What is the best way for developers to learn new software tools? An empirical comparison between a text and a video tutorial

Tutorials for software developers are supposed to help them to adapt to new tools quickly. While in the early days of computing, mostly text tutorials were available, nowadays software developers can choose among a huge number of tutorials for almost any popular software tool. However, almost no research was conducted to understand how text tutorials differ from other tutorials, which tutorial types are preferred and, especially, which tutorial types yield the best learning experience in terms of efficiency and effectiveness.

To evaluate these questions, we converted a “proven” video tutorial for a novel software tool into a content-equivalent text tutorial. We then conducted an experiment in three groups where 42 undergraduate students from a software engineering course were commissioned to operate the software tool after using a tutorial: the first group was provided only with the video tutorial, the second group only with the text tutorial and the third group with both.

Surprisingly, the differences in terms of efficiency are almost negligible: we could observe that participants using only the text tutorial completed the tutorial faster than the participants with the video tutorial. However, the participants using only the video tutorial applied the learned content faster, achieving roughly the same bottom line performance. We also found that if both tutorial types are offered, participants clearly prefer video tutorials for learning new content but text tutorials for looking up “missed” information. So while it would be ideal if software tool makers would offer both tutorial types, we think that it is more efficient to produce only text tutorials – provided you manage to motivate developers to use them.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

What Is the Best Way For Developers to Learn New Software Tools?

An Empirical Comparison Between a Text and a Video Tutorial

Verena Käfer, Daniel Kulesz and Stefan Wagner

Software Engineering Group, Institute of Software Technology, University of Stuttgart, Germany

ABSTRACT

Tutorials for software developers are supposed to help them to adapt to new tools quickly. While in the early days of computing, mostly text tutorials were available, nowadays software developers can choose among a huge number of tutorials for almost any popular software tool. However, almost no research was conducted to understand how text tutorials differ from other tutorials, which tutorial types are preferred and, especially, which tutorial types yield the best learning experience in terms of efficiency and effectiveness.

To evaluate these questions, we converted a "proven" video tutorial for a novel software tool into a content-equivalent text tutorial. We then conducted an experiment in three groups where 42 undergraduate students from a software engineering course were commissioned to operate the software tool after using a tutorial: the first group was provided only with the video tutorial, the second group only with the text tutorial and the third group with both.

Surprisingly, the differences in terms of efficiency are almost negligible: we could observe that participants using only the text tutorial completed the tutorial faster than the participants with the video tutorial. However, the participants using only the video tutorial applied the learned content faster, achieving roughly the same bottom line performance. We also found that if both tutorial types are offered, participants clearly prefer video tutorials for learning new content but text tutorials for looking up "missed" information. So while it would be ideal if software tool makers would offer both tutorial types, we think that it is more efficient to produce only text tutorials – provided you manage to motivate developers to use them.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

1 INTRODUCTION

Typical developers have to work with many different tools every day. While many developers get frustrated when the tools' user interfaces do not match their expectations [15], most developers manage to come to terms with what they get. However, as technology evolves, developers are expected to get along with new tools quickly. Thus, the learnability and understandability of a software tool is an important success factor [7, 8]. Some software tool producers believe to tackle these issues appropriately by providing tutorials for the software. However, as Martin et al. have shown, tutorials are often unavailable, incomplete or focusing on the wrong aspects [18].

While in the early days of computing mostly text tutorials were available, advances in technology made it possible to easily produce video tutorials and even interactive tutorials. Van Loggem's research indicates that the "classic" written (and printed) manuals usually are not the first choice – at least for most tool end-users today [17]. Instead, most tool users prefer interviewing colleagues or searching for tutorials in online sources, and especially the latter offer many different kinds of tutorials to choose from. Furthermore, many tutorials are not crafted by the original makers of the software or hired tutorial producers but by other tool users who post such tutorials on blogs or social media platforms.

1.1 *Problem Statement*

After reviewing the literature we were surprised that there is almost no insight available on how text tutorials differ from other tutorials, which tutorial types are preferred by developers and especially which tutorial types yield the best learning experience in terms of efficiency and effectiveness. This is bad because developers are overburdened with too many choices while software tool makers do not know on which type of tutorials they should spend their (typically) limited resources.

1.2 *Research Objective*

To address the issues described in our problem statement, we formulated the following research questions:

- RQ1: What kind of tutorial do developers prefer if both text and video tutorials are available?
- RQ2: Which tutorial takes learners less time?
- RQ3: Which tutorial is more effective?

1.3 *Context*

The study was conducted using students from an undergraduate software engineering course. We used an experimental software for testing spreadsheets which is implemented as an add-in for Microsoft Excel [19]. Since both the tool and its underlying approach are novel, learning and understanding them requires adequate tutorials. Therefore, results obtained from this study should be considered especially in contexts where developers are confronted with new software tools which are not directly mappable to previous experience.

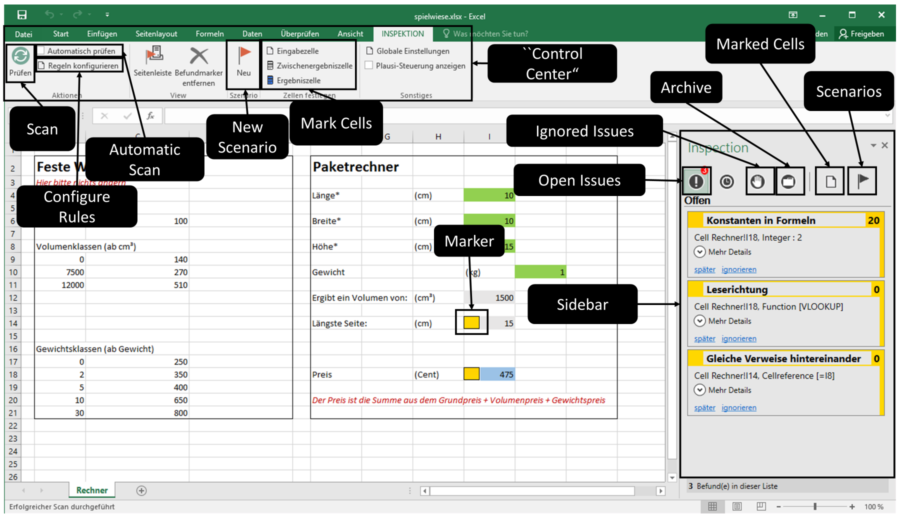
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

This paper follows the structure proposed by Jedlitschka et al. [10] with a slight deviation (we describe the procedure of the experiment earlier than proposed by Jedlitschka et al. and merged the alternative technologies and the related studies): after describing the background of the study, we explain the plan of the experiment and its deviations, we present the obtained results and discuss them before drawing a final conclusion and outlining future work. The graphical representations are based on the recommendations by Tufte [21].

2 BACKGROUND

2.1 *Technology under Investigation*

The *Spreadsheet Inspection Framework* (SIF) [12] is a software application for detecting faults in spreadsheets which has been developed at the University of Stuttgart and already has been evaluated in a number of previous studies [11, 14, 13]. It is operated through a plug-in for Microsoft Excel which allows end-users to automatically scan spreadsheets for "bad smells" (e.g. formulas referencing the same cell twice) or violations of design rules (e.g. formulas with constants). Apart from running these pre-defined scans, it also allows end-users to specify their own test scenarios which are comparable to unit tests for normal software applications. While SIF is primarily targeted at end-users, we expected developers to understand its concepts faster as they are similar to static analysis and unit tests in traditional software development.



Feste W
Hier bitte nichts ändern

	100
Volumenklassen (ab cm³)	
0	140
7500	270
12000	510
Gewichtsklassen (ab Gewicht)	
0	250
2	350
5	400
10	650
30	800

Paketrechner

Länge*	(cm)	10
Breite*	(cm)	10
Höhe*	(cm)	15
Gewicht	(kg)	1
Ergibt ein Volumen von:	(cm³)	1500
Längste Seite:	(cm)	15
Preis	(Cent)	475

Der Preis ist die Summe aus dem Grundpreis + Volumenpreis + Gewichtspreis

Konstanten in Formeln 20
Cell Rechner!I18, Integer : 2
Mehr Details
später ignorieren

Leserichtung 0
Cell Rechner!I18, Function [VLOOKUP]
Mehr Details
später ignorieren

Gleiche Verweise hintereinander 0
Cell Rechner!I14, Cellreference [=I8]
Mehr Details
später ignorieren

3 Befund(e) in dieser Liste

Figure 1: Main user interface of the Spreadsheet Inspection Framework

A screenshot of the main user interface of SIF is provided in Fig. 1. It shows that SIF adds a new tab on the ribbon bar which acts as SIF's "control center". Here, the user can configure the rules to be used for testing spreadsheets, create new test cases and start the automated scans of the spreadsheet. The sidebar on the right side provides the user with an overview about open, postponed, ignored and solved findings. Furthermore, the issues reported in the sidebar are synchronized with marker symbols in the corresponding cells of the spreadsheet. Although none of these elements of the user interface is overly complex, learning how to use them all in just a few minutes can be challenging for end-users – making this setting a proper ground for evaluating the efficiency and effectiveness of different tutorials.

For the purpose of another (yet unpublished) study, the second author has produced an introduction video and two video tutorials (educational screencasts) which explain the ideas behind SIF and how to use it. Since we regard comparing tutorials with different (depth of) content as not being fair, the first author simply converted the video tutorials into content-equivalent text tutorials, making them directly comparable. This also ensured that the videos had the same origin and thus were comparable to "official tutorials" produced by an original software maker (unlike tutorials made by hired tutorial producers or third parties).

2.2 *Related Studies*

Overall, there are only few studies which investigated tutorials for software. They can be divided into two groups: studies which investigated game tutorials and studies which focused on "usual" software applications.

In the first group, Andersen et al. investigated game tutorials and their effect on different aspects in games, such as the effect on game learnability or the effect of optional challenges [3, 1, 2]. Their results show that tutorials only have an effect for more difficult games but not so much for easy games. Additionally, optional challenges that players may or may not fulfill cause more harm than benefit whereas music or sound effects have no influence. Animations, on the other hand, make the users play more.

Several studies by Backer et al. investigating tutorials for application software support the thesis that dynamic visual representation is better than traditional static text [5, 6]. Additionally, studies by Harrison indicate that visual online help is more effective than non-visualized texts and that written help is preferred over spoken instructions [9]. Van Loggem adds that most users prefer other tutorial types over text tutorials [17].

The study described in this paper borrows its design and task descriptions from another study conducted in early 2016 by the second author (the study has not been published yet). However, it is not a pure replication: while the original study only used video tutorials to explain SIF, this study also used content-equivalent text tutorials which were produced for this very study. Furthermore, the participants were asked more questions about their perceptions on the tutorials they consumed, while this aspect was not targeted in-depth in the original study.

2.3 *Practical Relevance*

The better users can learn software tools, the faster they can start actually using it and the more efficiently they can later work with it. Thus, good learnability is a vital success factor for software tools. Therefore, knowing with which type of tutorial learnability can be increased has an immense practical relevance.

Furthermore, producing tutorials also requires effort. In our own experience, we perceived producing video tutorials to be far more difficult than producing text tutorials (the second author worked about 80 hours to produce three video tutorials with a total length of less than 25 minutes). Therefore, knowing if video tutorials actually have advantages in terms of learnability is also relevant for practice.

3 EXPERIMENT PLANNING

3.1 *Goals*

We broke the main research objective down into three smaller goals (derived from the original research questions) to compare text and video tutorials with each other.

- Goal 1: Analyze video and text tutorials
  For the purpose of understanding which tutorial type is preferred
  With respect to the preference rate of the users for each type
- Goal 2: Analyze video and text tutorials
  For the purpose of comparing their efficiency
  With respect to the time required by users to complete the tutorials
- Goal 3: Analyze the quality of the video and text tutorials
  For the purpose of comparing their effectiveness
  With respect to counting how often users looked something up, which percentage of the learning content they understood right away, the number of wrong actions they took and with respect to measuring the perceived difficulty level of the tutorials

3.2 *Participants*

To successfully pass one of the undergraduate software engineering lectures, all students were required to take part in one study, having the choice between three other studies and ours. We decided to use this population primarily because it was easily accessible. There were 42 participants and they were about 20 years old. 9 of them were female and 33 were male.

It is remarkable that all slots for our study were booked out 30 minutes after the start of the booking time, while the other studies had open slots even several days after the start – especially when taking into account that our experiment had the longest duration of the four (120 minutes compared to 90 and 60 minutes). We therefore assume that the participants had a higher motivation than just being forced to take part in this experiment as (a) they had the choice and actively decided to take our experiment and (b) because the experiment was booked out so quickly.

We gave the participants the choice between 14 time slots for doing the experiment, reserving them on a "first-come-first-serve" basis. Once three participants reserved a time slot, the time slot was full because we ran three parallel experiments in each time slot (nevertheless, the experiments were strictly separated as described in section 3.3).

Each participant had to fill out an online questionnaire a few days before the experiment. Based on the completion order of this questionnaire we assigned each participant a unique anonymous ID.

To ensure confidentiality, each participant had to sign an agreement at the beginning of the experiment about not disclosing any details of the experiment to other participants. Additionally, we stressed this point at the beginning of each experiment and explained the motivation behind this. All participants were allowed to abort the experiment at any time if they wished but no participant used this opportunity.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

3.3 Procedure

Figure 2 provides an overview of the procedure in our study. The actual steps where participants use tutorials are highlighted in light gray. However, we simplified the "final test" in this figure to make it more concise – in reality, some parts of the questionnaire were dependent on the type of experiment the participant had.

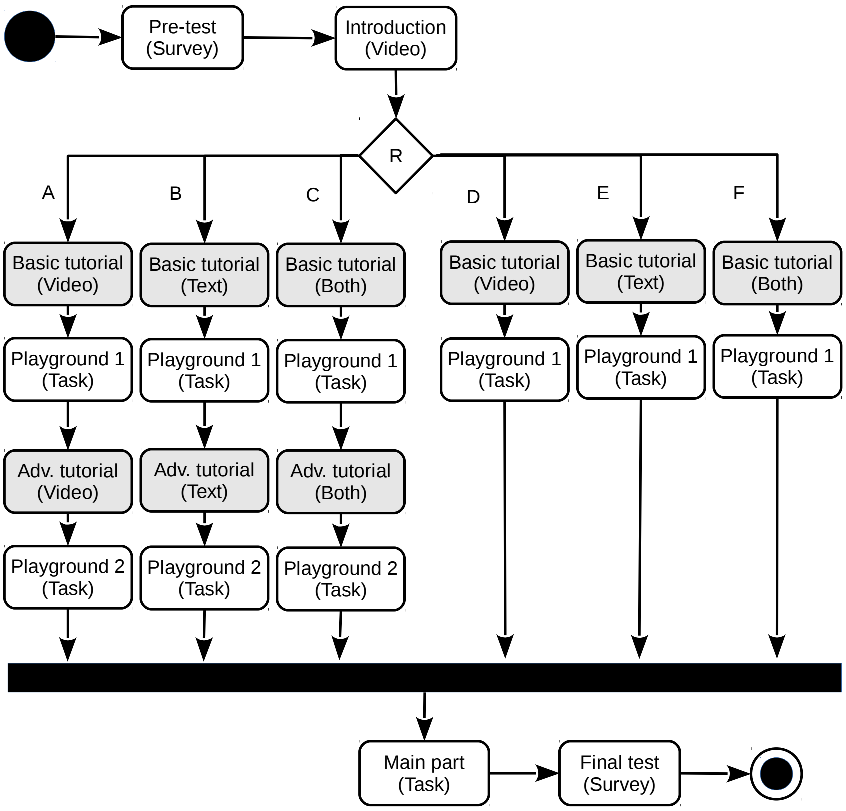


Figure 2: Overview of the experiment procedure (slightly simplified)

As it can be seen, we randomly divided the participants into six groups. Every group contained seven participants.

- A: with scenarios, video only
- B: with scenarios, text only
- C: with scenarios, video and text
- D: no scenarios, video only
- E: no scenarios, text only
- E: no scenarios, video and text

All participants had to watch the introduction video. After using their particular tutorial, all participants solved the first task (Playground 1). After finishing the first task, the participants of groups A, B, and C directly continued with the main task. By contrast, participants of the other groups first

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

had to do another round of tutorials in which they learned the (advanced) scenario testing technique and solved simple tasks with it (Playground 2) before continuing with the main task.

Since each of our time slots had the capacity for three participants, we either executed experiments A, B and C or D, E and F (the idea behind this was not to give participants the feeling that they are slower than others if they see that others who started at the same time had already finished). The particular assignment to A, B or C or D, E or F respectively was based on the order of arrival of the participants.

Before the actual start of the experiments, we explained to each group of three participants the whole setup and what they had roughly to do, not giving them any details except “you will get an introduction where everything will be explained”. The experiments took place in one of our computer pools with air-conditioning, on standard desktop computers with 23” monitors with a full HD resolution. The “work-places” of the participants were close to each other but we put separating walls in-between as shown in figure 3, so the participants could only see their individual screen.

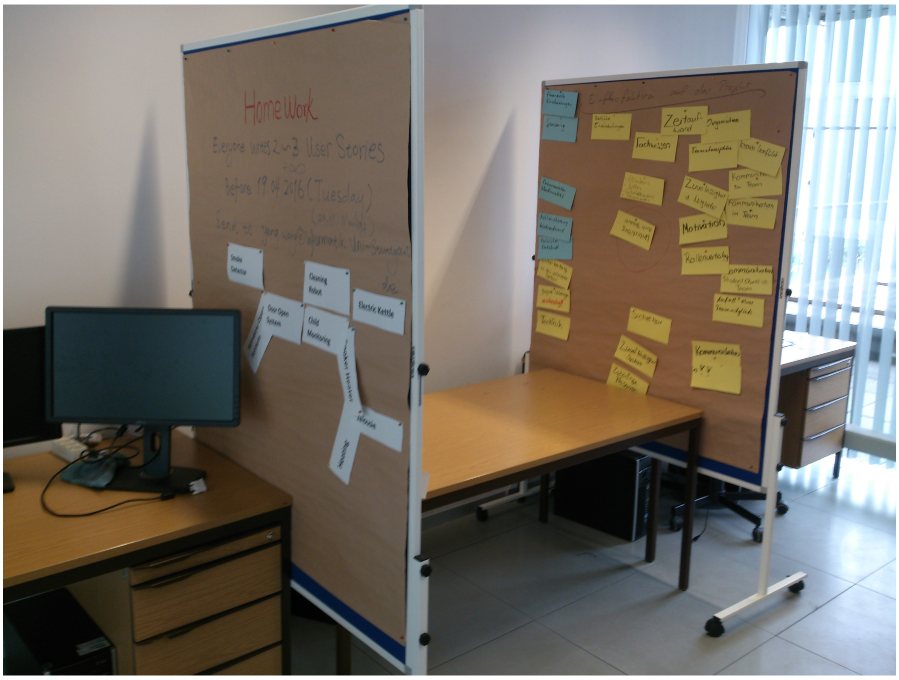


Figure 3: Our computer pool where the experiments took place

For later examination, we recorded all experiments using a screencast recorder (due to legal privacy restrictions we recorded only the computer screen and not the participants). Also, we watched the participants’ screens over VNC connections during the experiment thus avoiding to directly look over their shoulders so as to not make them feel uncomfortable. This was a tough trade-off, as it also made it impossible for us to count how many times the participants looked something up in the text tutorials ourselves.

There were some common issues which most participants encountered, such as not knowing how the VLOOKUP formula works in Microsoft Excel. In such cases we gave the participants “meta hints” like ”maybe you could try to look this up on the Internet?”. SIF also has several known bugs and we were aware of them already before the experiments (they are not straight forward to fix). If one of the participants encountered such a bug we helped them get around it. We took extreme care to not disturb the other participants in such cases as we did not want to influence them.

3.4 *Experimental Material*

As described in section 2.1, the participants of our experiment had to learn how to use the *Spreadsheet Inspection Framework* (SIF). In Fig. 12 in appendix A a frame of the video explaining basics of the SIF can be seen. Like in this frame, the video uses graphical annotations such as blue borders and arrows for highlighting particular parts of the user interface. The text tutorial counterpart for this excerpt of the video is illustrated in Fig. 13 in appendix A. As it can be seen, similar highlighting techniques have been used in the text to explain which actions the user has to take to reach a certain functionality (in the shown excerpt: clicking on a button in the ribbon bar to open the sidebar).

In the pre-test, we asked several questions using an online questionnaire to investigate the participants' experience and opinions. The pre-test was made up of a questionnaire which contained questions about prior knowledge of Microsoft Excel and what tasks the users typically solve using it, if any. During the experiment, each participant received printed instructions on what to do next (obviously, the instructions differed depending on which experiment the participant had). The instructions also contained questions about the last task they completed, so the participants worked their way through the instructions, alternating between reading the instructions, doing activities on the computer and writing down answers on the instruction sheets. An example of the instructions is shown in Fig. 4.

Apart from the instructions, we also provided the participants with the (printed) text tutorials, the tutorial videos or both (depending on the group). After finishing all practical tasks, each participant was asked to fill out a paper questionnaire. Here, we asked more questions about their background, what they liked and disliked about SIF and, of course, about the tutorials.

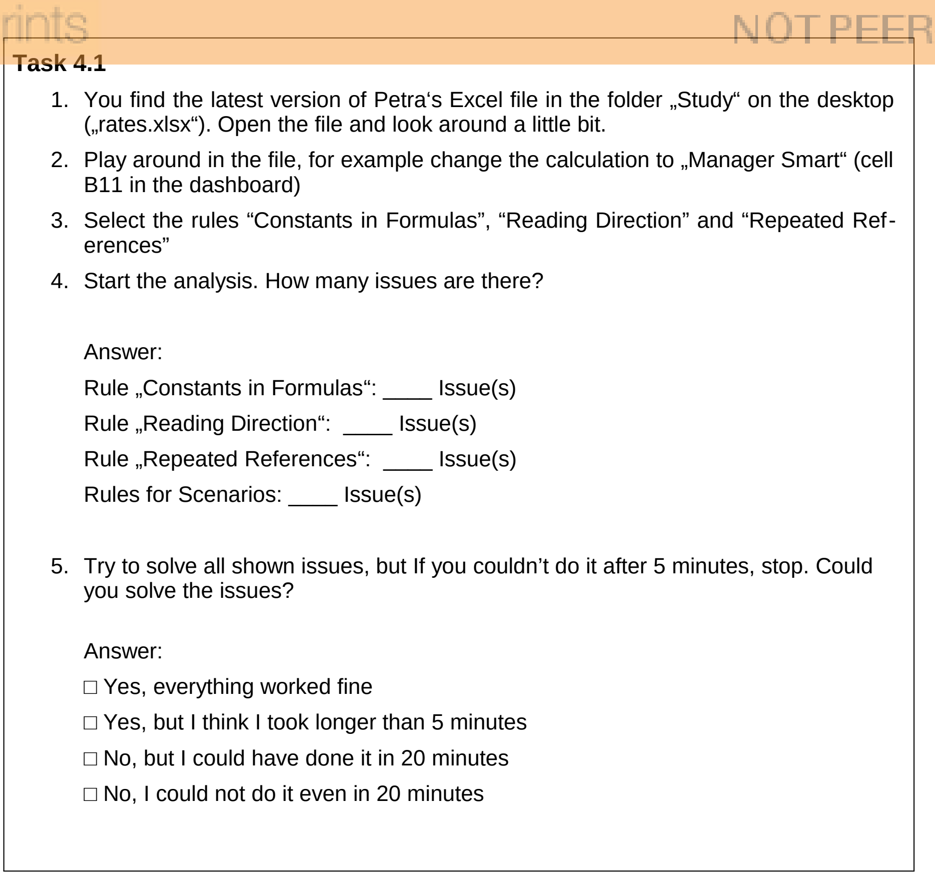
Last but not least the participants were provided with either two or three spreadsheets (depending on the group) which contained a number of seeded errors in order to solve the actual tasks.

3.5 *Tasks*

Depending on the group of the participants they had to perform either two or three bigger tasks during the experiment. These bigger tasks were split up into a series of small sub-tasks.

The first bigger task (Playground 1) was to apply the basic functions of the SIF. First, the participants were asked to open a spreadsheet, to activate several static rules and to initiate a scan. For this we prepared a spreadsheet that could be used to calculate the price of a parcel based on its size and weight. If done right, SIF reported for this spreadsheet an issue with the reading direction which could be solved by moving the content in one of the cells. The second issue reported by SIF was that one formula referred to the same cell multiple times using the MAX function (The formula was: MAX(I4;I6;I8;I8;I8;I8;I8;I8;I8). This issue could be solved by removing the obsolete references (it makes no sense to refer to the same cell multiple times in a formula which is composed only of the MAX function).

The second task (Playground 2) was only performed by participants in the groups that had received the advanced training where they were taught how



Task 4.1

1. You find the latest version of Petra's Excel file in the folder „Study" on the desktop („rates.xlsx"). Open the file and look around a little bit.
2. Play around in the file, for example change the calculation to „Manager Smart" (cell B11 in the dashboard)
3. Select the rules "Constants in Formulas", "Reading Direction" and "Repeated References"
4. Start the analysis. How many issues are there?

   Answer:

   Rule „Constants in Formulas": ____ Issue(s)

   Rule „Reading Direction": ____ Issue(s)

   Rule „Repeated References": ____ Issue(s)

   Rules for Scenarios: ____ Issue(s)

5. Try to solve all shown issues, but If you couldn't do it after 5 minutes, stop. Could you solve the issues?

   Answer:

   ☐ Yes, everything worked fine

   ☐ Yes, but I think I took longer than 5 minutes

   ☐ No, but I could have done it in 20 minutes

   ☐ No, I could not do it even in 20 minutes

Figure 4: Translated instructions for the final task

to use the SIF scenario testing technique. The participants were first asked to create a new scenario on their own and then to use it to find errors in a new spreadsheet. Again we used the example of the parcel price. The create scenario reported an issue because for the final price calculation the weight of the parcel was subtracted. The participants solved this by repairing the formula.

The last and lengthiest task was to extend a given spreadsheet by adding new data and features. The given spreadsheet calculated the monthly bill for using a mobile phone based on a user's consumption of minutes and text messages (with different rates based on the destination) for different tariffs (which had different rates for these minutes and text messages). It also featured a dashboard where monthly bills could be compared between the tariffs to find the cheapest one. The participants first had to add a new rate and then add the costs for on-net texts for every rate.

3.6 *Hypotheses*

We wanted to test the following hypotheses:

- H_{0_1}: There is no difference in how often the video tutorial or the text tutorial are chosen.
- H_1: The video tutorial is chosen more or less often than the text tutorial when both are provided.

- H_{0_2a}: There is no difference in the time needed to complete the tutorials.
- H_{2a}: The three tutorial groups differ in the time they need to complete the tutorials.
- H_{0_2b}: There is no difference in the time the three tutorial groups need to complete the tutorials and the following tasks.
- H_{2b}: The three tutorial groups differ in the time needed to complete the tutorials and the following tasks.
- H_{0_3a}: There is no difference in the amount of items the participants look up in the tutorials.
- H_{3a}: The participants look up more or less items in the video tutorial than in the text tutorial.
- H_{0_3b}:The participants differ in their understanding of the video and the text tutorial at the first attempt.
- H_{3b}: There is no difference in the understanding of the video and the text tutorial at the first attempt.
- H_{0_3c}: There is no difference in the number of wrong answers given during the tasks.
- H_{3c}: The three tutorial groups differ in the number of wrong answers given during the tasks.
- H_{0_3d}: There is no difference in how difficult the participants thought the tutorial was.
- H_{3d}: There is a difference in how difficult the participants thought the tutorial was.

  There is a difference in how difficult the participants thought the tutorial was.

3.7 *Analysis Procedure*

To analyze the results, we examined the screen recordings and the questionnaires. In the screen recordings, we measured when a participant started a task (when he or she opened the Excel file) and when the participant finished the task (closing the file). The time between one file and the next one was the time used for the tutorial. We measured the duration in seconds and combined the groups with the same used tutorial for the SIF part.

For evaluating the correctness of the results, we designed unit tests with input and expected output values – one unit test for each playground and two unit tests for the final task. We had two unit tests for the final task because the final task included two tasks – the the new rate and then the onnet texts. We then filled in the input values and checked if the actual output values matched the expected values. We judged a unit test for a spreadsheet to be correct if and only if *all* actual output values matched their expected values.

For the questionnaires, we first used LimeSurvey [16] to produce an electronic version of the questionnaires. Then we exported them to R [20] where we evaluated them.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

As the sample groups were quite small, we used t-tests to see if the group results were significantly different. For this, we chose a significance level of 0.05. Additionally, we measured the effect size with Cohen's d and the absolute mean difference MD.

To measure the normal distribution in section 5.3.1 we used a Shapiro-Wilk test due to a small sample size with a significance level of 0.05.

4 EXECUTION

4.1 *Preparation*

Before the participants arrived, we prepared the computers and started the screen broadcast over the VNC connection. For the participants, there was no particular preparation as all required information was provided during the experiment via the instruction sheets and the particular tutorial. Therefore, we did not ask the participants to prepare themselves in advance.

4.2 *Deviations*

There were only two deviations from the plan:

- By mistake, we used one of the computer accounts twice, so that the data of the previous participant was overwritten. However, thanks to the screen recording we were able to redo every single action taken by this user in the experiment, recreating the lost spreadsheet.
- One of the participants had issues with his e-mail account so he did not get our invitation for his particular time slot. Therefore, we ran only the remaining two experiments in this time slot. Then, we simply added a new time slot where this participant did the missed experiment alone.

5 ANALYSIS

5.1 *Descriptive Statistics*

5.1.1 *Goal 1 - Usage of the Tutorials*

For this goal we asked the groups with both tutorials which tutorial they used more often. Fig. 5 shows the results. As can be seen, from the 13 participants in group C and F who answered this question, six participants used (nearly) only the video tutorial, four used both equally and only three used (nearly) only the text tutorial.

5.1.2 *Goal 2 - Duration of Tutorial Usage*

Fig. 6a shows that for the SIF tutorial there were only small differences between the time needed by the video and the text tutorial. The participants with both tutorials took more time to complete the tutorial. The standard deviation of all three groups is close around the mean with the deviation of the *video and text* group slightly larger. (σ_{Video}=117.361, σ_{Text}=139.764, σ_{Both}=188.432).

Furthermore, it can be seen in Fig. 6b that there were only small differences in the overall time needed for the tutorial and the following task.

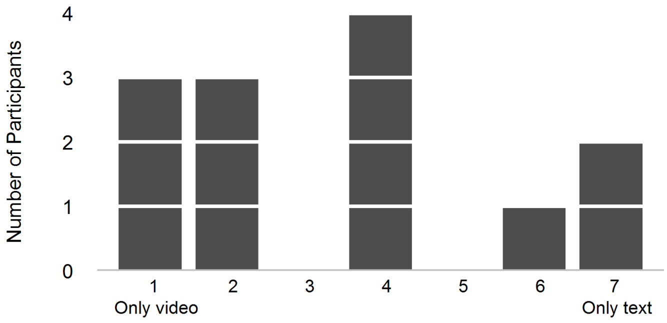


Figure 5: The participants with both tutorials stated which tutorial they used more often

The values spread wider around the mean (σ_{Video}=362.353, σ_{Text}= 450.317, σ_{Both}=344.713).

For the second tutorial the results were slightly different. Fig. 7a shows that this time the text tutorial was much shorter than the other two tutorial types. The values are much closer around the mean than for the SIF tutorial (σ_{Video}=362.353, σ_{Text}= 103.714, σ_{Both}=54.705). But again there are close to no differences when comparing the tutorial and the following tasks as can be seen in Fig. 7b. Again, the values are further away from the mean (σ_{Video}=396.333, σ_{Text}= 357.612, σ_{Both}=126.839).

5.1.3 *Goal 3 - Effectiveness of the Tutorials*

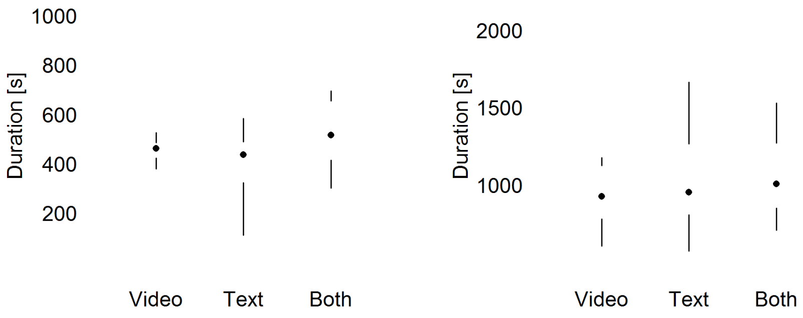
Fig. 8a shows how many times something was looked up in a tutorial by the two groups with text **or** video tutorial. The figure shows that the participants seldom looked something up in the video tutorial whereas they looked up things many times in the text tutorial.

Fig. 8b shows how many times the participants looked something up if they had **both** tutorials. Some participants used only one of the two tutorials (the other one was *not used*). Again, they looked up more things in the text tutorial.

Fig. 9 shows how much the participants understood of the tutorials at the first attempt as assessed by themselves. There is nearly no difference between the text and the video tutorial.

In Fig. 10 it can be seen how many correct and wrong answers the result files of the three tasks contained. The files from the first two tasks (Playground1 and Playground2) contained almost no errors. Conversely, the correctness of the files in the main task was very poor. As mentioned, we evaluated two unit tests for this task. One for the first part (the participants added a new rate to the existing ones) and one for the second part (for every rate the participants added the costs for on-net texts). As Fig. 10c shows, there were no significant differences between the three groups regarding the correctness of the first unit test. Also, when taking into account the second unit test the overall picture remains unchanged as shown in Fig. 10d.

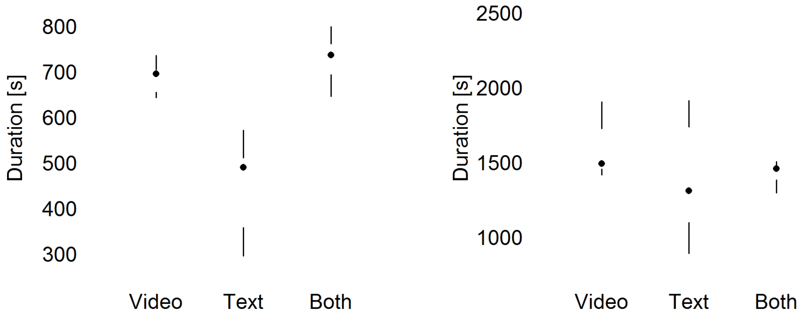
Finally, Fig. 11 shows that there are small differences between the groups in how easy or hard they perceived the whole experiment to be. There is



(a) Duration of using the SIF tutorial in seconds

(b) Duration of using the SIF tutorial and the following task in seconds

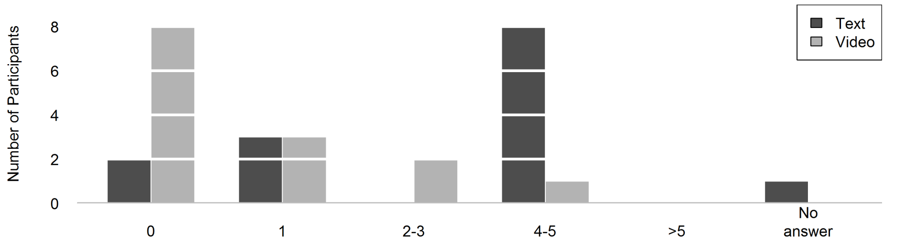
Figure 6: Duration of using the SIF tutorial and using the SIF tutorial plus the following task in seconds



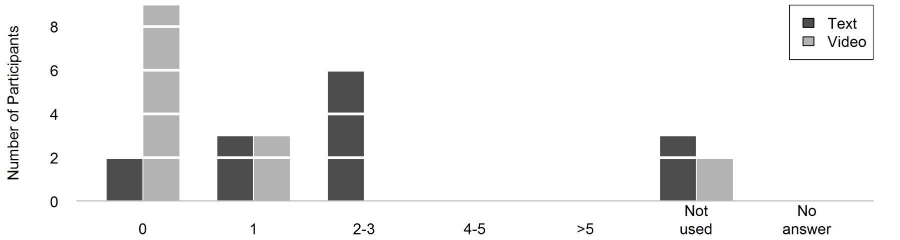
(a) Duration of using the scenario tutorial in seconds

(b) Duration of using the scenario tutorial and the following task in seconds

Figure 7: Duration of using the scenario tutorial and using the scenario tutorial plus the following task in seconds



(a) The participants with **only one** tutorial stated how many times they looked something up in the used tutorial



(b) The participants with **both** tutorials stated how many times they looked something up in which tutorial

Figure 8: How many times was an item looked up in which tutorial?

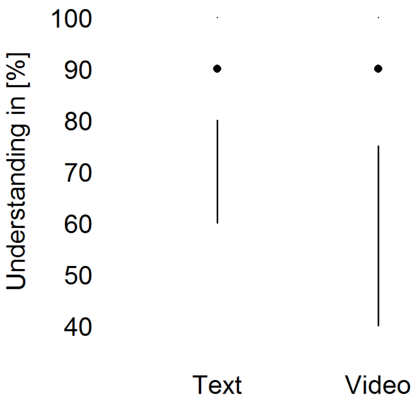
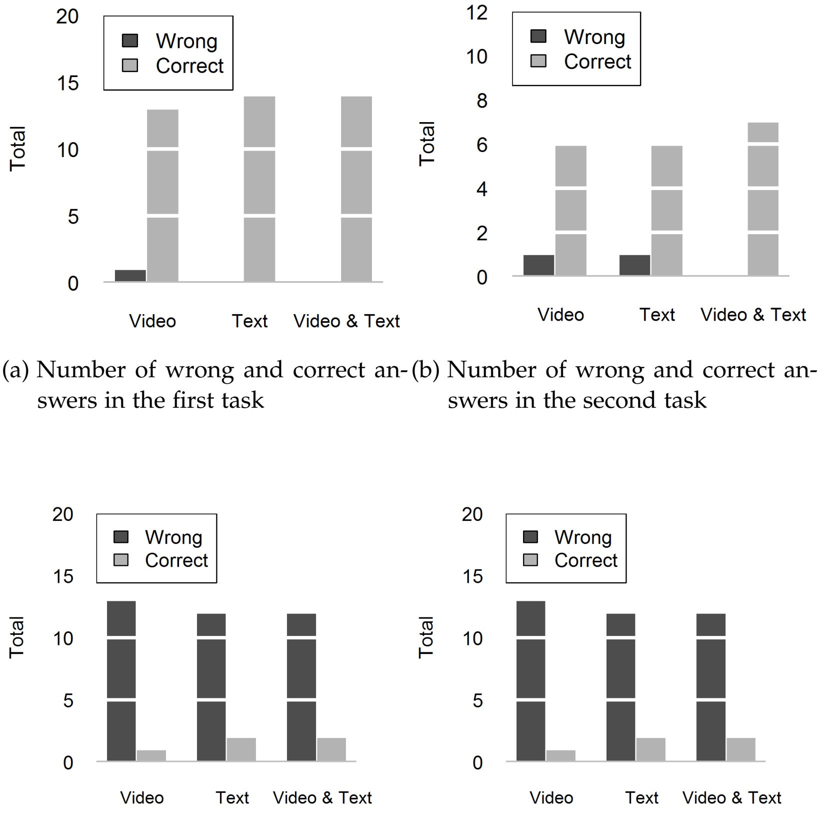


Figure 9: Percentage of understanding of the tutorial at the first attempt



(a) Number of wrong and correct answers in the first task

(b) Number of wrong and correct answers in the second task

(c) Number of wrong and correct answers in the final task after inserting the new rate

(d) Number of wrong and correct answers in the final task after inserting the on-net texts

Figure 10: Number of wrong and correct answers in the tasks

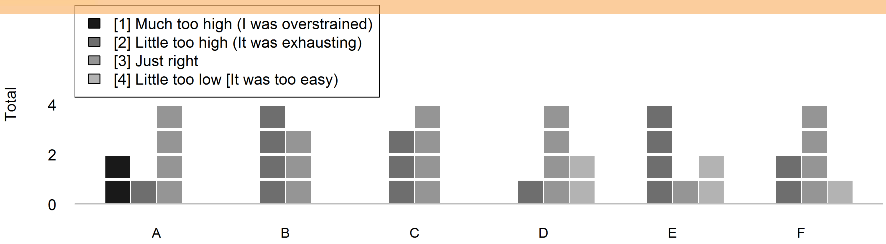


Figure 11: Difficulty level of our study perceived by the participants

no clear tendency that one group had more or less challenges when solving the tasks in our study.

5.2 *Data Set Preparation*

To reconstruct the overwritten data mentioned in section 4.2, one of us watched the screen recording and in parallel repeated the clicks from the video.

Depending on the goal, we merged the different results from the questionnaires. For example, for the question which tutorial was used, we merged the two groups with both tutorials, no matter if with or without the scenarios.

One problem was the question for groups C and F how many times they looked up information. Many used the option *Not used* as used zero times and not as the intended *I did not use this tutorial at all*. We found out about this because many participants stated that they used both tutorials equally in one question and then stated that they did not use one at all in the described question. We solved this problem by correcting the *Not used* answers to *0 times* when the participants did not state *only video* or *only text* in the other question.

5.3 *Hypothesis Testing*

5.3.1 H_1 - *Usage of the Tutorials*

The bar chart in Fig. 5 looks like there might be a tendency towards the video tutorial. We conducted a Shapiro-Wilk test and saw that the data is normally distributed (p=0.065) and there is no preference for videos. We cannot reject the null hypothesis, both tutorials were chosen equally often.

5.3.2 H_{2a} - *Duration of Tutorials*

As Fig. 6a shows, the *video and text* tutorial group took longer than the other two groups to complete the SIF tutorial. A t-test confirmed this. The video tutorial group was significantly faster than the *video and text* tutorial group. There is a statistically significant difference and a medium effect size(p=0.049, Cohen's d=0.781, r=0.364, MD=122.571). Also, the text tutorial group was significantly faster than the *video and text* tutorial group with a strong effect size (p=0.026, Cohen's d=0.919, r=0.417, MD=152.407). There was no difference between the video tutorial group and the text tutorial group (p=0.552).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

The results in Fig. 7a show that for the scenario tutorial, there were significant differences as well. This time the video tutorial took longer than the text tutorial with a very strong effect size (p=0.00053, Cohen's d=3.14, r=0.843, MD=242.571). Also, the video tutorial group took longer than the video and text tutorial group with a very strong effect size (p=3.221e-05, Cohen's d=3.441, r=0.865, MD=285.286). In this case there was no difference between the video tutorial group and the mixed group (p=0.119). The null hypothesis that all three groups need the same time can be rejected in both cases.

5.3.3 H_{2b} - *Duration of Tutorial and Task*

Concerning the time needed for the tutorial and the following task, the results are different. For the SIF tutorial and the task there are no significant differences between the three groups ($p_{\text{Video}-\text{Text}}$=0.627, $p_{\text{Video}-\text{Both}}$ =0.439,$p_{\text{Text}-\text{Both}}$=0.838). The same result shows for the scenario tutorial and the following tasks ($p_{\text{Video}-\text{Text}}$=0.206, $p_{\text{Video}-\text{Both}}$=0.199,$p_{\text{Text}-\text{Both}}$ =0.724). Therefore, the null hypothesis that there are no differences between the three groups cannot be rejected.

5.3.4 H_{3a} - *Looking Things Up*

A t-test confirmed that the values in Fig. 8a are significantly different with a very strong effect size (p=0.003, Cohen's d=1.274, r=0.547, MD=1.907). The participants looked up more things in the text tutorial than in the video tutorial. The same applies for Fig. 8b with a very strong effect size (p=0.001, Cohen's d=1.7, r=0.648, MD=1.15). We can reject the null hypothesis that the participants look things up in both tutorials equally often. They looked up more items in the text tutorial.

5.3.5 H_{3b} - *Understanding*

Fig. 9 already indicates that there is no difference in how much of the tutorial the participants understood at the first attempt. A t-test confirmed this result (p=0.4566). We can reject the null hypothesis. There is no significant difference between the two tutorial types.

5.3.6 H_{3c} - *Error Quota*

As Fig. 10 shows, there are no significant differences between the three groups regarding the number of correct and wrong answers for all four tasks (the final task was divided into two parts) ($p1_{\text{Video}-\text{Text}}$=0.336, $p1_{\text{Video}-\text{Both}}$=0.439, $p1_{\text{Text}-\text{Both}}$: data is constant), ($p2_{\text{Video}-\text{Text}}$=1, $p2_{\text{Video}-\text{Both}}$=0.337,$p2_{\text{Text}-\text{Both}}$=0.356), ($p3_{\text{Video}-\text{Text}}$=0.189,$p3_{\text{Video}-\text{Both}}$=0.453, $p3_{\text{Text}-\text{Both}}$=0.574), ($p4_{\text{Video}-\text{Text}}$=0.189, $p4_{\text{Video}-\text{Both}}$=0.453, $p4_{\text{Text}-\text{Both}}$ =0.574). The null hypothesis cannot be rejected, there are no differences between the three groups.

5.3.7 H_{3d} - *Level*

We checked with a t-test if there are significant differences in therms of the level between the three tutorial types or between the groups with the scenario task and the groups without it. Against our expectations, there are no significant differences (p_{A-B}=0.735, p_{A-C}=0.502, p_{B-C}:0.626, p_{D-E}=0.354,

p_{D-F}=0.454, p_{E-F}:0.753, p_{A-D}=0.078, p_{B-E}=0.502, p_{C-F}:0.404). We cannot reject the null hypothesis.

6 DISCUSSION

6.1 *Evaluation of Results and Implications*

6.1.1 H_1 - *Usage of the Tutorials*

Contrary to our hypothesis, both types of tutorial were chosen equally often. Apparently, the participants had no significant preference for one tutorial type. From what we could see during the experiments, many participants started with the video tutorial and simultaneously flipped through the text.

6.1.2 H_{2a} - *Duration of Tutorial Usage*

For the SIF tutorial, the groups with both tutorials took longer to get through the tutorial. This might be because they had two sources and therefore had a look into both which, of course, takes longer.

There was no difference between text and video in this tutorial, which might be because the whole program was new to the participants and, therefore, the participants with the text tutorial took their time to work through the text and fully understand everything.

For the scenario tutorial, again the group with both tutorials took longer than the text tutorial group, for the same reason. Additionally, for this tutorial the video group also took longer. This can have two reasons: either the text tutorial group was faster because they could easily skip uninteresting parts or because the scenario video took longer than the previous video and now the time difference was bigger.

All in all, we conjecture that the readers are faster because they can choose how fast they read. The video is always of the same length. The longer the video, the bigger the difference between video and text.

6.1.3 H_{2b} - *Duration of Tutorial and Task*

The results of this hypothesis were quite unexpected. Although the text tutorial group was faster during the tutorial itself, they were not faster overall, including the following task. We have one explanation for this. As the participants looked up more things in the text tutorial, our conclusion is that the text tutorial group might not have understood as much as the other groups and therefore needed more time during the task to look up information. So the time saved due to fast reading is compensated for by looking up things more often.

6.1.4 H_{3a} - *Looking Things Up*

Overall, most participants looked up more things in the text tutorial. This might be because it is easier to find things in a text than in a video. Maybe the video was better in explaining and, therefore, the participants did not have to look things up, though this is unlikely, as the participants stated to have understood both tutorial types equally well. Another explanation could be that the participants skipped some parts while reading. Both interpretations are possible.

6.1.5 H_{3b} - *Understanding*

We first created the videos and then transcribed them into a text. That is why the given explanations are the same, just the format is different. This leads to the conclusion that it does not matter how an explanation is given, as long as the explanation itself is good. A bad explanation will not get better just because one uses a different format.

6.1.6 H_{3c} - *Error Quota*

For the practicing tasks as well as for the final task all three groups made the same number of errors. Contrary to our hypothesis, no tutorial taught the program better, which may have led to a better result in general. As the tutorials were equally good, there is also no difference in the result oft the task. This leads to the conclusion that the format of a tutorial is less important than the content. As the content was exactly the same, the results were the same as well.

6.1.7 H_{3d} - *Level*

Quite unexpected, having to go through the advanced tutorials with additional information to learn had no impact on how easy or difficult the participants perceived the experiment to be. We see two possible explanations here: either the format of the tutorial really makes no difference or the tasks and the technology under evaluation were too easy or too hard for a difference to become evident.

6.2 *Threats to Validity*

We have identified a number of threats to validity for our study and split them into three groups: construct validity (CV), internal validity (IV) and external validity (EV). We discuss them in the following:

- (CV) We took the opinion of the participants directly as a measurement. Still, it might be that some participants misjudged their usage of the tutorial. Nevertheless, this should balance out due to the number of participants.
- (CV) As the participants had the choice between three different experiments, the results cannot be mapped to computer science students from our university in general since probably students who are interested in Microsoft Excel actively decided to take our experiment.
- (IV) The groups varied only by the kind of tutorial they received. Although we matched the participants and the tutorial groups completely randomly, the sample size might not have been sufficient to guarantee true randomness.
- (EV) Instead of using experienced software developers we used software engineering students in our study. However, if even rather unexperienced students get along with a tutorial, we assume that the tutorial should be suitable for experienced professional developers as well.
- (EV) While for the purpose of our experiment it was productive to have a text tutorial which is equivalent to a video tutorial contentwise, in practice a combination with one tutorial type covering basic

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

concepts supplemented by another tutorial type covering advanced topics might work even better.

- (EV) Although we had a total of 42 participants, the sample size is not sufficient for generalizing our findings. Yet, we could see several statistically significant differences with medium to strong effect sizes.

- (EV) The study only investigated short-term learning effects, since the participants had to apply the learned content directly after consuming the tutorial. However, there could be significant differences when comparing long-term learning effects.

- (EV) The spreadsheets we provided to the participants contained seeded errors which is problematic as Panko explains [4]. While this might be regarded as a general threat for studies, we do not see a negative impact in the context of this concrete study.

- (EV) We only used content-equivalent representatives for two text and video tutorials which explain a single software tool and which were produced by the same authors. To further generalize our findings, it would be necessary to investigate more content-equivalent tutorials.

Overall, the answers we found seem reasonable to us. We confirmed them by the aforementioned statistical tests (t-test and Shapiro-Wilk test) and found no contradictions. Nevertheless, we encourage other researchers to replicate our experiment to further confirm our results.

6.3 *Lessons Learned*

All in all, we observed no remarkable differences between text and video tutorials – at least in our case, where they contained the same information. It seems that developers consume text tutorials faster than video tutorials, however, this balances out as consumers of text tutorials tend to look up more things later. It is remarkable that although there was no difference in the results, the majority of our participants had a personal preference for video tutorials.

7 CONCLUSIONS AND FUTURE WORK

7.1 *Summary*

In this study we wanted to investigate differences in educational effects between text and video tutorials. Surprisingly, the learning effects of using content-equivalent text and video tutorials seem to be almost identical – in total, it takes the same overall time to consume and apply each tutorial type. Our major findings are that developers prefer video tutorials in the first place, but when looking something up *after* consuming a tutorial they prefer text tutorials.

This leads to the conclusion that education-wise it would be the best for software makers simply to provide developers with both text and video tutorials. As we assume that authoring text tutorials takes less effort than authoring comparable video tutorials, one could argue that text tutorials provide a much better value for the price. However, most developers prefer watching videos in the beginning instead of reading text. Therefore, software makers that only provide text tutorials will need to spend more effort in order to motivate developers to actually consume their text tutorial.

7.2 *Future Work*

There are many more different tutorial approaches than just text and video tutorials (though most of them are for games and fewer for application software). For further research it would be interesting to see how these tutorial types compare to video and text tutorials. Just because videos and text tutorials seem to be most prominent, this does not mean the lesser known and therefore lesser used methods have to be less effective.

Also, replicating our study with more participants or with typical end-users without a background or previous knowledge in software engineering could help to generalize our results. As it is commonly known, "people from IT" approach software applications differently from typical end-users. Therefore, it would be interesting to see how users from other fields work with unknown software, what problems they have and if they generally have a different approach towards tutorials than developers.

Another interesting point would be to investigate why developers seem to look things up in text tutorials more often than in videos and if this differs from typical end-users' behaviors. One might assume that "finding the right place" in a video causes a higher mental load than finding it in a written tutorial because one can only see single frames when winding through videos and so has to recognize the wanted explanation by a single frame. Therefore, providing more aids when looking things up in video tutorials could be interesting.

A further aspect might be the actual activity at which a tutorial type is targeted. Is it targeted just at teaching the basics or rather advanced topics?

REFERENCES

[1] Erik Andersen et al. "On the Harmfulness of Secondary Game Objectives". In: *Foundatios of Digital Games*. ACM, 2011, pp. 30–37. DOI: 10.1145/2159365.2159370.

[2] Erik Andersen et al. "Placing a Value on Aesthetics in Online Casual Games". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1275–1278. DOI: 10.1145/1978942.1979131.

[3] Erik Andersen et al. "The Impact of Tutorials on Games of Varying Complexity". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 59–68. DOI: 10.1145/2207676.2207687.

[4] Salvatore Aurigemma and Raymond R. Panko. "The Detection of Human Spreadsheet Errors by Humans versus Inspection (Auditing) Software". In: *Proceedings of EuSpRIG 2010 Conference Practical steps to protect organisations from out-of-control spreadsheets* (2010), pp. 1–14. URL: http://arxiv.org/abs/1009.2785.

[5] Ronald Baecker. "Showing Instead of Telling". In: *Proceedings of the 20th annual international conference on Computer documentation*. SIGDOC '02 (2002), pp. 10–16. DOI: 10.1145/584955.584957.

[6] Ronald Baecker, Ian Small, and Richard Mander. "Bringing Icons to Life". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '91. ACM, 1991, pp. 1–6. DOI: 10.1145/108844.108845.

[7] Alan Dix et al. *Human-Computer Interaction*. 3rd ed. Prentice-Hall, Inc., 2004. URL: http://fit.mta.edu.vn/files/DanhSach/__Human_computer_interaction.pdf.

[8] Tovi Grossman, George Fitzmaurice, and Ramtin Attar. "A Survey of Software Learnability: Metrics, Methodologies and Guidelines". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. ACM, 2009, pp. 649–658. DOI: 10.1145/1518701.1518803.

[9] Susan M. Harrison. "A Comparison of Still, Animated, or Nonillustrated On-line Help with Written or Spoken Instructions in a Graphical User Interface". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '95. ACM Press/Addison-Wesley Publishing Co., 1995, pp. 82–89. DOI: 10.1145/223904.223915.

[10] Andreas Jedlitschka, Marcus Ciolkowski, and Dietmar Pfahl. "Reporting Experiments in Software Engineering". In: *Guide to Advanced Empirical Software Engineering*. Springer London, 2008, pp. 201–228. DOI: 10.1007/978-1-84800-044-5_8.

[11] Daniel Kulesz. "From Good Practices to Effective Policies for Preventing Errors in Spreadsheets". In: *Proceedings of EuSpRIG 2011 Conference Spreadsheet Governance Policy and Practice"*. EuSpRIG, 2011. URL: https://arxiv.org/abs/1111.6878.

[12] Daniel Kulesz. *Spreadsheet Inspection Framework*. URL: http://www.spreadsheet-inspection.org/index.html.

[13] Daniel Kulesz, Jonas Scheurich, and Fabian Beck. "Integrating Anomaly Diagnosis Techniques into Spreadsheet Environments". In: *2014 Second IEEE Working Conference on Software Visualization (VISSOFT)*. 2014, pp. 11–19. DOI: 10.1109/VISSOFT.2014.12.

[14] Daniel Kulesz, Fabian Toth, and Fabian Beck. "Live Inspection of Spreadsheets". In: *Proceedings of the 2nd Workshop on Software Engineering Methods in Spreadsheets*. 2015. URL: http://arxiv.org/abs/1505.02428.

[15] Jonathan Lazar, Adam Jones, and Ben Shneiderman. "Workplace user frustration with computers: an exploratory investigation of the causes and severity". In: *Behaviour & Information Technology* 25.3 (2006), pp. 239–251. DOI: 10.1080/01449290500196963.

[16] *Limesurvey*. URL: https://www.limesurvey.org.

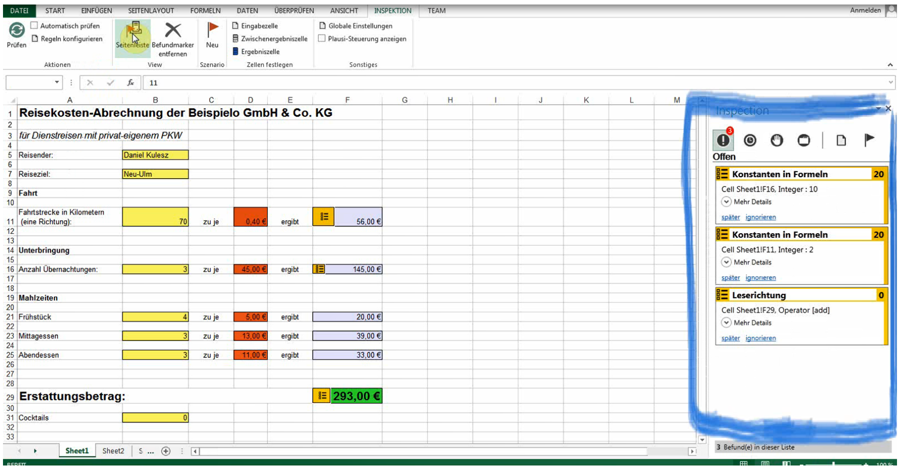
[17] Brigit van Loggem. "'Nobody reads the documentation': true or not?" In: *Proceedings of ISIC: the information behaviour conference*. 2014. URL: http://www.informationr.net/ir/19-4/isic/isic03.html#.V76fOeB8uUk.

[18] Andrew P. Martin et al. *Exploring the Persistent Problem of User Assistance*. Technical Report IS-TR-2005-08-01. Information School, University of Washzington, 2005. URL: https://digital.lib.washington.edu/researchworks/handle/1773/2079.

[19] Microsoft. *Excel 2013*. Software.

[20] *R*. URL: https://www.r-project.org/.

[21] Edward R Tufte. *The Visual Display of Quantitative Information*. Vol. 2. 2001. DOI: 10.1198/tech.2002.s78.



11

	A	B	C	D	E	F
1	**Reisekosten-Abrechnung der Beispielo GmbH & Co. KG**					
2						
3	*für Dienstreisen mit privat-eigenem PKW*					
4						
5	Reisender:	Daniel Kulesz				
6						
7	Reiseziel:	Neu-Ulm				
8						
9	**Fahrt**					
10						
11	Fahrtstrecke in Kilometern (eine Richtung):	70	zu je	0,40 €	ergibt	56,00 €
12						
13						
14	**Unterbringung**					
15						
16	Anzahl Übernachtungen:	3	zu je	45,00 €	ergibt	145,00 €
17						
18						
19	**Mahlzeiten**					
20						
21	Frühstück	4	zu je	5,00 €	ergibt	20,00 €
22						
23	Mittagessen	3	zu je	13,00 €	ergibt	39,00 €
24						
25	Abendessen	3	zu je	11,00 €	ergibt	33,00 €
26						
27						
28						
29	**Erstattungsbetrag:**					**293,00 €**
30						
31	Cocktails	0				
32						
33						

Inspection

Offen

Konstanten in Formeln 20
Cell Sheet1!F16, Integer : 10
Mehr Details
später ignorieren

Konstanten in Formeln 20
Cell Sheet1!F11, Integer : 2
Mehr Details
später ignorieren

Leserichtung 0
Cell Sheet1!F29, Operator [add]
Mehr Details
später ignorieren

3 Befund(e) in dieser Liste

Figure 12: Screenshot of the video tutorial for SIF

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2413v1 | CC BY 4.0 Open Access | rec: 2 Sep 2016, publ: 2 Sep 2016

4 Die Seitenleiste

Über die Schaltfläche *Seitenleiste* kann man sich eine Seitenleiste einblenden (Abbildung 10), die unter anderem die Befunde übersichtlich auflistet. Die Schaltfläche dafür befindet sich ebenfalls in der Inspektionsleiste (Abbildung 9).

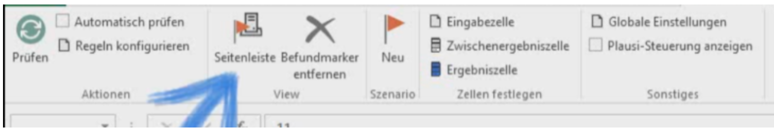
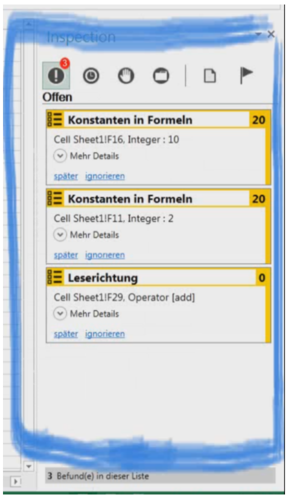


Abbildung 9: Schaltfläche zum Einblenden der Seitenleiste

Eine kleine Zahl in einem roten Kreis informiert darüber, wie viele Befunde noch nicht angeschaut wurden. Wie bei einem E-Mail-Programm wird ein Befund, sobald er gelesen wurde, weniger prominent dargestellt.



Konstanten in Formeln 20
Cell Sheet1!F16, Integer : 10
Mehr Details
später ignorieren

Konstanten in Formeln 20
Cell Sheet1!F11, Integer : 2
Mehr Details
später ignorieren

Leserichtung 0
Cell Sheet1!F29, Operator [add]
Mehr Details
später ignorieren

3 Befund(e) in dieser Liste

Abbildung 10: Die Seitenleiste

Figure 13: Content-equivalent excerpt of the text tutorial for SIF