

Methods for Evaluation of Tooltips

Helene Isaksen^{1(✉)}, Mari Iversen¹, Jens Kaasbøll¹, and Chipo Kanjo²

¹ University of Oslo, Oslo, Norway
{helenis, mariive, jensj}@ifi.uio.no

² University 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 is a subcategory of 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 of 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. [6] 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 [6] 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 [14] 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 [18].

When designing the tooltip, an important aspect is to not to overload the screen with extraneous information [9]. 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 [15] 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 [12, 13] 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 [5] 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 [19] or are not capable of [8] 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., [17] 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., [17] 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 [14] 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. [17], 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 Fig. 1.

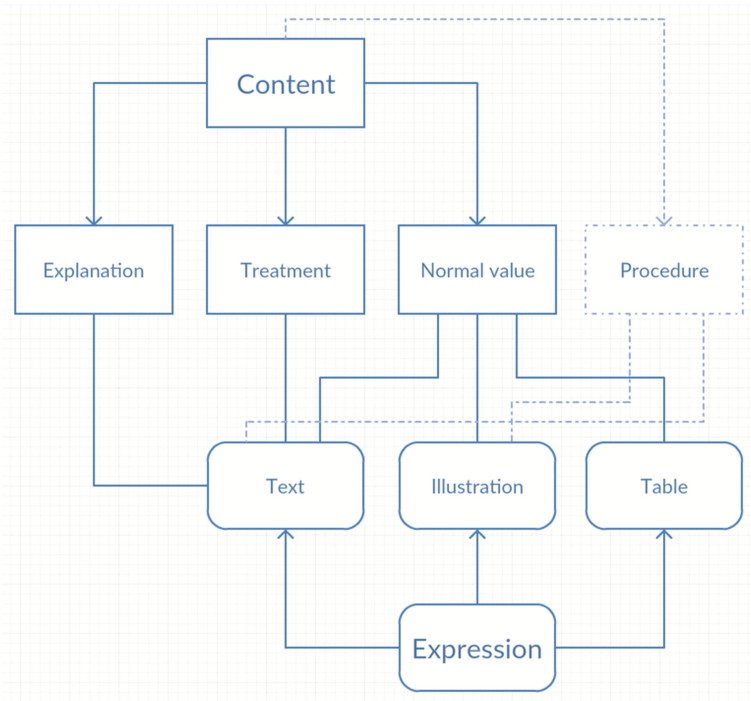


Fig. 1. Combinations of content and expression types in questionnaire

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 next to the boxes in the lower right corners were not included in the questionnaire but added here for clarification.

Statistically significant differences were found from 58 respondents, see Isaksen et al. [10].

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. [17], 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.

Fundal Height

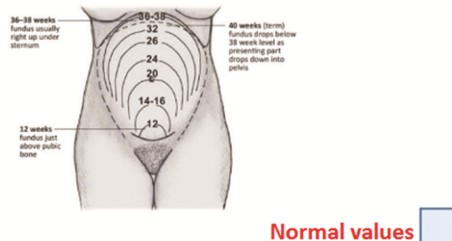
<p>Measurement from the pubic bone to the top of the uterus</p>	<p>If measurement is abnormal, please do extensive examination and tests, or refer the patient to a specialist.</p>
<p>Normal fundal height measurement: 20 weeks = 17-20 cm 28 weeks = 25,5-28,5 cm 36 weeks = 33-35 cm 40 weeks= 36-38 cm</p>	

Fig. 2. An example from the questionnaire

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 [6], 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. [6] 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 [6].

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 [4]. 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.



Fig. 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 [16]. 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. [17] 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 [7].

Kirkpatrick level 3 evaluates behavioral change. In an experiment [10] 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 2/3 of the videos were recorded.

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

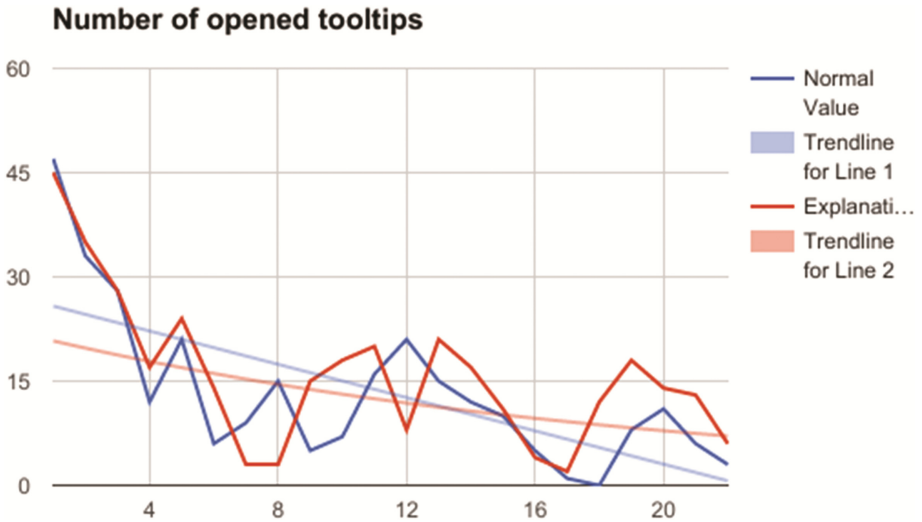


Fig. 4. Number of opened tooltips throughout the cases.

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 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.

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

Significant improvements and significant difference between the normal value and explanation group and interview results, made Isaksen et al. [11] 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 [12] 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. [6] 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, 17] 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 [12] 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 [12], 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.

Acknowledgement. This research has been supported by QU Horizon 2020 “mHealth4Afrika - Community-based ICT for Maternal Healthcare in Africa” (project 668015, topic ICT-39-2015), Norwegian Centre for International Cooperation in Education “Scholarly Health Informatics Learning” (UTF-2016-longterm/10032) and Norwegian Agency for Development Cooperation “Support to the Health Informations System Project - HISP” (QZA-14/0337).

References

1. Carroll, J.M.: *The Nurnberg Funnel: Designing Minimalist Instruction for Practical Computer Skill*. MIT Press, Cambridge (1990)
2. Dai, Y., Karalis, G., Kawas, S., Olsen, C.: Tipper: contextual tooltips that provide seniors with clear, reliable help for web tasks. In: CHI 2015 Extended Abstracts, pp. 1773–1778 (2015)
3. DHIS2 (2017). <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. In: *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services*, Helsinki, Finland, pp. 181–186 (2006)
5. Gregor, S.: The nature of theory in information systems. *MIS Q.* **30**(3), 611–642 (2006)

6. Grossman, T., Fitzmaurice, G., Attar, R.: A survey of software learnability: metrics, methodologies and guidelines. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, pp. 649–658 (2009)
7. Grossman, R., Salas, E.: The transfer of training: what really matters. *Int. J. Train. Dev.* **15**, 103–120 (2011)
8. Hadjerrouit, S.: Using a learner-centered approach to teach ICT in secondary schools: an exploratory study. *Issues Informing Sci. Inf. Technol.* **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: Proceeding 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. In: HCI International (2017)
12. Kirkpatrick, D.L.: Techniques for evaluating training programs. *J. Am. Soc. Train. 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*, vol. 295. Wiley, West Sussex (2010)
15. Messick, S.: Standards of validity and the validity of standards in performance assessment. *Educ. Meas. Issues Pract.* **14**(4), 5–8 (1995). doi:10.1111/j.1745-3992.1995.tb00881.x
16. Ormrod, J.E.: *Human Learning*. Merrill, Englewood Cliffs (2012)
17. Petrie, H., Fisher, W., Weimann, K., Weber, G.: Augmenting icons for deaf computer users. In: CHI 2004 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria, pp. 1131–1134 (2004)
18. Shroyer, R.: Actual readers versus implied readers: role conflicts in office 97. *Tech. Commun.* **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. *J. Bus. Commun.* **38**, 285–314 (2001)