Linebreaking Formulae – An Eye-Tracking Study

Andrea Kohlhase[1][0000−0001−5384−6702] and Michael Kohlhase[2][0000−0002−9859−6337]

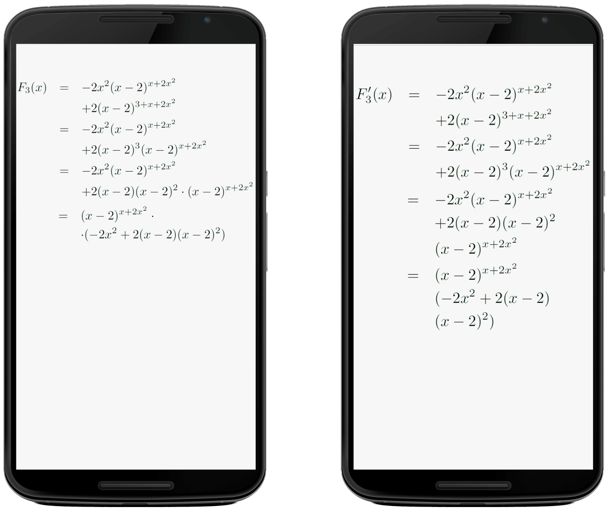
[1] Information Management, University of Applied Sciences Neu-Ulm
[2] Computer Science, FAU Erlangen-Nürnberg

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 investigate the effect of linebreaking in mathematical formulae for reading efficiency using eye-tracking experiments.

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



$$\begin{aligned} F_3(x) &= -2x^2(x-2)^{x+2x^2} \\ &\quad +2(x-2)^{3+x+2x^2} \\ &= -2x^2(x-2)^{x+2x^2} \\ &\quad +2(x-2)^3(x-2)^{x+2x^2} \\ &= -2x^2(x-2)^{x+2x^2} \\ &\quad +2(x-2)(x-2)^2 \cdot (x-2)^{x+2x^2} \\ &= (x-2)^{x+2x^2} \cdot \\ &\quad \cdot(-2x^2+2(x-2)(x-2)^2) \end{aligned}$$

$$\begin{aligned} F_3'(x) &= -2x^2(x-2)^{x+2x^2} \\ &\quad +2(x-2)^{3+x+2x^2} \\ &= -2x^2(x-2)^{x+2x^2} \\ &\quad +2(x-2)^3(x-2)^{x+2x^2} \\ &= -2x^2(x-2)^{x+2x^2} \\ &\quad +2(x-2)(x-2)^2 \\ &\quad (x-2)^{x+2x^2} \\ &= (x-2)^{x+2x^2} \\ &\quad (-2x^2+2(x-2) \\ &\quad (x-2)^2) \end{aligned}$$

Fig. 1. Two variants of a formula on a smartphone

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 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[3], 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 along the operator tree. A **gestalt** is a cognitive template that holistically combines layout and operator information. We conjec-

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

tured 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 refine this intuition by an eye-tracking experiment that concentrates on the effects of font size, linebreaking, and indentation on the formula reading efficiency We present an experimental design, which balances the influence between a nominal task that ensures attention and effect-neutrality of the test subjects with the effects to be studied. One of the results of running this experiment with 18 participants is that – contrary to common intuition – font size is only a secondary effect of linebreaking and structural effects are more important.

Overview Section 2 discusses the experimental setup and the ensuing Section 3 presents its results. We discuss them in Section 4 and establish hypotheses based on them. Section 5 summarizes the outcome and concludes the paper.

2 The Experiment

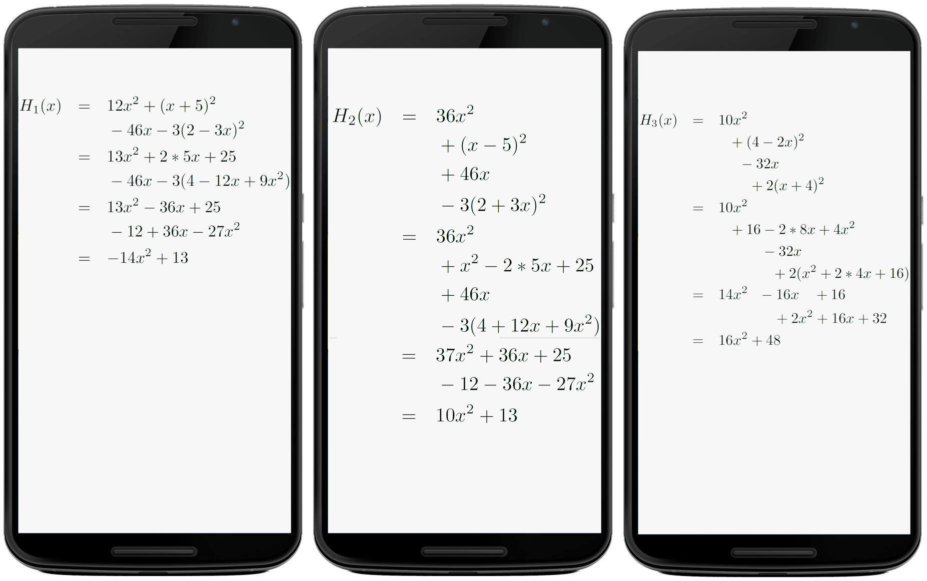
As there is a demonstrable correlation between what a participant attends to and where she is looking at – see for example [Ray98] for an overview, the eye-tracking methodology is an interesting angle of attack. **Eye-tracking**, i.e., the observation of eye movements, allows to get a better understanding 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 font size, number of required lines and indentation after linebreaks in formulae by setting up an eye-tracking experiment. Our goal is to compare the reading efficiency across several linebreaking variants of a mathematical expression.

2.1 The Conceptual Design

For such an eye-tracking study we need (longish) mathematical expressions where linebreaks make sense. Also, participants had to be motivated to look at those closely and not superficially. Moreover, the mathematical expressions used with different representations in terms of linebreaking had to be basically display-equivalent.

The design of the experiment turned out to be more difficult than we expected: In a previous experiment, we had decided on a task with a function $\mathcal{F}$ 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 cognitively overloading our participants: To ensure that they read the presented formulae carefully while being able to gather 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.

This time, we decided on a much more basic nominal task using just a little bit of mental arithmetics. To be precise, we decided that participants had to assess the correctness of a given equation. The simplifications in these equations contained expanding binomial identities, summing up terms and integer-multiplication. Concretely, we used a "**series**" of equation systems H_i shown in Figure 2 that consists of three isomorphic laddered[4] **equation systems** in three linebreaking variants:



$$\begin{aligned} H_1(x) &= 12x^2 + (x+5)^2 \\ &\quad - 46x - 3(2-3x)^2 \\ &= 13x^2 + 2 * 5x + 25 \\ &\quad - 46x - 3(4 - 12x + 9x^2) \\ &= 13x^2 - 36x + 25 \\ &\quad - 12 + 36x - 27x^2 \\ &= -14x^2 + 13 \end{aligned}$$

$$\begin{aligned} H_2(x) &= 36x^2 \\ &\quad + (x-5)^2 \\ &\quad + 46x \\ &\quad - 3(2+3x)^2 \\ &= 36x^2 \\ &\quad + x^2 - 2 * 5x + 25 \\ &\quad + 46x \\ &\quad - 3(4 + 12x + 9x^2) \\ &= 37x^2 + 36x + 25 \\ &\quad - 12 - 36x - 27x^2 \\ &= 10x^2 + 13 \end{aligned}$$

$$\begin{aligned} H_3(x) &= 10x^2 \\ &\quad + (4-2x)^2 \\ &\quad - 32x \\ &\quad + 2(x+4)^2 \\ &= 10x^2 \\ &\quad + 16 - 2 * 8x + 4x^2 \\ &\quad - 32x \\ &\quad + 2(x^2 + 2 * 4x + 16) \\ &= 14x^2 \quad - 16x \quad + 16 \\ &\quad + 2x^2 + 16x + 32 \\ &= 16x^2 + 48 \end{aligned}$$

H_1 H_2 H_3

Fig. 2. The H-Series of Distinct Layouts

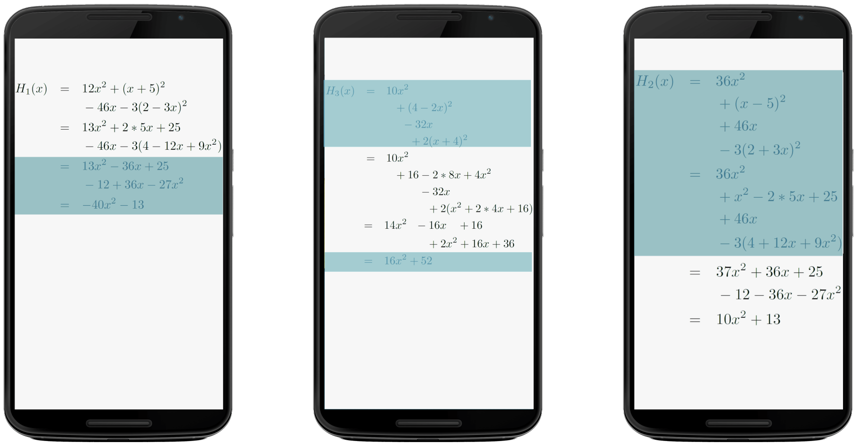
1. **simple break**: H_1 (on the left of Figure 2) breaks after half of the summands of the right hand side,

[4] 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.

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

Note that the H_i also vary coefficients, signs or literals to keep the participants from noticing the structural invariants just described. As the equation systems H_i are isomorphic up to these changes, we also speak of **layouts** H_i.

For each we masked all but one equation system fragment in blue to focus the attention of the participants on the white fragment (see Figure 3). 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. 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). 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, included small calculation errors into the equations



$$\begin{aligned} H_1(x) &= 12x^2 + (x+5)^2 \\ &\quad - 46x - 3(2-3x)^2 \\ &= 13x^2 + 2*5x + 25 \\ &\quad - 46x - 3(4 - 12x + 9x^2) \\ &= 13x^2 - 36x + 25 \\ &\quad - 12 + 36x - 27x^2 \\ &= -40x^2 - 13 \end{aligned}$$

$$\begin{aligned} H_3(x) &= 10x^2 \\ &\qquad + (4-2x)^2 \\ &\qquad\quad - 32x \\ &\qquad\qquad + 2(x+4)^2 \\ &= 10x^2 \\ &\qquad + 16 - 2*8x + 4x^2 \\ &\qquad\qquad - 32x \\ &\qquad\qquad\quad + 2(x^2 + 2*4x + 16) \\ &= 14x^2 \quad - 16x \quad + 16 \\ &\qquad\qquad\quad + 2x^2 + 16x + 36 \\ &= 16x^2 + 52 \end{aligned}$$

$$\begin{aligned} H_2(x) &= 36x^2 \\ &\quad + (x-5)^2 \\ &\quad + 46x \\ &\quad - 3(2+3x)^2 \\ &= 36x^2 \\ &\quad + x^2 - 2*5x + 25 \\ &\quad + 46x \\ &\quad - 3(4 + 12x + 9x^2) \\ &= 37x^2 + 36x + 25 \\ &\quad - 12 - 36x - 27x^2 \\ &= 10x^2 + 13 \end{aligned}$$

H_1^1 H_3^2 H_2^3

Fig. 3. Exemplary Masked Equation Variants

To further obfuscate the invariants, we showed the H_i^j to participants interspersed with other masked equation systems like either one in Figure 1 – though in systematic order within each equation system.

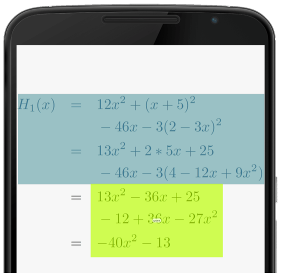
Note that strictly speaking the nominal task of assessment only measures the "grading efficiency". However, we posit that for mathematical texts and formulae, reading, understanding, and assessing the correctness are equivalent: none of them can be done without the others.

2.2 The Concrete Study

We presented participants with static images of a smartphone with various equations masked as described above (see Figure 3) and gathered gaze data with Tobii's X3-120 [TX3] eye-tracker. To ensure that the participants gave full attention to all aspects of the equations, we instructed them to "grade" the white parts of the equation systems, seeking errors. We also instructed 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 defined an **Area of Interest (AOI)**, i.e., an area 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 4).

Among several AOI metrics we selected the following three as the most meaningful:



$$\begin{aligned} H_1(x) &= 12x^2 + (x+5)^2 \\ &\quad - 46x - 3(2-3x)^2 \\ &= 13x^2 + 2 * 5x + 25 \\ &\quad - 46x - 3(4 - 12x + 9x^2) \\ &= 13x^2 - 36x + 25 \\ &\quad - 12 + 36x - 27x^2 \\ &= -40x^2 - 13 \end{aligned}$$

Fig. 4. Our Standard AOI

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

We ran the eye-tracking test for 18 participants. Table 1 gives an overview of the variables and their distribution.

3 Results

To interpret the results of the eye-tracking study we analyzed 18 recordings[5] with the Tobii Pro Full Lab suite; unfortunately, only 13 recordings provided valid gaze data (gaze samples $> 70\%$), probably caused by the participants' tendency to again and again lean forward when concentrating on finding errors.

age group	**20 < 30**	**30 < 40**		
	14	4		
gender	**male**	**female**	**diverse**	
	16	1	1	
math affinity	**none**	**weak**	**ok**	**strong**
	0	0	9	9
formula experience	**none**	**weak**	**ok**	**strong**
	1	3	3	11

Table 1. Participant Distribution

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.

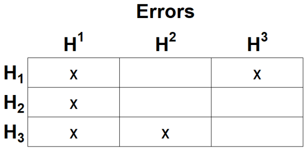
[5] The raw eye-tracker data and interpretation spreadsheets are available upon request to the authors.

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.

In the experiment we had to introduce a nominal task ("grading") which involved finding errors, which are independent of layout – we made them so – but influence these metrics, so we have to normalize for this before we can interpret reading efficiency.

3.1 Error Normalization Factors

To encourage our test subjects to look closely at each of the three equations H^j in each equation system H_1, H_2, and H_3, we included small calculation errors. Figure 5 shows the distribution of these errors. We have to assume that the presence or absence of errors will have an effect in our metrics, which we have to normalize for. Note that in equation H^1 all variants had (isomorphic) errors, so here the equation variants are directly comparable. But in H^2 and H^3 we have to normalize for them.



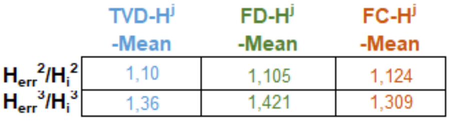
Errors

	H¹	H²	H³
H_1	x		x
H_2	x		
H_3	x	x	

Fig. 5. Error Pattern

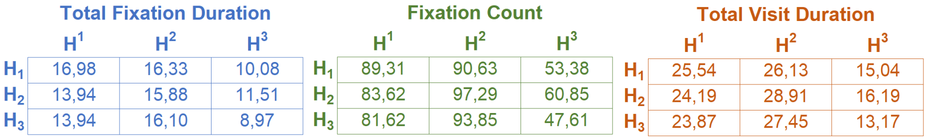
Concretely, we build the mean of the ratios H_3^2/H_i^2 for equation H^2 and likewise with H_1^3/H_i^3 for equation H^3 for the TVD-, FD-, and FC-metric data resulting in the factors in Figure 6.

We observe that the specific influences of errors vary with the equation, which – given the different equation sizes and structures – is plausible. But they are quite stable over the different measures. This supports the assumption that we are looking at an intrinsic effect of error handling by the participants and not an artefact of the experiment.



	TVD-H^j -Mean	FD-H^j -Mean	FC-H^j -Mean
H_{err}^2/H_i^2	1,10	1,105	1,124
H_{err}^3/H_i^3	1,36	1,421	1,309

Fig. 6. Normalization Factors for Errors



Total Fixation Duration

	H¹	H²	H³
H_1	16,98	16,33	10,08
H_2	13,94	15,88	11,51
H_3	13,94	16,10	8,97

Fixation Count

	H¹	H²	H³
H_1	89,31	90,63	53,38
H_2	83,62	97,29	60,85
H_3	81,62	93,85	47,61

Total Visit Duration

	H¹	H²	H³
H_1	25,54	26,13	15,04
H_2	24,19	28,91	16,19
H_3	23,87	27,45	13,17

Fig. 7. Error-Normalized Eye-Tracking Results for the H-Series

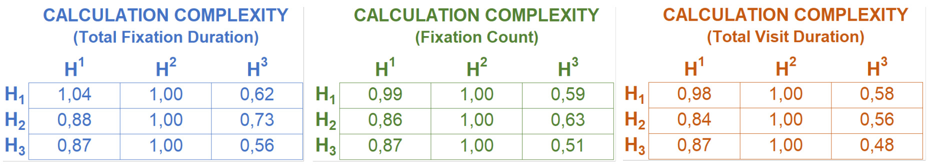
Applying these six error normalization factors to the eye-tracking results give us the values shown in Figure 7 for further processing.

3.2 Calculation Complexity

As we are interested in the trade-off between distinct layout factors, such as font size and number of rows within an equation, the next influence we want to understand is the difficulty of performing the calculations that are constitutive for a particular equation H^j: the **calculation complexity**. Note that – even though that would be interesting in itself – we do not want to predict the calculation complexity, but only normalize for it.

Qua design neither the entire equation systems H_i nor their equation variants H_i^j (for a fixed i) differ in calculation complexity, but each equation H^j differs from the others. Concretely, the first equation always involved two applications of the binomial identity, the second equation summing up corresponding monomials and multiplying out, and the third only summing up monomials.

We observe that all the metrics in Figure 7 deliver within any equation H^j, i.e., in each column, roughly the same values. In particular, H^2 clearly is the hardest to compute, followed by H^1 and with an obvious bigger gap H^3. If we normalize the gaze data in Figure 7 with respect to one column, we have a better grasp on this relation for all metrics, thus we normalize these with respect to its second equation by computing the ratios $C(H_i^j) = H_i^j/H_i^2$ - giving us Figure 8. With this normalization we can now compare the values in each row with each other. We could theoretically have chosen any column to normalize for except for the first as equation variant H_1^1 was the very first equation variant the test subjects looked at, so they had to get used to the format of the experiment and particularly the masking, which resulted in biased gaze data for H_1^1.



CALCULATION COMPLEXITY
(Total Fixation Duration)

	H¹	H²	H³
H_1	1,04	1,00	0,62
H_2	0,88	1,00	0,73
H_3	0,87	1,00	0,56

CALCULATION COMPLEXITY
(Fixation Count)

	H^1	H^2	H^3
H_1	0,99	1,00	0,59
H_2	0,86	1,00	0,63
H_3	0,87	1,00	0,51

CALCULATION COMPLEXITY
(Total Visit Duration)

	H^1	H^2	H^3
H_1	0,98	1,00	0,58
H_2	0,84	1,00	0,56
H_3	0,87	1,00	0,48

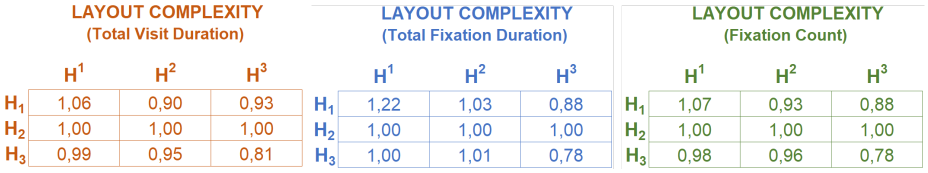
Fig. 8. Calculation Complexity $C(H_i^j)$ in each Equation System H_i for all Metrics

3.3 Layout Complexity

Recall that the purpose of our experiment is to study the influence of formula layout on reading efficiency. Ideally, we would be able to model reading efficiency as a function from primitive layout factors like font size, number of linebreaks, indentation, etc. or composite factors like the "penalty" in the MathML3 linebreaking algorithm.

For the moment, we will just compute the **layout complexity** $L(H_i^j)$ of our nine equations, and relate them qualitatively to the different layout factors. We leave the modeling task to future work, but remark that the methodology and data presented here will facilitate modeling and can be used to evaluate any models.

For understanding the layout complexity, we observe that the equation variants H_i^j for a fixed j differ exactly in their layout. Therefore we compute the ratios $L(H_i^j) := H_i^j / H_2^j$ for the values in Figure 7[6] resulting in Figure 9. The most efficient equation variant is the one with the lowest value.



LAYOUT COMPLEXITY (Total Visit Duration)

	H¹	H²	H³
H_1	1,06	0,90	0,93
H_2	1,00	1,00	1,00
H_3	0,99	0,95	0,81

LAYOUT COMPLEXITY (Total Fixation Duration)

	H^1	H^2	H^3
H_1	1,22	1,03	0,88
H_2	1,00	1,00	1,00
H_3	1,00	1,01	0,78

LAYOUT COMPLEXITY (Fixation Count)

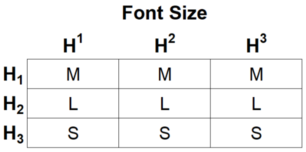
	H^1	H^2	H^3
H_1	1,07	0,93	0,88
H_2	1,00	1,00	1,00
H_3	0,98	0,96	0,78

Fig. 9. Layout Complexity $L(H_i^j)$ in each Equation H^j for all Metrics

4 Discussion

First, we will take a closer look at several aspects with respect to the reading efficiency of our H-series, which will be used later on to discuss calculation and layout complexity and other results of our experiment.

Font Size The font size in the H-series varies from small (S), middle (M) to large (L), where the difference between middle and large is more notable than between middle and small (see Figure 10). The font size itself does not change within a layout H_i, in particular, this pattern only varies between rows.

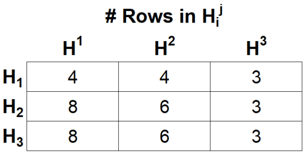


Font Size

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

Fig. 10. Size Pattern

Linebreaks: How many? One difference between the layouts H_i consists of the number of linebreaks used. Figure 11 gives us an overview. Therefore, the number of lines to check for errors by the participants vary. Within a layout this number decreases as each expansion does not change the number of lines, but each simplification by summarizing terms does.

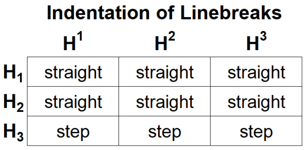


Rows in H_i^j

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

Fig. 11. Row Pattern

Linebreaks: Format Another difference is the formatting of the linebreaks. 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. Note, we didn't want to establish the optimal design, just to understand whether it is an influence factor.



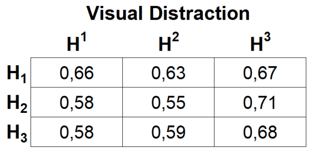
Indentation of Linebreaks

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

Fig. 12. Form Pattern

[6] Again, the normalization is arbitrary; we chose H_2 for consistency.

Visual Distraction Let us have another look at the absolute gaze data in Figure 7. 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).



Visual Distraction

	H^1	H^2	H^3
H_1	0,66	0,63	0,67
H_2	0,58	0,55	0,71
H_3	0,58	0,59	0,68

Fig. 13. TFD/TVD Pattern

4.1 Layout Complexity

Now let us discuss the numbers in Figure 9 with respect to the layout patterns detailed above.

Font size effects Any effect of the font size (see Figure 10) on the reading efficiency will show in the distribution of the layout complexity values $L(H_i^j)$ in the columns of Figure 9 in the form:

(I) $L(H_2^j) \leq L(H_1^j) \leq L(H_3^j)$ *would indicate that a smaller font size has a negative effect on reading efficiency*, whereas ↯

(II) $L(H_2^j) \geq L(H_1^j) \geq L(H_3^j)$ *would indicate that a smaller font size has a positive effect on reading efficiency.*

Recall that the equation variant H_1^1 cannot really be taken into account. Once we disregard H_1^1 though, we still do not have all columns uniformly supporting any of hypotheses *(I)* or *(II)*. In all metrics, equation H^3 (maybe surprisingly) supports hypothesis (II) with a difference of 10% respectively between small and middle and large font size. Note that this effect should be the most distinct in equation H^3 as the number of rows (and thus linebreaks) is identical in the variants of this equation but not in the others (see Figure 11). In all metrics and equations the largest font size turned out to be the least efficient one[7] in terms of reading.

But there is another possible explanation: We observe that the second line of H_3^3 swaps the constant and the square monomial (with a linear one in-between) compared to H_1^3. In NL discourse, such a mutation would have substantially changed the cognitive complexity of the variants, since a parallelism constraint is violated in H_1^3; see e.g. [GK97] for a discussion. As the difference between the variant in question is not larger than the one of the two equivalent variants H_1^3 and H_2^3, we assume that the equivalence of H_3^3 in terms of calculation complexity was not seriously compromised. We may have another instance here, where formula understanding differs from discourse understanding; we leave studying this to future work.

Turning to equations H^1 and H^2 we observe that the font size effect might have canceled out with the semantic indenting and the number of rows effect.

[7] There is one exception for TFD in variant H_1^2, but it is a very small surplus of $\sim 3\%$.

Especially in equation H^2, fewer, but longer fixations occur with an overall faster visit when being presented with a mathematical expression using a middle-sized font. Longer fixations indicate a slower perception, whereas fewer fixations hint at a more effective cognitive activity. Note that this might also be due to the structure of equation H^2 as the one with the most rows.

All of this runs counter to the intuitions of the math typesetting folklore, so there may be other effects at play here.

Effects of linebreaking Recall that operators are the orientation points for understanding formulae (see [KKF17]). So the main effect of linebreaks might be in adding structure to an equation rather than in allowing larger font sizes as often assumed.

In Figure 11 we note that the equation with the least linebreaks is H^3 having only 3 rows. Equation H^2 contains either 4 or 6 rows and H^1 either 4 or 8 rows. If we look at the gaze data in Figure 7 we find that in all metrics the higher the number of rows the higher the values (except for H^1_1, which we again discount). That means that linebreaks are expensive in terms of reading efficiency. Interestingly, when we discussed this with the subjects, most came up with the counter hypothesis or rejected it when the interviewer asked about it.

Remember that the font size with only 3 lines showed, that the smaller font size takes a 10% advantage, but when the equations had 6 lines the advantage shrinks down to 4% and for 8 lines it even reduces to 2%. This yields the hypothesis that the larger the number of linebreaks in a formula the less relevant the font size.

Semantic Indentation The indentation (see Figure 12) is most pronounced in H^1 on both sides of the equation, followed by H^2 mostly on the lefthand side and finally, a less notable one on the lefthand side in equation H^3. In particular, here we expect to see either a property like $H^1_i \ddagger H^2_i \ddagger H^3_i$, where $\ddagger$ is the relation $\leq$ or $\geq$ in both cases in Figure 7 or $C(H^1_i)\ddagger C(H^2_i)\ddagger C(H^3_i)$ in Figure 8 for each i. Or we would expect a property like $H^j_1 \ddagger H^j_2 \ddagger H^j_3$ in Figure 7 or $L(H^j_1)\ddagger L(H^j_2)\ddagger L(H^j_3)$ in Figure 9 for each j. Again we have to drop H^1_1 from the interpretation, but nevertheless no similarity according to either of these patterns can be recognized. As the effect in equation H^3 is rather weak, we only look at the equation variants H^1_2, H^2_2, H^1_3, and H^2_3. Even this reduced matrix does not conform to the pattern. Maybe an analysis of the heat maps would show less pronounced hot spots or gazeplots would show more precision because of more structure via semantic indenting, but this is beyond the scope of the paper.

Visual Distraction Comparing the values in Figure 11 for the number of lines with the visual distraction numbers in Figure 13, we can observe that the lower the number of lines the more time is left for distraction. Even though distraction has a negative bias, it may also be interpreted as understanding the terms in context, which indeed was observable in the gazeplots. We don't know whether this context gazing as an action was satisfied, comfortable, confirming, desperate, or simply confused, but this would be an interesting topic for future research:

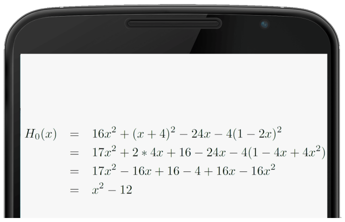
what are the emotional aspects about context and how important are they for example with respect to math affinity? Should context be shown in an eLearning system for math together with the solution? What aspects of the context are relevant?

Calculation Complexity The values for the TVD-, TFD-, and FC-metrics in Figure 8 for the second equation must be 1 for all layouts H_i (as this is our relative factor), so this doesn't give as any information. As the expansion of two binomial identities was done in purely symbolic form, it is plausible that the simplification to be assessed with 4 terms against 9 terms in H^1 is easier done than in H^2, where 9 terms have to be checked against 6 monomials. The last equation H^3 is even easier as the assessment only refers to 6 against 2 monomials.

It is also obvious from Figure 8 that all our selected metrics behave similarly, which indicates that all are valuable for observing calculation complexity.

4.2 The (Failed) Gold Standard H_0

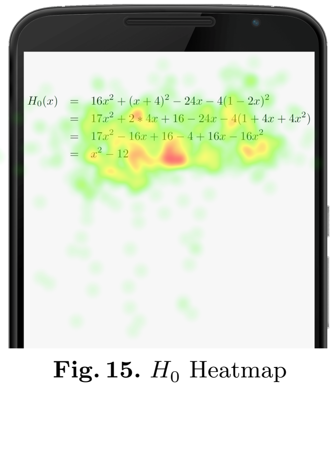
Our version of the H-series that did *not* have linebreaks on the right hand side of the equation was called H_0 (see Figure 14). It was designed to be isomorphic to the equation systems of the H-series. As the font size had to be tiny to fit the smartphone screen, we did not mask the distinct equations. 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.



$$\begin{aligned} H_0(x) &= 16x^2 + (x+4)^2 - 24x - 4(1-2x)^2 \\ &= 17x^2 + 2 * 4x + 16 - 24x - 4(1 - 4x + 4x^2) \\ &= 17x^2 - 16x + 16 - 4 + 16x - 16x^2 \\ &= x^2 - 12 \end{aligned}$$

Fig. 14. H_0 .

We planned to analyze the gaze data with the help of the gazeplots, which would indicate at what time which equation was being analyzed and we would have gathered the data for each recording for that individual time slot to obtain the $L(H_0^j)$ for the layout complexity table in Figure 9. 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 16. This meant that we were not able to create comparable data, so we did not have a gold standard to compare our other data against. Therefore, we did not include the H_0 data into our results discussed above.



$$\begin{aligned} H_0(x) &= 16x^2 + (x+4)^2 - 24x - 4(1-2x)^2 \\ &= 17x^2 + 2 * 4x + 16 - 24x - 4(1 + 4x + 4x^2) \\ &= 17x^2 - 16x + 16 - 4 + 16x - 16x^2 \\ &= x^2 - 12 \end{aligned}$$

Fig. 15. H_0 Heatmap

Nevertheless, these are interesting observations for interpretation. 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. This observation indicates that a mobile layout solution for large formulae should not consist of just shrinking the picture. Note that in Table 2 the equation system encounters the most hits for no assessment achieved.

With respect to the surprising finding (b) we can visualize this with the heatmap for H_0 in Figure 15: it shows the hot spots of fixation for all participants. Figure 16 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.

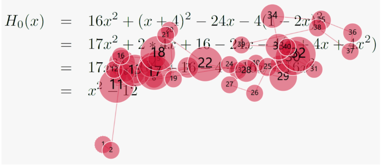


Fig. 16. A Typical Gazeplot for H_0

4.3 Error Assessment

In our experiment we asked participants to decide whether the equation presented was correct or contained an error. 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. Therefore we not only asked the participants to be as fast as possible, but also we provided an automatic cut-off of the presentation of every equation variant H_i^j after $40s$ and of an equation system (without masking) like H_0 after $60s$.

	# true	**# false**	**# none**
H_0	8	5	4
H_1^1	7	9	1
H_1^2	8	9	0
H_1^3	11	6	0
H_2^1	9	8	0
H_2^2	15	2	0
H_2^3	16	1	0
H_3^1	12	5	0
H_3^2	14	3	0
H_3^3	15	0	2 (invalid)

Table 2. Error Results

The error assessment by our "graders" can be seen in Table 2[8]. Whenever a participant told us, either that the seen formula had an error or it was correct, we added 1 to the respective first column "# true", if this statement was false we increased the respective value in the "# false" column, and if he could not decide we increased the "# none" column by 1.

Still, we were surprised by the outcome. Note that the distribution of correct and incorrect answers indicate a very high assessment error factor. This is probably not an artefact of the deadline we imposed, otherwise we would have many "do-not-know" comments in the third column. On the contrary, most participants had formed a stable intuition about the correctness. The equation variant H_3^3 stands out: everyone assessed

	Err	Rows	Font	Indentation	Distr	$L(H_i^j)$	$C(H_i^j)$
H_3^3	–	3	small	straight	max	min	min
H_2^3	–	3	large	straight	–	–	–
H_2^2	–	6	large	straight	min	–	max
H_3^2	x	6	small	steps	–	–	max

Table 3. Possible Influences on Error Assessment

[8] The data for one participant is lost, so the total sum of statements is 17.

the correctness right, closely followed by H^3_2, H^2_2, and H^2_3 which also were predominantly assigned the correct error status.

So, is there something special to these equation variants? We observe that they vary widely with respect to the aspects given in Figures 5 and 8 to 13 – Table 3 summarizes the situation. The only properties they share are that they are neither the start point for checking for errors, i.e., do not belong to equation H^1, nor do they belong to a special linebreaking layout, i.e., do not belong to layout H_1. We can cautiously phrase the hypotheses, that

- most difficulties in understanding an equation system happens when first engaging with it (H^1)
- the quality of mental arithmetics is supported by a less compact, that is a more structured, layout (H_1).

5 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 investigated reading efficiency of mathematical expressions on small screens, in particular the effect of distinct linebreaking scenarios.

An exploratory eye-tracking experiment using equational systems that varied in font size and linebreaking variants sheds some light on this. The main general findings – they are stable under all metrics in the experiment – include:

1. *Font size matters*: but the seemingly obvious “bigger glyphs make equations easier to read” could not be confirmed in the font-size range of the experiment. Tiny equations however were considered a nuisance.
2. *Overview matters (more)*: linebreaking adds structure, and this seems to have a stronger effect than the font size.

Unfortunately, our experiment does not suggest a metric version of these findings, i.e., expressing reading efficiency as a function of font size and some geometric invariant of the layout. Such a function would be needed for cognitively justified automatic linebreaking algorithms. To get this we would have have to gather much more data.

Our experiment also tried to evaluate *semantic indenting*, but the data is inconclusive. A surprising finding that was unforeseen in the experimental design was that in the tiny equation systems, many participants *grade equation systems starting from the back*, perhaps given the overview they afforded they started with the simplest subtask.

Given the above, an equally important – and non-trivial – contribution of our work is the experimental design in this paper. It was informed by the earlier (failed) experiment described in the beginning of Section 2, where the computation load drowned out the effect – the layout complexity – we wanted to measure. The nominal task of the experiment reported here – “grading equations” – does not seem to suffer from this flaw.

Our discussion in Section 4 shows what the possible influences on layout complexity might be, how the parameters can be varied, and what traps have to be avoided. This points the way to a larger follow-up study which varies the influence variables independently (and not in combination as sometimes in the current experiment).

There are a couple of additional aspects or presenting mathematical formulae on mobile devices a larger study should also take into account: These include

- MathML and thus HTML5 allow to specify `overflow="scroll"` for the "linebreaking regime" on a formula. This has the effect that the formula is rendered unbroken, but comes with an in-text scroll bar that allows to view the formula via scrolling.
- The stationary eye-tracker we used in our experiment could not simulate the practice of holding a smartphone near to the eyes to be able to decipher very small formulae. For this we would need a head-mounted eye-tracker and a real smartphone.

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).

[GK97] Claire Gardent and Michael Kohlhase. "Computing parallelism in Discourse". In: *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*. Ed. by Martha E. Pollack. Nagoya, Japan: Morgan Kaufmann, 1997, pp. 1016–1021. URL: `http://kwarc.info/kohlhase/papers/ijcai97.pdf`.

[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`.

[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`.

[TX3] *Tobii Pro X3-12 - Screen-based eye tracker*. URL: `https://www.tobiipro.com/product-listing/tobii-pro-x3-120/` (visited on 03/09/2020).