters or symbols) to keep the observant from looking at less important model parts and therefore reducing cognitive workload of *reading and understanding* (A) as well as *finding and comparing* (B). Further research needs to be conducted, in combining both elements in one process model to verify these conclusions.

4.4. Visual literacy and PM comprehension

We could not find significant differences in cognitive solution patterns between VL experts and novices. Thus, understanding and "solving" process models does not seem to depend too much on visual literacy as defined in this study. Apparently, comprehending the logic behind IF and OR gates as well as recognising pathways is crucial to follow the information flow in PMs. Even though the PMs are presented in a visual form, the ability to "interpret, analyse or appreciate visual media" does not seem to help understanding the "logical structure" of the PM. This result is useful with respect to other VL assessment items in terms of discriminatory validity. Given the small observed mean differences, it seems reasonable to hypothesize that the capacity for solving PMs does not contribute to the distinctiveness of visual literacy, which brings up an important distinction between logical models and other forms of visual information (e.g. parts/details of pictures (Vogt & Magnussen, 2007)). Regarding the eye tracking indicators, we also did not find significant differences between VL experts and novices on fixation duration between relevant and irrelevant PM parts. If VL had a substantial influence on PM comprehension, we would assume longer fixations on relevant AOIs and shorter on irrelevant AOIs, as indicated by Gegenfurtner et al. (2011). On the contrary, it seems that the search for subjective factors impacting PM comprehension (favoured by Recker and Dreiling, 2011) should not address primarily visual competence but cognitive capacities.

VL experts spent more time looking at the four statements below each model, and therefore took more time reading or thinking about the given statements. It would be interesting to see if artistic model features like colours or fonts would facilitate or distract specifically VL experts in following the logical character of PM. In further studies, longer linear models (requiring the exclusion of more nodes as "irrelevant") could help to distinguish between the workload emerging from actively omitting irrelevant facts from the workload necessary to draw logical decisions. That way the effect of verbal contribution on the distribution of cognitive load could be differentiated independently from the influence of logical gateway symbols.

4.5. Practical applications and future investigations

Eye tracking allows for a multitude of interesting experiments on analysing visual perception (Holmqvist & Andersson, 2017). Many other studies try to find differences in eye-movements between experts and novices. Experts in their field may faster distinguish relevant from irrelevant information than novices do (Gegenfurtner et al., 2011). For example, it can be shown that expert chess players are able to use their parafoveal vision (complete field of vision) to extract information that is relevant for the solution of the tasks better than novice players (Charness, Reingold, Pomplun & Stampe, 2001; Reingold, Charness, Pomplun & Stampe, 2001; Sheridan & Reingold, 2014). Higher cognitive functions like this holistic perception of a scene require perceptive as well as memory processes. Whether or not the VL experts in our study profit from their greater experience with visual stimuli or whether they were able to perceive relevant details more holistically, should not be decided on our novel setting, because the perceptive part of the visual tasks may be mantled by necessary logical reasoning.

There are implications that could lead to practical progress e.g. in teaching software engineering. The video recordings of participants gaze behaviour on target stimuli can be used as an educational tool, to show and teach novices when and where

to look at (e.g. in information retrieval from medical images; Gegenfurtner, Lehtinen, Jarodzka & Säljö, 2017). Combining eye-movement modeling examples (known as EMME) with other learning systems used for training in process model comprehension (e.g., a step-by-step assistant that teaches a complete and correct comprehension of process models) can be developed accordingly, thus enabling especially novices a better initiation to working with process models (see Jarodzka et al., 2017 for further proposals in using eye tracking in educational context).

The identification of latent classes with differing solution profiles helps to provide learners with useful feedback on adequate strategies how to improve their decisions. Assessment of visual competence might be helpful to address different target groups among apprentices while preparing specific learning materials (Andrà et al., 2009). We encourage further research on process model comprehension by means of eye tracking. Moreover, in the context of Industry 4.0, process models serve as an enabler for automatization. Because process models used for this purpose often are very complex and thus hard to read and comprehend, the methodology introduced by this article might contribute to enable further studies with high relevance for the field of organizational research (Meißner & Oll, 2017).

4.6. Limiting factors

Some limiting factors of our study need to be addressed. (1) We assume the same latent classification from Sample I (high-school students) to be present in Sample II (VL expert and novice group). However, it is possible that through age differences and recruitment outside a classroom context, different underlying classifications might be more appropriate. (2) When looking at AOIs from a narrow and dynamic visual angle, the risk for error prone AOI-fixation detection increases (Orquin & Holmqvist, 2018). Our AOIs were therefore drawn more conservatively (larger) and included multiple activities and pathways to compensate for eye tracking inaccuracy. Using remote devices with constant lightning conditions and steady head position (minimizing Pupil foreshortening effect, Hayes & Petrov, 2016) in future studies could avoid this imprecision and also allow pupillometric analyses. (3) The generalizability of the typology of cognitive solution patterns to other PMs is difficult, if not impossible, due to the different features of the PM that we used to operationalize complexity. Increasing model complexity was based on the guidelines from Becker et al. (2000): PM1 was constructed as a linear model, PM2 had one prominent parallel pathway and one loop, and PM3 had multiple inclusive and exclusive pathways in combination with a higher number of total activities. Whether this selection of demand characteristics is representative for the whole universe of possible model complexities cannot be decided from our data.

(4) Potential effects of various statistical aspects of our study: sequence of model presentation and/or of comprehension test items cannot be excluded due to their uniform ordering corresponding to their complexity. (5) The selection of valid eye tracking indicators: At this point, we could not deduce a single variable as major study endpoint because of lacking theoretical foundation, and also could not construct a combined scaled measure of the correlated variables in use due to the limited sample size of study II. (6) The semantic language structure of the four statements presented below each PM was also not varied systematically: PM1 only included questions regarding sequence (e.g. A follows B), PM2 included questions regarding sequence, conditional activities and loops, and PM3 included questions on sequence, on conditional activities as well as a statement on all activities in the model (PM3, Q3). Therefore, it is difficult to identify a specific model feature or statement as exerting the main influence on the solution patterns. Future studies should systematically vary the cognitive workload that results from the logical structure, labels or comprehension statements.

(7) Finally, if done in more detail, the latent structure of solution patterns could be analysed using more sophisticated psychometric models than a “simple” latent class analysis. Though LCA seems appropriate for the comparisons in this study, it is conceivable that different subgroups of high-school students (or adults) share different PM features for comprehension. Mixed Rasch Models or so called “hybrid models” (Rost & Langeheine, 1997) may be applied to test these patterns of responses in PM comprehension tasks. As in research on intelligence, one could also speculate on the existence of second order abilities (dominated from the subject’s characteristics) and first order task-specific latent classes.

5. Conclusion

To conclude, the present study demonstrates an association between problem solving behaviour as measured through eye tracking and the comprehension of PMs. Specific solution patterns could be revealed, depending on the structure and complexity of PMs. The condition of how PMs are presented (i.e., letter, sentence, or pseudo-sentences) displayed significant influence on the answering patterns and the time spent on each model. PMs cannot be interpreted solely based on their graphical nature, but their semantic structure plays an important role for their comprehension as well. Specifically, the use of single letters for model activities resulted in a faster and more precise understanding of the models. Experts in VL could not be shown to outperform novices with respect to PM comprehension. It seems worthwhile to focus on the cognitive mechanisms and less on visual competence of subjects when assessing their PM comprehension.

From a methodological point of view, eye tracking demonstrated a fruitful path into analysing the comprehension of graphical logical models like PMs. Fixation duration on different parts of a model enabled scrutinizing effects of verbal model features on attention distribution and cognitive workload. In future studies, relevant and/or difficult to comprehend parts in a process model may be extended with other visual features for effective guidance through a PM. Due to the restricted variation of characteristics of (Business) PMs, further research needs to include a wider range of model formulations.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be interpreted as a potential conflict of interest.

Author contributions

UF and KR were responsible for the conceptualization and aquired funding as well as project administration. MW, MR, and RP contributed to methodology (designed PMs) MT and MG were responsible for methodology of the eye tracking.. MW and MT wrote the software, under supervision of MR and RP, and implemented it on the tablets. MT led the investigation (field work) and data curation. MT and MG performed the formal analysis of the eye tracking measurements. MT and UF performed the formal analysis of Study I and Study II (all psychometric and other statistical analyses). MT, UF and KR wrote the original draft of the manuscript. All authors validated the manuscript, reviewed and edited it critically. All authors listed thus have made a substantial, direct and intellectual contribution to the work, and approved it for publication. We would also like to thank Kenneth Holmqvist for his helpful comments on an earlier version of this paper.

Funding

This work was supported by the German Ministry of Education and Science (grant number: 01JK1606A).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2019.06.032.

References

Aguilar-Saven, R. S. (2004). Business process modelling: Review and framework. *International Journal of Production Economics, 90*(2), 129–149.
Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123–140.
Andrà, C., Arzarello, F., Ferrara, F., Holmqvist, K., Lindström, P., Robutti, O., et al. (2009). How students read mathematical representations: An eye tracking study. In *Proceedings of the 33rd conference of the international group for the psychology of mathematics education: 2* (pp. 49–56). ThessalonikiGreec: PME.
Andrews, K., Zimoch, M., Reichert, M., Tallon, M., Frick, U., & Pryss, R. (2018). A smart mobile assessment tool for collecting data in large-scale educational studies. *Procedia Computer Science, 134*, 67–74.
Arendasy, M. E., & Sommer, M. (2013). Reducing response elimination strategies enhances the construct validity of figural matrices. *Intelligence, 41*, 234–243.
Asparouhov, T., & Muthén, B. (2012). Using Mplus TECH11 and TECH14 to test the number of latent classes. *Mplus Web Notes, 14*, 22.
Avgerinou, M. D., & Pettersson, R. (2011). Toward a cohesive theory of visual literacy. *Journal of Visual Literacy, 30*, 1–19.
Bačić, D., & Fadlalla, A. (2016). Business information visualization intellectual contributions: An integrative framework of visualization capabilities and dimensions of visual intelligence. *Decision Support Systems, 89*, 77–86.
Barthet, M. F., & Hanachi, C. (1991). What kind of interface for expert systems? *Expert Systems with Applications, 2*(2–3), 195–200.
Becker, J., Rosemann, M., & Von Uthmann, C. (2000). Guidelines of business process modeling. In *Business process management* (pp. 30–49). Berlin, Heidelberg: Springer.
Bednarik, R., & Tukiainen, M. (2006). An eye-tracking methodology for characterizing program comprehension processes. In K.-J. Räihä, & A. T. Duchowski (Eds.), *Eye tracking research & applications. ETRA '06*. San Diego, CA: ACM.
Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8*, 205–238.
Bharathi, S., Chervenak, A., Deelman, E., Mehta, G., Su, M. H., & Vahi, K. (2008). Characterization of scientific workflows. In *2008 third workshop on workflows in support of large-scale science* (pp. 1–10). IEEE.
Boy, J., Rensink, R. A., Bertini, E., & Fekete, J. (2014). A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics, 20*, 1963–1972.
Brill, J. M., & Maribe Branch, R. (2007). Visual literacy defined – the results of a Delphi study: Can IVLA (operationally) define visual literacy? *Journal of Visual Literacy, 27*, 47–60.
Brumberger, E. (2011). Visual literacy and the digital Native: An examination of the millennial learner. *Journal of Visual Literacy, 30*, 19–47.
Bucher, H.-J., & Schumacher, P. (2006). The relevance of attention for selecting news content. An eye-tracking study on attention patterns in the reception of print and online media. *Communications, 31*, 347–368.
Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & Cognition, 29*(8), 1146–1152.
Chen, Y.-C., & Yang, F.-Y. (2014). Probing the relationship between process of spatial problems solving and science learning: An eye tracking approach. *International Journal of Science and Mathematics Education, 12*, 579–603.
Dayton, C. M., & Macready, G. B. (2006). Latent class analysis in psychometrics. In C. R. Rao, & S. Sinharay. (Eds.), *Handbook of statistics* (pp. 421–446). New York: Elsevier.
Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2013). *Fundamentals of business process management*. Heidelberg: Springer.
Dumas, M., La Rosa, M., Mendling, J., Mäesalu, R., Reijers, H. A., & Semenenko, N. (2012). Understanding business process models: The costs and benefits of structuredness. In J. Ralyté, X. Franch, S. Brinkkemper, & S. Wrycza (Eds.), *International conference on advanced information systems engineering*. Gdansk (PL): Springer.
Figl, K. (2017). Comprehension of procedural visual business process models. *Business & Information Systems Engineering, 59*, 41–67.
Figl, K., Mendling, J., & Strembeck, M. (2013). The influence of notational deficiencies on process model comprehension. *Journal of the Association for Information Systems, 14*(6), 312.
Formann, A. K., Waldherr, K., & Piswanger, K. (2011). *Wiener Matrizen-Test 2: WMT-2. Ein Rasch-skalierter sprachfreier kurztest zu erfassung der intelligenz*. Hogrefe.
Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*, 523–552.
Gegenfurtner, A., Lehtinen, E., Jarodzka, H., & Säljö, R. (2017). Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis. *Computers & Education, 113*, 212–225.
Gruhn, V., & Laue, R. (2006). Adopting the cognitive complexity measure for business process models. In *2006 5th IEEE international conference on cognitive informatics: 1* (pp. 236–241). IEEE.
Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods, 48*(2), 510–527.
Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision, 11* 10-10.
Hogrebe, F., Gehrke, N., & Nüttgens, M. (2011). Eye tracking experiments in business process Modeling: Agenda setting and proof of Concept. In EMISA (pp.). In M. Nüttgens, O. Thomas, & B. Weber (Eds.), *Enterprise modelling and information systems architectures (EMISA 2011)*. Hamburg: Gesellschaft für Informatik e.V.
Holmqvist, K., & Andersson, R. (2017). *Eye tracking: A comprehensive guide to methods, paradigms, and measures*. Lund, Sweden: Lund Eye-Tracking Research Institute.
Jarodzka, H., Gruber, H., & Holmqvist, K. (2017). Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research, 10*(1), 1–18.
Lai, M. L., Tsai, M. J., Yang, F. Y., Hsu, C. Y., Liu, T. C., Lee, S. W. Y., et al. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review, 10*, 90–115.
Laurence, P. G., Mecca, T. P., Serpa, A., Martin, R., & Macedo, E. C. (2018). Eye movements and cognitive strategy in a fluid intelligence test: Item type analysis. *Frontiers in Psychology, 9*, 380.
Mayer, R. E. (1996). Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. *Educational Psychology Review, 8*, 357–371.
Mayer, R. E. (1999). Designing instruction for constructivist learning. *Instructional-design theories and models. A New Paradigm of Instructional Theory, 2*, 141–159.
Mccutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, London, New Delhi: Sage.
Meißner, M., & Oll, J. (2017). The promise of eye-tracking methodology in organizational research: A taxonomy, review, and future avenues. *Organizational Research Methods, 8*, 1094428117744882.
Mendling, J., Strembeck, M., & Recker, J. (2012). Factors of process model comprehension—Findings from a series of experiments. *Decision Support Systems, 53*(1), 195–206.
Moody, D. L., Sindre, G., Brasethvik, T., & Sølvberg, A. (2002). Evaluating the quality of process models: Empirical testing of a quality framework. In *International conference on conceptual modeling* (pp. 380–396). Berlin, Heidelberg: Springer.
Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*, 387.
Omg, O. M. G. (2011). OMG specification, object management Group.). *Business Process Model and Notation (BPMN) version 2.0*. OMG Group [Online] https://www.omg.org/spec/BPMN/2.0/ Accessed November 2018.
Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavior Research Methods, 50*(4), 1645–1656.
Petrusel, R., & Mendling, J. (2013). Eye-tracking the factors of process model comprehension tasks. In C. Salinesi, M. C. Norrie, & Ó. Pastor (Eds.), *International conference on advanced information systems engineering CAiSE 2013*. Springer.
Recker, J. C., & Dreiling, A. (2011). The effects of content presentation format and user characteristics on novice developers' understanding of process models. *Communications of the Association for Information Systems, 28*(6), 65–84.
Recker, J. C., & Dreiling, A. (2007). Does it matter which process modelling language we teach or use? An experimental study on understanding process modelling languages without formal education. ACIS 2007 proceedings, 45.
Reggio, G., Ricca, F., Scanniello, G., Di Cerbo, F., & Dodero, G. (2015). On the comprehension of workflows modeled with a precise style: Results from a family of controlled experiments. *Software & Systems Modeling, 14*(4), 1481–1504.
Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science, 12*(1), 48–55.
Reingold, E. M., & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. *Oxford handbook on eye movements*: 528.
Roehm, T., Tiarks, R., Koschke, R., & Maalej, W. (2012). How do professional developers comprehend software? In *Proceedings of the 34th international conference on software engineering* (pp. 255–265). IEEE Press.
Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). Process mining in healthcare: A literature review. *Journal of Biomedical Informatics, 61*, 224–236.
Rost, J. (1988). Test theory with qualitative and quantitative latent variables. In *Latent trait and latent class models* (pp. 147–171). Boston, MA: Springer.
Rost, J., & Langeheine, R. (1997). A guide through latent structure models for categorical data. *Applications of Latent Trait and Latent Class Models in the Social Sciences*, 13–37.
Schultheiss, L. A., & Heiliger, E. M. (1963). Techniques of flow-charting. *Clinic on Library Applications of Data Processing (1st: 1963)*.
Sheridan, H., & Reingold, E. M. (2014). Expert vs. novice differences in the detection of relevant information during a chess game: Evidence from eye movements. *Frontiers in Psychology, 5*, 941.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer. (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19–30). New York, NY: Cambridge University Press.
Sweller, J., & Sweller, S. (2006). Natural information processing systems. *Evolutionary Psychology, 4*, 147470490600400135.
Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In *Cognitive load theory* (pp. 71–85). New York, NY: Springer.
Ungan, M. (2006). Towards a better understanding of process documentation. *The TQM Magazine, 18*(4), 400–409.
Vakil, E., & Lifshitz-Zehavi, H. (2012). Solving the Raven Progressive Matrices by adults with intellectual disability with/without down syndrome: Different cognitive patterns as indicated by eye-movements. *Research in Developmental Disabilities, 33*, 645–654.
Van Der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., et al. (2010). Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology, 47*, 158–169.
Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*, 147–177.
Vansteenkiste, P., Cardon, G., Philippaerts, R., & Lenoir, M. (2015). Measuring dwell time percentage from head-mounted eye-tracking data–comparison of a frame-by-frame and a fixation-by-fixation analysis. *Ergonomics, 58*, 712–721.
Vogt, S., & Magnussen, S. (2007). Expertise in pictorial perception: Eye-movement patterns and visual memory in artists and laymen. *Perception, 36*, 91–100.
Wagner, E., & Schönau, D. (2016). *Cadre européen commun de référence pour la visual literacy-prototype – Common European framework of reference for visual literacy-prototype - Gemeinsamer Europäischer referenzrahmen für visual literacy-prototyp*. Waxmann Verlag.
Zimoch, M., Mohring, T., Pryss, R., Probst, T., Schlee, W., & Reichert, M. (2017). Using insights from cognitive neuroscience to investigate the effects of event-driven process chains on process model comprehension. In *International conference on business process management* (pp. 446–459). Springer.
Zimoch, M., Pryss, R., Layher, G., Neumann, H., Probst, T., Schlee, W., et al. (2018). Utilizing the capabilities offered by eye-tracking to foster Novices' comprehension of business process Models. , Cham. In X. Jing, M. Zhi-Hong, S. Toyotaro, & Z. Liang-Jie (Eds.), *International conference on cognitive computing. ICCC 2018* (pp. 155–163). Cham: Springer.
Zimoch, M., Pryss, R., Probst, T., Schlee, W., & Reichert, M. (2017). Cognitive insights into business process model comprehension: Preliminary results for experienced and inexperienced individuals. In *Enterprise, business-process and information systems modeling* (pp. 137–152). Cham: Springer.
Zimoch, M., Pryss, R., Schobel, J., & Reichert, M. (2017). Eye tracking experiments on process model comprehension: Lessons learned. In *Enterprise, business-process and information systems modeling* (pp. 153–168). Cham: Springer.