Heuristic Evaluation of persuasive systems: the case of educational programs

Eva L. Ragnemalm
Department of Computer Science,
Linköping University
Linköping, Sweden
eva.ragnemalm@liu.se

Magnus Bång
Department of Computer Science,
Linköping University
Linköping, Sweden
magnus.bang@liu.se

Ingrid Alin-Nilsson
Combitech AB
Linköping, Sweden
ingrid.alin-nilsson@combitech.se

ABSTRACT

The evaluation of persuasive systems is a time-consuming and expensive endeavor. In the area of human-computer interaction, Heuristic Evaluation was suggested as an inexpensive alternative method for evaluating the usability of software. This paper examines the use of Heuristic Evaluation as a means to evaluate the persuasive power of software systems. A heuristic evaluation of the persuasive power of two educational programs used in school was performed. The heuristics used were operationalized versions of well-known persuasive principles. The evaluation indicated that both programs have persuasive power. This result was in agreement with an expert's experience of the programs. The study shows that our heuristics can be used in a summative heuristic evaluation of persuasive power of educational systems.

Categories and Subject Descriptors

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous, K4.2: Computers and Society: Social Issues,

General Terms

Measurement, Design, Human Factors.

Keywords

Persuasive technologies, heuristic evaluation, educational software, heuristics, captology

1.INTRODUCTION

Many information technology systems and services today are being designed with an intention to change people's attitudes and behaviors. Fogg has defined persuasive technology as technology that is specially designed to change attitudes and behaviors of the users through persuasion and social influence, but not through coercion [3]. Typically, the systems that have been designed in this area of research and development are devoted to topics such as healthy living [6] and energy conservation [8], [2], [9]. Researchers have provided models of human behavior as well as persuasive principles [3], [15] that can be applied when designing such systems.

Evaluating the efficacy of the systems is often approached using a research approach, that is, using traditional statistical methods to evaluate if the persuasive strategies are effective. A typical study design comprises an intervention where variables of interest are measured before and after using the system for target and control groups. This approach to evaluate the persuasive power is quite resource-intensive and expensive in terms of the amount of work required to set-up the interventions, the methods for data collection and the subsequent statistical analysis. Moreover, an additional problem with this approach — in terms of system development — is that iterative evaluations cannot be done as part of the design process, since it requires a fully operational system.

The field of human-computer interaction has approached similar problems by developing usability inspection methods. One such method is heuristic evaluation, where a system is inspected by an expert using a set of heuristics, that is, general rules of thumb that indicate usability [11], [13], [14]. However, well-defined, inexpensive and pragmatic methods to analyze the persuasive power of computer programs — in the design or evaluation phases — are scarce (c.f. [17]). Recently, Kientz and colleagues suggested the use of heuristic evaluation to inspect persuasive technologies [7]. They developed a set of heuristics intended to find problems in persuasive technologies that would affect persuasive elements, adoption, or long-term effectiveness of the technologies. This means that their purpose is *formative*, that is, focused on finding problems. Heuristic evaluation can also be used in a *summative* manner, that is, evaluating the level of usability of a program.

In this paper, we explore the use of summative heuristic evaluation to evaluate persuasive power. Based on persuasive principles presented by Fogg [3], we have developed a set of specialized heuristics to evaluate educational systems used in school. We have evaluated two systems intended to teach reading and language awareness with the goal of measuring their persuasive power. The paper is organized as follows; first, we present heuristic evaluation of usability. Second, we discuss how our analysis was performed, including the choice of programs to test and operationalization of principles into heuristics. Third, we present the results of the analysis. Finally, we discuss the results and present some directions for future work.

2. HEURISTIC EVALUATION

Usability inspection methods were developed in the 1990s and are procedures for evaluating usability that can be applied to implemented programs as well as prototypes and even sketches or specifications of a system [11], [13]. These approaches require the systematic inspection or walkthrough of the user interface of a program (i.e. running it, trying out tasks and observing the function and visual impression of the program or sketch) and this inspection is based on a collection of design principles, or heuristics, that govern the usability of programs. One such method is Heuristic Evaluation [13], [14]. When using this method, several usability experts, armed with identical sets of usability

heuristics, first familiarize themselves with the program or specification, then systematically inspect the entire program, looking for features, properties, sequences of actions that do not fulfill the usability principles. These are termed usability problems and are carefully described in a document. After each expert has completed the inspection, the results from the different experts are accumulated, often usability problems are rated as to severity and expected frequency of occurrence and sometimes remedies are suggested. The result is then returned to the developers of the computer program or specification. This type of evaluation is useful during the design process, since the object to be inspected need not be a working program but may be a paper prototype or a written specification. When used in this manner, it is called *formative evaluation*. Heuristic evaluation can also be used for *summative evaluation*, which is performed in a similar manner as described previously, but the usability problems found are rated as to severity and documented, and the result of the evaluation is the number of problems or some accumulated numerical level of usability. The initial studies of the Heuristic Evaluation method focused on the summative use of the method [14]. An example of summative use is the evaluation of nursing equipment by four evaluators reported by Graham et. al. [4] where pairwise inter-rater agreement was found to be between 52-60% (expert evaluators were used and a 5-point scale).

Several sets of principles that can be used in Heuristic evaluation have been developed. One of the best-known collections is that of Nielsen [12], which was extracted by factor analysis of seven previous sets of heuristics applied to a set of known usability problems.

The unit of evaluation may be thought to be different when evaluating persuasive technology and usability. Many persuasive principles take the context of use into consideration while most usability principles focus on the system itself. For example, the principle of inherited trustworthiness - suggested by Fogg - encompasses the context, that is, whether the source of the program is trustworthy [3]. But some usability principles also take the context and the user into consideration, for example the principle of familiarity (i.e. that the terminology used in the system should be familiar to the user). Thus, the unit of analysis for heuristic usability evaluation also includes aspects of the context. Of course it is impossible to foresee for each and every user exactly which terms are familiar and which are not, so evaluations of this principle must always be based on the judgement of the analyst. The judgement is always based on the knowledge available at the moment of evaluation. Either the context is known, as in a summative analysis of a program in use, or it is not, and can not be analyzed. The remaining principles of persuasive technology or usability can still be analyzed and provide insight.

3. METHOD

To investigate the heuristic evaluation approach as a method for analyzing the persuasive power of computer systems, we analyzed two educational programs used in classroom settings in Sweden. The hypothesis was that when submitted to a summative heuristic evaluation such programs should rate high on a persuasive heuristics scale. The two selected programs were Klicker and Läseboken. Klicker (v1.01 published 1997 by Liber Multimedia, Sweden) provides language awareness exercises for pupils aged 5-7 in need of special tutoring For example, in the system, the pupil identifies different phonemes that makes up words or identifies vowels. Läseboken (v2.0 published by Allemansdata 1998, Sweden) is an educational program that provides reading training, fill-in-the-blank exercises, and words sorting. The target group is pupils 6-8 of age in need of special tutoring. Both systems are visually oriented, providing direct manipulation interaction. The keyboard is not often used in the exercises. These systems were selected on the basis of information from an experienced language teacher working in a K-6 school in Sweden [5]. The selection of educational software used for training in this particular school was based on recommendations from a central state agency providing educational materials to schools.

A common assumption is that learning is reflected by change in observed behavior. This is particularly true in school settings (for instance, pupils are expected to spell correctly after having learned to spell). Assuming that persuasion is seen as activities by one agent intended to change the behavior or attitude of another, human, agent, implies that persuasion is strongly related to classroom learning; both result in changes of behavior and are the intention of the teacher. We assume here that classroom learning can be seen as a subset of persuasion, due to the fact that the intention of the teacher is to convince the pupils that there is one correct way to spell (for instance) and pupils are thought to have learned when their behavior agrees with the goals of the teacher. Informal and unintentional learning is not included in this study.

Thus a computer program intended for educational purposes must have persuasive power; otherwise no change of behavior would result. Since the chosen programs were known to be effective (the children that use them do improve their reading skills, [5]), these programs must have persuasive properties.

3.1 Operationalization of the principles into heuristics

The principles for persuasive systems defined by Fogg [3] are formulated on a general level. Consequently, we were forced to reformulate the applicable principles as heuristics adapting them to be applicable for educational software that are used in a school environment. Some principles are formulated for mobile systems and web-based systems and these were deemed not relevant and omitted. In total, 22 statements, heuristics, were derived from the 24 principles [1]. Table 1 shows the persuasive principles and the corresponding heuristics.

Table 1. Persuasive principles and corresponding heuristics for educational software used in schools.

Persuasive Principle	Heuristic
Trustworthiness	Can students relate to and feel familiar with the context, images and figures that appear in the program?
Expertise	Does the system contain the knowledge to be taught?
Presumed credibility	Not included. Assumed to be high due to context
Surface credibility	Is the program visually attractive and appealing? Is the system dialogue adapted to the student's language and using appropriate language conventions? Is the dialogue with the student minimalistic and clear? Do the students have control and are there an option to cancel if the student so desires? Does the system provide adequate assistance to the student when problems arise?
Reputed credibility	Not included. Impossible to analyze without real context.
Earned credibility	Does the system behave consistently?

Near-perfect functional behavior	Does the system work without fuss and errors?
Reduction	Is the program easy to use and the tasks easy to perform with a small number of steps and keystrokes?
Tunneling	Does the program show clearly which task and action to perform next?
Tailoring	Is the information provided, text and sound, customized to the student, set at the right level of knowledge and adapted to the pupil's age?
Suggestion	Are suggestions and information about how the tasks should be addressed offered in a timely fashion?
Self-monitoring	Can students learn about how he/she solved the tasks on previous occasions when the system has been used?
Surveillance	Are students aware that the teachers can observe and see the results?
Operant conditioning	Is positive and negative feedback given to the learner, such as displaying the result of how a task was performed and / or suggestions on how to solve it?
Cause and effect	Is the result of actions (i.e., the immediate feedback when pressing a dialogue box) clearly shown.
Virtual rehearsal	Can students perform tasks more than once?
Virtual rewards	Do student get rewards and praise when a task is performed correctly?
Simulation in real-world contexts	Are the tasks realistic and applicable to familiar problems?
Attractiveness	Identical to the heuristic for surface credibility
Similarity	Identical to the heuristic for trustworthiness
Praise	Identical to the heuristic for virtual rewards
Reciprocity	Is the program perceived as helpful?
Authority	Is the program perceived as having a role as a teacher and a complement to the teacher?

All of the original principles were not applicable since the aim is to inspect the program itself using heuristic evaluation. This does not mean that the disregarded principles are unimportant, only that they can not necessarily be evaluated using this method. For instance the principle of trustworthiness of the program itself is comprised by three aspects: the credibility of the source of the program, whether the program or the source is perceived as arguing against own interest and if there is a feeling of similarity with the program. The source of the program in this context is the teacher, who should reasonably be afforded a relatively high credibility. Whether the teacher providing the program is arguing against his or her own interests is deemed to be impossible for the child to detect, thus the principle is not relevant The final aspect, if the child can experience familiarity or similarity to the program was expressed in terms of relating to the visual content of the system:

> *Is it possible for the pupil to experience a relation to, or a feeling of familiarity with, the pictures, figures and backgrounds provided in the program?*

The principles of reputed credibility and presumed credibility were also disregarded, since they relate to aspects beyond the design of the system — the source of the program (discussed above), and other people's opinions of the system (which is impossible to asses without a real context) respectively.

Other principles were analyzed and adapted. The principle of surface credibility is based on the apparent qualities of the program, and was thought to contain both visual and behavioral aspects, thus implying that the program must have high usability to be persuasive. Therefore some aspects of usability were thought to be relevant. This principle resulted in several statements to be evaluated:

> *Is the program designed to be visually attractive?*
> *Is the dialogue with the pupil expressed in the pupil's terms, are the pupil's language conventions followed?*
> *Is the dialogue with the pupil expressed in a minimalist manner without unnecessary information?*
> *Is the pupil in control, can the pupil stop the program if he/she wishes?*
> *Is relevant help provided at need?*

In order to create a self-explanatory scale from non-compliance to full compliance, each heuristic was provided with a negative version, and a 5-grade scale from -2 (never) to +2 (always) was provided in between, where half-way grading was allowed (thus in reality creating a 9-grade scale). The values and statements were arranged so that the negative values indicate low persuasive power and the positive values indicate high persuasive power (compliance with the principles). Each heuristic was also provided with space for notes by the evaluator. Figure 1 shows how a sample heuristic, in this case derived from the principle of virtual reward (providing reward for desired behaviour) was designed.

3.2 Analysis

Two evaluators analyzed the programs, separately and one program at a time, taking as much time as needed. One of the evaluators was the original constructor of the heuristics (and is one of the authors) while the other had only a brief introduction to the persuasive principles. Both evaluators were (at the time) final

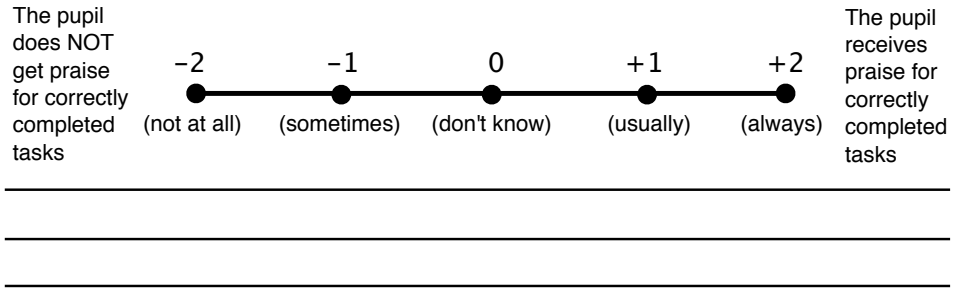

Figure 1. Sample heuristic in the form used (translated)

year master students in cognitive science whose experience of heuristic evaluation of usability was limited to educational settings.

After both evaluators had analysed both programs, the results were summarized. For each heuristic, the mean rating was calculated, then added into a total heuristic rating (the range of possible values for this being -44 to +44), where +44 indicates full compliance with the heuristic, and -44 would be the exact opposite.

4. RESULTS

The he accumulated summative value for the first program, Läseboken, was 13.75 and for the second, Klicker, 14.75. This is more than 60% of the possible score, which should indicate significant persuasive power.

All heuristics could be identified and scored and most of the heuristics were complied with to some extent in both programs (indicated by neutral or positive values). Only one heuristic was scored with strong negative values (mean of -1.75 and -2.0) in both programs. This was the heuristic based on the principle of reward, indicating that both programs lack the concept of reward.

The second evaluator (who was not very familiar with the area of persuasive systems) misunderstood two heuristics and did not evaluate them as the constructor had intended. The first heuristic was "Does the system provide timely suggestions and information on how to proceed?", which is based on the principle of suggestion. This principle stresses that information (suggestions) should be provided at the appropriate time, and the second evaluator disregarded the time aspect and evaluated only IF information was provided. The second case was "Is the result of an action, such as a dialogue box being pressed, clearly indicated?" This heuristic is based on the principle of cause-and-effect, which stresses that the effect of an action should be visible. This principle was interpreted with two heuristics, the other being "Is positive and negative feedback provided the pupil, such as the result of a task being shown?". Here the other evaluator interpreted both heuristics as pertaining to showing the result of tasks performed, whereas the intention of the constructor was that the first heuristic should relate to interaction with any interface objects, not just the completion of tasks. Thus 2 out of the 22 heuristics were misinterpreted. This may be a result of the second evaluator being less familiar with the principles used. It is known that the skill of the evaluator in the area of usability is a significant factor for the results of a heuristic evaluation of usability [10].

A comparison of the results of the two evaluators was also performed. Given the 9-point scale used to rate the compliance with heuristics, exact agreement is expected to be very rare. In Klicker, the evaluators agreed exactly on 6 heuristics out of 22 (27%), and had a difference of 1 or less (on the scale from -2 to +2), on 16 heuristics (73%). In Läseboken the evaluators agreed exactly on 9 heuristics (41%), and had a difference of 1 or less on 19 heuristics (86%). This can be compared to the pairwise inter-rater agreement for heuristic usability evaluation reported by Graham [1], who found 52-60% inter-rater agreement when using a 5-point rating scale. Our result suggest that different evaluators understand and use the heuristics consistently.

5. DISCUSSION

The summary evaluation performed on the two programs did indicate a persuasive power for both educational programs, which is to be expected since the programs were chosen because they were known to be effective. The relatively high rate of agreement between the two evaluators also indicates that the method is reliable. This implies that it is possible to use heuristic evaluation for a summary evaluation of persuasive power and that the heuristics used can be applied with a high degree of consistency.

The fact that both programs did not use the principle of reward is an interesting observation. It would be interesting to analyze more educational programs to see if this is a common phenomenon.

The fact that one evaluator misunderstood some heuristics is notable but not unexpected. It is well known from heuristic evaluation of usability that less skilled evaluators will find fewer errors. One reason for this might be that they misunderstand or misinterpret the heuristics supplied. Thus this may be more a reflection of the evaluator's skill than a problem with the method. In usability engineering the problem is handled by using several evaluators and collating or averaging their results. The same method could reasonably be used when evaluating persuasive power.

The conversion of the abstract principles for persuasion into more concrete heuristics shows that it is possible to express the principles in a manner more suited to evaluation by inspection, and while the heuristics used here are not necessarily ideal expressions of the principles (and also limited to educational software used in schools), they provide a start on an attempt to create heuristics for persuasive systems.

6. FURTHER WORK

A follow-up study would be to analyze programs that do not intend to persuade (for instance a game or a tool like Word) and see what level of persuasive power can be thought as "neutral" or not persuasive when using this measure. It is not at all certain that -44 on the scale from -44 to +44 means no persuasive power, since four heuristics were derived from usability heuristics. This means that a usable program will have a minor persuasive effect even if no other persuasive principles are fulfilled. Whether or not any persuasion takes place when using such a system is yet another issue, both philosophical and practical - can there be persuasion where there is no intention (not according to the definition), but what about unintended changes of behaviour or attitude?

Another important focus of future research is formulating the heuristics at a more general level, or formulating heuristics for other types of systems. It is not clear to what extent the heuristics formulated for this analysis can be used with other types of systems in other types of settings. Comparing these heuristics with those used in the study by Kientz et. al. [7] is also necessary. The method used by Nielsen [12] to generate the list of 10 usability heuristics that explain most of the usability problems in a known set might be possible to adapt to persuasive systems. Undertaking this analysis is a major challenge.

ACKNOWLEDGEMENTS

We would like to thank our evaluator, Sona Arabloui.

REFERENCES

[1] Alin-Nilsson, I. (2009). *Lärprogram som övertygande resurs — en heuristisk utvärdering av den övertygande förmågan hos lärprogram*. Master thesis LIU-IDA-KOGVET-A--09/005--SE, Linköping University, Sweden. In swedish.

[2] Bang, M., Torstensson, C., Katzeff, C.: The PowerHouse: A persuasive computer game designed to raise awareness of domestic energy consumption. In: PIJsselsteijn, W.A., de Kort, Y.A.W., Midden, C.J.H., Eggen, J.H., van den Hoven, E.A.W.H. (eds.), *PERSUASIVE 2006*. LNCS, 3962, New York, Springer Verlag (2006) 123-132

[3] Fogg, B. J. (2003) *Persuasive Technology, Using Computers to Change What We Think and Do*. Morgan Kaufmann Publishers, San Francisco

[4] Graham, M.J., Kubose, T.K., Jordan, D., Zhang, J., Johnson, T.R., Patel, V.L. (2004). Heuristic evaluation of infusion pumps: implications for patient safety in Intensive Care Units, International Journal of Medical Informatics, Vol 73, Issues 11-12, November 2004, Pages 771-779, ISSN 1386-5056, DOI: 10.1016/j.ijmedinf.2004.08.002. (http://www.sciencedirect.com/science/article/B6T7S-4DFNFCM-3/2/85d694477cdfebd993e4b1b426859bc3)

[5] Jacobsson, I. (2008). Personal communication, interview in 2008. Experienced language teacher and tutor for pupils in need of extra support in a K-6 school (mainly working at K-1 levels).

[6] Khaled, R., Fischer, R., Noble, J., and Biddle, R. A Qualitative Study of Culture and Persuasion in a Smoking Cessation Game, In the *Proceedings of the Third International Conference on Persuasive Technology for Human Well-Being*, PERSUASIVE 2008, 2008.

[7] Kientz, J. A., Choe, E. K., Birch, B., Maharaj, R., Fonville, A., Glasson, C. and Mundt, J. (2010) Heuristic evaluation of persuasive health technologies. In Proceedings of the *1st ACM International Health Informatics Symposium* (IHI '10), Veinot, T. (Ed.). ACM, New York, NY, USA, 555-564.

[8] Midden, C. and Ham, J. (2009) Using negative and positive social feedback from a robotic agent to save energy. *Persuasive '09*: Proceedings of the 4th International Conference on Persuasive Technology, April 2009

[9] McCalley, T., Kaiser, F., Midden, C., Keser, M., Teunissen, M. (2006) Persuasive Appliances: Goal Priming and Behavioral Response to Product-Integrated Energy Feedback. In Proceedings of PERSUASIVE 2006: 45-49.

[10] Nielsen, J. (1992) Finding usability problems though heuristic evaluation. Proceedings of CHI'92. May 3-7 1992. pp373-380.

[11] Nielsen, J. (1994a) Usability Inspection Methods, Proceedings of CHI ’94, Boston, Massachusetts ,USA, April 24-28,1994.

[12] Nielsen, J. (1994b) Enhancing the explanatory power of Usability Heuristics. Proceedings of CHI ’94, Boston, Massachusetts ,USA, April 24-28,1994.

[13] Nielsen, J. (2005) Heuristic Evaluation. Online at http://www.useit.com/papers/heuristic. Accessed feb 9 2011.

[14] Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. Procedings of CHI '90.

[15] Oinas-Kukkonen Harri & Harjumaa Marja. 2008. A Systematic Framework for Designing and Evaluating Persuasive Systems. In *Proceedings of Persuasive Technology: Third International Conference*, pp. 164-176.

[16] Sutcliffe, A. (2001) Heuristic Evaluation of Website Attractiveness and Usability in *Interactive Systems: Design, Specification, and Verification*, proceedings of 8th International Workshop, DSV-IS 2001 Glasgow, Scotland, UK, June 13–15, 2001. Springer. ISSN 0302-9743 (Print) 1611-3349 (Online). Pages 183-198.

[17] Torning, K. and Oinas-Kukkonen, H. 2009. Persuasive system design: state of the art and future directions. In *Proceedings of the 4th International Conference on Persuasive Technology* (Persuasive '09). ACM, New York, NY, USA.