

Towards Usability Evaluation of Interactive Theorem Provers*

Bernhard Beckert

Sarah Grebing

Florian Böhl

Karlsruhe Institute of Technology (KIT)

{beckert, sarah.grebing, florian.boehl}@kit.edu

Abstract: The effectiveness of interactive theorem provers (ITPs) has increased in a way that the bottleneck in the interactive process shifted from effectiveness to efficiency. Proving large theorems still needs a lot of effort for the user interacting with the system. This issue is recognized by the ITP-communities and improvements are being developed. However, in contrast to properties like soundness or completeness, where rigorous methods are applied to provide evidence, the evidence for a better usability is lacking in many cases. Our contribution is the application of methods from the human-computer-interaction (HCI) field to ITPs. We report on the application of focus groups to evaluate the usability of Isabelle/HOL and the KeY system. We apply usability evaluation methods in order to a) detect usability issues in the interaction between ITPs and their users, and b) to analyze whether methods such as focus groups are applicable to the field of ITP.

1 Introduction

Motivation. The degree of automation of interactive theorem provers (ITPs) has increased to a point where complex theorems over large formalizations of real-world problems can be proven effectively. But even with a high degree of automation, user interaction is still required on different levels. On a global level, users have to find the right formalization and have to decompose the proof task by finding useful lemmas. On a local level, when automatic proof search for a lemma fails, they have to either direct proof search or understand why no proof can be constructed and fix the lemma or the underlying formalization. As the degree of automation increases, the number of interactions decreases. However, the remaining interactions get more complex as ITPs are applied to more complex problems. Thus, the time has come to shift the focus from the effectiveness of automated proof search to the efficiency of user interaction by increasing usability of ITPs and providing a better user experience. Soundness of ITPs can be formally proven and the effectiveness of automated proof search can, e.g., be measured with benchmarks. But, in the area of usability of ITPs, objective and reproducible experiments are rare and often replaced by anecdotal evidence or hand-waving. Here, we report on two experiments applying the focus group method to two different ITPs: The tactical theorem prover Isabelle/HOL [11] and the interactive program verification system KeY [3]. The focus group method is a structured group discussion guided by a moderator. The main goal of our experiments was twofold: Firstly, on the “meta-level,” we wanted to see if focus groups can be used to evaluate the usability of ITPs and what impact the specific characteristics of ITPs have on the setup and the results of focus groups. Secondly, on the concrete level, our aim was to compare the two ITP systems w.r.t. their usability.

Related Work There have been some attempts to address the usability of theorem provers with structured usability evaluations, e.g., applying questionnaire based methods [12, 4, 2, 7], qualitative methods such as co-operative evaluations [6] or analyzing recordings of user errors [1].

2 Evaluation

Evaluation Method. Focus group discussions are a qualitative method to explore opinions of users about specific topics or products, e.g., in market research. In the field of human-computer interaction (HCI) they are used to explore user perspectives on software systems and their usability in an early stage of the usability engineering process [5, 10]. Based on their results, (prototypical) mechanisms for improving usability can be developed, which can then be evaluated with methods such as usability testing and user questionnaires to quantitatively measure increases in usability. While focus groups explore the subjective experience of users, they are designed to eliminate experimenter’s bias and to provide more objective results. The number of participants required (five to ten participants) to get significant results is much smaller than for quantitative evaluations, which makes focus groups well-suited for the relatively small user base of ITPs. The duration of the discussion groups is around one to two hours and it is guided by a moderator who uses a script to structure the discussion. Focus groups have three phases: Recruiting participants, performing the discussion and post-processing. In the following we will briefly describe how we conducted the focus groups for Isabelle/HOL and the KeY system.

Participants. The participants, mostly Master or PhD students, were recruited using personal contacts. We ensured that each group included novice, intermediate, and expert users in different proportions. The KeY group had seven and the Isabelle group five participants.

Script for the discussions. The main questions and tasks in the script were the same for both discussions as we

*The work presented here is part of the project Usability of Software Verification Systems within the BMBF-funded programme Software Campus.

wanted to compare the results. Adaptations of the questions and mock-ups to the specifics of the two systems were the main differences. The full scripts for our experiments are available at formal.iti.kit.edu/grebing/SWC. As warm-up task, we asked about typical application areas of the systems and about their strengths and weaknesses related to the proof process. In the main part of the discussion, we had two topics: (1) Support during the proof process and (2) Mechanisms for understanding proof states. For the cool-down task, we asked the participants to be creative and imagine their ideal interactive proof system.

Topic 1: Support during the proof process. In this topic we address the question “Does the tool give sufficient support during the proof process?”. We divided the discussion for this topic into two parts, namely the global proof process (finding the right formalization and decomposing the proof task) and the local proof process (proving a single lemma or theorem). For each part, participants were asked to describe the typical proof process and discuss where the prover gives support and where support is missing. Time-consuming actions should be pointed out as well.

Topic 2: Mechanisms for understanding proof states. For the second topic, we initiated a more focused discussion by presenting mock-ups for mechanisms not yet built into the tools. This included (a) a mechanism for tracing formulas, terms, and variables that are generated during proof construction back to the original proof goal (for both tools), (b) a visual support for proof management that shows which lemmas contribute to a proof (for Isabelle), and (c) a mechanism for highlighting local changes between two adjacent nodes in the proof tree (for KeY). Thus, we used focus groups to get a first assessment of new features.

For all presented mechanisms we had the same line of action and questions. First, the participants were asked to describe what they think the mechanism does (i.e., the mechanism was not explained by the moderator). This was done to avoid bias introduced by the moderator and to see if the mechanism is intuitive. Then, the participants should express opinion on the usefulness of the mechanism. The planned duration for both groups was 2 hours. Due to lively discussions, the actual duration was 2.5 resp. 3 hours.

Analysis and Results We transcribed the recorded discussion to use Qualitative Content Analysis [9] to analyze and structure the results to draw conclusions from the evaluation. We gained insight into strengths and weaknesses of the two systems, which mostly can be generalized for ITPs, e.g., missing comprehension about the automatic strategies of the tools. Also technical issues, which are annoying for the user and in their opinion compromising for the efficiency, were mentioned, e.g., unstable loading mechanisms or a slow user interface. These point out where the systems could be improved in particular. By showing mock-ups of improvements we gained lively feedback and opinions about the presented mechanisms, allowing us to improve our mechanisms and prototypically implement them in the future. The full details of the presented mechanisms, the evaluation and the results will be presented in the poster.

3 Conclusion and Future Work

Our experiments show, that focus groups can be used to get insight into ITPs and explore where to improve the systems in order to ease the interactive verification process for the user. The first results already show, that focus groups can help to determine which mechanisms might or might not satisfy the needs of the users and how to improve the presented mechanisms. A full analysis and interpretation of the recorded and transcribed material is currently being done. This will result in a detailed report on desirable features for interactive theorem provers. The mechanisms that attracted interest during the discussions need to be further developed and prototypically implemented. By using usability testing we ensure that the mechanisms suit the user’s needs and evaluate the impact on the usability of the systems. We will apply the User Experience Questionnaire method [8] to assess the usability of the KeY system quantitatively. Here, we will determine, whether such general-purpose questionnaires are helpful for evaluating the usability of ITPs, or whether more adaptable solutions are needed.

References

- [1] J. S. Aitken and T. F. Melham. An analysis of errors in interactive proof attempts. *Interacting with Computers*, 12(6):565–586, 2000.
- [2] B. Beckert and S. Grebing. Evaluating the usability of interactive verification system. In *Proceedings of COMPARE 2012*, CEUR Workshop Proceedings 873, 2012.
- [3] B. Beckert, R. Hähnle, and P. H. Schmitt, editors. *Verification of Object-Oriented Software: The KeY Approach*. LNCS 4334. Springer-Verlag, 2007.
- [4] J. Cheney. Project report – theorem prover usability. Technical report, 2011.
- [5] X. Ferré, N. J. Juzgado, H. Windl, and L. L. Constantine. Usability basics for software developers. *IEEE Software*, 18(1):22–29, 2001.
- [6] M. Jackson, A. Ireland, and G. Reid. Interactive proof critics. *Formal Aspects of Computing*, 11(3):302–325, 1999.
- [7] G. Kadoda, R. Stone, and D. Diaper. Desirable features of educational theorem provers: A Cognitive Dimensions viewpoint. In *Proceedings of the 11th Annual Workshop of the Psychology of Programming Interest Group*, 1996.
- [8] B. Laugwitz, T. Held, and M. Schrepp. Construction and evaluation of a user experience questionnaire. In A. Holzinger, editor, *HCI and Usability for Education and Work*, LNCS 5298, pages 63–76. Springer, 2008.
- [9] P. Mayring. *Einführung in die qualitative Sozialforschung – Eine Anleitung zu qualitativem Denken (Introduction to qualitative social research)*. Psychologie Verl. Union, 1996.
- [10] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, 1993.
- [11] T. Nipkow, L. C. Paulson, and M. Wenzel. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*. LNCS 2283. Springer, 2002.
- [12] V. Vujosevic and G. Eleftherakis. Improving formal methods’ tools usability. In *2nd South-East European Workshop on Formal Methods*. South-East European Research Centre (SEERC), 2006.