Conceptual Design for an Eye-Tracking Experiment on Formula Linebreaking

Andrea Kohlhase[1], Michael Kohlhase[2]

[1]*Information Management, University of Applied Sciences Neu-Ulm*
[1]*Computer Science, FAU Erlangen-Nürnberg*

Abstract
Traditionally, technical documents have been designed for print delivery in letter, A4, or similar sizes. Even the change to digital delivery using PDF has not changed the basic layout strategy and desktop screens can cope well. With the advent of mobile connected devices, it becomes natural to read technical documents (like everything else) e.g. on smartphones, which may demand other layout tradeoffs.

The document components most affected by this are diagrams and formulae, which – unlike text – cannot simply be reflowed to a new screen size. In this paper, we discuss an experimental study design that helps the investigation of the effect of linebreaking in mathematical formulae for reading efficiency using eye-tracking experiments.

Keywords
tbd

1. Introduction

In the age of "mobile first", how should we show technical documents to readers using smartphones? The standard reflex – "let's ask our users" does not work.

For instance, to obtain information about linebreaking in formulae we showed Figure 1 and asked *"Which one do you like better?"* Almost all test subjects chose the right one. Why? Because the font size was much bigger and thus presumably more readable. But when asked *"And if you want to decide whether the calculation is correct?"*, the answer often flipped. Why? Because then they wanted to have a better overview. Obviously, there is a tradeoff between font

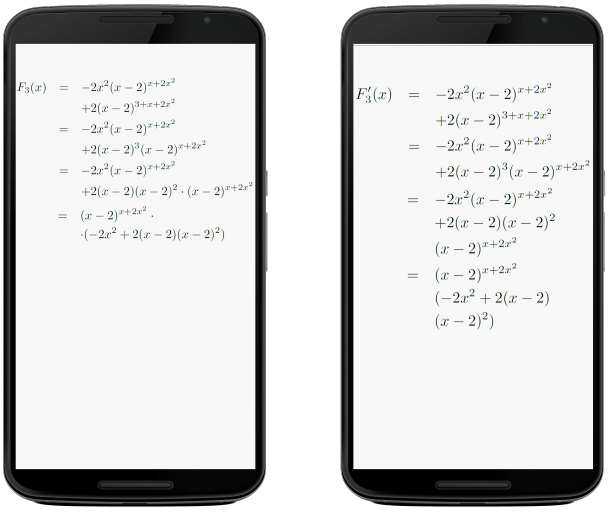


Figure 1: Two variants of a formula on a smartphone

MathUI 2021
andrea.kohlhase@hnu.de (A. Kohlhase); michael.kohlhase@fau.de (M. Kohlhase)
https://kwarc.info/kohlhase (M. Kohlhase)
0000-0001-5384-6702 (A. Kohlhase); 0000-0002-9859-6337 (M. Kohlhase)

size and overview, and in the extremes – tiny font size or extremely fragmented layout – legibility and readability suffer.

But can we do better? What are the relevant parameters/causes/effects?

Related Work Traditionally, formula linebreaking has been a task for scientific copy editors and experienced copy-editors and typesetters who were led by their experience and aesthetic intuitions. The introduction of TEX/LATEX, in the 1980s put typesetting, formula layout, and linebreaking into the hands of the authors and dedicated copy-editing of formulae has all but disappeared. This led to the development of explicit "rule books" for formula layout and linebreaking – see e.g. [Swa] Sections 3.2 to 3.4 – and LATEX packages that automate some of this. The `breqn` package is the most advanced example; see also Section 14 of [DHR] for a linebreaking "rule book" facilitated by the `breqn` infrastructure. In a nutshell, these rules give a set of constraints on linebreaking loci – and indentation of the subsequent line – that intend to make decoding the structure and meaning of formulae no more difficult than in the unbroken case.

Note that all of the above target paper or digital print media – usually via PDF nowadays – which have a paginated layout determined and fixed during typesetting. Interactive media with flexible page/screen sizes need to move page rendering (and thus formula layout and linebreaking) from the editing workflow to the display time, which calls for a much higher level of automation and makes hand-tweaking of layouts impossible because they are too brittle. The main representatives for interactive media for technical documents are web pages, web applications, and electronic books, all of which use some variant of HTML5 as the representation format and images, TEX/LATEX (via MathJax) or MathML for formulae. But

1. images do not allow re-layouting by nature,
2. MathJax [Mat] inherits fixed linebreaking from TEX/LATEX[1], and
3. the MathML3 Recommendation [MML310] specifies attributes for automated and manual line breaking, and sketches an algorithm for automated formula linebreaking based on minimizing a "penalty" computed from various factors; but current browsers do not implement it (yet).

While there is an established set of best practices for linebreaking in mathematics and a set of mathematic/semantic intuitions why these practices might be "best", there have not been any scientific investigations into the cognitive effects of formula linebreaking onto reading efficiency and effectiveness.

The main mechanism underlying the "best linebreaking practices" and algorithms seems to be that if we consider a formula as an operator tree (which encodes the meaning of the formula), then line breaks should be placed as high up in the tree as possible, so that the normal layout of subformulae corresponding to the subtrees are kept intact and thus intelligible. Indentation can be used to visualize nesting levels in the operator tree and to align subformulae corresponding to sibling subtrees, this is a form of **semantic indentation**.

This "semantics first" strategy is consistent with our findings in [KKF17], which describes formula understanding as a recursive process of establishing a gestalt tree and proceeding

[1] MathJax lists automated linebreaking as "high on the list for inclusion in a future release", but has not implemented it.

along the operator tree. A **gestalt** is a cognitive template that holistically combines layout and operator information. We conjectured that the acquisition of a suitable set of gestalts is an important aspect of acquiring mathematical literacy in a particular domain. Indeed if that is true, then the best linebreaking and indentation practices can be seen as the practices of not disturbing the gestalt of the subformulae.

Contribution In this paper, we want to discuss and outline the design principles of an eye-tracking experiment that concentrates on the effects of distinct layout properties of formulae on the formula reading efficiency. The requirements for such an experimental design balance the influence between a nominal task, that ensures attention and effect-neutrality of the test subjects, with the effects to be studied.

Overview Section 2 discusses the experimental setup considering the basic eye-tracking requirements in 2.1, the requirements on the experiment due to human factors in 2.2, the layout properties of formulae in 2.3, and the best common eye-tracker metrics for this task in 2.4. Section 3 summarizes the experimental set-up and concludes the paper with an outlook.

2. The Conceptual Design

As there is a demonstrable correlation between what a participant attends to and where she is looking at – see for example [Ray98] for an overview, the eye-tracking methodology is an interesting angle of attack. **Eye-tracking**, i.e., the observation of eye movements, allows to get a better understanding of visual attention. The "eye-mind hypothesis" [HWH99] even claims a correlation between the cognitive processing of information and the person's gaze at the specific location of the information. Therefore, it is sensible to look into the trade-off between properties like font size, number of required lines and indentation after linebreaks in formulae by setting up an eye-tracking experiment. In the following we will discuss how such an experiment could and should not be set up based on our experiences in previous experiments.

2.1. Basic Eye-Tracking Set-Up

Our goal is to compare reading efficiency across several linebreaking variants H_i of a mathematical expression.

As the screen area of a mobile phone is rather small and its position in real use rather flexible – this hampers the use of eye-tracking equipment – we suggest to use a normal computer screen showing images of a mobile phone containing these H_i variants to the user. This way the eye-tracker has a much better chance to collect *valid gaze data*.

In a typical eye-tracking experiment each participant is introduced to a *scenario* and is given a specific *task* to achieve. The scenario should be as plausible as possible, that is, a description of a familiar situation, and the task should be natural in that scenario. To make the data gathered in the experiment comparable, each of the H_i should differ from the others in one (critical) aspect.

Unfortunately, that is more difficult than it seems at first glance, as

- linebreaks seem only natural if there is a meaningful purpose for the break, or there is no space left on the right hand side of the mathematical expression,
- this depends on the chosen font-size, which in turn
- depends on the selected vertical or horizontal use of the phone's screen.

So, a mathematical expression has to be found that has sensible and distinct linebreak loci and allows for isomorphic variants for a given task in a natural scenario. The screen orientation can easily be fixed on the computer screen, but the font-size corresponds to the credibility of the linebreak: if the font-size is big, then linebreaks are in play, but if the font-size is small, then linebreaks are close to superfluous. We are also interested in the participants' behavior, if no linebreaks were present, so one variant H_0 without linebreaks should be included.

To create suitable H_i under the above conditions, several aspects with respect to human factors have to be taken into account.

2.2. Human Requirements

For an eye-tracking study as envisioned above we need (*longish*) mathematical expressions where linebreaks make sense. Also, they have to be *interesting* enough (given a specific task) that participants have to be motivated to look at those closely and not only skim them superficially. Moreover, the mathematical expressions used with different representations in terms of linebreaking have to be basically *display-equivalent* to be able to sensibly compare the collected gaze data on them.

In a previous experiment concerned with linebreaking, we had decided on a task with a function expression $\mathcal{F}$ in two variables consisting of a sum $\sum_{i=0}^{1}$ or a product $\prod_{i=1}^{2}$ over simple arithmetic expressions with fractions, products, and (simple) summations in these variables. The participants were supposed to recursively calculate points like $\mathcal{F}(0,1)$. Even though – in the end – most of the terms in the sum vanished, this experiment failed due to cognitive overload on the part of our participants: To ensure that they read the presented formulae carefully while we gathered gaze data, we asked them to do this computation without external tools like pen and paper. Doing so, we gathered a lot of gaze data, but – because of all the restarts due to short term memory failures – the recorded data were much too complex to conclude any hypotheses. In other words, the computational load induced was so large, that it drowned out the signal – the influence of the layout – we were looking for. What if we had provided pen and paper in this experiment? Unfortunately, then the gaze data would have been biased as predictably most of the eye-movements would have taken place on the paper not on our display.

Therefore not only the task given to participants needs to be simple (e.g., using just a little bit of mental arithmetics) but the mathematical expression itself needs to be simple enough to be able to focus on it and to contain sensible linebreaks. A good choice of a mathematical expression that satisfies these conditions seems to be an *equation system*, as it is often presented in mathematical documents with linebreaks and a lot of unattended empty space.

The equations itself need to contain simple math, so that a task can be created that does not lead to cognitive overload. For example, it can consist of rather *simple polynomials*. Then simplifications in these equations can for instance involve expanding binomial identities, summing up terms and integer-multiplication, each of which creates comparable cognitive actions even if

not identical expressions are used.

The variation of coefficients, signs or literals in each equation system H_i is necessary to keep the participants from noticing the structural invariants just described. Moreover, the participants' short term memory of achieving tasks with former H_i can bias its achievement in latter H_i. Observations in [MP14] suggest that additions and subtractions should be considered as different processes with respect to spatial-attentional processing.

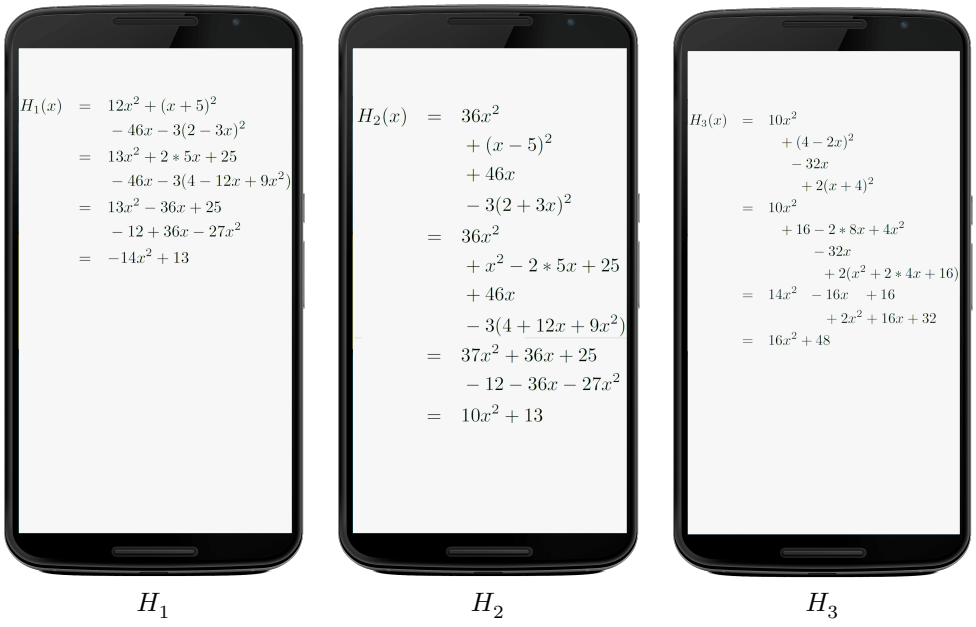


Figure 2: The H-Series of Distinct Layouts

Consider the "series" of equation systems H_i as shown in Figure 2 as an example. It consists of isomorphic laddered[2] **equation systems** in three linebreaking variants:

V1 **Simple Break**: H_1 (on the left of) breaks after half of the summands of the right equation,

V2 **Terms Straight**: H_2 (Figure 2 middle) uses a separate line for every one of the initial summands and keep that linebreaking for the results or computing with them, and

V3 **Terms Step**: H_3 (Figure 2 right) varies that by indenting subsequent lines semantically – called "step layout" in the `breqn` package.

So we have now three equation systems H_i with three equations H^j, where the equation variants H_i^j (representing the j^{th} equation in the i^{th} equation system) differ in terms of linebreak and font size, but are semantically isomorphic.

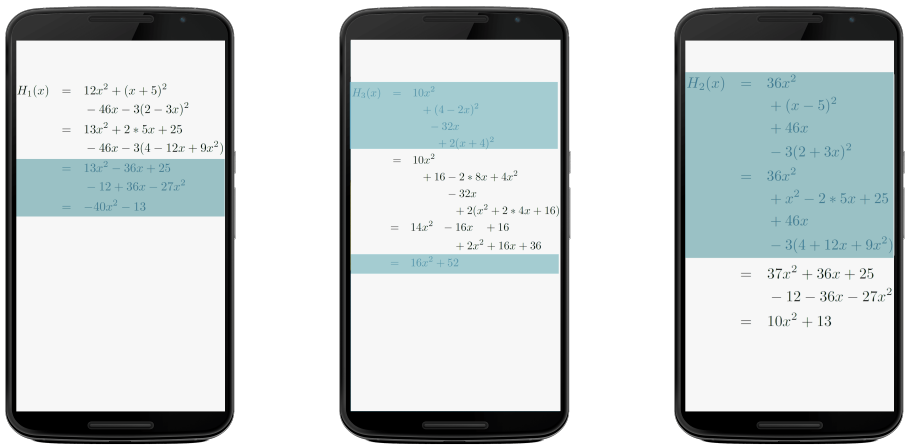
[2]We adopt the nomenclature of the `breqn` package that calls an equation system **laddered**, iff it is layed out as a three-column array with the left hand side on the first line of the first column, the equation operands in the second, and the subsequent equands – i.e., the arguments of equality in the equation system – in the third column.

A natural and cognitively rather undemanding task could then be the assessment of correctness of such an equation: in such a scenario the participants imagine themselves to be *graders*. Note that strictly speaking the nominal task of assessment only measures the "grading efficiency". However, we posit that for mathematical texts and formulae, reading, understanding, and assessing the correctness are equivalent: none of them can be done without the others.

To keep up the pretext of correctness checking and to encourage our test subjects to look closely at the three equations in each equation system H_1, H_2, and H_3, we include small *calculation errors* into (some of) the equations. Whenever we ask participants to decide whether the equation presented was correct or contained an error, there is a potential for biasing the data. For example, in an earlier experiment we observed that some participants took the nominal task very seriously and spent considerable time to finish - yielding compromised AOI values. A first solution could consist of asking the participants to be as fast as possible. Ambitious participants are not stopped by this in our experience. Therefore the provision of an *automatic cut-off* of the presentation of every equation variant H_i^j after a suitable threshold is suggested. This set-up was accepted by the participants in a previous study as they were asked in the experiment scenario to imagine themselves to be "under very hard time pressure".

Participants will try to make sense out of the experiment and therefore they are keen to solve this mystery. Therefore, if we were to show them just the variants H_i, then they still can easily deduce the linebreak variants in the math expression. To obfuscate these, we suggest to intersperse the math expressions relevant to the experiment by *others with random linebreaking behavior*.

We found out in a previous experiment that participants had to get used to the assessment task. In particular, they had a different pattern of looking at the first equation system than on the following ones. Therefore *the very first equation system* to be shown to participants should not be one of the H_i variants.



H_1^1 H_3^2 H_2^3

Figure 3: Exemplary Masked Equation Variants

To focus the attention of the participants on each equation, we mask all but one in blue (see

Figure 3). In Figure 3, for instance, we see the focus on the equation variants H_1^1 (1st equation in 1st layout), H_3^2 (2nd equation in 3rd layout), and H_2^3 (3rd equation in 2nd layout).

Another issue we also learned the hard way from a previous experiment. Our version H_0 of the H-series in Figure 2, that did *not* have linebreaks on the right hand side of the equation, had to use a very small font-size to fit the screen (see Figure 4). It was designed to be isomorphic to the equation systems of the H-series as seen in. As the font size had to be that tiny, we did not mask the distinct equations. Indeed, the blue shields would have refocused the participants to these as the formulae would have not been perceived at first glance because of its size.

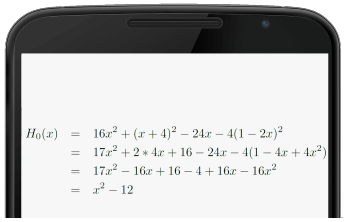


Figure 4: H_0

But we had not accounted for the tendency of participants to

(a.) bend forward and squint at the equation system H_0, and

(b.) start from the rear, that is, checking the correctness of the single equations starting with the last, see e.g. Figure 5.

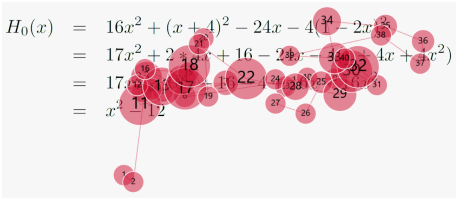


Figure 5: A Typical Gazeplot for H_0

Nevertheless, these are crucial observations to learn from for future set-ups. The observed body movement (a.) was very often accompanied by a sigh and it was clearly considered to be a nuisance to look at such a small equation. Our best guess is, that even if we could have tested direct smartphone use in this situation, it would probably have been still a hinderance to move the smartphone closer to the eyes to enlarge the formula.

With respect to the surprising finding (b) we can visualize this with the heatmap for H_0 in Figure 6: it shows the hot spots of fixation for all participants. Figure 5 shows the order of fixations in a gazeplot of a typical test subject. Our best guess for why participants started reading at the end is the human tendency to solve simple problems before difficult ones.

To being able to use the same masking as with the other variants to solve both issues (a) and (b), consider a different, but natural switch of screen orientation for H_0.

Figure 6: H_0 Heatmap

2.3. Layout Properties

First, we will take a closer look at several aspects when analyzing the data on the H-series. For this, we built matrices that visualize the properties for the equation variants H_i^j as used in a previous experiment with the math expressions in Figure 2.

Font Size We already discussed the font size as a property of the math expression that interferes with the credibility of the linebreaking variant used. It would be best, if the font size in all H_i were the same. If this cannot be achieved, then the font size should be

Font Size	H^1	H^2	H^3
H_1	M	M	M
H_2	L	L	L
H_3	S	S	S

Figure 7: Size Pattern

systematically distributed among the variants so that an analysis is possible. For instance, the font size in the H-series in Figure 2 varies from small (S), middle (M) to large (L). In the matrix seen in Figure 7 shows us where to search insights with respect to the font-size: compare the metrical data among the rows.

Linebreaks: How many? Another difference among the equation systems due to the font-size/linebreaking consists of the resulting number of rows. In the introduction we already indicated that the "overview" quality is also used by people when looking at math expressions. So, the gathered data can also be studied having this in mind. In particular, we assume that people get a better overview over a math expression it it stretches above less lines. For Figure 2 we get the matrix shown in Figure 8. Within a layout this number decreases as each expansion does not change the number of lines, but each simplification by summarizing terms does.

 # Rows in H_i^j

	H^1	H^2	H^3
H_1	4	4	3
H_2	8	6	3
H_3	8	6	3

Figure 8: Row Pattern

Linebreaks: Format Another difference within each H_i is the formatting of the linebreaks. In Figure 2 we distinguished linebreak variants into “simple break" (V1), "terms straight" (V2), and "termsStep" (V3) as specified above. Each of the variants has an indentation pattern as a consequence which is visualized in Figure 9 for the H-series displayed in Figure 2. The indentation could influence the overview quality of a math expression.

The first two layouts H_1 and H_2 start the content of the line after the linebreak ‘straight’ (that is, straight plus ϵ) aligned towards the beginning of the broken mathematical expression in the line before. The third layout H_3 follows a steps design, where the content of the line after the linebreak starts with a notable indentation.

Indentation of Linebreaks

	H^1	H^2	H^3
H_1	straight	straight	straight
H_2	straight	straight	straight
H_3	step	step	step

Figure 9: Form Pattern

When we run the experiment with the H-series displayed in Figure 2, the results did not show a clear trend: on the one hand because there were not enough valid data, but on the other hand because the analysis is hard as the different aspects visualized in the matrices above are design consequences of the requirements above. It would be a much better design to find equations that only varied e.g. in the linebreak format. But how to do it eludes us for the moment. Even if we solved it, the experiment was again almost on the verge of cognitively overloading the participants. That is, we cannot simply add more independent equation systems to deconstruct the implicit dependencies.

2.4. Eye-Tracker Metrics

Participants are presented static images of a smartphone with various equations masked as described above (see Figure 3). To ensure that the participants give full attention to all aspects of the equations, we instruct them to “grade” the white parts of the equation systems, seeking errors. We also should instruct the participants to do this as fast as possible to keep them from re-checking errors multiple times – otherwise we would (again) run the risk of drowning out the signal.

For each equation variant we use an analysis feature called **Areas of Interest (AOI)** supplied by the analysis subsystem. AOIs are areas in the stimulus for which the gaze data can be independently analyzed with several metrics, covering the area to be checked for errors on the right-hand side of the equation symbol (see Figure 10). Note that these AOIs are different from MOIs (math objects of interest) introduced by Greiner-Petter et al. in [GP+20] as the latter describe a semantic property rather than a screen area property of math expressions.

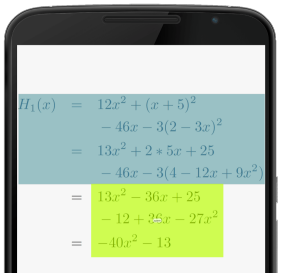


Figure 10: Standard AOI

Among several standard eye-tracker metrics on AOIs we find the following three as the most meaningful:

1. **Total Fixation Duration** (TFD) the overall time a user fixated points in the AOI,
2. **Fixation Count** (FC) the number of fixations in the AOI by the user, and
3. **Total Visit Duration** (TVD) the overall time a user spent on it.

We want to analyze the influence of the formula layout, i.e., the arrangement and sizing of visual elements, onto the reading efficiency, which we measure in terms of TFD, FC, and TVD. The type face is largely regulated by convention and color is usually standardized to the text color in mathematics, so we will disregard them here.

Visual Distraction The total fixation duration is naturally lower than the total visit duration. The difference indicates how long the participants spend within an AOI without fixating long enough to make our threshold for fixation or leaving the AOI for fixations elsewhere. Therefore, the TFD/TVD ratio gives us an indicator how busy the participants were with their visual attention elsewhere, that is, a measure for visual distraction (and correspondingly therefore cognitive distraction).

Error Assessment In the experiment we introduce a nominal task ("grading") which involves finding errors, which are independent of layout. As the experiment is not a test of correctness decisions, one could argue that the error assessment cannot give us insights. Note though that a cluster of wrong assessments compared with the linebreaking style or one of the layout properties matrices can indicate a correlation. If, for example, more errors accumulate for the equation-system using a large font, then we can establish a hypothesis, that a larger font increases legibility, but reduces readability.

	# true	**# false**	**# none**
H_0			
H_1^1			
H_1^2			
…			
H_3^2			
H_3^3			

Table 1: Error Results

Therefore we suggest to collect the correctness decisions in Table 1 as follows: Whenever a participant indicates either that the seen formula has an error or it is correct, we add 1 to the respective first column "# true" in Table 1, if this statement is false we increase the respective value in the "# false" column, and if s/he can not decide we increase the "# none" column by 1.

3. Conclusion and Future Work

Our long term goal is to better understand how technical documents (which prominently contain formulae) can best be presented on mobile devices. Concretely, we have tried to investigate reading efficiency of mathematical expressions on small screens in several experiments, in particular the effect of distinct linebreaking scenarios, but we have failed so far to find the right experimental set-up.

In this paper we tried to summarize our learning process with respect to necessary conditions on the experimental set-up for an eye-tracking study at least to avoid traps, at most fit to show insights into the best presentation of math expressions on mobile phones.

Before we really do the larger follow-up study with more independencies when varying the influence variables than before, we like to discuss the general approach with the interested community.

References

[DHR] Michael J. Downes, Morten Høgholm, and Will Robertson. *The breqn package*. URL: http://mirrors.ctan.org/macros/latex/contrib/breqn/breqn.pdf (visited on 02/25/2020).

[GP+20] André Greiner-Petter et al. "Discovering Mathematical Objects of Interest—A Study of Mathematical Notations". In: *Proceedings of The Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, 2020, 1445–1456. DOI: 10.1145/3366423.3380218.

[HWH99] John M. Henderson, Phillip A. Weeks Jr., and Andrew Hollingworth. "The effects of semantic consistency on eye movements during complex scene viewing". In: *Journal of Experimental Psychology: Human Perception and Performance* 25.1 (1999), pp. 210–228. DOI: 10.1037/0096-1523.25.1.210.

[KKF17] Andrea Kohlhase, Michael Kohlhase, and Michael Fürsich. "Visual Structure in Math Expressions". In: *Intelligent Computer Mathematics (CICM) 2017*. Ed. by Herman Geuvers et al. LNAI 10383. Springer, 2017. DOI: 10.1007/978-3-319-62075-6.

[Mat] *MathJax: Beautiful Math in all Browsers*. URL: http://mathjax.com (visited on 09/27/2010).

[MML310] Ron Ausbrooks et al. *Mathematical Markup Language (MathML) Version 3.0*. Ed. by David Carlisle, Patrick Ion, and Robert Miner. 2010. URL: http://www.w3.org/TR/MathML3.

[MP14] Nicolas Masson and Mauro Pesenti. "Attentional Bias Induced by Solving Simple and Complex Addition and Subtraction Problems". In: *Quarterly Journal of Experimental Psychology* 67.8 (2014). PMID: 24833320, pp. 1514–1526. DOI: 10.1080/17470218.2014.903985.

[Ray98] Keith Rayner. "Eye Movements in Reading and Information Processing: 20 Years of Research". English. In: *Psychological Bulletin* 124.3 (1998), pp. 372–422.

[Swa] Ellen Swanson. *Mathematics into Type*. updated edition. AMS. URL: https://www.ams.org/publications/authors/mit-2.pdf.