Pupillary measurement during an assembly task

Bækgaard, Per; Jalaliniya, Shahram; Hansen, John Paulin

Published in:
Applied Ergonomics

Link to article, DOI:
10.1016/j.apergo.2018.09.004

Publication date:
2019

Document Version
Peer reviewed version

Citation (APA):
Pupillary Measurement During an Assembly Task

Per Bækgaard\textsuperscript{a},*, Shahram Jalaliniya\textsuperscript{b}, John Paulin Hansen\textsuperscript{c}

\textsuperscript{a}DTU Compute, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby.
\textsuperscript{b}Department of Computer Science, IoTaP Research Center, Malmö University, Sweden
\textsuperscript{c}DTU Management Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby

Abstract

We conducted an empirical study of 57 children using a printed Booklet and a digital Tablet instruction for LEGO\textsuperscript{®} construction while they wore a head-mounted gaze tracker. Booklets caused a particularly strong pupil dilation when encountered as the first media. Subjective responses confirmed the booklet to be more difficult to use. The children who were least productive and asked for assistance more often had a significantly different pupil pattern than the rest. Our findings suggest that it is possible to collect pupil size data in unconstrained work scenarios, providing insight to task effort and difficulties.

Keywords: assembly manual, eye tracking, pupil dilation

1. Introduction

Mental workloads are commonly measured by subjective ratings (Reid and Nygren, 1988), which risk being rationalizations made in hindsight, or using dual-task paradigms (Wickens, 1991) that interfere with execution of the primary task. EEG provides another source of data for studying continuous cognitive load (Klimesch, 1999), but placing electrodes on a participant’s head can be a challenge, and subjects wearing them may feel awkward. Likewise, Galvanic Skin Response (GSR) and heart rate measurements require sensors to be attached to the subject, which may hinder mobility.
Pupillometry holds potential as an unobtrusive way to measure the cognitive effort associated with a given task. Gaze tracking is becoming low-cost and is likely to be integrated with future displays. However, in real task situations, one of the main challenges is to distinguish the rather large pupil reactions caused by unknown and uncontrollable changes in the ambient light from the minuscule dilations that reflect changes in cognitive effort. Another challenge is the large individual differences in pupil size and in people’s reactions to task difficulties. This paper seeks to clarify whether pupillary measures can be used to analyse user experiences in-the-wild. The main research question addressed is whether noisy pupil data can provide information about the user experience and task engagement for a given population.

2. Theory

2.1. Pictorial assembly instructions

The preferred medium for step-by-step instructions has been print on paper. In recent years, new digital forms of instruction material have emerged. Apps with interactive building instructions have been provided by LEGO® for some years. When downloaded to a PC, Tablet or smartphone they offer instructions on how to build a model. Sequential navigation is done by touching a forward or a backward button located in the corners. Details can be examined by zoom and rotation; animations show how to place a component and a forward/rewind slider helps find a particular event in the construction sequence (cf Fig. 1).

We expect digital instructions to become more common in the future because they offer easy updating, extended explanations (e.g. more steps and animations) and may be less costly to produce as a large amount of products are routinely modelled in 3D for design, production, and marketing.
2.2. Previous work

A number of human factor studies of building instructions have been reported (Table 1). Common interest was the impact of different media and display forms (Tang et al., 2003; Henderson and Feiner, 2011; Alexander, 2013; Wille et al., 2014; Funk et al., 2015) and how to best present the pictorial instructions (Pillay, 1998; Rodriguez, 2002; Richardson et al., 2004; Martin and Smith-Jackson, 2008). A few studies (Martin and Smith-Jackson, 2008; Gupta et al., 2012) looked at the influence of individual factors like age, gender,
Table 1: Experimental studies of assembly tasks with pictorial instructions

<table>
<thead>
<tr>
<th>Author (Year)</th>
<th>N</th>
<th>Independent variables</th>
<th>Dependent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pillay (1998)</td>
<td>40</td>
<td>Orthographic drawing; iso-metric drawing; physical model</td>
<td>Task time; errors; correct completed models; number of looks; study time.</td>
</tr>
<tr>
<td>Rodriguez (2002)</td>
<td>40</td>
<td>Diagrammatic instructions; text and diagrams</td>
<td>Task time; errors</td>
</tr>
<tr>
<td>Tang et al. (2003)</td>
<td>75</td>
<td>Printed manual; computer,; HMD⁰¹; AR</td>
<td>Errors; mental effort (NASA TLX rating)</td>
</tr>
<tr>
<td>Richardson et al. (2004)</td>
<td>64</td>
<td>Number of items, symmetrical planes, fastening points, fastenings, novel assemblies and assembly steps</td>
<td>Subjective rating of assembly complexity (from just reading the instructions); viewing time</td>
</tr>
<tr>
<td>Martin and Smith-Jackson (2008)</td>
<td>24</td>
<td>Toy instructions (LEGO⁰⁰, KNEX⁰⁰, Lincoln Logs⁰⁰); age; gender; building experience</td>
<td>Task time; errors; usability problems found; gaze time; subjective rating of ease and fun</td>
</tr>
<tr>
<td>Henderson and Feiner (2011)</td>
<td>6</td>
<td>Computer; AR; HMD¹</td>
<td>Task time; head movements; subjective preferences</td>
</tr>
<tr>
<td>Gupta et al. (2012)</td>
<td>16</td>
<td>Spatial competence; gender; educational level, building experience; feedback guidance (tracking); figure-based guidance</td>
<td>Task time; errors per step</td>
</tr>
<tr>
<td>Alexander (2013)</td>
<td>28</td>
<td>Print; video</td>
<td>Errors, comprehension, subjective rating of satisfaction; preferences</td>
</tr>
<tr>
<td>Wille et al. (2014)</td>
<td>20</td>
<td>HMD¹ (Google Glass); Tablet PC</td>
<td>Task time (per step); dual task performance; subjective rating of visual and mental fatigue</td>
</tr>
<tr>
<td>Funk et al. (2015)</td>
<td>15</td>
<td>In-situ projections; print</td>
<td>Task time; errors</td>
</tr>
</tbody>
</table>

education and building experience. The measurements most often taken were task time and errors; while five studies (Tang et al., 2003; Richardson et al., 2004; Martin and Smith-Jackson, 2008; Alexander, 2013; Wille et al., 2014) included subjective ratings of experienced difficulty, mental effort, fatigue and/or satisfaction and two (Henderson and Feiner, 2011; Alexander, 2013) asked for the participants’ preferences. Interestingly, three studies (Pillay, 1998; Richardson et al., 2004; Martin and Smith-Jackson, 2008) included a manual record of visual behaviour (in terms of "number of looks", "study time", "viewing time" and "gaze time") but without the use of gaze tracking equipment.
The potential benefit of using eye activity to measure cognitive workload in tasks has been addressed in several previous studies. For instance, Van Orden et al. (2001) presented a display with target density as the workload variable. Blink frequencies, fixation frequency and pupil diameter showed strong correlation with the density of targets. The changes in pupil size from 1 to 9 targets were highly significant, even though the average change in actual size (calculated as moving estimates of means over a 2 second window) was less than 1 mm. Ahlstrom and Friedman-Berg (2006) found no significant effects of task conditions on subjective workload ratings, but significant effects on blink durations, which became shorter when conditions were difficult. The mean pupil diameter increased from 2.4 mm to 3.9 mm for the least responsive subject, while the most responsive had an increase from 2.62 mm to 4.39 mm. The correlation between number of objects (i.e. aircraft) to be supervised (in a simulated air traffic controller operation) and the pupil dilations was rather high (i.e. \( r^2 = 0.7 \)). They concluded that measurements of eye activity provide a more sensitive measure of workload over task time than subjective ratings, and suggest this to be particularly relevant when trying to identify display components that cause workload changes which are not reported by operators themselves. In a study (Bhavsar et al., 2015) the variations of operators’ pupil size were found to be a reliable indicator of the perceived mental workload during a simulated plant emergency; subjects with low, moderate and high task loads showed distinguishably different dilation patterns across the incident.

Kiefer et al. (2016) found significant differences in mean pupil diameter between map tasks conducted by subjects. The authors interpret the results as an indication of low cognitive load when exploring maps freely, while route planning and focused search entailed a high cognitive load.

Dehais et al. (2008) conducted a study with six pilots flying under low-light (nightfall) conditions with a remote gaze tracker mounted in the cockpit. They observed a large dilation

\[ ^1 \text{HMD = Head Mounted Display} \]
after a simulated engine failure compared to the pilots’ pupils during normal flight. In a study presented by Palinko et al. (2010) performance data recorded in a driving simulator (i.e. variances of steering wheel angle and lane position) showed a high correspondence to changes in pupil size, even without explicit control of the lighting conditions.

Orlandi and Brooks (2018) found that marine pilots had higher pupil scores for difficult berthings compared to easy berthings and a light correlations between a self-assessment Likert scale and pupil dilation ($r=0.243$).

Recent studies by Čegovnik et al. (2018) and Hansen et al. (2018) has found that even a low-cost 30 Hz remote gaze tracker may provide reliable pupil measures that corresponds well with changes in task load during simulated driving and assembling of construction toy, respectively.

2.3. The pupil as an indicator of cognitive effort

The size of the pupil varies from 2mm and 8mm across different subjects and light conditions (Walker et al., 1990). Large variations (3mm) within subjects are caused by changes in light levels controlled by the pupillary light reflex, as reported by e.g. Ellis (1981).

However, Hess and Polt (1964) in a now famous experiment reported that pupil dilation could also be used as an index of mental activity during multiplication problems. This was based on earlier work (Hess and Polt, 1960) relating pupil size to emotional interest in presented material. Kahneman and Beatty (1966) subsequently confirmed this finding in a separate study, further suggesting that pupil responses are indicative of memory and processing load. This effectively (re)introduced pupillometry as a discipline (Beatty and Lucero-Wagoner, 2000; Stanners et al., 1979; Laeng et al., 2012).

Pupillary dilations in response to cognitive processes can be small (Beatty and Lucero-Wagoner, 2000) in less than extreme test conditions, often only around 0.5mm, which in typical conditions may amount to around 15% of the baseline pupil size. It is therefore
common to record several responses from repeated presentations of the same stimuli and conditions, and calculate the average of these in order to filter out noise from other processes. These are referred to as Task-Evoked Pupillary Responses (TEPR), a term probably coined by Beatty (1982) and Ahern and Beatty (1979).

Marshall (2002, 2007) has proposed a (patented) technology, the Index of Cognitive Activity, based on filtering abrupt pupil dilations using wavelet transforms to index and identify cognitively induced phasic responses from constrictions and dilations caused by the pupillary light reflex. In the present paper, we instead focus on the simpler concept of averaged means of the pupil size; a concept which reflects not only phasic responses but rather a combination of the current tonic level (arousal) and any phasic activations that take place. It may therefore serve as an index into a combined level of cognitive effort similar to what Hyönä et al. (1995) refers to as global processing load or the pupillometric estimate of mental load. Similar ideas have also been proposed elsewhere, e.g. Iqbal et al. (2004) percentage change in pupil size or Palinko et al. (2010) mean pupil diameter change.

3. Materials and Methods

3.1. Participants

57 children (54 boys and 3 girls) between 8 and 10 years-old (average 8.3 years) were recruited for the experiment at a LEGO® exhibition. Two were Swedish, the remaining 55 were Danish. Only one participant used glasses; none used contact lenses. All except five were frequent users of smartphones or tablets, having used one within the last week; 38 had used one the very same day. All participants had built a LEGO® model before, half within the last month. A majority of 51 had never used digital building instructions before; 6 had used them once or a couple of times. One of their parents signed a consent form, and after the experiment the child was given a LEGO® figure as appreciation for their participation. Non-invasive experiments on healthy subjects are exempt from ethical committee processing
by Danish law (Den Nationale Videnskabsetiske Komite, 2014). The study was conducted under the ethical standards of LEGO® A/S.

3.2. Equipment

LEGO® instructions served as our study material because they are sufficiently feature-rich and visually compelling to run naturalistic, engaging tests. The digital LEGO® application we used for our study showed model #42031, (Fig. 1), which has a total of 99 steps. Details can be examined by zoom and rotation; animations indicate how to place a component, and a forward/rewind slider helps find a particular event in the construction sequence. The Booklet used for comparison was the insert from the box, featuring 84 steps on 31 pages. This LEGO® model has a total of 163 pieces. Since there was a difference in steps between the Booklet and the Tablet we converted all Booklet steps into corresponding Tablet steps for the analysis of productivity (see below). An Apple Ipad®, model Air 2, with a "LEGO® Building Instructions" app turned on, provided the digital instructions.

A head-mounted tracker (Tobii model Glasses2 RU, 60 Hz, binocular) was used to record eye tracking data, and the Tobii Analyzer software, version 1.36.1430 (x64) was used to extract fixation data with corresponding pupil sizes. Visit times (e.i. view times) on manuals were scored manually, while the visit times on the Tablet were mapped automatically by the analyzer software. For automatic annotation of the videos, a snapshot image of the Areas of Interest (AOIs) are defined in the analyzer software. The automatic annotation did not work well for the Booklet so we used manual annotation in the Analyzer software. We reviewed the output of automatic annotation by the Analyzer software and fixed manually a few missing annotations. For both manual and automatic annotation, gaze fixations were detected automatically by the Analyzer software using the Tobii I-VT (attention) gaze filter (with a minimum fixation length of 60 ms). The pupil sizes reported by the tracker, as the diameter of the internal pupil in mm, is based on a 3D model of the eye and its position, and
is, according to the manufacturer\cite{footnote2}, a robust measure that does not require any additional calibration except for what is done for gaze points.

3.3. Procedure

Upon arrival, the children were seated in groups of four around one big table. Each child was randomly assigned to start building the LEGO® model with either a digital instruction on an Apple Ipad (N=28) or the Booklet that comes with the model used for the experiment (N=27). We told the child that the task was to build a model of LEGO® wearing a pair of glasses that would record what he or she was looking at. We emphasized that this was not any kind of competition – the gift would be given no matter how much of the model was completed by the end – and the child was told to ask for assistance if there were any problems.

A pair of Tobii Pro gaze tracking glasses was put on, and a one-point calibration procedure performed. Those starting with a Tablet version were given a short introduction to the main functions of it. Those starting with a Booklet were told to open the box and take out the building instructions. For both media we stressed that the child could go forwards and backwards as they pleased.

After 5 minutes of building, the adult assistant would note how far the child had progressed. When building using the Tablet, the child would at this point be shown how the slider feature for animation worked. Under the Booklet condition, the assistant would just ask the child to continue. After 10 minutes, the child was asked to pause for a short while for us to change the instructions media; a child starting with a Booklet would now try the Tablet version (given a short introduction) and a child starting with the Tablet would switch to the Booklet. The assistant made sure that the new instructions showed the step currently

in progress. If assistance was provided at any time during the building process, the assistant made a note of this.

The collection of eye data (gaze points and pupil size) continued uninterrupted throughout the procedure and between experimental blocks (booklet, tablet) for each participant.

At the end of the session, the participants were asked which of the two instructions they preferred and why. Also, they were asked whether they thought each of these were easy or difficult, slow or fast, and fun or boring (presented as forced choices).

3.4. Experimental design

Three independent variables were considered in the experiment:

Figure 2: A child wearing gaze tracking glasses while building LEGO® with Tablet instructions.
• Media: [Booklet / Tablet]

• Time: [First 10 minutes / Last 10 minutes]

• Novelty: [New: First five minutes with a given media / Known: Last five minutes with the same media presentation]

This creates in total 8 blocks.

When comparing data between identical sets of participants, within-group comparisons are used (e.g. related samples t-test or Wilcoxon signed-rank test), otherwise between-group comparisons are used (independent samples t-test or Mann-Whitney rank test).

Four dependent variables were examined:

• Productivity: Number of building steps produced

• Assistance: Number of times an adult assistant was called

• Gaze time: The proportionate time (in %) of the build time

• Pupil data: The mean pupil size (left and right eyes)

4. Results

We will first present some general results regarding performance and preferences and then a detailed analysis of gaze and pupil data.

4.1. Performance

Children were not significantly different in their productivity (steps progressed) between the Booklet (mean=18.4; SD=9.4) and the Tablet (mean=18.6 ; SD=10.1); p=0.47, Wilcoxon signed-rank Test. There were large individual differences, indicated by the high standard deviations. The fastest child built 71 steps in total; the slowest only 13 steps.
Productivity did not increase from the first 10 minutes (mean=18.6, SD=9.1) to the last, (mean=18.3, SD=10.4); p=0.30, Wilcoxon signed-rank Test.

There were no significant differences in the number of times children were assisted with the Tablet (mean=0.7, SD=0.9) and the Booklet (mean=0.9); p=0.25, Wilcoxon signed-rank Test). Assistance was not requested more often during the first 10 minutes (mean=0.9, SD=1.1 ) compared to the last 10 minutes (mean=0.7, SD=0.9); p=0.35, Wilcoxon signed-rank Test. There was a moderate statistically significant negative correlation between the number of times a child was assisted by an adult and their productivity, R= -0.65, p<0.001.

4.2. Preferences

When asked during the post session interview which of the two instruction forms the child preferred, 38 (66%) favoured the Tablet and 19 (33%) the Booklet. Table 2 shows how the children rated entertainment, ease and efficiency.

<table>
<thead>
<tr>
<th>Was it...</th>
<th>Tablet</th>
<th>Booklet</th>
</tr>
</thead>
<tbody>
<tr>
<td>fun or boring ?</td>
<td>55 (97 %)</td>
<td>36 (63 %)</td>
</tr>
<tr>
<td>easy or difficult ?</td>
<td>53 (93 %)</td>
<td>44 (77 %)</td>
</tr>
<tr>
<td>fast or slow</td>
<td>41 (72 %)</td>
<td>27 (47 %)</td>
</tr>
</tbody>
</table>

Table 2: Two-alternative forced choice subjective evaluation of Tablet and Booklet, N = 57

4.3. Gaze fixations

We recorded reliable gaze recordings of 48 of the 57 children for both the Booklet and the Tablet sessions, of whom 22 started with the Tablet and 26 started with the Booklet. Nine recordings failed due to extreme looking angles, hair dropping down and covering the eyes, or participants moving the glasses. Only full data sets are included in the analysis.

The children spent on average 52% of the time (SD=12%) studying the Booklet and 64% (SD=12%) studying Tablet instructions. The difference is significant (p<0.0001, Wilcoxon
Figure 3: Differences between fraction of time (in percent) with gaze at instructions between media (Booklet and Tablet; far left) and further split into first and second block with that particular media (middle), with the red line denoting one standard error of the mean to each side. There is a statistically significant difference (Wilcoxon signed-rank test, p<0.001) between the media types, as indicated by the ‡, and also between the first and second block for the Booklet (p<0.05), but not for the Tablet. There is also a significant difference (p<0.001) between the second block of Booklet vs Tablet. Also shown (right) is the average step time (seconds) for Booklet and Tablet. The difference is statistically significant (p<0.05).

signed-rank Test). The child who studied the Booklet least spent 18% of his time; the child who looked most spent 81%; the corresponding range for the Tablet was 34% to 86%. Comparing visit time per step produced showed a similar pattern, with Booklet being studied for 23.2s in each step (SD=18.3s) and the Tablet studied 27.7s (SD=19.9s); again this difference was significant; p=0.02, Wilcoxon signed-rank Test.

There was a low but statistically significant correlation between productivity and the time spent per building step looking at the Booklet (R= -0.37, p=0.01), and a high and statistically significant correlation for the Tablet (R= -0.71; p<0.0001, Wilcoxon signed-rank Test), meaning that the more the child had to look at the pictures for each step to decode them, the less was built by the end of that session.

Participants spent relatively more time looking at the Booklet instructions during the first 5 minutes than during the last 5 minutes (mean 54%, SD=15% versus mean 49%,
This difference, however, was not found for the Tablet condition (mean 64%, SD=15% versus mean 63%, SD=13%); p=0.71, Wilcoxon signed-rank Test. For the last 5 minutes, the difference between Booklet and Tablet was still highly significant, p<0.0001, Wilcoxon signed-rank Test).

We conducted an analysis on the fixation distribution for the Tablet condition, with a predefined set of areas of interest. The main attractor was the model with 66% of the proportional visit time during the first 5 minutes (SD=16%) and 64% (SD=16%) during the last five minutes.

4.4. Pupil data analysis

When the eye tracker reported fixations on the relevant instruction media, pupil data was collected for left and right eye and averaged over blocks of 5 minutes, corresponding to New (first 5 min) and Known (last 5 min) conditions for each of the two media, Tablet and Booklet. Subjects were divided into two groups according to the order of presentation, Tablet First or Booklet First. This creates in total 8 data blocks, with each subject contributing to 4 of these. The correlation between left and right eye was calculated for each subject, and subjects who had very low correlation, R<0.3, in any of the blocks were completely discarded (the resulting median correlation was R=0.85 across all remaining subjects). In addition, any subjects for whom we did not have reliable gaze data were discarded. This resulted in 43 remaining subjects, 23 having Booklet first and 20 having Tablet first.

The pupil data was initially made comparable between subjects and groups by normalizing to the relative deviation from a baseline of each subject’s overall mean, across all 4 blocks the subject was exposed to. This baseline was chosen as all subjects were exposed to the same set of media, only in different orders, and hence no single block could be assumed similar between the groups; using e.g. the first block as baseline would have meant that any (expected) differences between groups would then be carried over into later blocks.
Table 3: The average relative pupil size deviation for each block, in percent, with the standard error of the mean indicated in parenthesis. The same data is also shown graphically in Fig. 4.

<table>
<thead>
<tr>
<th></th>
<th>0-5 min</th>
<th>5-10 min</th>
<th>10-15 min</th>
<th>15-20 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Booklet</td>
<td>1.9 (SEM=0.3)</td>
<td>0.8 (SEM=0.4)</td>
<td>-0.2 (SEM=0.5)</td>
<td>-0.8 (SEM=0.6)</td>
</tr>
<tr>
<td>Tablet</td>
<td>0.5 (SEM=0.4)</td>
<td>0.5 (SEM=0.4)</td>
<td>-1.3 (SEM=0.3)</td>
<td>-0.9 (SEM=0.3)</td>
</tr>
</tbody>
</table>

Table 4: ANOVA model A and model B, with the dependent variable being the mean pupil size during each of the 4 · 5 minute blocks for each participant. Model A includes as independent variables time (first or last 10 minutes of the experiment), media (Booklet or Tablet) and interaction between media and novelty (first or last 5 minutes with the media). Model B includes additionally interaction with the performance.

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>F</th>
<th>p</th>
<th>$\eta^2$</th>
<th>$\omega^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Time</td>
<td>251.8</td>
<td>1</td>
<td>F(1,339)=39.1</td>
<td>0.000</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>Media</td>
<td>45.5</td>
<td>1</td>
<td>F(1,339)=7.1</td>
<td>0.008</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>Media:Novelty</td>
<td>36.2</td>
<td>2</td>
<td>F(2,339)=2.8</td>
<td>0.061</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>2183</td>
<td>339</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Time:Performance</td>
<td>361.1</td>
<td>2</td>
<td>F(2,336)=30.5</td>
<td>0.000</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td>Media:Performance</td>
<td>204.8</td>
<td>3</td>
<td>F(3,336)=11.5</td>
<td>0.000</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>Media:Novelty</td>
<td>36.1</td>
<td>2</td>
<td>F(2,336)=3.0</td>
<td>0.049</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>1991.8</td>
<td>336</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On average, data was captured for fixations that occurred from 49% (SD=14%) to 64% (SD=15%) of each 5 minute interval, when the participants were looking at the instructions; see Fig. 3 for additional details. The overall mean of the pupil size across participants was 5.4 mm (SD=0.7 mm). Table 3 and Fig. 4 shows the average relative deviation for each of the 8 blocks.

A three-way ANOVA analysis of the effect on the mean pupil size for each 5-minute block of the three independent variables Time, Media and Novelty (with interactions) was performed. There were statistically significant ($p<0.05$) effects of Time [$F(1,339)=39.1$, $p<0.001$] and Media [$F(1,339)=7.1$, $p<0.008$], with the effect of Media:Novelty not being statistically significant [$F(2,339)=2.8$, $p=0.061$]. Table 4 summarizes the results (Model A).³

The slowest group of the 43 participants, which had achieved less than 25 steps and had received assistance at least twice, were picked out and marked as slow performers. Similarly,

³It was tested whether individual preferences for either Booklet or Tablet stated after the experiment had an impact on the mean pupil size differences, but no effect was found [$F(1,338)=0.228$, $p=0.633$].
those that built more than 50 steps and received assistance at most once were marked as fast performers. This resulted in 11 fast performers and 12 slow performers, leaving 20 as normal performers.

A further four-way ANOVA analysis was performed that also included the effects of being slower to build on the mean pupil size. There were statistically significant effects of Time:Performance \([F(2,336)=30.5, p<0.001]\) (the effect of Time differed between the performance groups), Media:Performance \([F(2,336)=11.5, p<0.001]\) (the effect of Media differed between the performance groups). Media:Novelty barely had a statistically significant effect \([F(2,336)=3.0, p=0.049]\). Table 4 summarizes the results (Model B).

Comparing the two Models A (the simpler model) and B (which also includes slow performance) further reveals differences in the effect \(\eta^2\) caused by the independent variables. The total explained variance, \(R^2\) of Model A is 0.138 (adjusted \(R^2\) 0.128) whereas Model B has an explained variance of 0.214 (0.197). The AICs for the two models are 1622 and 1596, indicating Model B may be the better fit even if it has less degrees of freedom. Redoing the same two models A and B using the absolute pupil dilation (mm) from the baseline instead of the relative (percentage-point) gives very comparable model results and a total explained variance of A: 0.141 (0.131) and B: 0.204 (0.188). For this reason, we have reported only the relative (percentage-point) pupil dilation. Note that with the reported average pupil size of 5.4 mm, 1 percentage-point corresponds to an absolute pupil dilation of approximately 0.054 mm.

The residuals resemble a normal distribution, although with some outliers and a slightly higher kurtosis. For this reason, the more conservative non-parametric tests will be used for the Post-Hoc tests\(^4\).

Non-parametric Post-Hoc tests were performed between the blocks where it was meaningful. The results are shown overlayed the descriptive statistics in Fig. 4, where the predictions

\(^4\)The statistical significance levels are however very comparable if using parametric t-tests.
Figure 4: The average relative pupil size deviations for each block and Post-Hoc tests between these. The bars show the average relative pupil size deviation for each block, colour coded according to the two groups of participants that had Booklet first (blue) and Tablet first (green). The red vertical lines indicate one standard error of the means to each side. The predicted value is additionally shown as a dashed line. The Post-Hoc between the means of the data from each of the blocks of the ANOVA regression is indicated with corresponding significance levels marked: Each line above the bars indicate where there are statistically significant differences, with solid lines indicating within-group and dotted lines between-groups. The markers α indicate uncorrected significance, h corrected with Holm-Bonferroni factor, and b the most conservative with Bonferroni correction. The levels are at p<0.05 unless marked with a † or ‡, that indicates 0.01 and 0.001 respectively.

of Model A are also shown. When the two blocks contain the same set of subjects, within-group (Wilcoxon signed-rank) tests were used, otherwise between-group (Mann-Whitney rank) tests were used.

Looking at the underlying regression model further illustrates the effects of Time, Novelty and Media on the mean pupil size for Model A and B (divided into slow and not-slow performers). See Table 5.

For Model A, the mean pupil dilation is 1.71 percentage-points lower in the second half of the experiment (the probability \( p(H_0) \) of the effect being 0 is \( p<0.001 \)) compared to the first half. The last 5 minutes (Known) of using the Booklet has a mean pupil dilation of 0.89 (\( p=0.022 \)) less than the first 5 minutes (New); there is a smaller effect of 0.22 the other
Table 5: Correction terms (generated by the regression) for Model A and Model B. The numbers are percentage-points that can be summed when estimating the model mean pupil size for a given block of the experiment. The * indicates that the term differs from 0 at a significance level of $p<0.05$. The ‡ indicates a $p<0.001$ significance level. Note that Model B, when compared to Model A, should also include an additional factor of -0.08 to be added to the baseline for the slowest performers (not shown in the table).

way round for the Tablet, but this is not statistically significant. Finally, comparing the use of a Booklet to the Tablet shows a mean pupil dilation of 1.28 ($p=0.001$) less for the Tablet.

For Model B, that includes the effect of being a slow performer, there are differences between how the mean pupil dilation changes over conditions: For the slow performers, there is a small and statistically insignificant difference between the first 10 minutes and the last 10 minutes of the experiment, but for the rest (normal and fast performers) there is a decrease in mean pupil dilation of 2.41 ($p<0.001$) percentage-points. The Novelty effect is similar to the above (-0.89 ($p=0.017$) and insignificantly 0.22 respectively), but the differences between the use of a Booklet and Tablet are different: For the slowest performers, the Tablet has a 3.02 ($p<0.000$) percentage-point smaller mean pupil dilation compared to the Booklet, whereas the rest of the subjects have a smaller (and insignificant) difference of only 0.72 ($p=0.077$).

Further, comparing the measured mean pupil dilation of the slowest performers to the normal performers revealed a statistically significant ($p<0.05$ after Bonferroni correction) difference of approximately +4% between the levels during the 10 min on the Booklet after the Tablet.

Fig. 5 further illustrates the overall trends of the mean pupil size across subjects for
Figure 5: Filtered trend of how the mean pupil size develops over time for the two experimental groups of participants that had Booklet first (blue) resp. Tablet first (green). The graph is the result of filtering the data for each participant using a 2nd order polynomial (i.e. a heavy low-pass filtering) and then calculating the means over all of these. The shaded area represents one standard error of the mean to each side.

the two groups Booklet-first and Tablet-first. The graph has been created by fitting a 2nd order polynomial to each subject’s relative pupil dilation and averaging across subjects. The Booklet-first group starts around 3% higher than the mean value, and declines down to around 1% less than the mean. The Tablet-first group shows a more flat curve.

5. Discussion

We found statistically significant differences in pupil dilation between the eight blocks of the experiment (Table 4) and different trends depending on which media was used first (Fig. 5). Post-hoc analysis (Fig. 4) and modelling (Table 5) showed modest but statistically significant effects caused by the independent variables Time, Media and interaction between Media:Novelty. This was further modulated by being a slow performer.

Separating the effects, we found that the pupil constricts close to 2% from the first to the second half of the experiment, although the slowest performers did not show such an
effect. We also saw a difference between the two media, Booklet and Tablet, with the Tablet leading to an approximately 1% smaller pupil size, although interestingly we found that the slowest performers had a larger difference of close to 3% between the two media.

With a careful balanced experimental design, this suggests that even in "in-the-wild" conditions like here, where it is not possible to control for variations in light levels, it may be possible to use pupil size as a measurement of cognitive effort: Even if it cannot be ruled out that different luminosity levels impact the pupil size, the identification of novelty effects (changes between first and second block of using the Booklet but not significantly so for the Tablet) and the differences in effect size of Time and Media between slow performers and the rest of the group suggests that the simpler explanation is that these differences are linked to levels of cognitive effort: 1) Those that experience a slow progress keep investing a higher level of effort and attention compared to the rest of the group, and 2) they also appear to benefit more from using the Tablet instead of the Booklet.

Our main motivation for conducting the analysis of the children’s pupillary response during the experiment was to investigate whether this could yield a deeper insight to the experiences they had while building with the Booklet and the Tablet. The majority of the children (i.e. 67%) preferred the Tablet to the Booklet, and it was generally considered more fun, easy and fast to use (cf. Table 2). The verbal explanations they gave for their preference for the Tablet were that they liked the rotations and animations, that it was something new, and that it was so easy to get to the next step with no need to turn pages. Those who liked the Booklet explained that it was a format familiar to them and the Tablet was annoying to use because it required them to press buttons all the time. Performance data gave no indications of this difference in preferences, as productivity and the number of times the children were assisted by an adult were not significantly different for the two types of instructions.

Preferences for one of the two forms of instructions did not reveal itself in different pupil
patterns, c.f. section 4.4 on page 14. However, the pupil trend curve (Fig. 5) suggests that children starting with the Booklet had to show relatively more effort in the beginning of the experiment than those who started with the Tablet. This is further confirmed when comparing the first 5 minutes and last 5 minutes of each block (Fig. 4). Across the full session, when participants became more acquainted with the building task, this difference diminished. The group of children who had started with the Tablet did not show any noteworthy increase in pupil dillation when they later encountered the Booklet. Being deeply involved in the building task at this time, they would just continue at the same level of effort as with the Booklet. One group, however, did not follow this general pattern, namely the slow builders starting on a Tablet. They began the experiment with a comparatively smaller relative pupil size, but then, when the Booklet was encountered, they showed an increase in pupil size that was significantly different from the other groups. Slow builders starting with a Booklet and then trying the Tablet, did not show this increase in pupil size. This finding suggests that monitoring changes in pupil size may be used in the evaluation of a learning process, c.f. (Kaklauskas et al., 2013) and may help identifying tools for learning that support challenged students best.

The participants looked for a relatively longer time in total at the Tablet instructions (i.e. 64%) than the Booklet (i.e. 52%). This difference may have be explained by the fact that the Tablet version showed slightly more individual building steps (i.e. 99) than the Booklet (i.e. 84), but the difference was also significant when analysing the relative attention time per normalized step. Another explanation might have been that children had to look for the forward button on the Tablet whenever they wanted to go to the next step. Folding a page every now and then in a Booklet, which showed 4 steps per page-turn, would require less attention to doing this. However, an analysis of visit time on the Tablet showed that only 3 % of time was spent looking at the forward and backward buttons. Manual inspections of gaze recordings revealed that most of the button clicks were performed when the child
was looking at the model area, i.e. in the fringe of the vision. Presumably, the main reason for them to look more at the Tablet was the time it took for it to play animations and the time children spent rotating, zooming and sliding the model. The time spent on doing this was not counter-productive, and it may have contributed to the positive impression of the Tablet.

If future displays of digital instructions - on tablets or head mounted displays - get eye tracking capabilities, we imagine that this would make it possible for designers at companies like LEGO® or IKEA® to have a test panel of e.g. 100 people, who would try out various versions of instructions and various new models at home. Logging their use of the instructions would allow for measurements of time spent on each step of a task, regressions, visit time on AOIs in the instructions, time-to-first-fixation on AOIs, dwell time, saccade lengths and fixation patterns. In addition, the experiment reported in this paper suggests that pupil trend curves may reveal patterns in mental effort across the task. When recording pupil variations from a large population, a normal model prediction could be built for every product, identifying e.g. those that were unusually demanding. LEGO®, for instance, could consolidate its recommendation of the age group suitable for that particular model. Furthermore, if future displays become able to record the single areas looked at for every image shown, the pupil analysis may become even more precise, taking into account the brightness of the fixated area. Most importantly, this would allow for an analysis of the within-task variation in task demands that single pieces of information might cause.

6. Conclusion

We compared digital and paper instructions for LEGO® and found significant differences in children’s preferences and gaze behaviour for the two formats, but no difference in their task performance between the two media. Pupil measurements were in accordance with the participants’ impression of the digital version being easy to use, indicated by a smaller
relative dilation in pupil size when encountering the Tablet as the first media, compared to the dilation found for those encountering the Booklet first. Children with low performance had a different pupil pattern associated with the the Booklet than the rest, while all children showed a decrease in pupil size across the experiment.

Acknowledgment

The authors would like to thank the Danish Innovation Foundation, who provided part of the funding, and Olav Gjerulfsen and Thomas Visby Snitker, LEGO®, Henning Boje Andersen, Diako Mardanbegi, Lars Yndal Soerensen, Emil Andersen, DTU and Henrik Skovsgaard, The Eye Tribe.

Bibliography

URL http://www.sciencemag.org/cgi/doi/10.1126/science.472746


URL http://www.tandfonline.com/doi/abs/10.1080/14640749508401407


URL http://dx.doi.org/10.1007/BF00994048


