

User Preference and Effectiveness of Tooltips for Registering Health Data

*A Field Study Exploring Content Types for
Tooltips for Undereducated Health Staff
in Developing Countries*

Helene Isaksen & Mari Iversen



Master Thesis
Informatics: Design, Use and Interaction
60 credits

Department of Informatics
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

May 2017

User Preference and Effectiveness of Tooltips for Registering Health Data

*A Field Study Exploring Content Types for
Tooltips for Undereducated Health Staff in
Developing Countries*

“Kafukufuku”

© Helene Isaksen and Mari Iversen

2017

User Preference and Effectiveness of Tooltips for Registering Health Data – A Field Study
Exploring Content Types for Tooltips for Undereducated Health Staff in Developing Countries

Helene Isaksen and Mari Iversen

<http://www.duo.uio.no/>

Print: University Print Centre, University of Oslo

Abstract

Many health professionals in developing countries lack the proper competence for the work they are doing. Tooltips have previously proven to be effective, and may assist such health professionals in their daily routines entering data to health information systems. By using qualitative and quantitative methods, this thesis aim to research two aspects of tooltips; finding user preference for content types for tooltips and evaluations methods to find the most effective tooltips, in terms of helping users enter correct data to the system. The target group is health workers in African countries, and especially those who lack the proper competence for the work they are performing.

This research has included participants from three African countries, Malawi, South Africa and Ethiopia. Most of our findings are based on Malawian health workers, as they represent a greater share of our participants. We have used questionnaires, a modified question-suggestion approach and a quasi-experiment to explore preferences of tooltips, what tooltips would be effective, what methods one should use to find effective tooltips and whether tooltips have an effect or not.

Firstly, we found that most of our participants prefer a range of normal values, instead of explanations, as content type for tooltips. Secondly, we found that tooltips containing explanations outperform those with normal values, in terms of correctness of data entry. Thirdly, we found that low content validity evaluations, such as a questionnaire, could not replace high content validity evaluations, such as field experiments. Lastly, we found that tooltips do have an effect, since the correctness of data entry increased and several participants stated that they learned from the tooltips.

Acknowledgement

First of all, we would like to thank our supervisors Jens Kaasbøll and Chipso Kanjo for helping us and guiding us through this research.

Thanks to Cecilia Persson for helping us through the setup of our test-material and giving us valuable information on how to write understandable. Thanks to Yamikani Phiri and Lawrence Byson for helping us when our server crashed. Special thanks to Lawrence for assisting us when we were not able to conduct our post-interviews in Malawi. - Zikomo!

Thanks to all participants and other contributors to the research project in Malawi, South Africa and Ethiopia. Thanks to those who gave us valuable insight into medical terms, and how to use them.

Thanks to fellow master students on the 6th floor for drinking coffee and nagging over that the thesis will never be finished, and for the days when we were able to encourage each other.

Mari: Thanks to my mom and dad for being supportive through 17 years of education. Special thanks to my dad for slightly reviewing this thesis.

Helene: Thanks to my family for reviewing the thesis at the end and supporting me through a lifetime of education. Thanks to my supportive boyfriend for sticking out with my complaints for the past five years.

Helene Isaksen & Mari Iversen

April 2017

Table of Content

Abstract	V
Acknowledgement.....	VII
List of Tables.....	X
List of Graphs.....	XI
List of Figures	XII
Abbreviations	XIII
1 Introduction	1
1.1 Background.....	1
1.2 Context.....	3
1.3 Research Question	4
1.4 Usability.....	4
1.5 User Experience.....	6
1.6 Methods	7
1.6.1 Our Approach.....	7
1.6.2 Philosophical Assumption.....	8
1.6.3 Action Research	9
1.6.4 Participatory Design.....	9
1.6.5 Data Collection Methods.....	11
1.7 Structure of Thesis.....	14
1.7.1 Chapter 1 – Introduction	14
1.7.2 Chapter 2 – Methodology.....	15
1.7.3 Chapter 3 – Finding User Preference	15
1.7.4 Chapter 4 – Finding Effective Tooltips.....	15
1.7.5 Chapter 5 – Conclusion	15
2 Methodology	16
2.1 “Methods for Evaluation of Tooltips”	16
3 Identifying User Preference	34
3.1 “Design of Tooltips for Health Data”.....	34
3.2 Content Types – Results	44
4 A Field Experiment	46

4.1	“Design of Tooltips for Data Fields – A Field Experiment of Logging Use of Tooltips and Data Correctness”.....	46
4.2	Final Results from Field Experiment.....	65
4.2.1	Normal Values versus Explanations	66
4.2.2	Malawi versus South Africa.....	73
4.3	Did the Experiment Alter Their Preferences?	78
4.3.1	Malawi.....	78
4.3.2	South Africa	79
4.3.3	Changes at User Level.....	80
4.4	User Experience Questionnaire	81
4.4.1	Results	83
4.4.2	Helpful – Did the Participants Find the Tooltips Helpful?	85
4.4.3	Rewarding – Did the Tooltips Give the Participants New Knowledge?.....	85
4.4.4	Comparison of Contradictory Statements	86
4.4.5	Comparing Normal Values and Explanations	86
4.4.6	Improvements.....	86
5	Conclusion.....	88
5.1	First Research Question.....	88
5.2	Second Research Question	88
5.3	Third Research Question	89
5.4	Fourth Research Question	90
5.5	Understanding Users.....	91
5.6	Reflections	91
5.7	Recommendations	92
	References	94
	Appendix - Feedback to DHIS2 software developers.....	100

List of Tables

Table 1: Usability goals comparison with Tracker Capture and tooltips	6
Table 2: Preference of content type - Malawi vs. South Africa	44
Table 3: Data element example with different tooltips	69
Table 4: Correctness for the first and last seven cases	73
Table 5: Correctness for the first and last seven cases	77
Table 6: Average score for preference pre and post experiment.....	78
Table 7: Wilcoxon significant differences of Table 6.....	79
Table 8: Average score for preference pre and post experiment.....	79
Table 9: Indicating changes of preference at user level	80
Table 10: Results from UX questionnaire.....	84

List of Graphs

Graph 1: Average number of opened tooltips through the cases	67
Graph 2: Percent of correctness through the cases.....	68
Graph 3: Percent of correctness versus number of opened tooltips per user	70
Graph 4: Percent of successful tooltips through the cases	71
Graph 5: Average number of opened tooltips through the cases	74
Graph 6: Percent of correctness through the cases.....	75
Graph 7: Percent correctness versus number of opened tooltips per user.....	76
Graph 8: Percent of successful tooltips through the cases	77

List of Figures

Figure 1: Example from booklet	72
Figure 2: Screenshot of UX questionnaire	82

Abbreviations

ANC	Antenatal Care
AR	Action Research
CHW	Community Health Worker
DHIS2	District Health Information System version 2
HIS	Health Information System
HISP	Health Information System Programme
MDG	Millennium Development Goals
PD	Participatory Design
SA	South Africa
UN	United Nations
UX	User Experience
WHO	World Health Organization

1 Introduction

The aim of this thesis is to explore various aspects of tooltips in several developing countries.

1.1 Background

Africa constitutes about 16 % of the world population (Worldometers, n.d.), and suffer from a huge and growing healthcare crisis. According to a World Health Organization (WHO) report (2014), it is estimated that 22.8 skilled health workers per 10.000 population is needed to cover essential health interventions. Most of the countries in Africa are below this line, having 1.8 million health workers of the total 27.2 million in the world. This means that even though Africa has 16% of the world's population, they only have 6.6% of all skilled health workers.

Those working within the health sector in these countries usually range from community health workers (CHW) to medical doctors, though medical doctors are rare. In addition, most of the educated health workers tend to seek work in the bigger cities or different countries, due to better salaries (Sood et al., 2008). The rural clinics often lack health personnel with the right competence, both within health and computer skills (Oluoch et al., 2012). In most health facilities, especially in rural areas, nurses and midwives are the most educated personnel. However, the workload exceeds their capacity, and those without proper competence have to step in and do their tasks, which may lead to, amongst other, misdiagnosing of patients and wrong data capturing. Training and education is often too expensive or impossible, due to staff shortages. This means that there is a need for improving the knowledge among existing health staff, in a cheaper way.

Tooltips have previously proven to be effective, in several formats (Dai, Karalis, Kawas, & Olsen, 2015; Grossman & Fitzmaurice, 2010; Petrie, Fisher, Weimann, & Weber, 2004), which we will explain further in chapter 2, 3 and 4. The most common type will show information relevant to a given situation, and can be viewed either by hovering over or by pushing a button. Also, tooltips are a cheaper way of increasing knowledge of a specific system or domain, compared to for example training or workshops. Therefore, this thesis opt to explore various aspects of tooltips, such as preference in content and format, and effectiveness. By effectiveness we mean helping users enter correct data.

Checking tooltips is a self-initiated action. However, when people already think they know the answer to something, even if it is wrong, they tend to stick to it, unless challenged (Rourke & Kanuka, 2009). Thus, the effect of tooltips may only be achievable if people seek information on their own initiative. For example, if someone think they have knowledge of a medical term, they would most likely continue to believe in that knowledge, and not check the tooltip, meaning they may hang on to a misconception of the truth. The effect of tooltips may therefore be non-existent.

As of November 2015 approximately 830 women die “from preventable causes related to pregnancy and childbirth” every day (WHO: Maternal Mortality, 2015), and about 99% of these occur in developing countries. Even though the numbers have been reduced significantly over the past 25 years, they are still a lot higher than in other parts of the world.

According to the National Health Service, “Antenatal care is the care you receive from healthcare professionals during your pregnancy” (2015). The intention of ANC is to make sure that both the pregnant woman and the fetus get the care they need, in order to have a healthy pregnancy. WHO’s current guidelines on ANC recommend at least four visits during a woman's pregnancy (WHO: What matters to women during pregnancy, n.d.).

This thesis is a part of the mHealth4Afrika project, which is a three year collaborative research project focusing on community based maternal and newborn health care in four countries in Africa; Malawi, South Africa (SA), Kenya and Ethiopia. The project focuses on two of the eight United Nations millennium development goals (MDG), aimed to be reached by 2015 (United Nations, 2015). These goals are on MDG 4: reducing child mortality and MDG 5: Improve maternal health. Thus, antenatal care (ANC) is a focus point in our thesis.

The use of mHealth could, in the long term, support the delivery of high quality healthcare and enable more accurate treatment. Mobile health, or mHealth, is defined by Global Observatory for eHealth as “mHealth or mobile health as medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring device, personal digital assistants, and other wireless devices” (WHO, 2013).

This study is also part of Health Information System Programme (HISP) research. HISP is a global network with Department of Informatics at the University of Oslo in Norway as their main coordinator. The HISP project started in 1994 as an action research (AR) project targeting

health systems in post-apartheid South Africa by utilizing a participatory design (PD) approach. This led to the development of (DHIS2). HISP have many partners and are working on multiple projects all over the world, mainly focusing on countries in Africa and Asia (DHIS2: In Action, n.d.). Their main goal is to strengthen the Health Information System (HIS) of a country by utilizing DHIS2.

1.2 Context

This thesis has focused on utilizing an app from the DHIS2, to address the issue of wrong data entry. DHIS2 was set up with an ANC program in the Tracker Capture Android app, using tablets as hosts. However, this thesis does not revolve around DHIS2 itself, but it is rather used as a tool to our fieldwork.

Our research has taken place in three low to medium income countries, Malawi, South Africa and Ethiopia, focusing on the maternal care in the communities. Both South Africa and Malawi have adopted programs from the DHIS2 package (DHIS2: Deployment, n.d.), while Ethiopia is in the starting phase of adopting it.

All the aforementioned countries have clinics in rural areas, and in many situations the power grid does not reach the most rural health centers, and even if it does, the power-supply is characterized by being unstable. Some bigger institutions have fuel-powered aggregators, though, the possibility of the aggregators running out of fuel is present. Thus, IT-equipment requiring stable power are not optimal in such cases.

On the other hand, according to the International Telecommunication Union (2015), there are about 7.1 billion mobile phone subscriptions in the entire world, and more than 95% of the world's population is now covered with mobile signals. Furthermore, the coverage of mobile phone network in many low-income countries often surpasses other infrastructure. Thus, the interest for using mobile technology to address health related issues, has increased (Braun, Catalani, Wimbush, & Israelski, 2013).

1.3 Research Question

As mentioned, health workers in developing countries often have to perform tasks beyond their competence. Therefore, a health system should, to the extent possible, provide health workers with support for completing their tasks. Our goal has been to research how to increase the amount of correct data entry by providing effective tooltips.

In order to understand a health worker's needs and preferences when it comes to tooltips, we have developed the following research questions aimed at addressing this goal.

1. What content types for tooltips do health workers in developing countries prefer?
2. What content type for tooltips lead to more correct data entry among undereducated health workers?
3. What techniques can be utilized to answer research question 1 and 2?
4. Do tooltips have an effect?

1.4 Usability

A definition of usability is "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." (ISO 9241-11, 1998). Thus, usability refers to how usable the system is for the targeted user and how enjoyable it is for the targeted user to use (Rogers, Sharp, & Preece, 2011) . To ensure good usability one may conduct a usability testing session to identify any usability issues and to track the user's satisfaction with the product. This may constitute a set of usability goals which the usability engineer may follow and evaluate. These goals may be effectiveness, efficiency, safety, utility, learnability, memorability (Rogers et al., 2011). Other researchers have identified similar goals, such as Nielsen (1996), who states that usability normally is measured as five product attributes: learnability, efficiency, memorability, errors and satisfaction.

We chose to extract three usability goals based on the definitions of Rogers et al. (2011), these being effectiveness, memorability and learnability, though, our main focus lies with learnability.

- Effectiveness concerns how good the product is at doing what it is supposed to do.
- Memorability concerns how easy it is for the users to remember the functionality of the system when coming back from a period of inactivity.
- Learnability concerns how easy the system is for the user to learn by exploring. Michelsen, Dominick and Urban (1980) talk about evaluation of an interface using software engineering principles. They present several software characteristics which are necessary for the continuous use of the system. One of them being learnability, which is characterized as “the system should be easy to learn by the class of users for whom it is intended.”. Their paper also describe six factors that might indicate learnability; using new commands, increase complexity, decrease think time, user comments on learnability, decreasing errors and decreasing use of help commands.

These three usability goals were chosen based on our research questions. We have applied these to both the Tracker Capture app and to the tooltips, and made a comparison to see the connection between them (see Table 1). However, during our research we focused on the tooltips and therefore did not look into the Tracker Capture aspect. As the goal of tooltips is to help the user carry out tasks, we wanted to explore the effectiveness of the tooltips, and see how effective they are at supporting health workers in entering correct data. In addition, we want to explore if the tooltips are understandable and whether the health workers are able to learn and remember the information they contain, hence our focus on learnability and memorability.

Table 1: Usability goals comparison with Tracker Capture and tooltips

Usability Goals	Tracker Capture	Tooltips	
Effectiveness	How good is the product at doing what it is supposed to do?	Are the health workers able to enter health data to the system?	Do the tooltips support health workers in entering correct data to the system?
Learnability	How hard is it to learn the function by exploring?	Are the health worker able to navigate and learn the functions of the app by exploring?	Are health workers able to learn and understand the content of the tooltips?
Memorability	What support has been provided to help users remember how to carry out tasks?	Are the health workers able to remember the functionality of the different elements in the app?	Are the health workers able to memorize the tooltips?

1.5 User Experience

User experience (UX) is closely related to usability, and the two are often hard to distinguish (Rogers et al., 2011, p. 18). However, UX differs from usability by focusing on the user’s experience of a system, rather than its usefulness (Rogers et al., 2011). Hassenzahl (2008) defines UX as “a momentarily, primarily, evaluative feeling (good-bad) while interacting with a product or service.”. He further states that this does not “exclude summary retrospective”, meaning an evaluation done in retrospect of the action. This is similar to the evaluation we have conducted, as we did not have an UX evaluation until after the experiment was finished.

Many UX goals have been identified and articulated within the field of interaction design (Rogers et al., 2011, p. 23). We chose to focus on two relevant UX goals, helpful and rewarding (Rogers et al., 2011). These were applied, not to the technology itself (Tracker Capture), but to the tooltips within the system. As the main goals of tooltips are for them to be helpful and

rewarding in terms of giving the users knowledge, these goals were the natural choice of focus for us. A tooltip is helpful if the users are able answer correctly to data entry after reading the tooltips. We consider a tooltip to be rewarding if the users are able memorize it and enter correct information to the system next time they use it.

1.6 Methods

This section gives an overview over the methods used in the research, during two field trips; the first to Malawi and Ethiopia during September and October 2016, and the second to Malawi and South Africa during January and February 2017. These field trips are referred to as iterations for the remainder of the thesis. Details on how we used the following methods can be found in chapter 2, 3 and 4.

1.6.1 Our Approach

Our process was made up of two clear iterations, which will be explained in detail below.

First Iteration

Our first iteration consisted of modified/adapted question-suggestion sessions. The question-suggestion protocol is based on the “question-asking protocol” proposed by Kato in 1968 (as cited by Grossman, Fitzmaurice, & Attar, 2009, p. 652), and implies that participants may ask questions at any time during the use of a system. Grossman, Fitzmaurice and Attar (2009, p. 653) augmented the protocol, and suggested that “the expert can also freely provide advice to the user.”. We modified this further, and included interviews, observations and a walkthrough of the technology. The participants were also encouraged to ask us any questions they might have. In addition, a paper-based questionnaire and the tooltips added to the Tracker Capture, acted as prototypes. These also enabled both us and our users to discuss different opportunities with a mutual understanding of the purpose of the tooltips.

Second Iteration

For our second iteration we conducted a quasi-experiment where we included two groups and two measurements or conditions. An experiment is considered a quasi-experiment if it “involves multiple groups or measures but the participants are not randomly assigned to different conditions” (Lazar, Feng, & Hochheiser, 2010, p. 42). Our participants were not randomly assigned to the different conditions. We chose to have a between-group design to our experiment. Between-group design, often called between-subject design, involves that each of the participants in the experiment is only exposed to one experiment condition (Lazar et al., 2010). The number of participant groups directly corresponds to the number of conditions.

In our case, we had two conditions; tooltips containing normal values for medical terms and tooltips containing explanations of the medical term. The participants were divided accordingly. In this thesis, we will sometimes refer to these groups as the normal value group and the explanation group. Due to time constraints and workload, it was beneficial to let the participants be exposed to only one condition, as opposed to both, like in within-group design. By being exposed to only one condition, we could reduce the risk of, amongst other things, confounding factors such as fatigue and participants being frustrated. However, a disadvantage of between-group experiment is that we are comparing the performance of two groups and the results may be affected by individual differences, hence a large number of participants is beneficial. Mapping individual differences may be hard, so therefore one of the focus point in our research was to find people of the same cadre and with the same experience with technology.

1.6.2 Philosophical Assumption

According to Myers (1997), “All research (whether quantitative or qualitative) is based on some underlying assumptions about what constitutes ‘valid’ research, and which research methods are appropriate”. Positivist studies explore what can be researched in a structured way, with the goal of increasing the predictive understanding of a phenomena (Orlikowski & Baroudi, 1991). Research is classified as positivist when formal suggestions, quantitative measures and conclusions based on a sample population are present (Orlikowski & Baroudi, 1991, p. 5). Interpretive studies are constructed through language, shared meanings and consciousness, and focus on humans and their way of thinking and making sense (Myers, 1997). The aim is to understand reality through the subjective and intersubjective meanings people assign them (Orlikowski & Baroudi, 1997, p. 5). It is characterized by nondeterministic perspectives,

examination in natural settings from a user perspective and that the researcher do not impose his or her pre-understanding of the situation (Orlikowski & Baroudi, 1991, p. 5).

Our philosophical assumptions are not clear cut, though we adopt several aspects from positivist studies, as well as some from interpretive studies. One of the goals of the quasi-experiment was to be able to predict the effect of tooltips, and both our quasi-experiment and questionnaire were structured ways of research which produced quantitative data. Research question one, two and three attempt to give practical suggestions on how to design tooltips, while the fourth question attempts to draw a conclusion about the effect of tooltips, based on a population sample consisting of health workers. Through interviews, we have tried to understand why people preferred the different tooltips, and create intersubjective understandings of what content would be most effective, helpful and rewarding, and why. We also made sure to not impose our personal opinions regarding tooltips, in order to not affect the participants' understandings.

1.6.3 Action Research

“Action research aims to contribute both to the practical concerns of people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework.” (Rapoport 1970, p. 499). Action research (AR) is an important component of participatory design (PD), and it seeks to engage both the affected workers and the outside researcher in studying and remedying existing problems (Greenwood & Levin, 1998). One of the researcher's goals in AR is to identify a problem or issue and come up with a possible solution. It is based on a collaboration between the researcher and the group of people who are experiencing an issue. The process of AR is iterative, where the first step is to diagnose the problem, then plan and do the action, and afterwards evaluate. The last stage involves specifying learning (Baskerville, 1999). This cycle is then repeated.

1.6.4 Participatory Design

The origin of Participatory Design (PD) is the democratic ideal that those who will be using an artifact should be given the right to decide on its design: its functioning as well as its form, and through this gain more control over the use situation and achieve a larger space for action. (Joshi & Bratteteig, 2015)

PD is an approach which involves the end-user of a system or product (Simonsen & Robertson, 2013). The purpose of PD is to overcome the difference in understanding and knowledge between users and developers, through a practical, hands-on approach (Simonsen & Robertson, 2013).

According to Simonsen and Robertson (2013), design by doing is an essential aspect within PD. This may include the usage of prototypes and mock-ups, and may enable the users to utilize their skills and open up for a more robust participation and a shared understanding between the user and the designers (Simonsen & Robertson, 2013). Prototypes are great for giving users a firsthand experience with a product or a practice, and for leveling out different understandings (Simonsen & Robertson, 2013).

User participation is considered “the core of Participatory Design” (Simonsen & Robertson, 2013, p. 5). Ives and Olson (1984) suggest six degrees of user involvement, meaning “the amount of influence the user has over the final product” (p. 590):

1. No involvement
2. Symbolic involvement
3. Involvement by advice
4. Involvement by weak control
5. Involvement by doing
6. Involvement by strong control

Also, Mumford (as cited by Ives & Olson, 1984) suggests that there are three different types of participation; consultative, representative and consensus. Consultative involves that developers make all decisions, though user needs and satisfaction are considered. Representative means that the user group is represented in the design group. Consensus participation attempts to involve the entire user department, in some way, through the entire process. Consultative is the least direct form of user participation, representative is in the middle and consensus is the most direct type of participation. We focus on the consultative type of participation, with a 3. degree of user involvement.

1.6.5 Data Collection Methods

We have utilized a mix of qualitative and quantitative research methods to do both a methodological triangulation and a triangulation of data (Rogers et al., 2011).

Questionnaire

Questionnaires are a good technique for collecting demographic data and users' opinions (Rogers et al., 2011, p. 238). It is a great method to gather information about larger groups, and may be used in conjunction with other data collecting methods. Questions may be open or closed (Rogers et al., 2011, p. 238), and should be clearly defined to ensure good data quality. When designing one of the questionnaires, we got help from a professional writer who gave valuable input on formulations. In addition, we conducted two pilot-tests to ensure that the questionnaire was understandable.

As our research question emphasize preference among health workers, a questionnaire measuring user preference of content types for tooltips was created (see chapter 3 for more information). The alternatives for content types were created based on interviews with other researchers, and their experiences from collaborating with health workers in developing countries. This approach should be considered a consultative type of participation with a 3. degree of involvement (Ives & Olson, 1984).

In addition to interviews, an UX questionnaire was created to measure the user experience of the tooltips. Our questionnaire was a blend between UX and usability, as it addressed both helpfulness and learnability of tooltips. It was more of an evaluation of the participants' user experience of the usability of the tooltips. A Likert scale is used to measure opinions, attitudes and beliefs, as well as to evaluate user satisfaction with a product (Rogers et al., 2011) Hence, when measuring the UX of the tooltips, we created scales containing sets of statements that represented a range of possible opinions. For instance, one of our statements contained a scale ranging from strongly agree to strongly disagree.

Diaries

Diaries are a technique mainly used to document events in the participant's life at the time of occurrence (Alaszewski, 2006). One may record everything from simple activities in the participant's life to explanations or reflections (Lazar et al., 2010, p. 126). Within human to

computer interaction, diaries aim to fill the gap between observations in natural settings, surveys and observation in a fixed lab (Hyldegård, 2006). They are also good for understanding how participants utilize technology in non-controlled settings (Lazar et al., 2010). Hence, diaries are good for capturing data about technology use in real world settings.

Diaries can split into two groups, elicitation diaries and feedback diaries (Lazar et al., 2010). Elicitation diaries are mainly used for prompting, and when interviews take place at a later stage in the research, the participants are asked to expand on each data point in the diary. Feedback diaries often tend to have instructions for when the participant should make the diary entry. Hence, while feedback diaries often focus on events that are interesting for the researcher, elicitation diaries often focus on events that are interesting for the participants (Lazar et al., 2010).

For our experiment we created a booklet inspired by diaries with questions and keywords we wanted our participants to elaborate on after using the technology, hence our approach is a hybrid between the two groups.

Document analysis

“Document analysis is a systematic procedure for reviewing or evaluating document, both printed and electronic (computer-based and Internet-transmitted) material.” (Bowen, 2009, p. 27) The documents analyzed may be anything from agendas, attendance register to manuals, books, journals (Bowen, 2009). Other research also include videos and pictures as forms of documents (Lazar et al., 2010, p. 284). Document analysis is used to give some meaning and context around the asset topic.

Observation

Observations are used to gain information and empirical material both in natural settings and in laboratory settings. Observation techniques are often divided into two categories, participatory observation and passive observation, where the researcher will either participate or observe from a distance (Crang & Cook, 2009). As part of the question-suggestion protocol, we conducted a form of participatory observations, in which we sat together with the participants.

Automated data collection tools

Automated data collection tools, such as screen recording tools, enable the researcher to easily collect detailed information about user interaction with a system (Lazar et al., 2010). These kind of tools may increase the amount of data collected, as well as ease the workload for the researcher. Screen recordings of applications are often used to test usability and see how the users interaction with the given software. It is sort of a passive observation of user action in the system and may provide information about possible struggles the user might have.

Interviews

Direct feedback from users is fundamental within human to computer interaction (Lazar et al., 2010). Interviews are not naturally occurring, they are constructed by the researcher, and therefore they do not provide direct access to the experience of the people studied (Silverman, 1998). However, they may contribute to a deeper understanding of users and their way of thinking, and open for direct feedback and subjective meanings.

Interviews can be divided into three groups, structured, semi-structured and unstructured interviews (Rogers et al., 2011, p. 228-229; Crang & Cook, 2007, p. 60). We have utilized semi-structured interviews, using both closed and open questions as a part of a script of subjects to discuss with our participants (Rogers et al., 2011, p. 228-229; Crang & Cook, 2007, p. 60). The advantage of semi-structured interviews, as opposed to structured or unstructured, is that one is able to elaborate on interesting statements while also covering all intended subjects. However, one has to be careful not to get too carried away, and remember to stick to the basic script of subjects.

Field notes

Notes are a flexible way of recording data, and if handwritten, they may also be less intrusive than for example using a keyboard (Rogers et al., 2011, p. 227). A disadvantage with notes is that it may be tiring to write, while at the same time trying to observe and listen. Though, this can be solved through working with another person (Rogers et al., 2011, p. 227). As a part of our research, we kept notebooks with field notes, which we tried to fill in daily. The purpose of such notebooks are to keep record of what the researchers learn, make sense of (mis)understandings and/or settings, and to provide detailed descriptions for readers to “stand in their shoes” (Crang and Cook, 2007, p. 50).

Analysis techniques

Crang & Cook (2007) describe two ways of analyzing an interview, statistically and discursively, depending on the number of informants and what kind of information you are after. Statistical analysis collects quantitative data, while discourse collects more qualitative data. We had a mixed approach to this, using both statistical and discourse analysis. Examples could be whether they entered data or checked tooltips first (statistical), or why they checked the tooltips before entering any data (discourse).

Data for statistical analysis need to be processed and cleaned because it may contain errors (Lazar et al., 2011, p. 70). It is important to trace as many errors as possible in the collected data in order to minimize the negative impact caused by potential errors (Lazar, 2011, p.71). In addition, some data need to be coded into numbers before any statistical analysis can be done (Lazar et al., 2011, p.71). For example, when coding gender, female could be coded to 0, while male could be coded to 1. To analyze our data we coded the recordings while we watched them, and gave them values in an excel-document. We also used t-tests, Wilcoxon's signed rank test and Pearson's correlation to find significant differences and correlation between different data sets.

1.7 Structure of Thesis

This section explains the unconventional structure of our thesis, as it also consist of three research papers, which are to be published in 2017. The papers have been copied into the thesis in their entirety, as they are individual, and may be read independently of each other and the thesis. Common for all is that the main research and evaluations are done by us, Helene Isaksen and Mari Iversen. Chipo Kanjo has contributed with all participants in Malawi, as well as arranging for a lot of the implementation of the research.

1.7.1 Chapter 1 – Introduction

The first chapter has introduced the thesis by explaining the context of the research, our motivation and some literature related to our methods. It finishes off with this part, explaining how our thesis is put together, consisting of papers to be published and additional text covering what has not been explained in the papers.

1.7.2 Chapter 2 – Methodology

The second chapter consists of another paper, which is to be published at the International Conference on Human Computer Interaction 2017 in Vancouver, Canada. It explains our study from a methodological perspective, exploring the validity of our research and the power of the methods used. This paper has been written in cooperation with our supervisors, Jens Kaasbøll and Chipso Kanjo. Kaasbøll has been responsible for major parts of this paper, including the literature, as well as most of the discussion. All parts related to power of methods and validity, is written by him.

1.7.3 Chapter 3 – Finding User Preference

The third chapter contains a paper, which is to be published at the IST-Africa Conference 2017 in Windhoek, Namibia. It also has a section which explains other aspects that were not included in the paper, due to the submission deadline being before the research was concluded. The paper explores user preferences for content types and expression formats for tooltips, based on our results. This paper was also written in cooperation with our supervisors, Jens Kaasbøll and Chipso Kanjo. It is difficult to specify who wrote what, though three of the authors, Isaksen, Iversen and Kaasbøll, contributed equally.

1.7.4 Chapter 4 – Finding Effective Tooltips

The fourth chapter consists of third paper, which is also going to be published at the International Conference on Human Computer Interaction 2017 in Vancouver, Canada, and presents some of the results retrieved during the quasi-experiment. As not all the results are included in the paper, due to the deadline of submission, one of the sections presents the final results. The other sections present aspects of the second iteration which were not included in the paper. The paper in this chapter was written by us, Isaksen and Iversen.

1.7.5 Chapter 5 – Conclusion

The fifth and final chapter will revisit the most important findings and answer our research questions. It may also repeat some of the conclusions from the papers in order to give a full picture of the outcome of our research. Lastly, we will give some reflections of our research and some practical recommendations, based on our research.

2 Methodology

The following paper will be published at the International Conference on Human Computer Interaction 2017, in Vancouver, Canada in July.

2.1 “Methods for Evaluation of Tooltips”

Methods for Evaluation of Tooltips

Helene Isaksen^{1*}, Mari Iversen¹, Jens Kaasbøll¹ and Chipo Kanjo²

¹University of Oslo, Oslo, Norway

✉ helenis@ifi.uio.no

mariive@ifi.uio.no

jensj@ifi.uio.no

²University of Malawi, Zomba, Malawi

chipo.kanjo@gmail.com

Abstract. Tooltips are context-sensitive help aimed at improving learnability of a system. Evaluation of tooltips would therefore be a part of evaluation of documentation, which again is a category under evaluation of software learnability. Previous research only includes two evaluations of tooltips, both gauging learning outcome after initial training, while the purpose of tooltips is helping users whenever in doubt when using systems after training. The previous evaluations are therefore of a low content validity. This paper concerns data field tooltips aimed at improving correctness of data entry. It presents studies a scale of content validities. On the low level is a questionnaire on users' opinion, which is a cheap evaluation. The medium type of evaluation was an adapted question-suggestion test measuring learning outcome. The high content validity evaluation method was a field experiment over two weeks, which demonstrated improved performance caused by tooltips. If the cheap questionnaire came out with the same preferences as the costly experiment, doing the questionnaire could have replaced experiments. However, the experiment did not confirm the results from the questionnaire.

Keywords: Research methods. Usability evaluation. Learnability. Context-sensitive help. Content validity. Explanatory power. Predictive power.

1 Introduction

The case triggering this research is a patient information system for nurses in developing countries, which is also used by health personnel below the nursing level due to scarcity of nurses. It was observed that the lower level personnel struggled with entering medical data. The practical objective of this research is to bring the lower level health personnel up to the nurses' level at entering health data. Due to other means of training being too costly and other interface design too inefficient, tooltips were deemed the most feasible way to improve the health workers' performance. Due to lack of knowledge on contents and expression in tooltips, the research aimed at finding design criteria for these two aspects.

The main purpose of a tooltip is to provide additional help to the users who are unsure about what to do, such that they are more likely to complete their tasks successfully. Tooltips are therefore aimed at improving software learnability and should be evaluated accordingly.

Grossman et al. [7] suggested a taxonomy of learnability definitions, including the user's competence level, their ability to improve performance and the time period over which improvement is going to take place. This study concerns the ability to improve performance over specific intervals for users whose domain knowledge is below optimal. Two different time intervals are included; one hour and two weeks.

Tooltips are parts of user documentation, hence their evaluation belongs in the metrics based on documentation usage in Grossman et al.'s [7] categories of learnability metrics.

The case concerns the lower level cadres' ability to perform at the nursing level after some practice. An evaluation of their work sometime after initial learning would constitute an appropriate measure of what the tooltip intervention aimed at, and we will call such appropriateness of the measurement method high content validity. However, field tests in real life settings are in general costly. Thus, a simple method for zooming in on the more useful types of tooltips before embarking on the most expensive evaluation would be advantageous.

A sequence of usability evaluations from cheap and theoretical to cumbersome and realistic could be:

- Heuristic expert evaluation according to guidelines for design.
- Lab experiment with users, e.g., thinking aloud.
- Field evaluation of actual use.

Since no guidelines for tooltip contents seem to exist, a heuristic evaluation was impossible. A questionnaire to the target user group on their preference was chosen as a low cost alternative. The questionnaire did not measure the health workers' learning, hence being of low content validity.

Lazar et al.'s [15] textbook on HCI research methods brings up the validity of methods in the sense of applying well documented procedures. Surveys with questionnaires can be carried out with rigor, but that does not improve their content validity in our case.

Content validity and cost are important qualities when selecting evaluation method. Since no assessment of learnability evaluation methods with respect to these qualities seem to exist, this paper aims at filling these knowledge gaps. In addition, the power of research findings to explain or predict is a consequence of choice of method and will therefore also be considered.

The next session introduce tooltips. Thereafter the theoretical background for content validity and power of output are presented, and these qualities will be used for characterizing previous tooltip evaluations. The evaluation methods applied in this research will be presented and assessed on these qualities. The methods will be compared and a taxonomy of evaluation methods will be built. In the conclusion, evaluation of tooltips for domain data will be compared to tooltips for IT functionality and to other inline help.

2 Tooltips

In this paper, we stick to an understanding of tooltips as a small window with help, which appears besides a button or data field on mouse-over or by tapping particular places on a touch screen. The tooltip disappears when the button is tapped or when the user starts or completes entering data in the field. This definition excludes in-line help, which stays on the screen until removed by the user. It also excludes alerts which pop up after a particular user operation or seemingly by itself, as for instance the Office97 Clippy [21].

When designing the tooltip, an important aspect is to not to overload the screen with extraneous information [10]. Earlier research has shown that too much help information may confuse the user, and prevent them from gathering the information needed to do the task [1]. Therefore, it is important to allow the user to stay focused by excluding unnecessary information. Both the need for keeping the task visible on the screen and making help minimal imply that tooltips should be short. Thus, the main challenge is to identify the necessary information for the tooltip and the right delivery mechanism for the information.

3 Aspects of Research Methods

This section will present literature on research methods relevant to our purpose.

3.1 Content Validity

The term validity has been used for several qualities of research methods. The validity type of particular interest for assessment of methods in this study, is whether the method measures what it aims at. Measure will be taken in a broad sense to include qualitative as well as quantitative data.

Since this paper deals with learnability, validity concepts from educational science are adopted. In educational science, the quality of “measuring what it aims at” is called content validity [16] and this term will thus be used in this paper.

We assume that tooltips and any other interventions to improve learning amongst users aim at long term impacts like improved efficiency, effectiveness (including fewer mistakes), safety, satisfaction, etc. Methods for evaluating tooltips should therefore measure such impacts in order to reach the highest content validity. Impact evaluations would require evaluation of the possible impacts (for instance, fewer mistakes) some time after the introduction of the tooltips, and attribution of the impacts to the introduction of the tooltips. Randomized controlled trials with a control group receiving placebo tooltips would be the method of choice, but these studies are normally very expensive and ethically questionable, since they require surveillance and the control group may receive a less desirable outcome.

Kirkpatrick [13, 14] developed a four level model for evaluation of in-service training, where impact is the highest level and the lower levels have lower content validity and also normally lower costs:

- Reaction. Reaction is the participant's opinion of the training. The reaction can, e.g., be found through a questionnaire asking their opinion of the training material and teaching.
- Learning. This is an assessment of what the user has learnt from the training. A pre-test before and a post-test after training will gauge the learning outcome.
- Behavioral change. An investigation of people's use of their new competence when back in business. For example, ask the users about to what extent they use some IT functionality being taught in the training or observe their use.
- Impact. This is a measurement of changes in organizational performance, for example the number of mistakes being made.

Both in-service training and tooltips are interventions for improving performance at work, and there is nothing inherent in the overall description of Kirkpatrick's model above which

prevents it from also being used for other interventions than training.

Evaluations at any of these levels can illuminate tooltips. The extent to which a user opens tooltips would be a Level 3 measurement which could be found by observing or logging use. A multiple-choice test of users before and after being exposed to a series of tooltips explaining domain concepts (Level 2) could unveil whether they improved their conceptual understanding. A questionnaire concerning alternative ways of presenting tooltips would be a Level 1 evaluation.

While a Level 4 evaluation would have the highest content validity, combining it with evaluation at Level 1 or 2 can also bring insight into why certain impacts are reached.

3.2 Power of Methods

Gregor [6] characterizes four different outcomes of information systems research;

1. Analysis and description of constructs. Relationships and generalizability, but no causality. E.g., "all novel users open tooltips" would be a description with no bearing on learning.
- 2a. Explanation of why things happened, causality. E.g., the user tells that she opened the tooltip because she wanted to know what the data field was about.
- 2b. Prediction of what will happen in the future if conditions are fulfilled. Predictions could be based on statistical correlation between a before and an after situation without being able to explain the mechanism behind the change.
3. Prescription, like a recipe which will bring about the wanted result. This could be a set of all necessary predictions to bring about the result. A sequence of instructions for carrying out a task could be a prescription of what individual users do. However, many users refrain from [22] or are not capable of [9] following such prescriptions, hence no results can be guaranteed.

We would say that this list constitutes an increasing power of the results of the research. Since Power 3 seems unattainable for tooltip evaluations, powers 2a and b are desirable.

4 Previous Evaluations of Tooltips

After extensive search, we have only come across two scientific papers evaluating tooltips. The evaluation methods in these papers are presented below and characterized according to content validity, power and cost.

4.1 Questioning Users on Preferences for Tooltip Expressions

A study by Petrie et al., [20] identified four ways of expressing tooltips for deaf and hearing impaired users: Sign Language, Human Mouth, Digital Lips and Picture tooltips. The 15 informants used the tooltips (randomly ordered) in two tasks. Thereafter they were asked to rate understandability, satisfaction, order of preference and provide other comments. Results were statistically significant with a non-parametric test.

Petrie et al., [20] asked the participants on their opinion of the tooltips, hence their method was at Kirkpatrick level 1. We therefore do not know whether the preferred tooltips will have a higher impact than those disliked by the informants. To get up to level 3 or 4 in content validity, the research should have included a test at a later stage than the introduction, where use of tooltips should have been correlated with the outcome of the task tested.

The significant preference implies that the results are predictive in the sense that other people in the target group will respond similarly. The open questions yielded qualitative data on why the Human Mouth and the Digital Lips were inappropriate, hence the power is at level 2a and b.

Nothing is stated concerning the cost of this study. A lot of investment has probably been made in setting up the tooltips and the system used. An additional test to improve content validity could therefore have been a worthwhile extension of the study.

4.2 Pre- and Post-test and Interviews

Dai et al. [3] developed a tooltip software extending Google Chrome and tested it with seniors.

Five seniors were questioned concerning their understanding of five functions, yielding a total of 3 correct answers. Then they were shown the tooltips for these functions. Afterwards, the same questions were given as a post-test; now with a total score of 24 for all participants. The evaluation was at the level 2 on content validity. No statistics were shown, thus the test has no predictive power. One participant said in an open interview that the tooltips were instrumental for him being able to search the internet, hence 2a explanatory power was demonstrated. Again, the authors seem to have invested a lot in the tool without carrying out the test which could have provided content validity at levels 3 or 4.

Some of the help provided by their software consisted of step-by-step instructions for carrying out tasks. Since tooltips disappear after one operation, they are unsuitable for displaying sequences of instructions. We therefore interpret Dai et al.'s [3] series of instructions as in-line help which falls outside the tooltip concept.

In summary, no evaluation of tooltips at content validity levels 3 or 4 seem to exist. Our recent studies target also these levels, and our methods and experiences will be presented in the sequel.

5 Evaluations Carried Out

The tooltips in our research concern data fields for Antenatal Care (ANC) for health workers in African countries. Tooltips explaining data fields are of particular importance in low income countries where nursing work is often carried out by health staff with lower qualifications. Our user evaluations were carried out in Ethiopia (low income), Malawi (low) and South Africa (middle income country).

As indicated in the Introduction, we opted for evaluations at several levels.

5.1 Expert Evaluation

Heuristic review based on design [15] constituted our initial consideration. The only design criterion, as mentioned above, is that tooltips should be short, and we saw no need for an external usability expert to measure the length of the tooltips. The way of triggering the tooltips in the software was given and outside of our control. Hence, heuristic evaluation in the HCI sense was deemed useless.

The authors are IT experts, while the tooltips concern medical data for users being health personnel. Therefore, we had the contents of the tooltips checked by two nurses and one medical doctor. They responded with three comments leading to some changes in the tooltips.

Kirkpatrick's level 1 Reaction is the participants' opinion of the training. Extending the concept of the participant to include external evaluators of the training material, an expert evaluation can be considered at the lowest level of content validity. It has explanatory (2a) power but not predictive, and it has the obvious advantage of low cost.

5.2 Questionnaire with Subsequent Interview

We aimed at finding out the preferred contents and expression format for tooltips for data fields; the study is presented in (Isaksen et al., submitted for publication). For these objectives, we followed the approach from Petrie et al. [20], asking our users to rank different tooltips according to preference, and followed up with conversations/interviews based on their answers to the questionnaire.

We first interviewed researchers familiar with ANC systems in African countries, which lead to the following suggested tooltip content types:

- The formal medical definition, e.g., Fundal height is the distance from pubic bone to the top of the uterus.
- Normal values for the medical term, e.g., Normal fundal height measurement: 20 weeks = 17-20 cm, 28 weeks = 25,5-28,5 cm, 36 weeks = 33-35 cm, 40 weeks = 36-38 cm

- Treatment following danger signs, e.g., If measurement is abnormal, please refer the patient to a specialist.

These types were included in the questionnaire. After 28 responses, a fourth content type was also identified;

- procedures in order to find values.

Since this type came up late, we decided to keep the questionnaire with the three first content types.

Several delivery mechanisms or expression formats were also identified; text, illustration, videos and table. However, due to limitations we were not able to use video as expression format in the questionnaire. The tooltips in the questionnaire included the five combinations illustrated in Figure 1.

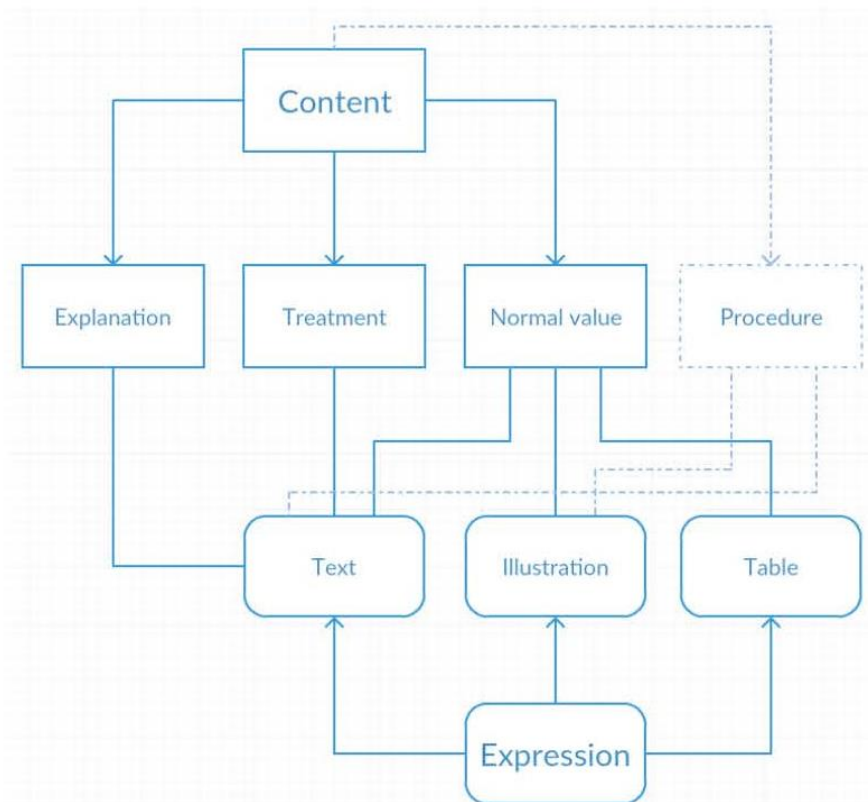


Figure 1: Combinations of content and expression types in questionnaire

The questionnaire consisted of three cases where the informants were supposed to rank the different options on a scale of 1 to 4, where 1 was the most preferred one. Figure 2 shows an example. The labels in red were not included in the questionnaire but added here for clarification.

Fundal Height

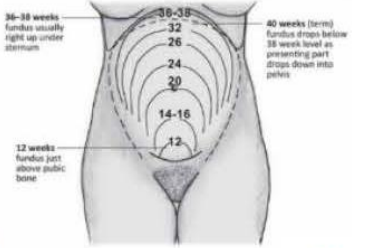
<p>Measurement form the pubic bone to the top of the uterus</p> <p style="text-align: right;">Definition</p>	<p>If measurement is abnormal, please do extensive examination and tests, or refer the patient to a specialist.</p> <p style="text-align: right;">Treatment</p>
<p>Normal fundal height measurement:</p> <p>20 weeks = 17-20 cm</p> <p>28 weeks = 25,5-28,5 cm</p> <p>36 weeks = 33-35 cm</p> <p>40 weeks= 36-38 cm</p> <p style="text-align: right;">Normal values</p>	 <p style="text-align: right;">Normal values</p>

Figure 2: An example from the questionnaire

Statistically significant differences were found from 58 respondents, see Isaksen et al. [11].

After the informants had filled the questionnaire, we asked them to elaborate on why they answered the way they did, or if they had any further comments or suggestions. Some referred to the textbooks they were familiar with as the cause of their preference, since the tooltip resembled the explanation in their textbook.

Corresponding to Petrie et al. [20], our method was at Kirkpatrick level 1, meaning we don't know the possible outcomes of using the tooltips. Also like Petrie et al. (2004), the qualitative interviews provided some explanations, such that the power is at level 2a and b.

Some of the informants were not fluent English speakers, even if they used English for patient recording. One of the researchers translated into local language. This observation concerned also the two evaluations following below.

One unexpected lesson early in the study, was that presenting the questionnaire before the informants had actually seen or used the tooltips, left them in limbo as to what tooltips were.

Hence, after five informants, we changed the order, doing the Adapted Question Suggestion (below) before the questionnaire. This provided the participants with some experience while filling the questionnaire, and increased their understanding of the task.

5.3 Adapted Question Suggestion

This evaluation aimed at finding out how users managed to open and understand the tooltips in an application. Thus we were aiming for more than Kirkpatrick level 1 and needed users to test the tooltips in our ANC system.

The applications were built within the District Health Information System [4], using its Tracker Capture app for Android devices. The informants were 17 health workers with different level of knowledge and experience in both domain and technology, and 11 students, a total of 28 informants in Ethiopia and Malawi. The informants were recruited either by showing up at their respective clinics and asking for their time, or by calling shortly ahead, asking for permission to visit them. This was a convenience sample, and all informants were recruited through local contacts, who also contributed with translations when needed.

Two testing programs were created, one for informants in Ethiopia and one for the informants in Malawi. The application used in Ethiopia was based on the Ethiopian community health information system program form for ANC, while the program for Malawi consisted of a selection of data elements in the Malawian health passport for pregnant women.

To structure the testing sessions, we developed cases, where the aim was to make the informants to go through the testing program and use the provided tooltips. We wanted to observe whether they were able to enter the information without any problems or issues and whether they opened the tooltips.

Our initial thought was to develop two cases of various difficulties, aiming to see how the different informants would cope. The first version used the same expressions in the case as in the data field title, aiming for an easy start. The second type of case challenged the informants by not using the same expression as the data field title, but rather using the terms which appeared in the tooltip. However, after trying out the cases on our first group of informants, we figured that one case was enough due to time constraints, so the simple case was abandoned.

A sentence from the case is:

During Manjula's first pregnancy, she lost her female child in the 36th week of pregnancy, before the onset of labour.

The correct data entry based on this sentence would be to tick the data field

Antepartum Stillbirth

Users who were unsure about where to tick could open the tooltip for Antepartum Stillbirth and find:

Birth of a fetus that shows no evidence of life. Occurring before the onset of labour.

The tooltip has expressions which match the case, hence the user could infer that this is the correct choice.

Evaluating the use of tooltips could be carried out in several ways.

Time to complete a case is a metric for learnability [7], but requiring that the user looks up tooltips on the way may be counterproductive, since tooltips should be accessed only when in doubt. In our study, correct data entry is more important than speed. One way of comparing the effects of tooltips would be to set up groups of users with the same system and cases but different tooltip contents. This might have been

achievable in a lab session lasting a couple of hours. However, at the time of setting up the test, we did not have the questionnaire response, and we were not able to gather a sufficient number of informants for testing five different type of tooltips and compare the outcome. Hence, we used the medical definitions, since this seems to be the common way of providing explanations.

Methods for evaluation of software usability have not targeted inline help, like tooltips. Grossman et al. [7] developed the question suggestion (QS) procedure which targets software learnability specifically. Since we would evaluate learnability, we took QS as a starting point. QS builds on the Thinking-Aloud protocol. It requires an expert sitting alongside the learner suggesting alternative ways of working, and this has unveiled 2-3 times as many learnability issues as thinking-aloud [7].

Our aim was not testing learnability of the software, but of the tooltips in the software. Distinct from Thinking-Aloud, QS could take our users past possible difficulties they may encounter with the system, and allow focusing on the actual use of tooltips, rather than the learnability or the natural use of the system. Without this adaptation, QS has the disadvantage of only making the user access some tooltips, otherwise tooltip suggestions may constitute obstacles for the learner.

We also switched QS from lab to field, since this cater for more reliable results [5]. This implied that we had to cater for the available group of informants, and could not assign one expert to one informant. With up to five informants, it was difficult for two expert to follow up all. At times, some of the informants held the tablet in front of them, disabling observation. Figure 3 shows a session in a health facility.



Figure 3: Adapted QS in the patients' waiting area. The back of two of the researchers.

The QS session started out with a short introduction and asking the informants some basic questions about their technological experience. We then proceeded to going through the case with the users, helping them if we saw them struggling with anything. We always ensured the users that asking questions was okay. We introduced them to tooltips by showing where to tap and explaining the purpose of the tooltips. We also asked them questions along the way and reminded them of the option of accessing the tooltip button if we saw them answering wrongly.

We tried to install software in the tablets to log use, but the software failed. We therefore only observed and noted what the informants did. As stated above, this was impossible at times.

In summary, the observations showed that nurses and midwives often knew which data to enter. However, informants with less education were often unsure about the match between the case information and the data fields, and were encouraged to look up the tooltips to answer correctly. In some cases, it helped them understand the titles, but many of them still answered wrongly. These results are at Kirkpatrick level 2, showing the learning outcome of the tooltips. No statistical data was collected, but the difference between health workers with and without nursing degrees was clear from the qualitative and partly quantitative observations, hence providing a modest predictive power.

Two nursing students in years 3-4 were actively using the tooltips and mostly entering correct data. They explained that the tooltips were used just to verify their own input to the system. This answer was surprising and provided a new insight into the learning effects of tooltips, as verification is a positive reinforcement of learning [19]. Second, it points to that learning effects of tooltips cannot be measured only by looking for users who look up tooltips before entering data. Third, it provided some explanatory power to the results, such that the experiment had some power at level 2a and b.

Similar to the studies of Petrie et al. [20] and Dai et al. [3], the cost of this experiment was relatively high without bringing about a higher content validity.

Since the questionnaire had come out with normal values as preferred tooltips with medical definitions significantly lower ranked, and since the investment of setting up the test could be reused in a test with higher validity, we decided to also carry out a test at Kirkpatrick level 3 or 4. This test is described in the next section.

5.4 Logging Use

Evaluations of in-service training at level 3 and 4 concern users' application of what they have learnt during training in their work. This is called transfer of training to work and is, counter to intuitive beliefs, normally unsuccessful [8].

Kirkpatrick level 3 evaluates behavioral change. In an experiment [11] tooltips were introduced in a training session similar to the adapted QS. We thus interpret behavioral change as users opening tooltips also for a prolonged time after the introduction. Level 4 concerns improvement of performance, and this would in our case be an increased percentage of correct data entered. To ensure a time distance from the introduction to use of the system and the tooltips, we let the users use the system for two weeks.

Transfer should be to work. Being a system under development, we had to substitute work with work-like, fake data. We developed a booklet consisting of 22 cases, where our informants had to, each day during a period of 11 days, use the cases and fill information into the app. The booklet also included open ended questions for the day. We followed the same style in the cases as we did during the adapted QS. In order to measure the learnability of the tooltips, during a period of time, similar cases appeared at different days. The aim was to see whether or not the tooltips provided were understandable.

Based on the results from the questionnaire, we chose to compare explanations and normal values as content type for tooltips. By using the existing testing program from Malawi, with a few minor changes, we created a copy of the program and changed most of the tooltips to normal values. Both programs were installed on 30 tablets, and given to 20 participants in Malawi and 10 in South-Africa.

The participants were again recruited by convenience, although we tried to avoid those who did the Adapted QS. All participants were given one tablet (including a SIM card and airtime), locked for all other use than the test program. In order to track the informants' progress we implemented the screen recording program "UXcam" on each tablet. UXcam enabled us to record and watch every touch points and gestures the informants made and analyze the outcome. Thus, we were able to see whether the informants opened the tooltips and whether they filled in the correct information and used the

correct data element. We emphasized to our participants that they should have internet connectivity whenever they use the program. We informed the participants that the screens were recorded and that they should never enter real patient data in the system, only the cases we provided.

In order to motivate all participants to fulfill the test, we told them that they could keep the tablet after the test and that we would open it for any use. Since we had no other plans for the tablets after the experiment and since paying cash to participants in foreign countries out of a university account is a bureaucratic process which has previously failed, we went for the gift option. The value of the tablet could correspond to half a month's salary for the health workers in Malawi, and we hoped that this would lead to all the participants to completing the experiment.

We handed out 22 cases of pregnant women to each participant and gave them the same open ended questions to fill daily. Over a period of about 2 weeks, they entered information from two cases a day in the system and answered the questions of the day. The two first authors watched the videos and entered data for opening of tooltips and correct data in Google sheets and also carried out all statistical analysis there.

After two weeks, we returned to the participants, and interviewed them on why they did as they did, and what they think of the experiment now that it's done.

At the deadline of paper submission, 15 of the participants had completed the experiment. Due to internet issues, only $\frac{2}{3}$ of the videos were recorded.

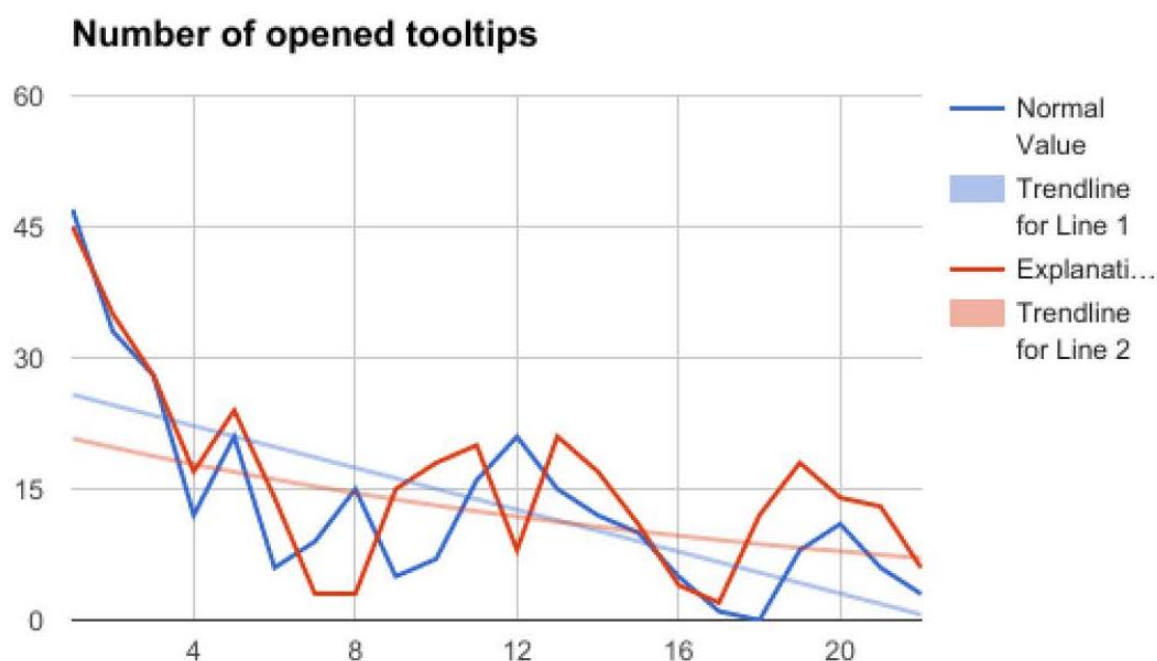


Figure 4: Number of opened tooltips throughout the cases.

Figure 4 shows the trend in opening tooltips less frequently over the cases. This gauges the behavioral change at Kirkpatrick Level 3.

A successful tooltip is when the user has opened the tooltip for the data field and entered correct data. Due to that some users opened the tooltip to verify data entered, we do not distinguish between opening the tooltip before entering data or vice versa.

In order to analyze differences over time, we compared the first third of the cases with the last third. The average number of successful tooltips during the first seven cases was 1.52, and in the last seven cases 0.62, which is a significant difference (T-test, two sided, paired, Google sheets, $p=0.02$). Thus the log has a predictive power (2a) concerning tooltip use.

The reason given by the participants in the interview was that after some time, they knew and didn't have to look it up more times. Thus explanatory power (2b) was added in the interviews. The booklets assisted the participants during the post-interviews, and they referred to it when they, for example, explained what they found confusing in the cases. It also contributed to further discussions, as we were able to ask them about things they might not have memorized.

Table 1: Results on correct data entry from logging use

	Average % correct first 7	Average % correct last 7
Normal values (n=7)	76	87
Explanations (n=6)	83	85
All participants	79	86

Table 1 summarizes results from the 15 participants on changes in performance. Due to videos not being recorded, only 13 users had traceable results both during the first and last seven cases. Pairwise statistically significant differences are marked in grey.

Significant improvements and significant difference between the normal value and explanation group and interview results, made Isaksen et al. [12] to conclude that tooltips caused impact on correctness, (Kirkpatrick level 4) with predictive (2a) power without being able to state the size of the improvement in correct data entry.

5.5 Summary of the Evaluations

The evaluations carried out by the authors are summarized according to the content validity levels as defined through Kirkpatrick's [13] model, see Table 2.

Table 2: Outcome of evaluation methods according to Kirkpatrick's four level model for evaluation of training

	Opening tooltip	Tooltip content and expression
1 - Reaction	Interviews (2b)	Questionnaire + interviews (2a+b)
2 - Learning	Adapted QS (2a+b)	
3 - Behavioral change	Logging use + interviews (2a+b)	
4 - Impact	Logging use + interviews (2a+b)	Logging use (2a)

A weakness in the logging at levels 3 and 4 was that the participants did not use the system as part of their job, but as a side activity for which they were rewarded.

The series of evaluations required about two years of work for the researchers. The 30 tablets cost USD 10 000, and travel costs are additional.

6 Discussion and Conclusion

Grossman et al. [7] categorized learnability metrics concerning use of IT. They identified documentation usage area as one out of seven categories, and the assessment methods in this paper concerns tooltips, which is within the documentation category.

Previous evaluations of tooltips [3, 20] gauged users' opinion and learning outcome of the tooltips. The three first studies carried out by the authors of this paper, expert evaluation, questionnaire and adapted QS, also measured opinion and learning outcome. All of these studies required a considerable amount of work for setting up the systems and creating the tooltips. The purpose of tooltips is to assist users learning about the system during use. Yet, all of these studies were only able to find users' opinion of tooltip contents or gauge the learning outcome at the end training, hence the studies did not measure precisely what they were supposed to. This is characterized as low content validity.

A model for evaluation of training [13] has come up with four levels of content validity, where the learners' opinion and their learning outcome are the two lowest levels. With heavy investments already done, in our case, 15 months, it was a pity not to follow up with a study at higher content validity level, where the informants used the system for some period for its normal purpose in a real or close to real setting. Our approach was to give the informants tablet PCs and cases to enter over a two weeks period where they worked on their own but could also consult colleagues. Their activities were logged. This last experiment consumed around 9 months of work.

The experiment was able to demonstrate that users opened tooltips after the initial training, and that their usage dropped because they learnt more of the system by means of the tooltips. Their opening of the tooltips is a behavioral change resulting from the training. Behavioral change is at level 3 of content validity in the training evaluation model [13], being more valid than the user opinions and learning outcomes.

Finally, the experiment also demonstrated that the tooltips worked as intended, in the sense that users entered more correct data as a consequence of opening tooltips. These findings explained what happened in addition to being able to predict that tooltips will help users enter more correct data. With no placebo tooltips included, it is impossible to conclude about the proportion of improvement caused by the tooltips.

A research question in the questionnaires and in the experiment was which type of contents of the tooltips that were superior. In the questionnaire, normal values for a variable was preferred over a medical explanation. If normal values also led to more correct data entry than the explanations in the experiment, this would have been an indication that future tooltip designers could do with a cheap questionnaire instead of setting up a costly experiment. Unfortunately, the explanations provided more correct already from the start, while the normal value group reached the same or a better level after having entered 17 cases in the system. Based on these findings, we cannot conclude that questionnaires can replace experiments.

This study consumed approximately two years of work, plus 30 tablet computers. Such an investment could not be justified for a system with a small user group. The point-of-care system studied could potentially have tens of thousands of users. For such a user base, improved tooltips could replace parts of costly training and possibly also reduce errors; the latter being crucial in health services.

This study also aimed at the more general research objective of finding out the better type of contents in tooltips. The explanation type yielded quicker improvement in performance, hence this type of tooltip could also be used for information systems in other domains until other research demonstrates otherwise. Also the study showed that tooltips help, meaning that other system developers should include the small effort of making the tooltips, even if they don't evaluate them.

7 References

1. Carroll, J. M.: *The Nurnberg Funnel: Designing Minimalist Instruction for Practical Computer Skill*. Cambridge, Mass., MIT Press. (1990).
2. Dai, Y., Karalis, G., Kawas, S., & Olsen, C.: *Tipper: Contextual Tooltips that Provide Seniors with Clear, Reliable Help for Web Tasks*. CHI'15 Extended Abstracts, 1773-1778 (2015).
3. DHIS2. (2017). Available from: <https://www.dhis2.org/>.
4. Duh, H. B. L., Tan, G. B. C., Chen, V. H. H.: *Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests*. Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services, Helsinki, Finland, 181-186 (2006).

5. Gregor, S.: The nature of theory in information systems. *MIS Quarterly* 30 (3), 611-642 (2006)
6. Grossman, T., Fitzmaurice, G., Attar, R.: A Survey of Software Learnability: Metrics, Methodologies and Guidelines. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston, MA, USA, 649-658 (2009).
7. Grossman, R., Salas, E.: The transfer of training: what really matters. *International Journal of Training and Development*. 15, 103-120 (2011).
8. Hadjerrouit, S.: Using a Learner-Centered Approach to Teach ICT in Secondary Schools: An Exploratory Study. *Issues in Informing Science and Information Technology*. 5, 233-259 (2008).
9. Instone, K.: Heuristics for the Web. <http://instone.org/heuristics>.
10. Isaksen, H., Iversen, M., Kaasbøll, J., Kanjo, C.: Design of Tooltips for Health Data. In *proceedings of IST-Africa* (2017).
11. Isaksen, H., Iversen, M., Kaasbøll, J., Kanjo, C.: Design of Tooltips for Data Fields: A Field Experiment of Logging Use of Tooltips and Data Correctness. *HCI International 2017*. (2017).
12. Kirkpatrick, D. L.: Techniques for evaluating training programs. *Journal of American Society of Training Directors*, 13, 21-26 (1959).
13. Kirkpatrick, D. L., Kirkpatrick, J. D.: *Evaluating Training Programs: The Four Levels*. Berrett-Koehler, San Francisco (2006).
14. Lazar, J., Feng, J. H., Hochheiser, H.: *Research Methods in Human-Computer Interaction*. John Wiley & Sons Ltd, West Sussex. 295 (2010).
15. Messick, S.: Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*. 14 (4), 5-8 (1995). DOI:10.1111/j.1745-3992.1995.tb00881.x.
16. Ormrod, J. E.: *Human Learning*, Englewood Cliffs, New Jersey, Merrill (2012)
17. Petrie, H., Fisher, W., Weimann, K., Weber, G.: Augmenting Icons for Deaf Computer Users. *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, Vienna, Austria, 1131-1134 (2004).
18. Shroyer, R.: Actual Readers Versus Implied Readers: Role Conflicts in Office 97. *Technical Communication*. 47, 2, 238-240 (2000).
19. Smart, K. L., Whiting, M. E. & Detienne, K. B.: Assessing the Need for Printed and Online Documentation: A Study of Customer Preference and Use. *Journal of Business Communication*. 38, 285-314 (2001).

3 Identifying User Preference

The following paper is written by Helene Isaksen, Mari Iversen and Jens Kaasbøll from the University of Oslo, and Chipso Kanjo from the University of Malawi. It is to be published at the IST-Africa Conference in Namibia May/June 2017.

3.1 “Design of Tooltips for Health Data”

Design of Tooltips for Health Data

Helene ISAKSEN¹, Mari IVERSEN², Jens KAASBØLL³, Chipso KANJO⁴
^{1,2,3}*University of Oslo, Oslo, Norway*
+47 22852410, {helenis,mariive,jensj}@ifi.uio.no
⁴*University of Malawi, Zomba, Malawi*
+265 1528775, chipso.kanjo@gmail.com

Abstract: Tooltips are regarded as beneficial methods for user to understand either a user interface element or tasks related to the system. However, little research has compared text, tables and illustrations or addressed the content of tooltips, thus this research aim to address this aspect. Through a question-suggestion approach, accompanied by interviews and questionnaires, we have looked at what actual users would prefer as expression format and content type for tooltips. We found that text is the preferred type of expression, while normal values are the preferred content type.

Keywords: Tooltips, ANC, DHIS2, design, health systems.

1. Introduction

Health workers in many African countries encompass several professions, ranging from community health workers to medical doctors. However, most facilities, particularly in rural settings, have nurses and midwives as the most educated personnel. Staff turnover may be high, and in such cases, those from lower cadres (often less educated) have to step in, doing tasks which according to protocol should have been done by those with more education. Like elsewhere in the world, remote rural places seem to be less attractive for highly educated staff.

While newly graduated staff in industrialized countries may have trouble understanding some medical terms, it is worse in situations where staff have to perform tasks that go beyond their formal qualifications. For example, in Malawi, a patient attendant may have to capture pregnancy history in place of a nurse or midwife. The pregnant woman may say that she lost a child at week 37 in her last pregnancy, and the patient attendant may register this as an abortion, while the correct entry would have been antepartum stillbirth. Such erroneous registrations may have consequences for the pregnant woman, and will cause distorted figures in the final statistics.

This study aims at reducing the problem of wrong data entries for undereducated health workers in Africa. Our results are going to be useful for the project mHealth4Africa [1], which concerns antenatal care (ANC), hence ANC is our domain. However, this will be applicable to other areas of care because it addresses the importance of identifying content types and expression formats for tooltips.

Training courses and guidance on the job constitute ways of improving the health workers' skill, but these may be expensive or impossible due to staff shortages. Therefore, we set out to provide help during data entry for health workers who are unsure about the interpretation of the data to be entered. A possible way of providing assistance could be to include a longer explanation in front of each data field, instead of only a short field title. However, longer texts would have cluttered the screen for expert users, and since

applications aim at efficient use by experts, the screen should be simple. Therefore, use of help functionality is preferred.

Studies indicate that users do not read manuals and seldom search for help [2] [3]. Users prefer context-sensitive help concerning their current location in the software, and tooltips have been found useful [4]. This research concerns optimal design of tooltips.

Tooltips aim towards helping users understand the user interface (UI) element, through short explanations of what the UI element is or can do, or how to perform different tasks in relation to the element. While health personnel also need to learn the functionality of the technology, this study focuses on the domain knowledge of the patient information system. The target group for the study are undereducated health personnel, as described above.

2. Tooltips

Tooltips normally appear as a box, when a UI element is hovered over, also called balloon help [5]. However, most smartphones and tablets do not offer hover functionality. Instead, on-click tooltips, triggered by users clicking an UI element, can be used. When the tooltip has been read, the user may need to close it, or it disappears when the cursor or finger are moved.

An important aspect when designing for usability is to identify users' needs and wishes for displaying information, and making the user understand the content in the most efficient way. We have identified five possible ways of displaying tooltips; plain text, tables, illustrations, photos and videos.

Textual tooltips are used widely. They consist of plain text, either explaining a concept or naming the UI element. Textual tooltips also enable skimming and re-reading, and research suggest that many users just skim the help text [6][7]. Users may not be able to understand certain terminology (ibid.), thus it is important to find familiar words and expressions. The text should only consist of the most essential information, and that will mainly depend on the system, the task to be done and the users themselves.

Table tooltips consist of a table and a short explanatory text or title. Using a table can minimize the number of data values, giving the user less to focus on and better access to the actual information they need [8].

An illustration may include multiple elements and can, e.g., show how to measure values or carry out tasks. Earlier research has concluded that an illustration accompanied by a textual explanation has improved the learnability of systems [9]. Visuals within tooltips may lead to the user performing tasks in less time and with fewer errors, as opposed to tooltips without visuals [10].

Photo normally have a lot more details than illustrations, allowing the user to access more information in a smaller space. However, this may also be a problem, as users may need to focus more to find exactly what they are after.

“Toolclips” include narrated video clips and in-depth textual documentation [4]. Research shows that informants using toolclips showing steps in procedures to accomplish tasks completed seven times more unfamiliar tasks than others [ibid]. However, we were not able to use videos in our tooltips due to technological limitations.

An important principle of information design and usability is not to overload the screen with extraneous or irrelevant information [11]. Too much information may confuse the users and prevent them from extracting the information needed for the task. Tooltips should be short, since users prefer doing, not reading [6]. Users often look for quick ways to make changes to their work, and may therefore be unwilling to risk the investment of time and effort to read lengthy texts and explanations [7].

3. Objectives

The objective for this research is to find optimal presentation of tooltips, such that more health workers will understand the health data they enter in a patient information system for ANC. Little research has compared text, tables and illustration or addressed the content of tooltips, thus our research questions are;

- do users prefer text, tables or illustrations in tooltips for data entry?
- which type of contents for tooltips in data entry do users prefer?

4. Methodology

Initially, we carried out three interviews to gain experience on which kinds of issues users struggled with and what kind of help they preferred. The informants were researchers who had worked with ANC information systems in African countries.

The field work for the main study was done in several sites in Malawi, Ethiopia and South Africa, focusing on the more rural clinics in Malawi, a central clinic in Ethiopia and a hospital in South Africa. These countries participate in the mHealth4Africa project [1].

4.1 Informants

During the fieldwork, a total of 58 informants, 44 in Malawi, 4 in Ethiopia and 10 in South Africa, were involved in the sessions. The original plan was to have a more equal deployment of informants. However, because of the taut situation in Ethiopia at the time of the study, we were not able to go to our contact located further north in the country. Also, due to communication and availability issues in South Africa, we only got 10 informants there. Instead, our contact in Malawi, was able to find more informants, in order for us to have a proper sample size. We included health professionals with various technological skills and domain knowledge.

The informants were recruited either by showing up at their respective clinics and asking for their time, or by calling shortly ahead, asking for permission to visit them. All informants were recruited with help from local contacts, who also contributed with translations when needed.

4.2 Document Analysis

We studied an existing Malawian ANC health passport, in order to design a more familiar system for testing in the sessions in Malawi and South Africa. This contributed to finding a selection of data elements and a good displaying sequence. We used familiar titles for the data elements, which can also be found in the health passports.

For our research in Ethiopia we used a community health information system programs form for ANC. This was a request from our contact in Ethiopia, as this form had been developed in co-operation with the Ministry of Health.

To get a better understanding of the health workers' knowledge and focus, we analyzed some of the maternal and neonatal training material used in Malawi. This gave us insight in their ways of thinking, in terms of procedures related to their occupation. We also got pointers on how we should formulate possible textual tooltips.

4.3 The Testing Programs

We created two testing programs and adapted for health workers in each of the countries, one for Malawi and South Africa, and one for Ethiopia. The testing programs were used to give the health workers some context on the tooltips' functionality, as well as when the

tooltips were supposed to be obtained by the users. During the study, one of the testing programs was modified, due to technical difficulties. The main difference between these two, was that one contained warnings with normal values, while the other didn't. Warnings appeared if abnormally high or low values were entered. We also made some changes to the previous pregnancy stage. Instead of creating new events for every child, the data fields got a tabular format, and enabled entering information about all previous pregnancies in the same event. This was done to make the system more familiar to health workers, as this is how they usually register previous pregnancies. Explanations as content type for the data elements were used in the testing programs.

4.4 Question-Suggestion method

We chose to have a modified question suggestion approach [12], which involved observing and sitting together with the informants, guiding them and suggesting alternative ways of working, if we noticed them being in doubt of anything. During the sessions, notes were taken on their use of the testing programs and their interpretation of the questionnaires.

In order to give the testing part of the sessions some structure and consistency with the different informants, we developed a use case (figure 1). All informants used this, imagining it would be a pregnant woman speaking and answering questions appearing in the app. The use case focused on entering information about the woman's previous pregnancy, without explicit telling the exact term or data element to be entered. The use case also included blood pressure (BP), which is normally entered as systolic/diastolic, e.g. 120/80. In our testing program BP is entered as two separate values. In order to check whether informants read and understood the data field title and the associated tooltip, we switched their order.

Registering a pregnant woman

Below are some information about a fictive person. Please use the Tracker Capture app to register this patient's information.

Date of visit: September 15th, 2016
Name: Manjula Chakila
Birthdate: March 12th, 2000
Mobile number: 88 77 44 55 66

Manjula's previous pregnancies

- During Manjula's first pregnancy, she lost her female child in the 36th of pregnancy, before the onset of labour. During her pregnancy, she suffered from abnormally high blood pressure and protein in her urine.

Figure 1: Example of use case

4.5 Semi-structured Interviews

The question-suggestion session also included semi-structured interviews, where we discussed subjects such as education and technical experience. Notes were taken during the interviews, but no audio recording. We also asked the informants about their thoughts while they were using the testing program, and the questionnaires created discussions.

4.6 Questionnaire

Questionnaires were developed to find out how the tooltips should be formulated and what kind of content the health workers would prefer. These were, as mentioned, used for creating discussions, both amongst the informants, and with us, regarding why one option would be better suited than the other.

It consisted of three examples of data elements; fundal height, hypertension and pre-eclampsia. Fundal height and hypertension had four alternatives for content, while pre-eclampsia had three alternatives. The labels next to the boxes in the corners were not included in the questionnaire but added here for clarification (see figure 2).

A T-test was utilized to find significant differences between the content types.

Fundal Height

Measurement from the pubic bone to the top of the uterus	If measurement is abnormal, please do extensive examination and tests, or refer the patient to a specialist.
Explanation	Treatment
Normal fundal height measurement: 20 weeks = 17-20 cm 28 weeks = 25,5-28,5 cm 36 weeks = 33-35 cm 40 weeks = 36-38 cm	
Normal value	Normal value

Figure 2: Example from the questionnaire

5. Technology Description

District Health Information System version 2 (DHIS2) is a generic software package for hierarchical organizations, enabling aggregation of statistics and tracking cases following specified processes [13] [14]. The DHIS2 is run through a web browser or Android apps, and stores data in a server. For this study, the Tracker Capture (TC) app was used for hosting our testing programs. TC is based on an Android operating system, and contains practically the same functions as the web-based TC. It can register and track people or objects over a period, and contain search and enroll functions, just like the web app.

Still, in contrast to the web app, the Android app can store data locally, meaning it is not dependent on network connection. Also, considering that it can be used on devices like smartphones or tablets, it is not dependent on neither stable power nor generators. This is a huge benefit when implementing electronic health systems in rural areas with low or no connectivity and unstable or non-existent power.

Configuration of TC is done in the web-version in DHIS2, and affects both the web app and the Android app. For this study, organization units, attributes, programs, stages, sections, option sets, program rules and data elements were created from scratch, in order to tailor the testing programs. The tooltips (figure 3) were also added and edited in the web-based version under “Description” in each data element. Only text or URLs can be entered here.

Description
 Measurement from pubic bone to the top of the uterus. This is done to assess how far into the pregnancy the woman is.

Form name
Fundal height

Domain Type (*)
Tracker

Figure 3: Example from the interface where tooltips are created

6. Developments

The tooltips in this study have been of the textual kind. Links to web pages which could include illustrations and tables were deemed unusable, since the app may be used offline in areas without internet connection.

In order to see these tooltips, the informants have to trigger them manually, by pressing an icon to the left of the data element in the app (see figure 3). Textual information about the data element then appears in a rectangular, white box in the middle of the screen, titled “Detailed information” at the top, and with an “OK” button at the bottom. The rest of the screen surrounding the box becomes somewhat darker in order to emphasize the detailed information box (see figure 4).



Figure 4: The button for tooltips



Figure 5: An example of tooltip in the app

7. Results

7.1 Initial Study

Based on the interviews from the initial study, three content types for tooltips were identified; formal definition, treatment and normal values (see figure 6).

We included formal definition as a content type, for more undereducated health workers to understand the terms. An example used in our research is «Pre-eclampsia occurs when the woman has high BP and protein in the urine. It can happen at any point after week 20 of pregnancy.»

A tooltip about treatment should guide the health worker into the correct procedure if there are any danger signs present. For instance, if the health worker answers yes to severe hypertension, a tooltip could be-, “give medicine for lowering BP”.

A range of normal values help detecting danger signs regarding the client. An example is BP, where the tooltips consisted of normal BP values for pregnant women.

After the analysis, we identified a fourth content type; how to measure the actual value or carry out a procedure. An example is how to measure the fundal height.

7.2 Main Study

None of the informants found the tooltips-button by themselves before getting introduced to them, indicating that they need to be more prominent and visible.

The use cases forced users to figure out the meaning of the data field titles, through the tooltips. Medically educated informants, such as nurses and midwives, did not always need to check these, as they already knew the answer. However, undereducated informants

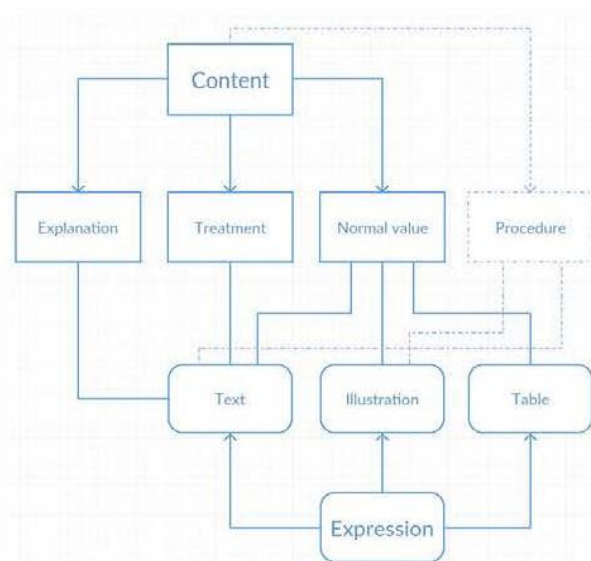


Figure 6: Illustration of how content and expression is connected

weren't always sure about the field titles, and were encouraged to use the tooltips to answer correctly. In some cases, it helped them understand the titles, but many of them still did not answer correctly.

Nurses seemed to be more confident in their medical knowledge and therefore did not use the tooltips as much as informants with lower education levels. The exception was two nursing students at the end of their studies, actively using the tooltip-buttons and entering the correct data in most places. They explained that the tooltips were used just to verify their own input to the system. They also stated they did not think that they would use the tooltips in real life, however, someone with less education may find them more useful in an everyday situation.

Concerning the switched order of BP, all users systematically entered the numbers in the sequence they were used to, hence got it wrong, even though explicitly encouraged to read the tooltips. This may indicate that tooltips are either not read or understood, possibly due to how they were expressed. This is a strong indication that tooltips cannot mediate a poor design where the conventional sequence of numbers are reversed.

A CHW suggested rewriting the data field title to "Does the woman have BP within..", to find out if the woman could be in the danger zone. He said that this would be of great use for him, and that if he answered this question with yes, he could refer the woman to a clinic for further examination and tests. This comment indicates that if health workers know what the normal values are, it will be easier for them to know when to refer a patient to further examinations.

As table 1 shows, normal values as text is significantly (based on the t-test) preferred over the other content types. Explanations and normal values as illustrations or tables have similar scores, while treatment is significantly less preferred than the others. A health worker suggested that by having easy access to normal values she could easily detect if something was abnormal. Another possible reason for this may be that the normal values provide examples. People seem to develop abstract models based on prototypical examples [15], such that a tooltip with normal values is a way of expressing the concept which resonates with users having a half-baked understanding. Since tooltips aim at such users, examples like typical values may be the best contents option. Normal values as text being preferred over illustrations or tables, contradicts Mayer's [9] findings that illustration with text is the best way of expressing an explanation. Possible reasons for this discrepancy could be that the explanations resembled those from the informants' medical training, or that the tooltip is a reminder instead of a first presentation of the medical concept.

Table 1: The results from the questionnaire

Type	Average, 1 (best) – 4 (worst)
1 Normal values, text	1.93
2 Explanation	2.17
3 Normal values, illustration and table	2.33
4 Treatment	3.02

An exception to the findings was that nurses in general preferred illustrations over normal values. Our semi-structured interviews provided possible reasons for the questionnaire responses. Two nurses had different preferences to whether a tooltip should be an illustration or text. After some discussion, we found that they had their medical training at different colleges with training material using illustrations in various degree.

The nursing students and the personnel with 0-2 years of health education followed the general sequence of preference as given in the table.

8. Conclusion

This study aimed at finding out whether users prefer text, tables or illustrations in tooltips for understanding of data, and included 58 informants in Ethiopia, Malawi and South Africa. The informants favored text format to illustrations, which contradict Mayer's [9] finding that illustration with text is the best way of expressing an explanation. One reason may be that Mayer studied understanding of systems with causality, while our tooltips concerned explanations of medical terms.

We also considered what type of contents health workers would prefer. The output from a study of documents included definitions and treatment as possible tooltips. Interviews with experts on patient information systems pointed to that health workers wanted to see the range of normal values or a medical measurement.

Normal values were the preferred content of tooltip from the questionnaire amongst the health workers. Explanation and illustration were ranked significantly less favorable, and treatment even lower. Little previous research has been found on this subject, and these results are to be considered as a new contribution to the research on tooltips.

We also found that the nurses did not need additional information in the tooltips, possibly because of their higher level of education. Most of the other informants, however, entered incorrect information, hence they did not read or understand the tooltips provided. Thus, the tooltips had less effect than video tooltips [4]. A possible reason is again the contents; the videos were showing steps in procedures to accomplish tasks, while the tooltips explained medical terms.

Tests demonstrated that tooltips were not able to make up for counterintuitive design, where data fields appeared in an unconventional order. Tooltips may help where users are not fully familiar with data to be entered, but poor design should be mended rather than trying to help the user through.

The findings from this study may be applied to other areas of medical care, as normal values, for instance, for medical terms also exist elsewhere. The plan further is to test the usage of the tooltips over several days to see which of the content types actually lead to more correct data entry. This should involve participants within the same target group, who have not yet been introduced to system, in order for them to have the same starting point.

References

- [1] About mHealth4Afrika [Internet]. Available from: <http://www.mhealth4afrika.eu/page/about>
- [2] Novick DG, Elizalde E, Bean N. Towards a More Accurate View of When and How People Seek Help With Computer Applications. Special Interest Group on Design of Communication 07. El Paso, Texas, USA; ACM; 2007. 95-102. DOI: 10.1145/1297144.1297165.
- [3] Rettig M. Nobody Reads Documentation. Communication of the ACM- Special issue on computer graphics; 1991 July; 19:24. DOI: [10.1145/105783.105788](https://doi.org/10.1145/105783.105788).
- [4] Grossman T, Fitzmaurice G. ToolClips: An Investigation of Contextual Video Assistance for Functionality Understanding. ACM Conference on Human Factors in Computing Systems 10. Atlanta, Georgia, USA ; ACM; 2010. 1515-1524.
- [5] Farkas DK. The Role of Balloon help. ACM SIGDOC Asterisk Journal of Computer Documentation. 1993; 17(3). DOI: 10.1145/154425.154425.
- [6] Carroll JM. The Numberg Funnel: Designing Minimalist Instruction for Practical Computer Skill. Cambridge, Mass., MIT Press. 1990
- [7] Huang J, Twidale MB. Graphstract: Minimal Graphical Help for Computers. ACM Symposium on User Interface Software and Technology 07. Newport, Rhode Island, USA; ACM; 2007. 203-212.
- [8] United nations economic commission for Europe (UNECE). Making data meaningful Part 2: A Guide to Presenting Statistics. United nations economic commission for Europe (UNECE); 2009.
- [9] Mayer RE. Models for Understanding. Review of Educational Research [Internet]. Spring;1989;59(1). DOI: 10.3102/00345543059001043.

- [10] Harrison SM. A Comparison of Still, Animated, or Nonillustrated On-Line Help with Written or Spoken Instructions in a Graphical User Interface. ACM Conference on Human Factors in Computing Systems 95. Denver, Colorado, USA; ACM; 1995.
- [11] Instone, K. Usability Heuristics for the Web [Internet]. 1997 [Cited 7 December 2016] Available from: <http://instone.org/heuristics>
- [12] Grossman T, Fitzmaurice G, Attar R. A Survey of Software Learnability: Metrics, Methodologies and Guidelines. ACM Conference on Human Factors in Computing Systems 09. Boston, MA, USA; ACM; 2009. 649-658.
- [13] DHIS2. Data managements and analytics [Internet] n.d [Cited 7 december 2016]. Available from: <https://www.dhis2.org/data-management>
- [14] DHIS2. Technology Platform [Internet] n.d [Cited 7 december 2016]. Available from: <https://www.dhis2.org/technology>
- [15] Ormrod JE. Human Learning. Merrill, Englewood Cliffs, New Jersey;2012;237-241.

3.2 Content Types – Results

The questionnaire for content types was used during the entire research effort, without alterations. We got a total of 58 responses during this research. As the Ethiopian health workers were underrepresented, they were not included as a group on their own.

As mentioned in the paper “Design of Tooltips for Health Data” (see section 3.1), the participants were asked to rank the options in the questionnaire on a scale from 1 to 4, with 1 being the most preferred. Afterwards, we calculated the average of the four different options, and considered the lowest score to be the most preferred. It should be noted that in SA, however, the users had more difficulties understanding this scale system, thus we switched the order to 4 being the most preferred. In our representation, the results have been converted to the original setup, with 1 being the most preferred.

This questionnaire only shows the participants’ preferences in content types for tooltips, and not which content type will give the best results or improve data quality. This will be further explained in chapter 4.

Table 2: Preference of content type - Malawi vs. South Africa

Content type	Malawi	South Africa
Explanation	2.06	1.92
Treatment	3.53	2.82
Normal Value	1.60	2.62
Illustration (normal value)	2.20	2.06

As seen in the table above, there is a slight difference between the preference of the Malawi group and the SA group. The Malawi group seemed to prefer normal values as content types, while the SA group seemed to prefer explanations. A possible reason for this finding may be that the participants from SA worked at a big hospital, and had access to equipment used for finding normal values of medical terms, hence the need for that (normal values) as a tooltips may have seemed less. The Malawian participants, on the other hand, were stationed at rural clinics with limited resources. For instance, we learned that at one of the clinics in Malawi the prospective mothers enrolled in the ANC program were asked to bring their own candles in case of power breaks during delivery.

4 A Field Experiment

The following paper is to be published during the International Conference on Human Computer Interaction 2017, in Vancouver, Canada in July. It only contains the results from half of the participants of the experiment, due to time constraints. Therefore, we recommend the reader to skip chapter 5 and 6 in the paper , as the results of the paper may not correspond with results given in section 4.2.

4.1 “Design of Tooltips for Data Fields – A Field Experiment of Logging Use of Tooltips and Data Correctness”

Design of Tooltips for Data Fields

A Field Experiment of Logging Use of Tooltips and Data Correctness

Helene Isaksen¹, Mari Iversen^{1*}, Jens Kaasbøll¹ and Chipo Kanjo²

¹University of Oslo, Oslo, Norway

helenis@ifi.uio.no

✉ mariive@ifi.uio.no

jensj@ifi.uio.no

²University of Malawi, Zomba, Malawi

chipo.kanjo@gmail.com

Abstract. Many health professionals in developing countries carry out tasks which require a higher level of education than they have. To help such undereducated health workers filling correct data in patient information systems, data fields were furnished with tooltips for guiding users. In a previous study with questionnaires and interviews, health workers preferred tooltip contents being normal values of the data with medical explanation as the second best. The experiment reported in this paper set out to test these content alternatives and also aimed at finding health workers' use of tooltips and possible effects on data correctness. In order to resemble the work setting, each of the 15 undereducated health workers participating was given a tablet PC with the patient information system and booklet of 22 cases to be entered over a period of two weeks. They were given a one hour introduction to the system. Their use of the tablet was recorded, and after completing, the participants were interviewed. The health workers opened tooltips frequently for the first cases, and thereafter the use dropped. Reasons given were that they learnt the data field during the first cases, and thereafter they did not need the tooltips so often. The number of correct data entries increased over time. The group with medical explanation tooltips performed better than the group with normal value tooltips, thus the preferred tooltip in the questionnaire gave a lower performance than the second alternative. While the experiment demonstrated that tooltips improved performance, it did not quantify the effect.

Keywords: Usability evaluation. Field experiment. Logging use. Learnability. Context-sensitive help. Tooltip contents. Normal data values. Formal definitions. Data quality.

1 Introduction

Health workers in developing countries are often assigned tasks meant for those of higher cadres. As an example, undereducated staff have to do the tasks of nurses [6]. Doing work-related tasks beyond one's competence may lead to wrong data capturing and may cause fatal decision making. Training and follow ups of undereducated are often unsuccessful due to lack of supporting staff and funding. In addition, IT systems are often designed for expert users, thus there is a need for providing information health workers can look up and use themselves.

There are several methods to provide additional information for users. These include users looking up information online, from external sources or by including inline information in the system. Adding inline additional information may be a solution, however, this research aim to test different content types for additional information, and to find the most effective type. Tooltips are the most common ones and have been shown several times to be effective [1, 9, 4]. Due to limitations in the software used for the experiment, textual tooltips are the basis for our research.

Our definition of tooltips is information that can be viewed when the user push a button. The information will disappear from the screen when a button is pushed, or when the user start or finish entering data into the field. The goal for tooltips, in our case, are for the users of the system to understand the medical terms and enter correct information.

Little previous research has addressed the identification of the most effective tooltips in terms of correctness of data entry. Some research has considered user-preference of expression format for tooltips. Petrie et al [9] identified four expression formats for tooltips and asked their participants to rate the different formats based on satisfaction, understandability and preference, however the research did not opt to find the most effective tooltips. One of the end goals for tooltips are for the user to use the system effectively, therefore, a decreasing usage of help commands or tooltips is seen as a sign of system learnability [8]. Dai et al. [1] developed a software consisting of step-by-step instructions for carrying out tasks. However, these instructions would not function with tooltips, as tooltips are unsuitable for displaying sequences of instructions, since they disappear once a single task is finished. Isaksen et al. [6] conducted a survey of preferences of content types of tooltips by lower cadre health workers. The health workers preferred tooltips expressed as normal values of the data to be entered. However, their study did not explore if the tooltips actually led to more correct data entry. Their findings constitute a basis for our study.

The objectives for this research is to compare two content types for tooltips and find out whether there is a difference between them in terms of correctness of data entry. We also wish to see if the tooltips actually affect the correctness. Our research is, therefore, an experiment to find out how often the users use the tooltips, and if they can be seen as successful. By successful tooltip, we mean that they have opened the tooltip, and that they enter the correct data.

2 Tooltip Contents

Through interviews with professionals within Antenatal care (ANC) systems, Isaksen et al [6] identified four content types for tooltips for medical terms. These content types were normal values, the formal definition, treatment, and procedure to find measurements. They found that normal values were the most preferred among health workers of different cadres, with formal definitions as the runner up. Therefore, this study will focus on these two alternatives.

Tooltips containing formal definitions, or explanations, explain medical terms. An example from the study is “Occurs when the woman has hypertension and proteinuria. It can happen at any point after week 20 of pregnancy.”, which is the explanation of pre-eclampsia.

Normal values in the tooltips provide either a range of normal values or signs of the given condition. For example, pre-eclampsia has the following normal value tooltips: “Signs: Diastolic blood pressure above 90 and protein in urine.”.

Below are some examples from the experiment, showing both versions of the tooltip.

Table 1: Examples of the two content types

Data element	Normal value	Explanation
Pre-eclampsia	Signs: Diastolic blood pressure above 90 and protein in urine.	Occurs when the woman has hypertension and proteinuria. It can happen at any point after week 20 of pregnancy
Diastolic blood pressure	Diastolic blood pressure should be between 60 and 80.	Diastolic blood pressure is the minimum blood pressure.
Fundal height	Normal fundal height measurement: 20 weeks = 17-20 cm 28 weeks = 25,5-28,5 cm 36 weeks = 33-35 cm 40 weeks = 36-38 cm	Measurement from the public bone to the top of the uterus. This is done to assess how far into the pregnancy the woman is.

3 Technology description

In order to conduct the experiment, we utilized a generic software package called District Health Information System 2 (DHIS2). The DHIS2 package can either be run through a web browser or through Android apps. For our study the Tracker Capture (TC) android app was used for hosting the testing program. The TC enables the end users to track people or objects over a period of time, and follow up each individual case. The TC can be tailored in the web version for different purposes, and one can create specific programs. For our research the two first authors created two shortened antenatal care programs, and added data elements, skip logics, tooltips and options sets. The data elements were chosen based on Malawian health passports. The programs used exactly the same data elements and order, but the tooltips had different content types.

In Malawian health passports, blood pressure is registered in a single field, labeled either “Blood pressure” or just “BP”, and is not marked diastolic and systolic. Therefore, we wanted to check the participants’ ability to cope with unusual order of data fields, and chose to list diastolic and systolic in the opposite order of how one usually writes them (see Figure 1).

Clinical examination	
✓ * Is LMP date known?	Find Option
✓ Fundal height	Enter number
✓ Diastolic blood pressure	80
✓ Systolic blood pressure	120
✓ Hypertension	Find Option
✓ Eclampsia	Find Option

Figure 1: Example of diastolic and systolic data elements in Tracker Capture

The data elements were assigned to stages, like “Previous pregnancies” and “First antenatal care visit”, and categories, like “Family history” and “Clinical examination”. “Previous pregnancies” stood out by being the only one which contained checkboxes for different data elements. This was done for the program to resemble the health passports, where information is entered for all previous pregnancies in one page, rather than separate pages for each pregnancy.

← Previous Pregnancies

Date of visit
2017-01-03

i / * Gravity
Enter integer

i / * Parity
Enter integer

i / Live born

i / Antepartum stillbirth

i / Intrapartum stillbirth

i / Stillbirth of unknown timing

i / Neonatal death

i / Abortion/termination of pregnancy

i / Spontaneous vaginal delivery (SVD)

i / Assisted vaginal delivery

i / Caesarean section/C-section

i / Pre-eclampsia
Find Option

i / Eclampsia
Find Option

Figure 2: Here is “Live born”, “Antepartum stillbirth”, “Stillbirth of unknown timing” and “Spontaneous vaginal delivery (SVD)” checked, meaning that the woman has experienced these in her previous pregnancies.

In order to register the informant's behavior in the system, an analytic tool called UXcam was utilized. UXcam is a tool used for improving user experiences in applications, through screen recordings, emphasizing the touches on the screen. The recordings are stored on UXcam’s server and are accessible through their web page. The tool was added to the TC code, enabling us to watch and analyze the informants behavior on the screen. The tablets could be traced by the tablet's own ID, as well as the profession of the participant using the tablet. This gave us an impression of their progress throughout the experiment. However, there were risks using this additional software, as we were dependent on the participants being connected to internet when doing their tasks. UXcam is only able to send recordings if connected to the internet,

meaning we were at risk of not getting all of the recordings. Thus we equipped each of the tablets with sim cards and preloaded internet bundles. To ensure that the internet bundle was only used for the experiment, an app called “Applocker” was installed, blocking the usage of all other applications.

For the study, 30 tablets were bought, one for each participant. The two first authors installed the TC on all the tablets, making sure the system was running.

4 Method

In order to get a better understanding of the health worker’s use of the tooltips, we decided to carry out an experiment. We chose to conduct the experiment in natural settings, as this could introduce issues which the participants would not encounter in a lab [2]. It was also important to test over time, in order to see their evolvement. We also wanted to see if they learned anything from the tooltips.

As mentioned, the tablets contained either a program with tooltips containing normal values, or explanations, and these were given to the participants randomly.

4.1 Informants

We chose participants of cadres lower than nurses and higher than community health workers, with ANC experience. A total of 30 people participated in this experiment, however, some of them turned out to be nurses of different degrees. The initial idea was to do 15 participants in South Africa and 15 in Malawi. However, due to misunderstandings and time constraints, the distribution was 20 in Malawi and 10 in South Africa.

This article will include results from the first 15 participants from Malawi, as the experiment extends past the deadline for final version of this paper. The participants in Malawi were recruited by the fourth author, either by appointments or by asking acquaintances and other participants if they knew anyone in the respective cadres.

4.2 Cases

To ensure that the participants used every part of the system and the provided tooltips, the two first authors created a total of 22 cases. Data from these cases was entered into the TC app by the participants over a period of eleven days, two cases a day.

The cases contained information about fictive pregnant women, often quite sick and having lost multiple children. However, it was not written straightforward, but was instead disguised as symptoms, or resembling the information the participants could find in the tooltips. Examples are “.. lost the child in week 38, before the onset of labour”, which indicates an antepartum stillbirth, or “.. has abnormally high blood pressure and protein in the urine”, which indicate pre-eclampsia.

Several of the cases contained similar information, and these were distributed evenly over the period. This was to see if the participants learned the different expressions from one day to another.

Enrollment date: Today's date

First name: Pika

Last name: Chula

Date of birth: 14th march 1985

Marital status: Single

Mobile number: 123 123 245

Previous pregnancies:

Date of visit: today's date

Pika has had two embryos removed, and has given birth to three babies. One of them was delivered through an incision in the abdomen, but, unfortunately, died before the onset of labour.

Pika doesn't remember much of it because she was in a coma. The two other were born in normal manners.

First visit:

Date of visit: today's date

Pika's father react to blinking lights and often get seizures, while her mother has a disorder of metabolism which makes her drink a lot of water and produce large amounts of urine. Pika herself often experience difficulties breathing due to spasms in the bronchi of the lungs, and is in addition allergic to antibiotics in general. She cannot remember her LMP, but her fundal height is 25 cm, her blood pressure is 120/90 and she has protein in her urine. She has been given malaria prophylaxis and iron supplements.

Figure 3: An example of a case from the booklet

4.3 Introducing the Experiment

The experiment started with a brief introduction about who we were, where we came from, and that we wanted to work on improving the usability of a system. We did not inform them about the testing of the tooltips, to make sure we wouldn't affect the results. We then introduced them to the tablets and the TC, explaining what the application did, using a modified question suggestion approach [6]. This included making them aware of the tooltips, informing them that they could use these if they were in doubt regarding what information to enter. We also presented them with the same example case, similar to the next 22 cases they would solve.

The participants in Malawi were situated in groups of three, four or five people, enabling them to cooperate and discuss the matter as they would have in a normal work situation. This also gave us the opportunity to observe what each of them did, and to evaluate their technical skills. The observation enabled us to adapt the information given during the introduction, and to give proper follow-up on each participant. Also, a lot of the explaining of the different elements and tasks was repeated in

Chichewa, the local language, by the fourth author. This seemed to increase their understanding of the experiment, the tasks and other unfamiliar expressions. At the end of the introduction they were given the same questionnaire as Isaksen et al used, capturing the preference of content types for tooltips.

4.4 The Booklets

For this experiment we created a booklet containing information about us, the experiment and 22 cases with tasks for each day. Diaries are used to collect data about user behavior and activities over a longer period of time, and may provide a contextual understanding of the usage of the system [3]. Thus, the booklets were inspired by a diary technique, where the task section would function as a diary. Here, the participants could write down when and where they entered the case, how they felt using the system, what data elements they used and thoughts on the cases. The goal of this was to make them reflect on their case, and to make it easier for them to discuss their thoughts and ideas during the post-interview. The participants were given the booklets after going through the example case

The booklet also contained information about who we were, and what they were supposed to do. Email contact information was also given in the booklet, allowing for the participants to contact us if they had any questions. In addition, they were also given a phone number to the fourth author, who functioned as a local contact, in case of urgent questions.

4.5 The Post-Interviews

After approximately two weeks we asked the participants for a semi-structured interview, aiming to get a better understanding of their use of the tooltips and general thoughts of the entire experience. The questions focused on opinions on the information in the tooltips, and whether they opened the tooltips before or after data entry, and why they did so.

We also collected the booklet and had the participants do the aforementioned questionnaire again to see whether the opinion remained the same or changed. In addition, an online questionnaire were created capturing the participants user experience of the tooltips (hereby UX questionnaire). In this article, we are only using the responses from the 15 participants mentioned above, as well as the responses Isaksen et al. used in their study.

4.6 Analysis

The recordings were structured and analyzed in a google sheet document. The participants were differentiated by having separate sheets, listing all data elements from the program. The first two authors registered whether the participants entered correct information, and if they opened any tooltips. The sheets were set up to calculate successful tooltips, if both data entry was correct and the tooltip was opened.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Data element Midwife nurse	Case 1			Case 2			Case 3			Case 4		
2	O=opened, C=Correct, S=successful tooltip. =1 if true	O	C	S	O	C	S	O	C	S	O	C	S
3	Gravidity	1	1	1	1	1	1		1	0	1	1	1
4	Parity	1	1	1	1	1	1	1	1	1	1	1	1
5	Live born	1	1	1	1	1	1	1	1	1			0
6	Antepartum stillbirth	1		0	1		0	1	0	0		1	0
7	Intrapartum stillbirth	1	1	1	1	1	1	1		0	1		0
8	Stillbirth of unknown timing	1		0	1		0	1	1	1		0	0
9	Neonatal death			0	1	1	1	1		0	1		0
10	Abortion	1		0			0	1		0	1	1	1
11	SVD	1		0	1	1	1	1	1	1			0
12	Assisted vaginal delivery	1		0			0			0			0
13	C-section		1	0			0			0		1	0
14	Pre-eclampsia	1	0	0	1	0	0	1		0	1	0	0
15	Eclampsia	1	1	1		1	0	1		0			0
16	Hypertension		0	0		1	0	1	1	1	1	1	1

Figure 4: Screenshot of the spreadsheet used to register opened tooltips and correct data entry

4.7 Motivation

In order to motivate the participants to take part of the experiment, they were told, at the end of the introduction, that if they did all their tasks, the tablet would be theirs to keep. This is probably part of the reason why everybody entered all cases, and gave feedback to the tasks. In addition, being aware of that their usage of the systems was being monitored, may also have resulted in a higher willingness to finish the tasks given. We did not start with introducing the reward, as we wanted to recruit somebody that were somewhat interested in the project.

5 Results

On average, there were 14 cases recorded per user, in addition we lost all recordings from one user and had one user where we only received eight recordings. This was probably due to connectivity issues, as we, during the post-interviews, found all 22 cases on their tablets.

After analyzing the information we received from the booklets and the interviews, we learned that the participants, on average, spent 20-25 minutes on each case, and it took them about 3 days to get comfortable with the system. However, many of the participants also stated that they wished they had more training with using the application, as for some of them, this was their first time using a touch screen.

Several informants requested more detailed cases, in order to diagnose the patients properly. They also stated that instead of camouflaging the information we should have written it straight forward, indicating that they were not fully aware of the goal of the experiments. This makes the results more trustworthy.

5.1 Tooltips

Below is a graphical presentation of the number of opened tooltips throughout the 22 cases. Normal Value represent the opened tooltips of normal values, while Explanation represent the opened tooltips of explanations. The x-axis shows the cases, while the y-axis represent the total number of opened tooltips for all participants. A trendline was added to better see the development from the first to the last case.

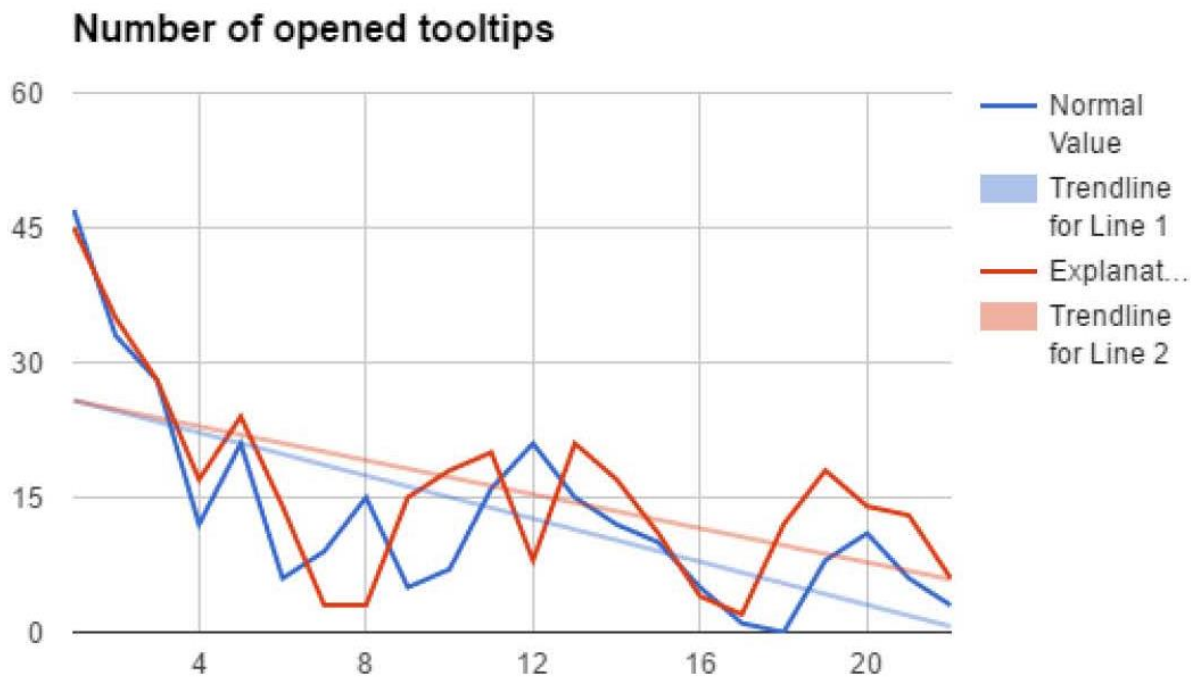


Figure 5: Graph displaying opened tooltips throughout the cases

The graph above shows that both normal values and explanation have a decrease in number of opened tooltips, normal values being slightly lower. This corresponds with what we learned from the post-interview, that the participants used the tooltips a lot in the beginning and less during the last cases. There is no significant difference between the two.

Through the post-interviews, we found that most of the participants confirmed that they used the tooltips less throughout the cases, because they had learned them by heart. This also corresponds with several of our results from the UX questionnaire, where the participants gave a 4.5 out of 5, on both “The need for opening the tooltips were less as the days went by” and “The tooltips helped me learn medical terms by heart”. One of them even quoted the tooltip about eclampsia, proving that she really had learned the term. Another said that she “check with the information I

got earlier”, and further explained that she kept learning the terms when she opened the tooltips, and eventually she knew what to answer, without using them. One participant said she used the tooltips frequently in the first cases, but “Not frequently in the last cases because they helped us understand what it was.”.

Another thing we noticed in the recordings, was that the tooltips were mostly used during the Previous Pregnancy stage, which may be because this is the first stage they enter information into. It may also be because pregnancies have different outcomes, and, therefore, it may be more difficult to differentiate between the different outcomes or delivery methods. Thus, it would require more of a need to consult with the tooltips. When we asked the participants during the interview what they found difficult in the system, the different stillbirths during previous pregnancies was mentioned several times. The difference between antepartum stillbirth, intrapartum stillbirth and stillbirth of unknown timing was confusing. Some also said that several of the terms used in the previous pregnancies stage, are terms that are more familiar to fully educated nurses and midwives, and might be difficult for people with less education to understand. Some also suggested that in order for non-medical personnel to understand what data to enter, signs and symptoms should be listed. This corresponds with the responses we received from the questionnaire regarding content types, that normal values is the most preferred content type.

The graph below show the percentage of successful tooltips from first to last case. The percentage was found by dividing number of successful tooltips with all opened tooltips. Its representation is mostly the same as the graph above, except from the y-axis, which represent the percentage of successful tooltips.

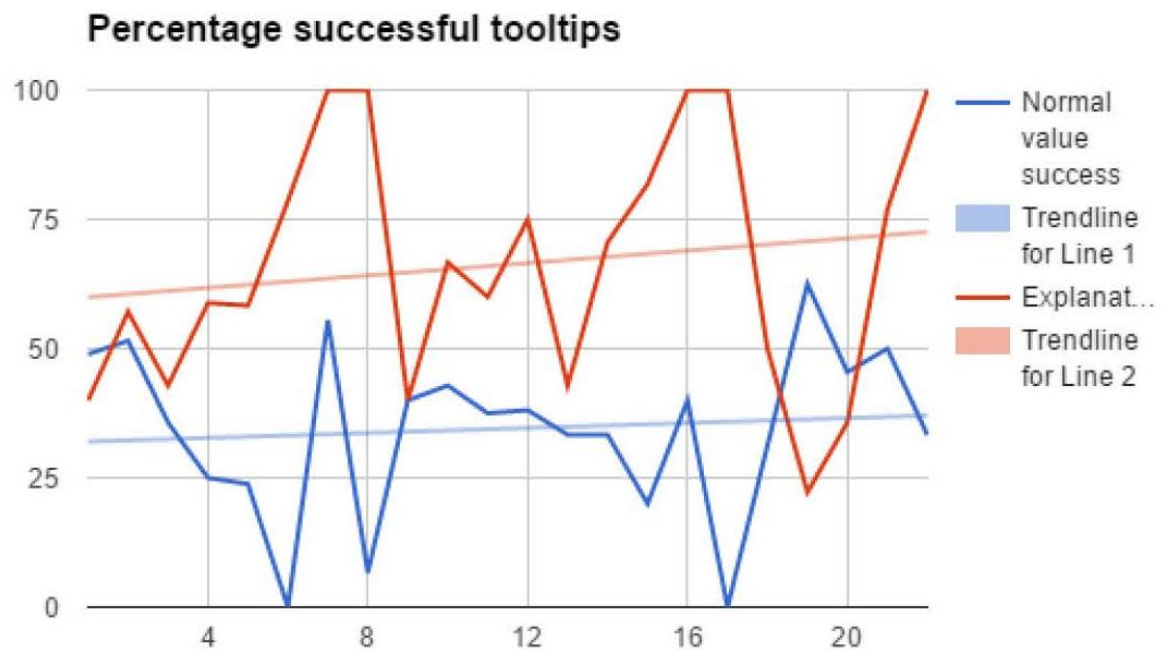


Figure 6: Graph displaying percentage of successful tooltips throughout the cases

The graph above show that the percentage of successful tooltips increase towards the last cases. Also, as seen, the tooltips containing explanations has both a higher percentage of successful tooltips, and a steeper increase through the cases, than normal values.

During the post-interviews we found out that eleven of the 15 participants claimed that they open the tooltips first, and then enter the information. The last four entered data first, and then used the tooltips to check the information they entered and to confirm their answer. We also found out that they had discussed with each other, and other colleagues, during the experiment, when solving the cases.

In addition to the interviews, we also used the booklet to find out what the participants thought. All of them wrote comments and thoughts for most of the cases, and also about the system and some of the tooltips they found useful. "I used the (i) to give me the meaning of the things or terms used" and similar comment are found in several of the booklets. A majority of the participants learned about gravidity and parity, and the different stillbirths. Especially did we notice that if the correct data entry was antepartum stillbirth, intrapartum stillbirth was quite often opened as well. "I learned the difference between antepartum and intrapartum stillbirth" one of the participants said. She often opened both tooltips to understand the difference between them. Also, we learned that ways of delivery contributed to learning. "The allow

guided me on breech delivery" is a quote from one of the booklets, saying that the "allows", meaning the tooltips, taught her about breech delivery, something we also discussed during the interview.

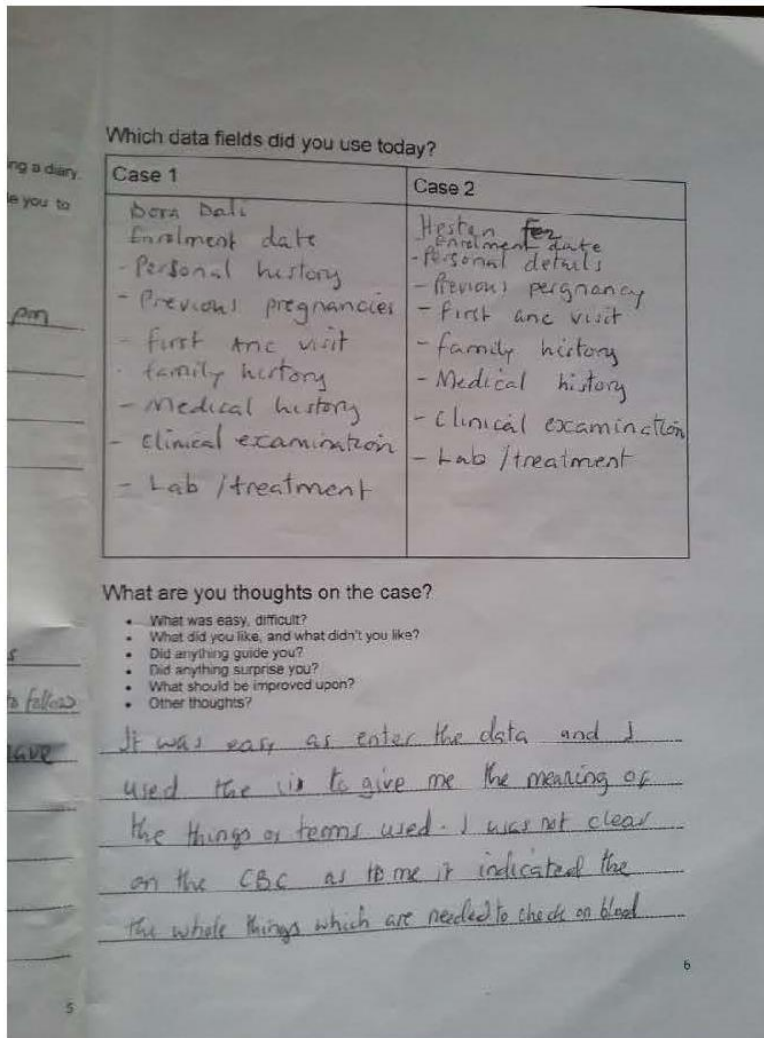


Figure 7: An example from the tasks in the booklet

Also, the tooltips for hypertension, pre-eclampsia and eclampsia were used more in the previous pregnancies stage. This was their first encounter with those tooltips during each case, and many of the participants found the terms confusing. We also found out that participants have different definitions of some terms, like for example pre-eclampsia. Some do not consider only protein in urine as a way of diagnosing pre-eclampsia, as it can indicate other diseases. Another interviewee said that "In our facility we don't have a lot of resources, so high BP means pre-eclampsia.", meaning they diagnose pre-eclampsia only based on high blood pressure. It is important to have formal definitions, however, it is absolutely vital to take into consideration the health facilities without the necessary resources for diagnosing certain conditions.

When analyzing the booklets and the post-interviews, several suggestion of improvement materialized. One participant suggested that we should add more vital

signs to the data elements, another stated “Add more information to the i’s. For example, can you have pre-eclampsia with only hypertension?”. A third participant suggested that we should “for instance giving the normal ranges for BP”. A fourth participant suggested signs and symptoms instead of formal definitions. She justified the statement by saying that non-medical personnel would not know what a condition is, based on the explanations. What is interesting is that all these participants had been using the testing program containing explanations as their content type for tooltips. These findings are also cohesive with the response from the UX questionnaire, where the following statements, “..should have provided more information..” and “..should have provided different information” received scores of 3.2 and 2.9 out of 5, indicating that they partly agree with the statements.

The chart below shows a scatter plot of the number of opened tooltips per user (x-axis) and % correct data (y-axis). Each dot represents a participant.

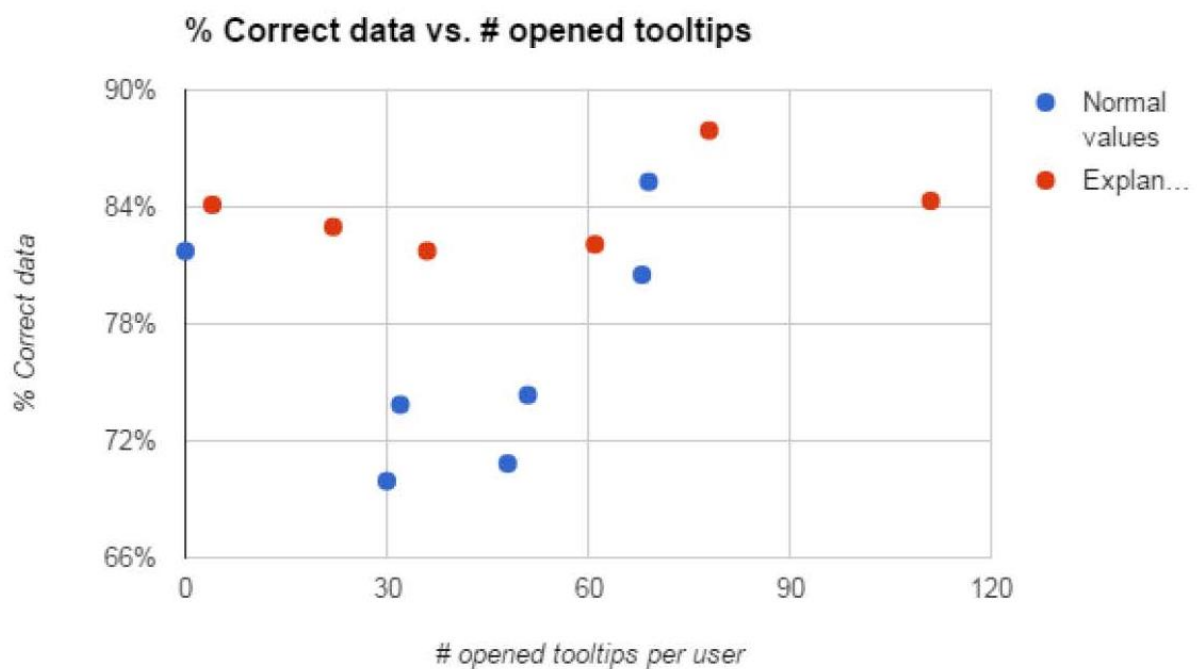


Figure 8: A scatter plot of number of opened tooltips per user and % correct data

There seems to be two users never or seldom opening tooltips who nevertheless enter data of with a high percentage of correctness (upper left). One of these was a nurse, who was sufficiently educated and outside the target group for the tooltips. Two other nurses participated.

The other participants were scattered more linearly. A weak correlation between the number of opened tooltips and correct data entry was found (Pearson, $r=0.26$). For successful tooltips correlated with correct data entry, $r=0.35$, hence a moderate correlation.

5.2 Normal Values versus Explanations

The graph below represents the correctness of data entry in all the cases. The x-axis is the same as in the graph under “Tooltips”, the cases, while y-axis is the correctness, measured in percent per case. Also here, a trendline was added in order to get a better view of the development from the first to the last case.

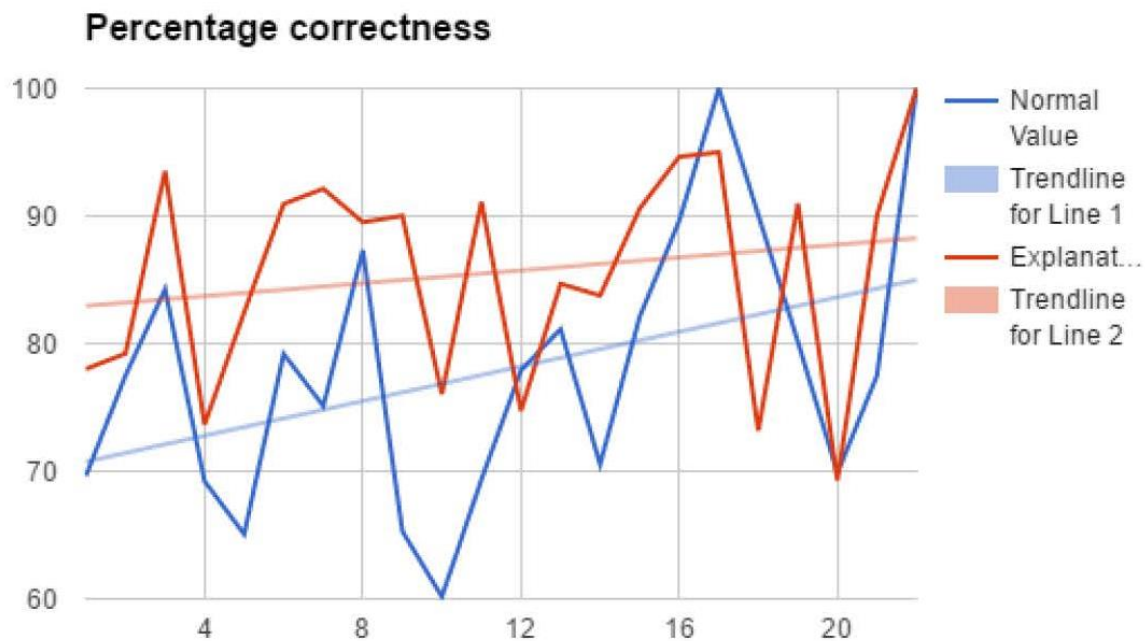


Figure 9: A graph displaying the percentage of correctness

The graph shows that explanations (red line) clearly start out with a higher percentage of correctness compared to normal values. However, if we look at the normal values (blue line), we can see that it increases faster than explanations. This may indicate that if the experiment had lasted over a longer period of time, normal values would have approached 100% correctness before explanations.

The scatter plot above shows that the users who received Normal value tooltips, performed less well than the Explanation group (means 77% vs 85%). Although the number of users is small, their individual scores are averaged over 22 cases. We therefore used the T-test (two-sided, two-sample), and it came out with a significant difference ($p=0.01$) between the two groups.

To check possible statistically significant change of performance over time, the 22 cases were divided into three portions; the first seven, the eight middle and the last seven cases. Then the number of correct data entries in the first seven were averaged per participant and also for the last seven cases. The table below shows the mean values of correct data entry.

Table 2: Results on correct data entry from logging use

	Average % correct first 7	Average % correct last 7
Normal values (n=7)	76	87
Explanations (n=6)	83	85
All participants	79	86

The difference in correctness between normal value tooltips (76%) and explanations (83%) is significant for the first seven cases (T-test, two-sample, equal variance) (yellow). Since the improvement for Normal values is stronger than for Explanations, the study cannot conclude about the long term effect.

The T-test (two sided, paired) shows a significant ($p=0.04$) difference between the first and the last seven for the normal values (grey). Thus, the normal value group had fewer correct data entries in the beginning, but in the end of the 22 cases, they were at an insignificantly higher level than the Explanation group. This may be because normal values started out with less correct answers than explanations, and may therefore have “more room to grow”.

There is also a significant difference between the first and last seven cases for the total group ($p=0.03$). Normally, people improve their performance through repetitions. Our study was not designed with a placebo to differentiate effects of tooltips vs. no tooltips. Therefore, we cannot state that a particular percentage of the improvements followed tooltip use.

However, the interviews indicate that some of these improvements are due to tooltips, which is also cohesive with the UX questionnaire. “The tooltips helped answer correctly to the tasks given” received a total of 4.7 out of 5, meaning that they strongly agree with the statement. Also, there was a low correlation between opening of tooltips and correct responses (Pearson $r=0.26$). The difference in performance between the Normal value and Explanation tooltips groups shows that the tooltips had effects. We therefore conclude that tooltips caused improvement in correct data entry.

Our usage of similar terms both in the cases and in tooltips containing explanations may have influenced the results of the experiment. This may be part of the reasons why the participants using the tooltips containing explanation had a higher correctness and higher percentage of successful tooltips, as they more easily could recognize the phrases used.

6 Conclusion and Further Research

The goal of this research was to find out whether tooltips helped users entering correct data and whether specific contents for tooltips were better than other. The study comprises an experiment with 30 users, where all their use of the software was logged and the participants were interviewed after completion. At the time of final paper submission, only 15 of the participants had completed the experiment, thus only the results for these 15 have been included in the paper. The results may therefore change after all participants have completed, and the final results will be presented during the conference.

Isaksen et al. [6] identified normal data values as the most preferred content type for tooltips for data fields. Formal explanations was the second most preferred type. Previous studies of tooltips [9, 1] have also come up with preferences and have not tested effects of long term use.

This study therefore compared the two types of tooltips during a two weeks experiment.

The user group which were given explanations in their tooltips had a higher percentage of successful tooltips, meaning instances of opening a tooltip and entering a correct value, possibly in the opposite sequence. The explanations group also had a steeper increase than normal values. We also found that, in terms of the correctness in data, explanations have a higher percentage. However correctness for normal values increase faster, and after two weeks, the normal value group was slightly ahead of the explanations on correctness. When comparing the first seven cases with the last seven, we found that tooltips containing normal values has a significant increase in correctness. The difference in correctness between explanations and normal values for the last seven cases is insignificant, as is the increase in the explanations group.

Thus, we see no correlation between user preference and the usefulness of the different content types. In addition, the UX questionnaire revealed that the participants found the tooltips both helpful and understandable.

Both normal values and explanation has a decrease in number of opened tooltips from the first to the last case. The difference between them is not significant. This is also consistent with what we learned through our post-interviews, as participants told us that they did not need the tooltips at the end of the experiment, as the information was learned by heart. This is consistent in the increase in the percentage of successful tooltips from first to last case.

An unexpected finding was that users also opened tooltips after they had entered the data. During post-interviews, they said that this was in order to check that they had entered data correctly. This way of learning from tooltips has not been mentioned in previous user studies of tooltips [9, 1].

We also learned that they used tooltips more during the previous pregnancy stage, which was probably due to it being the first encounter with the terms, difficulties in differentiating the pregnancy outcomes, or because the terms are more used by nurses and midwives.

In order to increase the validity of the experiment, we could have included a control group of participants. Here, the aim would have been to compare the effects of a system with tooltips and a system without tooltips. This is similar to research on

medication, where one group is given real medicine, while the other is given placebo medication. However, the comparison between the two groups would not have been symmetric, as one group would have been introduced to tooltips and the other group not. An alternative way could be create a testing program with some meaningless tooltips. This would have made the groups more symmetric, giving one group actual tooltips and the other group “placebo-tooltips”.

7 References

1. Dai, Y., Karalis, G., Kawas, S. & Olsen, C.: Tipper: Contextual Tooltips that Provide Seniors with Clear, Reliable Help for Web Tasks. CHI'15 Extended Abstracts, 1773-1778 (2015).
2. Duh, H.B.L., Tan, G.B.C., Chen, V.H.H.: Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services. Helsinki, Finland, 181-186 (2006).
3. Flaherty, K.: Diary Studies: Understanding Long-Term User Behaviour and Experiences. <https://www.nngroup.com/articles/diary-studies/>.
4. Grossman T, Fitzmaurice G.: ToolClips: An Investigation of Contextual Video Assistance for Functionality Understanding. ACM Conference on Human Factors in Computing Systems 10. Atlanta, Georgia, USA ACM. 1515-1524 (2010).
5. Isaksen, H., Iversen, M., Kaasbøll, J., Kanjo, C.: Design of Tooltips for Health Data. Submitted for publication.
6. Mayer, R. E.: Models for Understanding. Review of Educational Research. 59, 43-64 (1989).
7. Petrie, H., Fisher, W., Weimann, K., Weber, G.: Augmenting Icons for Deaf Computer Users. CHI '04 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria. 1131-1134 (2004).

4.2 Final Results from Field Experiment

This following section contains updated results, including all 30 participants, while the paper, “Design of Tooltips for Data Fields - A Field Experiment of Logging Use of Tooltips and Data Correctness”, only encompassed results from 15 of the 30 participants. Note that some sentences and paragraphs have been copied from the paper.

One of our intentions when setting up this experiment was to not include participants who had already been introduced to the system, so that all participants would have the same starting point. However, because we did not recruit enough participants in SA, we ended up including two people who had been introduced to the system a few months prior to the experiment. Still, we do not believe that this had any major impacts on the results, as they seemed to have forgotten the system by the time we conducted our study. We also tried to only include people with less education than nurses. However, due to miscommunications, some of the participants from Malawi were educated midwives, a specialization within nursing. Meaning, some of our participants had more education than intended for the research, which most likely has affected the results.

On average, there were 14 cases recorded per user. The reason we did not get all recordings from all participants may be due to poor connectivity, or participants may have accidentally turned off the internet on the tablets. However, we found all cases on the tablets at the end of the experiment. All participants in the Malawi group did all 22 cases and filled in the booklets, while 40% of the SA group did not complete all cases, and half of them wrote little to nothing in their booklets.

After analyzing the booklets, we found that it took about two days (2.25) before they felt comfortable using the system. Additionally, they spent, on average, 21 minutes per case. Some also stated during the interviews that they spent less time on the final cases, which may be a sign of increased learnability, according to Michelsen et al. (1980). This was something we also noticed in the recordings, considering they became shorter in the later cases, compared to the first ones.

There was, however, a noticeable difference between the average time of those from Malawi and those from SA. The Malawi group spent 19 minutes per case, while the SA group spent 29 minutes per case. This was unexpected, as SA is considered a more developed country where

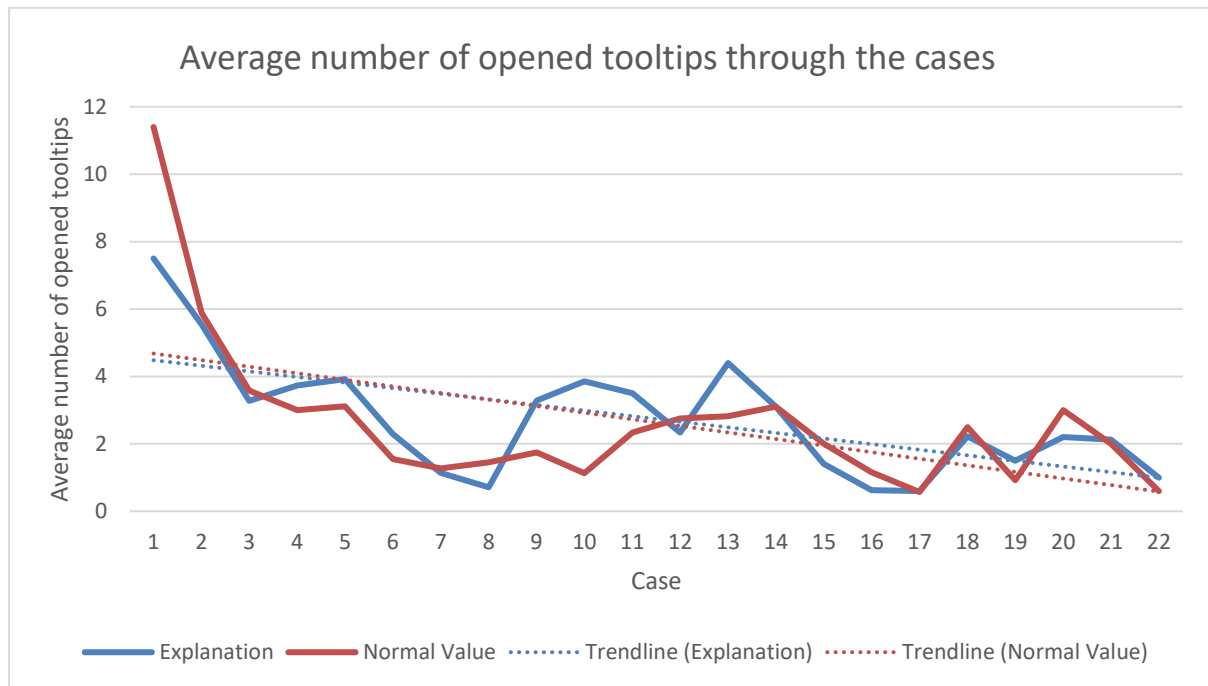
people are more used to technology, which we thought would be to their advantage. Though, the participants from Malawi were of different cadres, some more educated than others, which may have given them an advantage in the understanding of the cases. All participants from SA were, as mentioned, assistant nurses, and having education than some of the participants in the Malawi group. One participant from SA stated during the interview that “sometimes I don’t understand the story”, which may have been a part of why there is such a time difference between the two groups.

During the interviews we asked the participants whether they checked the tooltips before or after they started entering information. We found no patterns on when the participants did one or the other. Sometimes they checked before, and other times they checked after. Sometimes they checked both before and after, while other times they did not check them at all. The fact that some checked tooltips regardless of them knowing the answer or not, contradicts Rourke & Kanuka’s (2009) statement about people hanging on to their possible misconceptions until challenged.

4.2.1 Normal Values versus Explanations

Below we will compare the results of the normal value group and the explanation group to find out which lead to more correct data entry.

Opened tooltips



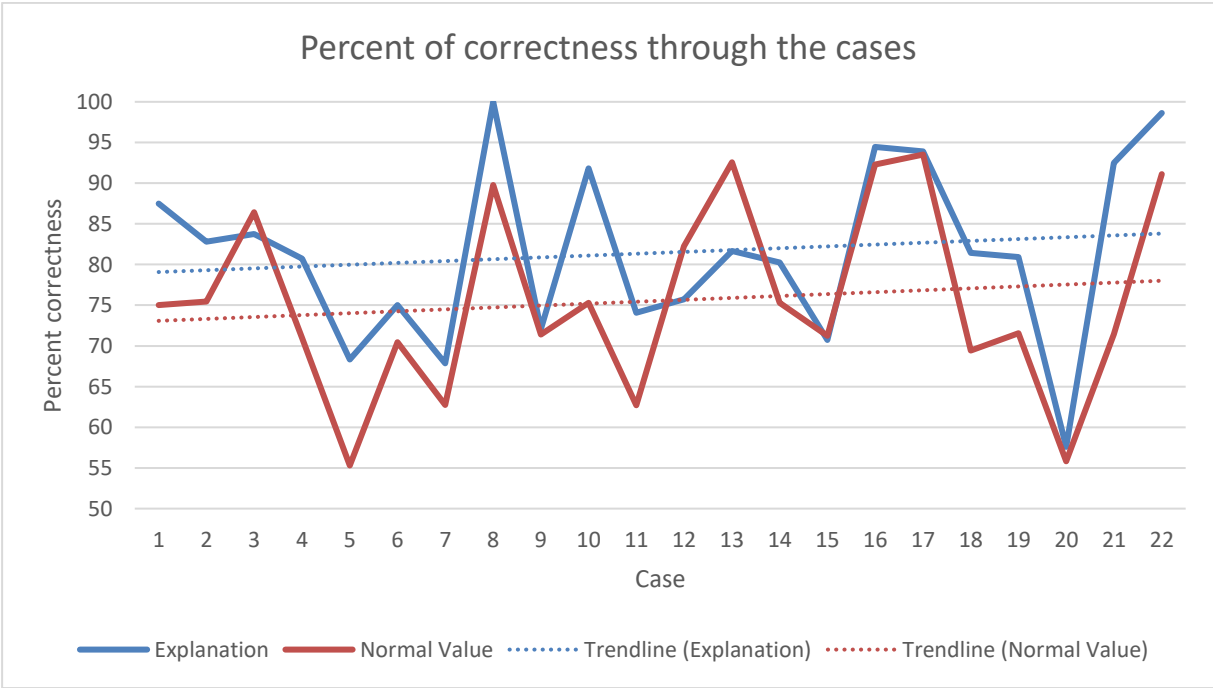
Graph 1: Average number of opened tooltips through the cases

Graph 1 illustrates that, as the days went by in the experiment, the need for tooltips declined. This also corresponds with what we learned from the post-interviews, that the participants used the tooltips more in the beginning than towards the end. This may also be seen as sign of learnability of medical terms, because of the decrease in usage of tooltips (Michelsen et al., 1980). However, as with other repetitive tasks, the willingness to fulfill it may go down as time passes, hence the number of opened tooltips would also decrease. Comparing it to correctness of data entry will therefore be beneficial.

Through the post-interviews, we found that most of the participants confirmed that they used the tooltips less throughout the cases, because they had learned them by heart. One of them even quoted the tooltip about eclampsia, proving that she really had learned the term. According to Michelsen et al. (1980), this may be a sign of learnability, due to the learnability-related content of the comment. Another said that she “check with the information I got earlier”, and further explained that she kept learning the terms when she opened the tooltips. Eventually she knew what to answer, without using them. One participant said she used the tooltips frequently in the first cases, but “not frequently in the last cases because they helped us understand what it was.”. This indicates that the users did learn something from the tooltips, as the need for opening them were not as high towards the end of the experiment as at the start.

Another thing we noticed in the recordings, was that the tooltips were mostly used during the “Previous Pregnancy”-stage, which may be because this is the first stage they enter information into. Also, pregnancies may have different outcomes, like for example antepartum stillbirth or intrapartum stillbirth. These may be hard to differentiate, as they sound quite similar, especially for someone who are not familiar with the terms. Thus, it would require more of a need to consult with the tooltips. When we asked the participants during the interview what they found difficult in the system, the different stillbirths during previous pregnancies were mentioned several times. They found the difference between antepartum stillbirth, intrapartum stillbirth and stillbirth of unknown timing was confusing. Some also said that several of the terms used in the previous pregnancies stage, were terms that were more familiar to fully educated nurses and midwives, which might have been difficult for people with less education to understand. Some also suggested that in order for non-medical personnel to understand what data to enter, signs and symptoms should be listed. This corresponds with the responses we received from the questionnaire regarding content types, that normal value is the most preferred content type.

Correctness



Graph 2: Percent of correctness through the cases

As shown in the graph above, there is an increase in correctness from the first cases to the last. According to Michelsen et al. (1980), this may be a sign of learnability, as there is a decrease in error rates. Both the explanation group and the normal value group seem to have approximately the same increase. However, the explanation group have a slightly higher percent of correctness, about 6%. This may contradict the assumption that the willingness to fulfilling the task decreased as time went by, and the decrease in number of opened tooltips is rather due to users learning the information in the tooltips.

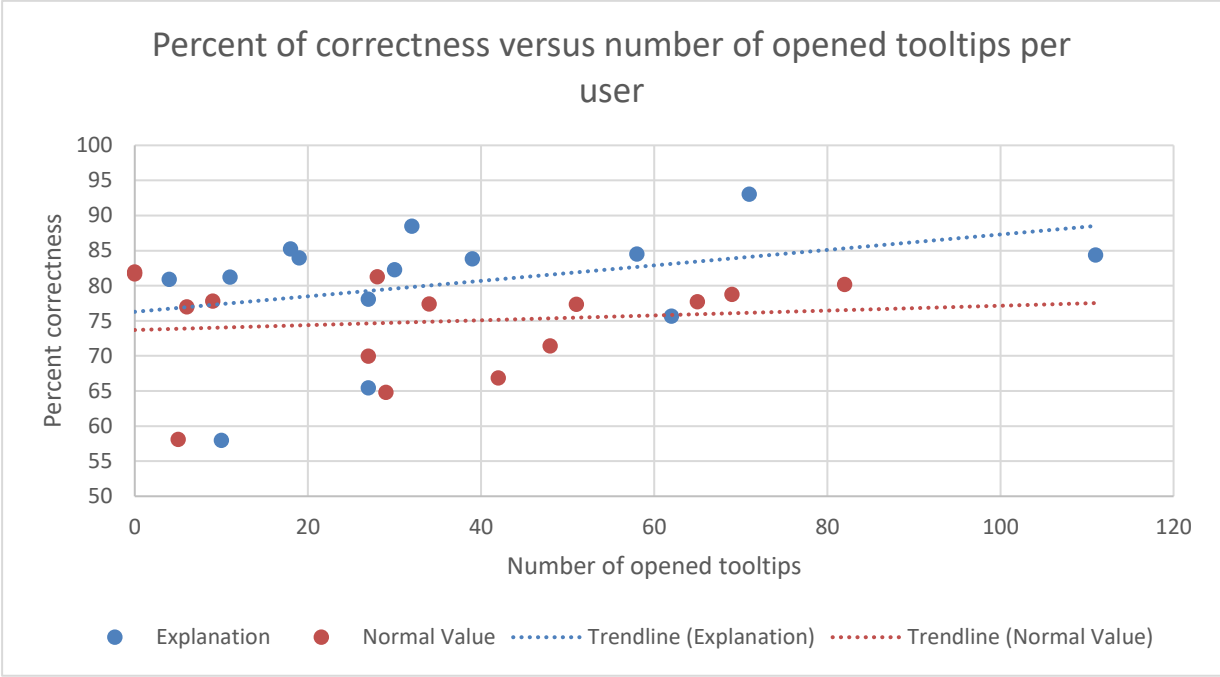
However, we need to take into consideration that the explanation tooltips contained sentences which were also present in the cases. This may have affected the results, as the participants in this group could have compared sentences and expressions from the case with the tooltips (see example in Table 3 below). Participants with normal value tooltips were not able to do such a comparison, as the normal values mostly did not appear in the cases. This may be a possible reason why the explanation group have a higher correctness, as they easier could recognize the phrases used.

Table 3: Data element example with different tooltips

Data element	Example from case	Explanation tooltip	Normal value tooltip
Fundal height	Her measurement from the pubic bone to the top of the uterus is 20 cm	Measurement from the pubic bone to the top of the uterus. This is done to assess how far into the pregnancy the woman is	Normal fundal height measurement: 20 weeks = 17-20 cm 28 weeks = 25,5-28,5 cm 36 weeks = 33-35 cm 40 weeks = 36-38 cm

Upon interviewing the participants and asking them whether they learned something from the tooltips, most stated that they did learn something and that they found them useful, which corresponds with their notes in the booklets. Most of the participants stated that the tooltips helped them enter correct information and guided them in the effort of doing so. The increase in correctness, combined with the responses from the interviews, is a good indication of tooltips actually providing necessary help for the health workers to enter correct data.

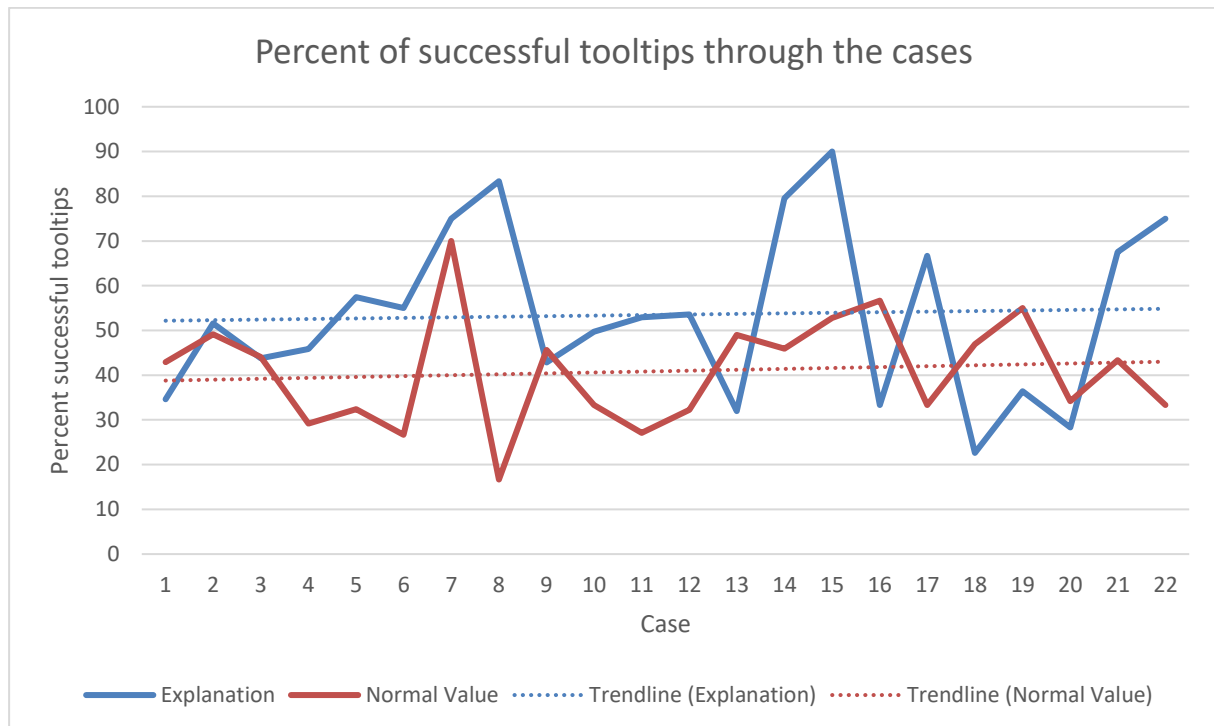
Correctness versus opened tooltips



Graph 3: Percent of correctness versus number of opened tooltips per user

In the graph above, we used Pearson’s correlation to identify possible correlations between opened tooltips, correctness and successful tooltips. We found a weak correlation between the number of opened tooltips and the correctness of data, at $r=0.27$. We also found a moderate correlation between successful tooltips and correctness, at $r=0.50$. These correlations indicate that tooltips have an effect, which also corresponds with both interviews and booklets.

Successful tooltips



Graph 4: Percent of successful tooltips through the cases

The percentages in the graph above were found by dividing the number of successful tooltips with the number of opened tooltips. Both the explanation group and the normal value group seem to have had a slight increase in successful tooltips. Though, it should be noted that the explanation group seem to have about 13% more successful tooltips, compared to the normal value group.

In addition to the interviews, we also used the booklets to investigate the participants' opinions on the cases, the tooltips and the system. All of them wrote comments and thoughts for most of the cases, and also about the system and some of the tooltips they found useful. "I used the (i) to give me the meaning of the things or terms used" (the (i) indicating the button for opening the tooltip) and similar comments were found in several of the booklets. A majority of the participants learned about gravidity and parity, and the different forms of stillbirths. We especially noticed that, if the correct data entry was antepartum stillbirth, intrapartum stillbirth was often opened as well. "I learned the difference between antepartum and intrapartum stillbirth" one of the participants stated. She further stated that she often opened both tooltips to understand the difference between them.

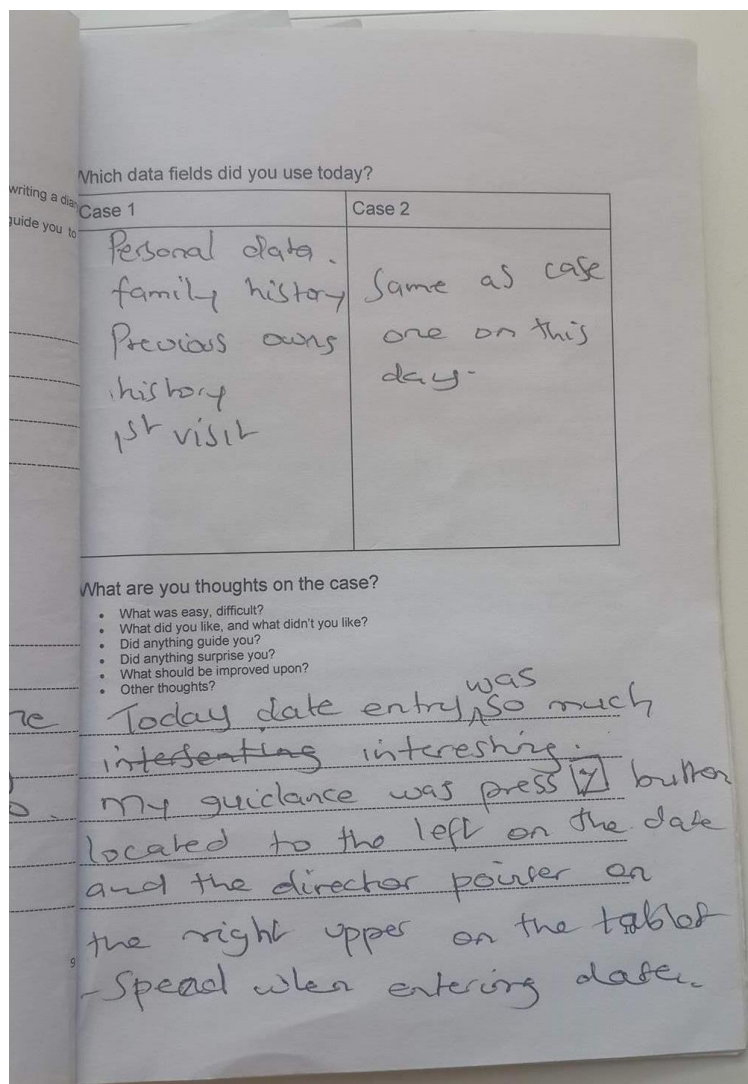


Figure 1: Example from booklet

Also, the tooltips for hypertension, pre-eclampsia and eclampsia were used more in the “Previous Pregnancy”-stage. This was their first encounter with those tooltips during each case, and many of the participants found the terms confusing. We also found that participants have different definitions of some terms, like for example pre-eclampsia. Some do not consider only protein in urine as a way of diagnosing pre-eclampsia, as it can also indicate other diseases. Another interviewee stated that “in our facility we don’t have a lot of resources, so high BP means pre-eclampsia”, meaning that they diagnose pre-eclampsia only based on high blood pressure. Even though it is important to have formal definitions, it is absolutely vital to take into consideration the health facilities that do not have the necessary resources for diagnosing certain conditions. When creating tooltips, one should consider both of these aspects, and additionally ensure that the tooltip can be effectively used by all clinics, independently of resources.

First and last seven cases

Table 4: Correctness for the first and last seven cases

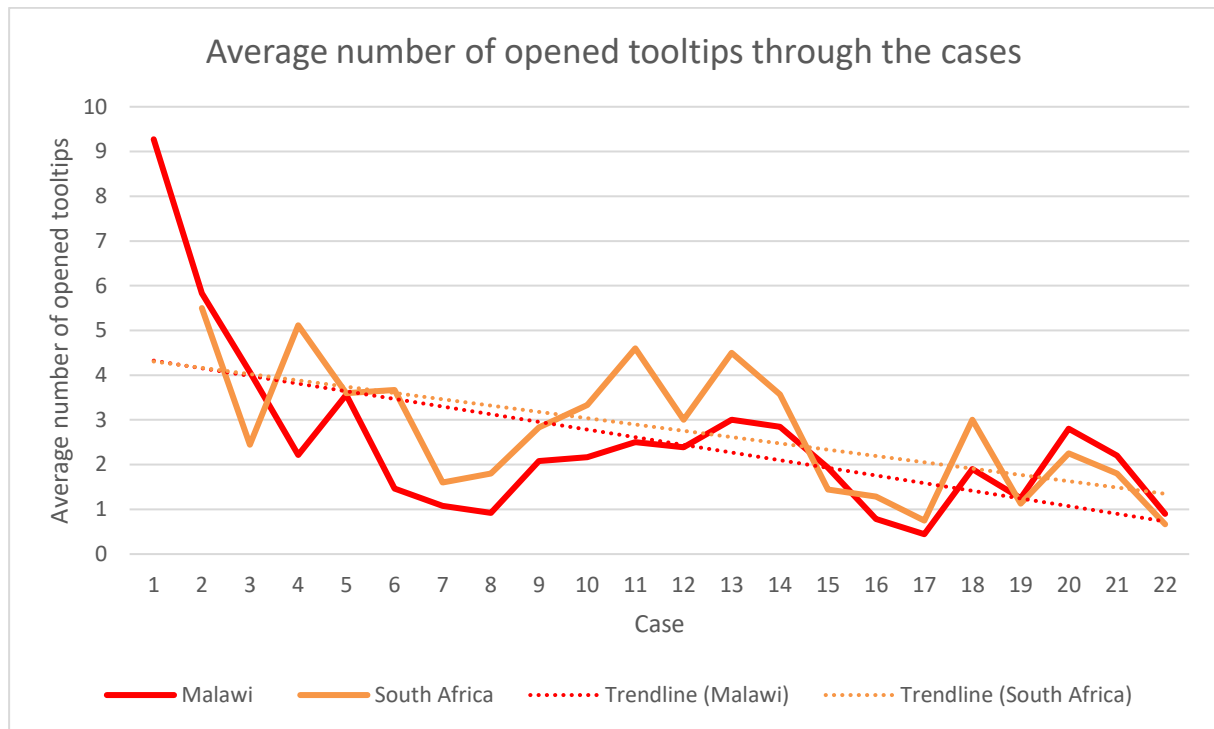
	First seven cases	Last seven cases	Increase
Normal value	70.9	77.9	+7.0
Explanation	78.0	85.6	+7.6
Total	74.5	81.8	+7.3

The table above shows that both groups had almost the same increase in correctness from the first seven cases to the last seven, explanations being slightly higher. Based on paired, two-tailed t-tests, there is a significant difference between normal values and explanations in the first seven cases ($p=0.01$) (lighter grey area in Table 4) and in the last seven cases ($p=0.03$) (darker grey area in Table 4). This corresponds with previous results that indicated that explanations have a higher correctness than normal values. Though, there was no significant difference between the first seven cases and the last seven cases, neither for normal values ($p=0.27$), nor explanations ($p=0.19$). Even though there is no significant difference from the first seven to the last seven cases, there still exists an increase in correctness.

4.2.2 Malawi versus South Africa

We wanted to compare Malawi and SA because of variations at national levels, such as the fact that SA is a more developed country than Malawi.

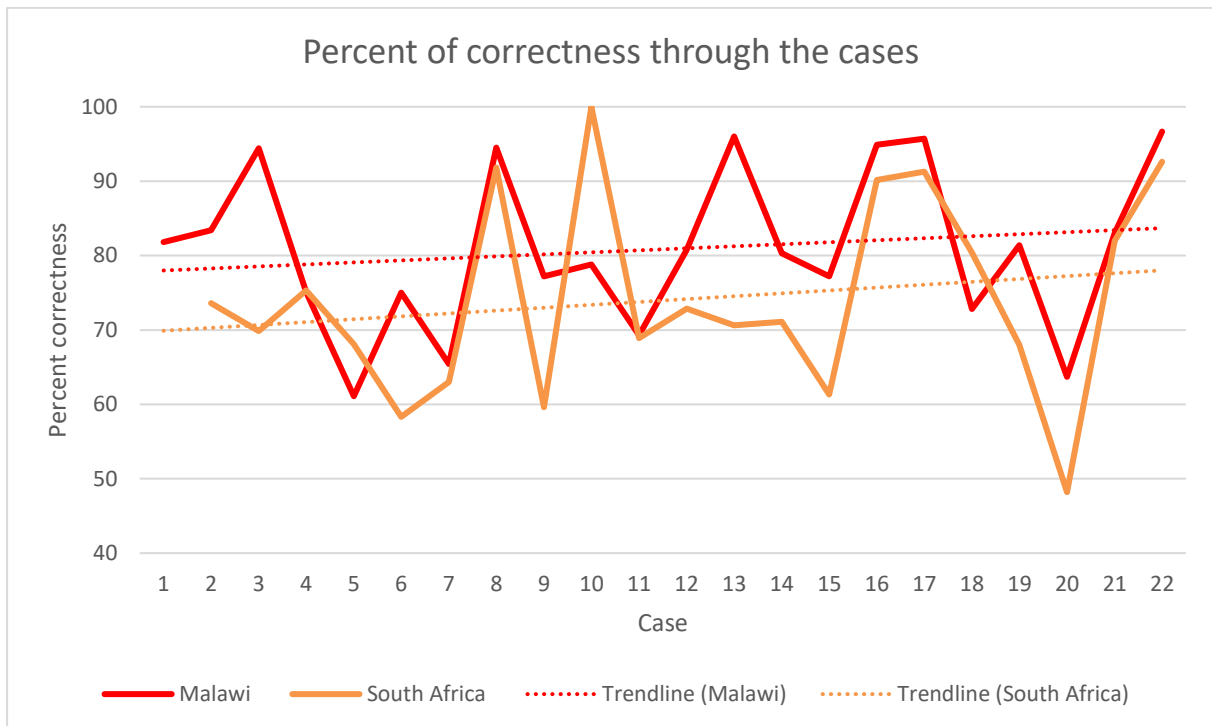
Opened tooltips



Graph 5: Average number of opened tooltips through the cases

As seen above, the number of opened tooltips decreases towards the later cases. This decrease was something we expected, due to that similar phrases and tasks were repeated in the cases towards the end, hence the participants should have already opened the tooltips earlier. This corresponds with what the participants in the Malawi group told us during the interview as well, that they used the tooltips less in the last cases. However, the majority of the participants from the SA group said during the interviews that they used the tooltips just as much at the end as in the beginning. This is not consistent with what the graph shows, as it shows a decrease in opened tooltips, and not a steady line. This may be due to the fact that what people say and what people do, is not always consistent.

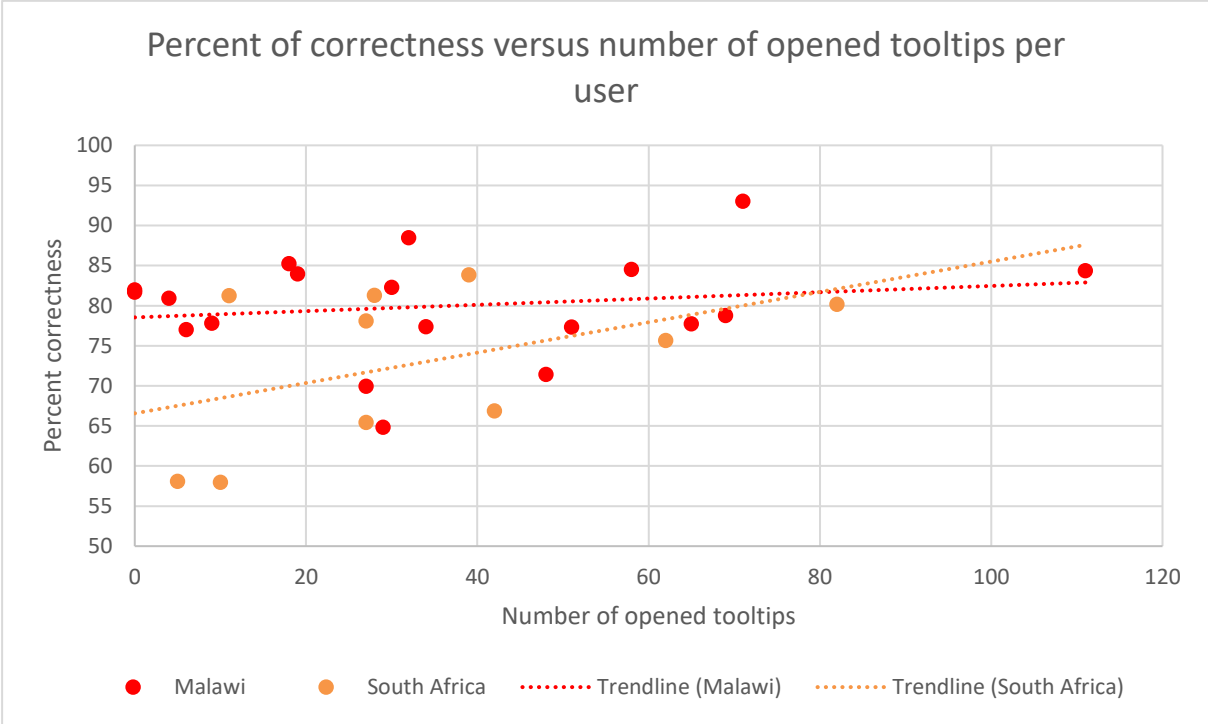
Correctness



Graph 6: Percent of correctness through the cases

Graph 6 shows no major differences, even though, the Malawians seem to have a slightly higher percentage of correctness. This may be because of the different and somewhat higher levels of knowledge between the participants from the Malawi group as compared to the SA group, where all participants were assistant nurses. Also, the involvement from the SA group may have affected their desire to answer correctly, as they were not as engaged as the Malawi group. In the post-interviews with the participants from SA, we experienced that we struggled to get information from them, as they were not as willing to talk and elaborate in the conversation as the Malawian participants.

Correctness versus opened tooltips

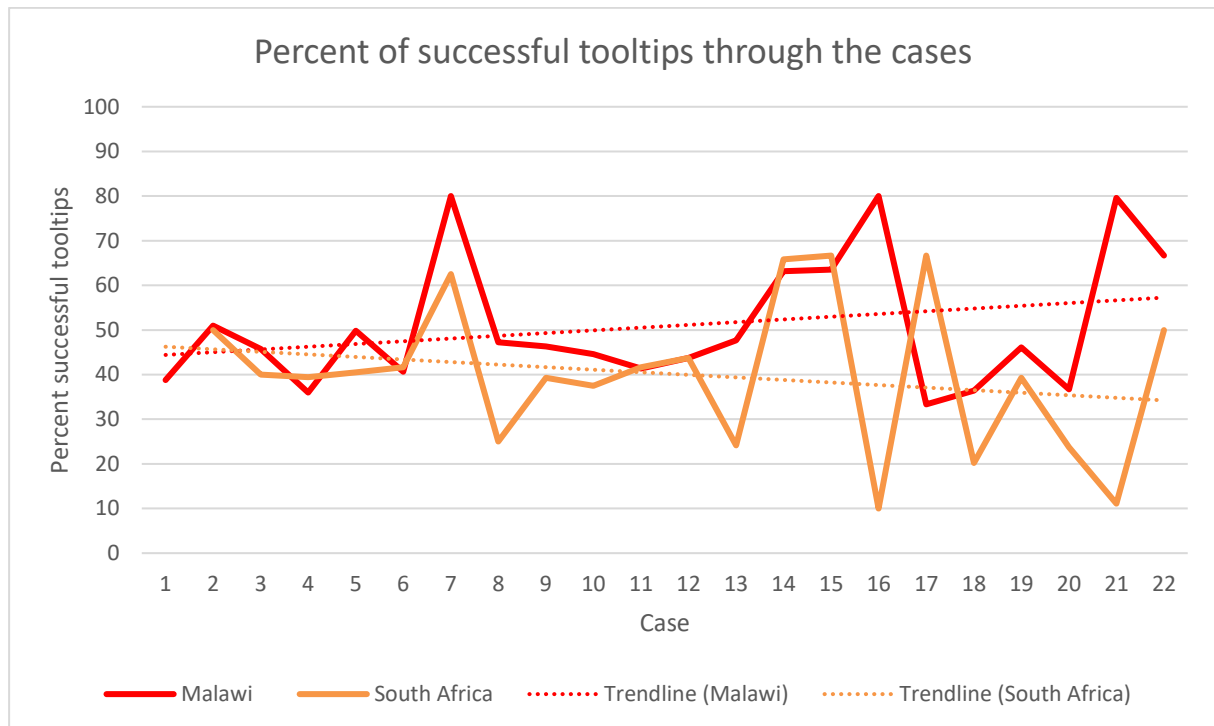


Graph 7: Percent correctness versus number of opened tooltips per user

As the scatter plot in Graph 7 show, the two participants that stood out by not performing well in terms of correctness were from SA. They had approximately 58 % correct answers and thus pulling down the overall correctness for the participants from SA. Though, it is worth noting that these two participants hardly opened any tooltips, which may be part of the reason for the low correctness.

Two of the participants from Malawi never opened the tooltips. Both of these participants were nurses, hence, they had appropriate training for ANC. Therefore, the need for opening tooltips were not present. What is interesting, is that even though several of the Malawian participants opened fewer tooltips than the SA participants, they still maintained a higher correctness. Though, this may be due to the difference in education within the Malawi group.

Successful tooltips



Graph 8: Percent of successful tooltips through the cases

As Graph 8 shows, there is quite a difference between the participants from Malawi, and the participants from SA. The Malawi group show an increase of successful tooltips, while the SA group have a clear decrease. The reason for this may be the same as mentioned earlier, less educated, less engaged to the experiment and more resistant to elaborate in discussion in post-interviews.

First and last seven cases

Table 5: Correctness for the first and last seven cases

	First seven cases	Last seven cases	Difference
Malawi	76.6	84.0	+7.4
South Africa	68.0	79.8	+10.9
Total	74.5	81.8	+7.3

As seen in the table, the SA group had a higher increase compared to the Malawi group from the first seven to the last seven cases. However, the results from the Malawi group are in general

better than those from SA, as they show a 5% higher correctness than the SA group. Still, there were no significant differences between any of the results.

4.3 Did the Experiment Alter Their Preferences?

When conducting the experiment, participants filled in the questionnaire for content types twice, once during the first introduction and once after the experiment, during the post-interviews. This was to explore whether the experiment altered their preferences or not.

4.3.1 Malawi

In the second study, the participants from Malawi preferred normal values both before and after the experiment. However, there was a significant change in their ranking of the other content types. Prior to the experiment explanations were ranked higher than illustrations, while after the experiment illustrations were ranked higher than explanations.

Table 6: Average score for preference pre and post experiment

Content type	Average score pre experiment	Average score post experiment
Explanation	2.03	2.25
Treatment	3.41	3.21
Normal value	1.72	1.83
Illustration (normal value)	2.28	2.06

Explanation and treatment showed a significant change from the questionnaire prior to the experiment, to the questionnaire post experiment. This was based on the Wilcoxon’s signed rank test, because of the low number of values (see Table 7 below).

Table 7: Wilcoxon significant differences of Table 6

Content type	Wilcoxon's signed rank test result	Conclusion
Explanation	0.01	Significant difference
Treatment	0.01	Significant difference
Normal value	0.21	Insignificant difference
Illustration (normal value)	0.05	Insignificant difference

4.3.2 South Africa

In contrast to the Malawians, the participants from SA preferred explanations over normal values, both before and after the experiment (see Table 8). Their preferences did not change significantly, based on Wilcoxon's signed rank test. This may be related to the fact that all participants from SA worked at the same hospital and had the same profession (assistant nurses), which may explain why their answers were consistent both before and after the experiment. In addition, compared to the Malawians, not all SA-participants completed the experiment, meaning that they did not all entered all cases, and only a few used the booklets.

Table 8: Average score for preference pre and post experiment

Content type	Average score pre experiment	Average score post experiment
Explanation	1.92	1.89
Treatment	2.82	2.99
Normal value	2.62	2.36
Illustration (normal value)	2.06	2.19

4.3.3 Changes at User Level

In order to look at changes at user level, we present the table below. It illustrates three possible indications that may be the case when the aforementioned questionnaire is presented prior to and after being exposed to technology containing tooltips. The letters A and B represent two different kinds of tooltips.

Table 9: Indicating changes of preference at user level

	Tooltips appearing in app	Tooltips preference before introduction to the app	Tooltips preference after introduction to the app	Indication
1	A	B	A	The user may feel like the tooltips in the app has been useful, hence the change in preference
2	A	B	B	This may be an indication that the users did NOT find the tooltips useful.
3	A	A	A	The tooltips in the app did not alter anything.

The table above only presents speculations, and the best way to understand a user’s possible change in preference would be to ask. These results are based on 20 out of the 30 participants from the experiment, all from Malawi. The participants from SA were not included because their pre and post-results were not possible to track to the individual participant.

After finishing the analysis of the questionnaires, we found that the first scenario (users may have found the tooltips useful and then changed preference) was the least occurring with only four occurrences. It is therefore deemed as unlikely. The second scenario (the users did not find the tooltips useful) was the second least occurring with only five occurrences, and is also deemed unlikely. The third scenario (the tooltips did not alter any preferences) was the most

occurring with a total of eleven occurrences. Hence, the third scenario applies best for our research.

4.4 User Experience Questionnaire

As briefly mentioned in “Design of Tooltips for Data Fields - A Field Experiment of Logging Use of Tooltips and Data Correctness”, we created an online UX questionnaire, filled in by all 30 participants in the experiment. Since we did not audio record the interviews, we were not sure to capture all info only through taking notes. The UX questionnaire was therefore more appropriate in capturing the participants’ evaluations of the tooltips, and to open up for comparisons with other techniques, which may lead to higher validity.

The questionnaire consisted of ten statements, which the participants were asked to rank from 1 to 5. 1 meant either “Strongly disagree”, “Not helpful at all”, “Not easy at all” or “Little correct information” , while 5 meant either “Strongly agree”, “Very helpful”, “Very easy” or “A lot of correct information” (see Figure 2).

	QUESTIONS	RESPONSES
	I already knew most of the medical terms used in the app	<p>1 2 3 4 5</p> <p>Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree</p>
	The provided tooltips helped me answer correctly on the tasks given.	<p>1 2 3 4 5</p> <p>Not helpfull at all <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Very helpfull</p>
	The information given in the tooltips were easy to understand	<p>1 2 3 4 5</p> <p>Not easy at all <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Very easy</p>
	The need for opening the tooltips were less as the days went by	<p>1 2 3 4 5</p> <p>Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree</p>
	The information provided in the tooltips were correct	<p>1 2 3 4 5</p> <p>Little correct information <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> A lot of correct information</p>
	The tooltips helped me learn medical terms by heart	

Figure 2: Screenshot of UX questionnaire

As a test in the UX questionnaire, we added paired statements claiming the opposite of one another, to see how the participants responded, and if their responses were coherent. The first pair was statement number 3 and 7, and the second pair consisted of statement number 8 and 10 (see Table 9).

The table in section 4.4.1, gives an overview of the statements used in the UX questionnaire, alongside with their average scores. As mentioned, participants were divided into two groups, one which got tooltips with explanations, and one which got tooltips with ranges of normal values. Table 9 show the average score for both explanations and normal values, and the total average score, from all 30 participants.

4.4.1 Results

The following sections will focus on comparing the UX questionnaire to the post-interviews, booklets and recordings. We will see how they fit to the two UX goals we are focusing on, helpful and rewarding. In the table below (Table 9), the scores are averages.

Table 10: Results from UX questionnaire

No.	Question/statement	Options	Total score	Normal value score	Explanation score
1	I already knew most of the medical terms used in the app.	1=Strongly disagree 5=Strongly agree	3.80	3.80	3.79
2	The provided tooltips helped me answer correctly on the tasks given.	1=Not helpful at all 5=Very helpful	4.63	4.87	4.43
3	The information given in the tooltips were easy to understand	1=Not easy at all 5=Very easy	4.47	4.40	4.57
4	The need for opening the tooltips were less as the days went by	1=Strongly disagree 5=Strongly agree	4.17	3.93	4.43
5	The information provided in the tooltips was correct	1=Little correct information 5=A lot of correct information	4.67	4.73	4.57
6	The tooltips helped me learn medical terms by heart	1=Strongly disagree 5=Strongly agree	4.47	4.33	4.57
7	The tooltips were difficult to understand	1=Strongly disagree 5=Strongly agree	1.86	1.50	2.07
8	The tooltips provided enough information for me to understand what I should enter to the system	1=Strongly disagree 5=Strongly agree	4.57	4.80	4.36
9	The tooltips should have provided different information	1=Strongly disagree 5=Strongly agree	2.63	2.47	2.93
10	The applications should have provided more information in the tooltips	1=Strongly disagree 5=Strongly agree	3.03	3.07	2.93

4.4.2 Helpful – Did the Participants Find the Tooltips Helpful?

During the post-interviews, one of our objectives was to find out whether the participants found the tooltips useful and helpful. Most of our users expressed that they found them helpful and the majority stated that when they were in doubt, they would “consult the tiny i’s”, the icon/button for opening the tooltips. This also corresponded with what most of the users wrote in the booklet. In addition, as the correctness of data entry increased over time it should be seen as a strong reason to believe that the participants found them helpful. To further argue for this belief, the UX questionnaire also revealed that they found them helpful, as most of the participants strongly agreed on the question about whether the tooltips helped them answer correctly. “The tooltips helped answer correctly to the tasks given” received a total of 4.7 out of 5, meaning that they strongly agree with the statement, which corresponds with the low ($p=0.27$) and moderate ($p=0.50$) correlation mentioned. Also, the fact that there is a difference between the explanation group and the normal value group, substantiates the indication of that tooltips have an effect. The second statement, “The provided tooltips helped me answer correctly on the tasks given”, got an average of 4.63 out of 5, where 24 of the 30 participants gave a 5. This means that they found the tooltips helpful to very helpful, which also conforms with the recordings in general. Many participants often chose a wrong option, opened and read one or more tooltips and then corrected their answer.

4.4.3 Rewarding – Did the Tooltips Give the Participants New Knowledge?

An objective we focused on in the UX questionnaire was whether or not the participants learnt something from the tooltips, and if this could indicate that the participants found the tooltips rewarding. Many of the participants stated during the post-interviews that they learned something, which corresponds with the result from the UX questionnaire. The statement “The tooltips helped me learn medical terms by heart” got an average of 4.5. In addition, the analysis of the video recordings revealed that the tooltips were opened less as the days went by, which corresponds with the results from the statement “The need for opening tooltips were less as the days went by” which got a 4.1. According to Michelsen et al. (1980), “Decreasing usage of help commands” may indicate that users are learning, which confirms the rewarding aspect of our research.

4.4.4 Comparison of Contradictory Statements

The result from statement 3, “The information given in the tooltips were easy to understand”, and statement 7, “The tooltips were difficult to understand”, were quite consistent, as the users agreed with first statement to a large degree, and disagreed to the latter statement. The results for statement 8, “The tooltips provided enough information for me to understand what I should enter to the system”, and statement 10, “The applications should have provided more information in the tooltips”, were not quite as consistent. The first statement got responses indicating agreement to strong agreement, while the second statement got more neutral responses, which we interpret to mean that the participants agreed to that the tooltips provided enough information. However, the participants also seems to partly think there should have been more information. This may indicate that they did not understand both of the questions, or that they found the tooltips adequate but they would not mind more information.

4.4.5 Comparing Normal Values and Explanations

There were no major differences between the results from the normal value group and the explanation group. However, there were some variations on some of the statements. The explanation group seemed to have less of a need for opening the tooltips as the days went by (statement 4), possibly indicating that they feel they may have learned more. The normal value group seemed to be more pleased with the amount and type of information (statement 8 and 9), based on the results above (Table 9).

4.4.6 Improvements

When analyzing the booklets and the post-interviews, several suggestions of improvements materialized. One participant suggested that we should add more vital signs to the data elements, while another stated “Add more information to the i’s. For example, can you have pre-eclampsia with only hypertension?”. A third participant suggested that we should “for instance giving the normal ranges for BP”. A fourth participant suggested signs and symptoms instead of formal definitions. She justified the statement by saying that non-medical personnel, by which she meant those with less education than nurses, would not know what a condition is, based on the explanations. This corresponds with previous research, as people tend to find it easier to understand new concepts through examples (Ormrod, 2012).

What is interesting is that the participants discussed above had been using the testing program containing explanations as their content type for tooltips. The fact that they suggested other types of information, correspond to the response from the UX questionnaire, where the following statements, “..should have provided more information..” and “..should have provided different information” received scores of 3.2 and 2.9 out of 5, indicating that the participants partly agree with the statements.

Statement 9 received an average score of 2.63, which indicates that many participants think the tooltips should have provided different information. What is interesting is that hardly anyone, except two Malawian participants, expressed this during the interviews. Some, however, suggested to further add more information. Also, there are no differences in the results between the participants from Malawi and SA, indicating more reliable results, as two groups are indicating the same.

5 Conclusion

5.1 First Research Question

Our first research question was “What content types for tooltips do health workers in developing countries prefer?”. Previous research in this area is limited, though Petrie et al. (2004) explored tooltips for hearing impaired participants, and identified Sign Language, Human Mouth, Digital Lips and Picture tooltips as the most preferred.

Similarly to Petrie et al., we also identified four types of tooltips adapted to our user group; explanation, normal value, treatment and illustration with normal value. The latter may be comparable to Picture tooltips (Petrie et al, 2004). In addition, they are both at Kirkpatrick’s level 1 (Kirkpatrick, 2006) and are characterized as low content validity (Gregor, 2006), due to only addressing participants’ preferences and opinions.

Based on the questionnaire presented during both iterations of this research, we found that the Malawian health workers preferred tooltips with normal values as content. The SA health workers, on the other hand, preferred tooltips with explanation as content type. Overall, normal values and explanations were most preferred of the four content types presented, thus these were used in the quasi-experiment. Of the two, normal value tooltips was significantly more preferred than the others, giving a predictive power.

5.2 Second Research Question

Our second research question was “What content type for tooltips lead to more correct data entry among undereducated health workers?”. Dai et al. (2015) developed a software that included step-by-step instructions, though it is difficult to compare this with our tooltips, as tooltips are not suited for displaying sequences of instructions, since they disappear once the button is tapped or when the user starts or completes entering data in the field.

Through our analysis of recordings from the experiment, we have found that explanations lead to more correct data entry among undereducated health workers. Explanations got significantly higher scores than normal value, indicating a predictive power.

Findings from our first iteration concluded that tooltips with a range of normal values was the most preferred one, however not the most effective in terms of correctness of data entry. A possible explanation for these findings could be that when we designed the cases, we used a lot of terms similar to the explanations in the tooltips. However, people tend to know what methods and techniques they learn best from, therefore the difference in preference and the actual effective tooltip was unexpected.

5.3 Third Research Question

Our third research question was “What techniques can be utilized to answer research question 1 and 2?”. We have found no previous research addressing this issue, though previous studies have utilized methods such as interviews, observations, surveys, ranking and pre- and post-tests (Petrie et al., 2004; Dai et al., 2015).

In order to find preference, we have learned that questionnaires, accompanied by interviews, observations and a question-suggestion approach to introduce the system, have been successful, and is thus recommendable. Though, one should be careful when designing the research, and always have in mind who is going to be a part of it and what their background is, as misunderstandings can easily arise.

An experiment is a beneficial way of finding what content type actually lead to more correct data entry, as one is able to have several sample groups, do a comparison and measure results. Using a screen recording tool has been an essential part of our results, and we therefore recommend that to be included as well. We have found that the combination of experiment and logging, encompasses Kirkpatrick’s level 3 and 4. Logging gives the opportunity of gathering statistical data which can be used to give results a predictive power. However, it requires an enormous amount of time, as a participant may spend a lot more time on one task than expected, hence creating long recording sessions. However, this issue may be abolished by utilizing a tool with opportunities for automation, such that only some user actions automatically get recorded. One also has to take into consideration the time it takes to transcribe the participant’s actions, as well as the interviews afterwards.

The conducted interviews in this research gave us a greater understanding of the user. Especially the post-interviews gave us insight into the participants’ reason making and thoughts on the tooltips, hence interviews should be a given part of any method, as it provides

explanatory power (at level 2a (Gregor, 2006)). We also found it useful to ask the participants to make notes/ write in the booklet each day as they could more easily remember thoughts and ideas that appeared during the process, leading to more productive interview sessions.

The questionnaire did not point to the most effective tooltips, hence having the first iteration seems pointless as the preference and effectiveness does not correspond. However, the questionnaires helped us narrow down the research as it would be very time consuming and expensive to do the experiment with all the different tooltip types. On the contrary, by excluding some tooltip alternatives from the experiment we might have missed on the opportunity to really find the most effective. For instance, we did not go into deeper exploration whether tooltips in picture format could have been a better option, which would have corresponded with Mayer's (1989) findings.

5.4 Fourth Research Question

Our final research question was “Do tooltips have an effect?”. Previous research has proven tooltips to be effective (Dai et al., 2015; Grossman & Fitzmaurice, 2010; Petrie et al., 2004;). Though, Rourke & Kanuka (2009) state that people hang on to their misconceptions until challenged. However, through our research we found that, despite their medical knowledge, the participants still opened the tooltips. We also found variations in correctness between different content types for tooltips. We have also found a low correlation between the number of opened tooltips and correctness ($p=0.27$) and a moderate correlation between successful tooltips and correctness ($p=0.50$). These correlations have a predictive power. In post experiment interviews, we found that participants think they learned from the tooltips, and that they helped them answer correct, which gives an explanatory power. In addition many of the participants stated both in the UX questionnaire and in the booklets that they found the tooltips useful and that the “i’s”(the tooltips) guided them in entering information to the system. Based on this, we conclude with that tooltips do cause improvements in correct data entry. The experiment has shown that there is a decrease in errors, as well as a decrease in use of help commands. Hence, this is an indication of learnability of the tooltips.

5.5 Understanding Users

Our research had a third degree of user involvement, because users' advice were acquired through interviews and questionnaires. User participation was of type consultative, meaning we took users' needs and preferences into consideration when designing tooltips.

We found that participants have different tolerance for asking questions. Some felt comfortable enough to ask a lot of questions when in doubt, others did not. Thus, when answering questions from the different participant, the results may have been affected as we could have answered in different manners, giving more information to some than others.

Also, in Malawi we had local contacts helping us answering and explaining the questions the participants had, ensuring that they understood the tasks. In SA, on the other hand, we did not have any local contact assisting with explanations. In addition, the introduction in SA included all ten participants simultaneously, and individual follow-up was difficult. This may be a possible reason for why the Malawi group performed better than the SA group.

We also found that some struggled to understand the Likert scale in the UX questionnaire. Based on interviews and observations it was not clear to everyone what the middle values were. A few participants asked about this, though most did not say anything. This may be a reason for why most of the results are either in the higher or the lower parts of the scale.

5.6 Reflections

In order to increase the validity of the experiment, we could have included a control group of participants. Here, the aim would have been to compare the effects of a system with tooltips and a system without tooltips. This is similar to research on medication, where one group is given real medicine, while the other is given placebo medication. However, the comparison between the two groups would not have been symmetric, as one group would have been introduced to tooltips and the other group not. An alternative way could be create a testing program with some meaningless tooltips. This would have made the groups more symmetric, giving one group actual tooltips and the other group "placebo-tooltips". Though, it may be difficult to disguise meaningless tooltips for users, as they might understand when a tooltip is not giving them any useful information.

Another thing we could do to improve the validity of the research would be to focus more on avoiding use of similar words and phrases in the cases and tooltips. Another possibility would be to include the same amount of phrases in the cases from both of the tooltips' contents. However, the cases were created from the pregnant woman's point of view, explaining her situation. This was an attempt to simulate a clinic visit, hence we found it strange to have the woman herself list medical condition values.

To minimize or eliminate the uncertainties around why explanations got a higher correctness than normal values, we could have created separate cases for the two groups which contained tailored concepts and tooltips. Though, this would have required a substantial amount of time, and would not be feasible within our time constraints. It would also be more difficult to compare different cases.

Another possible issue with our research was that the participants from SA were underrepresented. Hence, our original idea of comparing Malawi and SA was difficult. It might be beneficial to map only one country at a time, based on the fact that Malawi and SA preferred different content types. However, if we had been able to gather more participants in SA, we could have made a better comparison.

Most of our participants struggled with understanding the rating system on the questionnaire about preference for content types for tooltips, where the number 1 was the most preferred content type. Initially, when we first designed the questionnaire, 4 was the most preferred one. However, after conducting two pilot-tests of the questionnaire we switched the order, as suggested by our pilot participants. Most of our main participants did not have problem with adapting to our rating system after having it explained, though it contributed to confusion and time got lost explaining.

5.7 Recommendations

Based on our research, we recommend including tooltips in system because they are both effective and a cheap solution compared to other training materials. Though, it is important that people are aware of them and their function. During the introduction of a new system, tooltips should be visualized and demonstrated, and users should practice the tooltips. By doing this, they will likely understand when to use them in a real life situation.

References

- Alaszewski, A. (2006). *Using Diaries for social research*. London, United Kingdom, SAGE Publications Ltd.
- Baskerville, R. L. (1999, October). Investigating Information Systems with Action Research. *Communication of the AIS*, 2. Retrieved from http://delivery.acm.org/10.1145/380000/374476/a4-baskerville.pdf?ip=193.157.248.193&id=374476&acc=ACTIVE%20SERVICE&key=CDADA77FFDD8BE08%2E8BE0DFE7B528F835%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=920675229&CFTOKEN=44561038&acm_=1491467085_291753aa4338fce849abc4baf6085c83
- Braun, R., Catalani, C., Wimbush, J., & Israelski, D. (2013). Community Health Workers and Mobile Technology: A Systematic Review of the Literature. *PLOS ONE*. doi:10.1371/journal.pone.0065772
- Carroll, J. M. (1990). *The Nurnberg Funnel: Designing Minimalist Instruction for Practical Computer Skill*. Cambridge, Mass., MIT Press.
- Crang, M., & Cook, I. (2007). *Doing Ethnographies*. London, United Kingdom, Sage Publications Ltd.
- Dai, Y., Karalis, G., Kawas, S., & Olsen, C. (2015). Tipper: Contextual Tooltips that Provide Seniors with Clear, Reliable Help for Web Tasks. *CHI'15 Extended Abstracts*, 1773-1778. doi:[10.1145/2702613.2732796](https://doi.org/10.1145/2702613.2732796)
- DHIS2. (n.d.). Retrieved January 31st 2017 from <https://www.dhis2.org/>
- DHIS2. (n.d.). Data managements and analytics. Retrieved December 7th 2016 from <https://www.dhis2.org/data-management>
- DHIS2. (n.d.). Deployment. Retrieved April 6th 2017 from <https://www.dhis2.org/deployments>
- DHIS2. (n.d.). In Action. Retrieved April 6th 2017 from <https://www.dhis2.org/inaction>

- DHIS2. (n.d.). Technology Platform, Retrieved December 7th 2016 from <https://www.dhis2.org/technology>
- Duh, H.B.L., Tan, G.B.C., & Chen, V.H.H. (2006). Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. *Proceedings of Conference on Human-Computer Interaction with Mobile Devices and Services*, 8, 181-186.
doi:[10.1145/1152215.1152254](https://doi.org/10.1145/1152215.1152254)
- Farkas, D.K. (1993). The Role of Balloon help. *ACM SIGDOC Asterisk Journal of Computer Documentation*, 17(3), 3-19. doi:[10.1145/154425.154425](https://doi.org/10.1145/154425.154425)
- Flaherty, K. (2016, June 5th). Diary Studies: Understanding Long-Term User Behaviour and Experiences. Retrieved from <https://www.nngroup.com/articles/diary-studies/>
- Greenwood, D.J., & Levin, M. (1998). Action research, science, and the co-optation of social research. *Studies in Cultures, Organizations and Societies*, 4:2, 237-261.
doi:[10.1080/10245289808523514](https://doi.org/10.1080/10245289808523514)
- Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, 30(3), 611-642. Retrieved from <http://www.jstor.org/stable/25148742>
- Grossman, T., & Fitzmaurice, G. (2010). Toolclips: An Investigation of Contextual Video Assistance for Functionality Understanding. *ACM Conference on Human Factors in Computing Systems*, 10, 1515-1524. doi:[10.1145/1753326.1753552](https://doi.org/10.1145/1753326.1753552)
- Grossman, T., Fitzmaurice, G. & Attar, R. (2009). A survey of software learnability: metrics, methodologies and guidelines. *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, 27.
doi:[10.1145/1518701.1518803](https://doi.org/10.1145/1518701.1518803)
- Grossman, R., & Salas, E. (2011). The transfer of training: what really matters. *International Journal of Training and Development*. 15, 103-120. doi:[10.1111/j.1468-2419.2011.00373.x](https://doi.org/10.1111/j.1468-2419.2011.00373.x)
- Hadjerrouit, S. (2008). Using a Learner-Centered Approach to Teach ICT in Secondary Schools: An Exploratory Study. *Issues in Informing Science and Information Technology*, 5, 233-259. Retrieved from

http://s3.amazonaws.com/academia.edu.documents/30423624/iisitv5p233-259hadj424.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1491492862&Signature=w0iH1IH0HagF9zBJWsoMyAzfovI%3D&response-content-disposition=inline%3B%20filename%3DUsing_a_Learner-Centered_Approach_to_Tea.pdf

Harrison, S.M. (1995). A Comparison of Still, Animated, or Nonillustrated On-Line Help with Written or Spoken Instructions in a Graphical User Interface. *CHI '95 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 82-89. doi:[10.1145/223904.223915](https://doi.org/10.1145/223904.223915)

Hassenzahl, M. (2008). User Experience (UX): towards an experimental perspective on product quality. *IHM '08 Proceedings of the 20th Conference on l'Interaction Homme-Machine*, 20, 11-15. doi:[10.1145/1512714.1512717](https://doi.org/10.1145/1512714.1512717)

Huang, J., & Twidale, M.B. (2007). Graphstract: Minimal Graphical Help for Computers. *ACM Symposium on User Interface Software and Technology 07*, 203-212. doi:[10.1145/1294211.1294248](https://doi.org/10.1145/1294211.1294248)

Hyldegård, J. (2006). Using Diaries in Group Based Information Behaviour Research - A Methodological Study. *IiX Proceedings of the 1st international conference on Information interaction in context*, 153-161. doi:[10.1145/1164820.1164851](https://doi.org/10.1145/1164820.1164851)

Joshi, S., & Bratteteig, T. (2015). Assembling Fragments into Continuous Design: On Participatory Design with Old People. *Lecture Notes in Business Information Processing (LNBIP)*, 223. doi:10.1007/978-3-319-21783-3_2

Instone, K. Heuristics for the Web. Retrieved January 2017 from <http://instone.org/heuristics>

International Telecommunication Unit. (2015). Measuring the Information Society Report 2015. Geneva, Switzerland: ITU. Retrieved April 20th 2017 from <http://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2015/MISR2015-w5.pdf>

ISO 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability. Retrieved April 6th 2017 from <http://www.userfocus.co.uk/resources/iso9241/part11.html>

- Ives, B., & Olson, M.H. (1984). User Involvement and MIS Success: A Review of Research. *Management Science*, 30(5), 586-603. Retrieved from <https://www.jstor.org/stable/pdf/2631374.pdf>
- Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of American Society of Training Directors*, 13, 21-26.
- Kirkpatrick, D. L., Kirkpatrick, J. D. (2006). *Evaluating Training Programs: The Four Levels*. San Francisco, USA, Berrett-Koehler.
- Mayer, R. E. (1989). Models for Understanding. *Review of Educational Research*, 59(1), 43-64. doi:[10.3102/00346543059001043](https://doi.org/10.3102/00346543059001043)
- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8. doi:[10.1111/j.1745-3992.1995.tb00881.x](https://doi.org/10.1111/j.1745-3992.1995.tb00881.x)
- Michelsen, C. D., Dominick, W. D. & Urban, J. E. (1980). A methodology for the objective evaluation of the user/system interfaces of the MADAM system using software engineering principles. *ACM Southeast Regional Conference*, 18, 103-109. doi:[10.1145/503838.503847](https://doi.org/10.1145/503838.503847)
- Myers, M. D. (1997). Qualitative Research in Information Systems. *MIS Quarterly*, 21(2), p. 241-242. Retrieved from <http://www.qual.auckland.ac.nz/>
- National Health Service. (2015, January 8th). Your antenatal care - Pregnancy and baby guide - NHS Choices. Retrieved April 19th, 2017 from <http://www.nhs.uk/conditions/pregnancy-and-baby/pages/antenatal-midwife-care-pregnant.aspx#What>
- Nielsen, J. (1996). *Usability Metrics: Tracking interface improvements*. CA, United States: SunSoft.
- Novick, D.G., Elizalde, E., & Bean, N. (2007). Towards a more accurate view of when and how people seek help with computer applications. *ACM, SIGDOC*, 7. doi:[10.1145/1297144.1297165](https://doi.org/10.1145/1297144.1297165)

- Oluoch, T., Santas, X., Kwaro, D., Were, M., Biondich, P., Bailey, C., Abu-Hanna, A., & de Keizer, N. (2012). The effect of electronic medical record-based clinical decision support on HIV care in resource constrained settings: A systematic review. *International Journal of Medical Informatics* 81(10), e83–e92. doi:10.1016/j.ijmedinf.2012.07.010
- Orlikowski, W.J., & Baroudi, J.J. (1991). Studying Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research*, 2(1), 1-28. doi:[10.1287/isre.2.1.1](https://doi.org/10.1287/isre.2.1.1)
- Ormrod, J. E. (2012). Human Learning. *Englewood Cliffs, New Jersey, Merrill*, 237-241.
- Petrie, H., Fisher, W., Weimann, K., & Weber, G. (2004). Augmenting Icons for Deaf Computer Users. *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, 1131-1134. doi:[10.1145/985921.986006](https://doi.org/10.1145/985921.986006)
- Rapoport, R.N. (1970). Three Dilemmas in Action Research - With Special Reference to the Tavistock Experience. *Sage journals, Human Relations*, 23(6), 499-513. doi:[10.1177/001872677002300601](https://doi.org/10.1177/001872677002300601)
- Rettig, M. (1991). Nobody Reads Documentation. *Communication of the ACM- Special issue on computer graphics*, 35(7), 19-24. doi:10.1145/105783.105788
- Rogers, Y., Sharp, H., & Preece, J. (2011). *Interaction Design: beyond human-computer interaction (3rd ed)*. West-Sussex, United Kingdom: John Wiley and Sons, Ltd.
- Shroyer, R. (2000). Actual Readers Versus Implied Readers: Role Conflicts in Office 97. *Technical Communication*, 47(2), 238-240.
- Silverman, D. (1998). Qualitative research: meanings or practices?. *Information Systems Journal*, 8, 3–20. doi:10.1046/j.1365-2575.1998.00002.x
- Simonsen, J., & Robertson, T. (Ed.). (2013). *Routledge International Handbook of Participatory Design*. New York, USA and Oxon, Canada: Routledge.

- Smart, K.L., Whiting, M.E. & Detienne, K.B. (2001). Assessing the Need for Printed and Online Documentation: A Study of Customer Preference and Use. *Journal of Business Communication*, 38(3), 285-314. doi:[10.1177/002194360103800306](https://doi.org/10.1177/002194360103800306)
- Sood, S. P., Nwabueze, S. N., Mbarika, V. W. A., Prakash, N., Chatterjee, S., Ray, P., & Mishra, S. (2008). Electronic Medical Records: A Review Comparing the Challenges in Developed and Developing Countries. *Proceedings of the 41st Hawaii International Conference on System Sciences*. doi:[10.1109/HICSS.2008.141](https://doi.org/10.1109/HICSS.2008.141)
- United Nations. (2015, July). Millennium Development Goals Report 2015. Retrieved from [http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20\(July%201\).pdf](http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf)
- United nations economic commission for Europe. (2009). Making data meaningful Part 2: A Guide to Presenting Statistics. United nations economic commission for Europe (UNECE).
- Worldometers. (n.d.). Population Africa. Retrieved March 15th 2017 from <http://www.worldometers.info/world-population/africa-population/>
- World Health Organization. (2013). mHealth: New horizons for health through mobile technologies. Retrieved March 6th 2017 from http://www.who.int/goe/publications/goe_mhealth_web.pdf
- World Health Organization. (2014). A universal truth: No health without workforce. Retrieved March 6th 2017 from http://www.who.int/workforcealliance/knowledge/resources/GHWA-a_universal_truth_report.pdf
- World Health Organization. (2015). Maternal mortality. Retrieved April 19th, 2017 from <http://who.int/mediacentre/factsheets/fs348/en/>
- World Health Organization. (n.d.). What matters to women during pregnancy: a different approach to antenatal care. Retrieved April 19th, 2017 from http://who.int/reproductivehealth/topics/maternal_perinatal/anc/en/

Appendix - Feedback to DHIS2 software developers

The main focus of this research were to explore different aspects of tooltips, however some possible usability flaws were also detected. Below we present the most common feedback from the users as well as some observations we made during the research

Complete button

Many users find the complete button confusing. As the button appear on the first screen of a form, some user automatically think they should press it after entering information on that page.

Progress bar

The reality is that there are several screen in one form, thus there should be a progress bar indicating on which screen the user is and a possible solution is to put the complete button at the last page of a form.

More feedback

After pressing the complete button inside a form most of the users expected some sort of feedback. Either by the system providing a dialog box or by the system going back a screen to the client patient record

Visibility – Data fields

Many of the user struggled with understanding which data field they operated within, adding wrong data to wrong data fields. This indicates a need for higher visibility of which data field the user is working in, for instance by somehow highlight the data field to a larger degree.

Search function

Too many search function. Making it confusing for users to navigate.

Search fields should also catch misspellings as well

Arrows /discoverability

The arrows to be found on the top right of each form, used to navigate between the pages, are barely discoverable. Hardly anybody participating in this research found them without any help

Press on the heading to enter information

Several participants in this research pressed the heading of the data fields in hope to enter information.

Make the i\s / tooltips-button more prominent

None of the user noticed the i\s before getting introduced to them. In addition, one person thought they meant subtopics and did not consider them important

Patient record

Clarify whether patients are completed or still “active”

Clearer indication on which patient record the user is working with.

It should be possible to see a sort of summary document of the patient

Log in and log out

Enable offline login as people may accidentally log out. Without internet access they are not able to log in again, leading to no data capturing.

Different design/tailoring of the setup of fields for different cadres

Should not skip “birth date” when pressing Next downwards the page

The save-button should have some sort of text indicating its purpose

Switch between two languages in the app, depending on who is using it

Add support for pictures and/or videos

The date picker should be more intuitive, and not reset dates if one accidentally pushes a back-button

