The Concrete Evonne: Visualization Meets Concrete Domain Reasoning (User Studies Report)

Christian Alrabbaa[1], Franz Baader[1], Raimund Dachselt[2], Alisa Kovtunova[1], and Julián Méndez[2]

[1] Institute of Theoretical Computer Science, TU Dresden, Germany
[2] Interactive Media Lab Dresden, TU Dresden, Germany
{first name.last name}@tu-dresden.de, julian.mendez2@tu-dresden.de

1 Overview

Evonne[3] is a web application primarily designed to explain Description Logic (DL) entailments using an interactive visualization approach for proofs. Recently, an extension of Evonne to DLs with concrete domains (CDs) was introduced, enabling the formalization of concepts whose definitions involve quantitative information. Specifically, there are two extensions of the DL $\mathcal{EL}_\bot$: one with constraints formulated as linear equations ($\mathcal{D}_{\mathbb{Q},lin}$) and the other with difference constraints ($\mathcal{D}_{\mathbb{Q},diff}$). To assess the concrete domain explanations produced by Evonne we conducted two qualitative studies — one for $\mathcal{D}_{\mathbb{Q},diff}$ and the other for $\mathcal{D}_{\mathbb{Q},lin}$— using online structured interviews. This document provides a comprehensive report of these user studies, detailing the methodology, charts, key findings, and participant feedback.

Both studies followed an identical design — see Section 2 — and compared the classical proof trees (e.g., Figure 1) with their respective alternative CD explanation (e.g., Figures 2, 3). The studies was pre-registered [2], fully recorded after informed consent, and anonymized. The goal of our studies was to compare the effectiveness of our explanations and collect feedback to improve them. We present the results for $\mathcal{D}_{\mathbb{Q},lin}$ in Section 4, for $\mathcal{D}_{\mathbb{Q},diff}$ in Section 5, and dedicate the final section, Section 6, to participant feedback by including representative quotes.

2 Study Design

We employed a 2x2 factorial design with two independent variables: **representation** (i.e., *plot/cycle* or *proof*), and **task type** (i.e., identifying *unsatisfiability* or veryfying an *entailment*). The task type condition was necessary due to slight differences in plot and cycle representations between cases. Within each domain, four visual explanations of comparable difficulty were created with Evonne.

[3] The source code, proof examples, and further resources for Evonne are available at https://imld.de/evonne

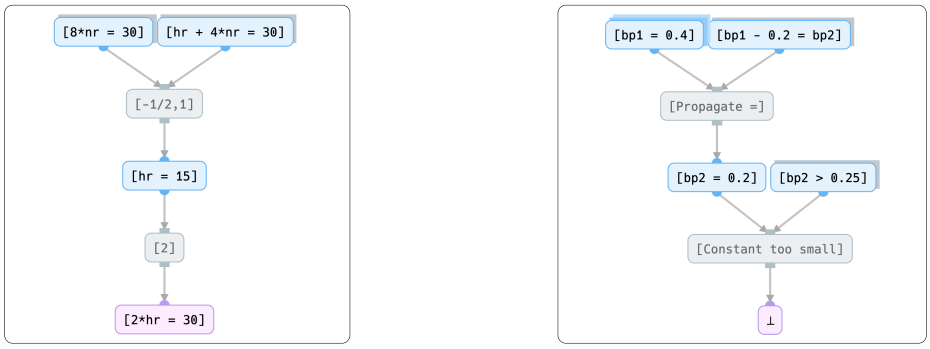
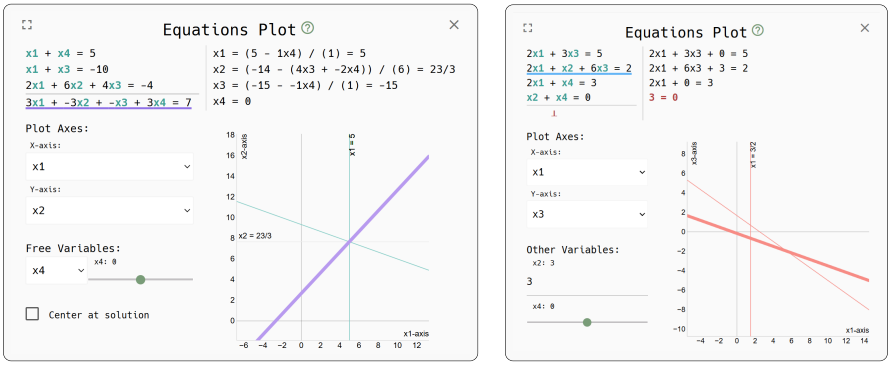


Figure 1: Examples $\mathcal{D}_{\mathbb{Q},lin}$ (left) and $\mathcal{D}_{\mathbb{Q},diff}$ (right) tree proofs in Evonne



(a) Implication of $3x_1 - 3x_2 - x_3 + 3x_4 = 7$ (b) Implication of $\bot$

Figure 2: Examples of explanations for $\mathcal{D}_{\mathbb{Q},lin}$ implications in Evonne

Each task in $\mathcal{D}_{\mathbb{Q},lin}$ involved 3-4 linear equations and, for $\mathcal{D}_{\mathbb{Q},diff}$, 4-6 difference constraints. We used a within-subjects design with a randomized order: all participants experienced all four condition combinations across four different explanations. The dependent variables include **subjective preference**, **ease of use** (measured using SEQ), and **user experience** (assessed with UEQ-S). Below, we provide a brief overview of the two latter measures.

The Single Ease Question (SEQ) [4] is a 7-point rating scale from 1 "very difficult" to 7 "very easy" to assess how difficult users find a task. The ratings of difficulty measured by SEQ are reliable and accurate [4]; they correlate with other ones like task-time and task-completion [5]. SEQ is also effective in competitive settings. It was administered immediately after a participant attempted to understand an explanation. Additionally, we collected immediate diagnostic information by asking users to briefly describe why they found the task difficult.

The User Experience Questionnaire (UEQ) [3] is a fast and reliable questionnaire to measure the user experience of interactive products. The short version of

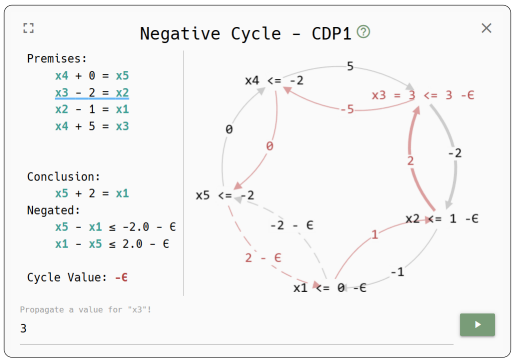


Figure 3: Example of an explanation for $\mathcal{D}_{\mathbb{Q},diff}$ implications in Evonne.

UEQ, UEQ-S [8], consists of 8 items: 4 of these items represent pragmatic quality and 4 hedonic quality aspects. Pragmatic usability focuses on the task-oriented nature of an experience, whereas hedonic usability reflects non-utilitarian aspects such as the appeal, originality, and joy-of-use. Calculated according to the handbook [6], values between -0.8 and 0.8 represent a neural evaluation of the corresponding scale and values above 0.8 represent a positive evaluation. In addition, items that belong to the same scale should show in general a high correlation. In our surveys, across both domains, the Cronbach-Alpha value [7], a measure for the consistency of a scale, is greater than 0.63, which is considered sufficient.

The dependent variables — subjective preferences, ease of use and user experience — are interrelated. As demonstrated in Sections 4-5, subjective preferences consistently aligned with the results from SEQ and UEQ-S, reinforcing the validity of our findings.

The online surveys were hosted through a LimeSurvey instance for Saxony universities [1]. If the $\mathcal{D}_{\mathbb{Q},lin}$ study took an average of 50 minutes, the $\mathcal{D}_{\mathbb{Q},diff}$ study took only an average of 35 minutes. In each study we included an introductory video explaining the structures of respective explanation representations.

3 Participants

The two studies included same 11 participants (9 males and 2 females), with 9 aged 25–34, two aged 35–44, and one aged 18–24. In terms of education, 7 participants held a Doctoral degree, 3 had a Masters degree, and 1 had a Bachelors degree. When self-assessing their experience with logic on a scale from 1 “no knowledge at all/ no experience” to 5 “expert/ a lot of experience”, 4 participants rated themselves as 4, while the remaining 7 rated themselves as 5. Screening criteria also included consent to record screen and voice during the experiment. Participants were informed about the studies objective and consented to anonymous data use for scientific purposes.

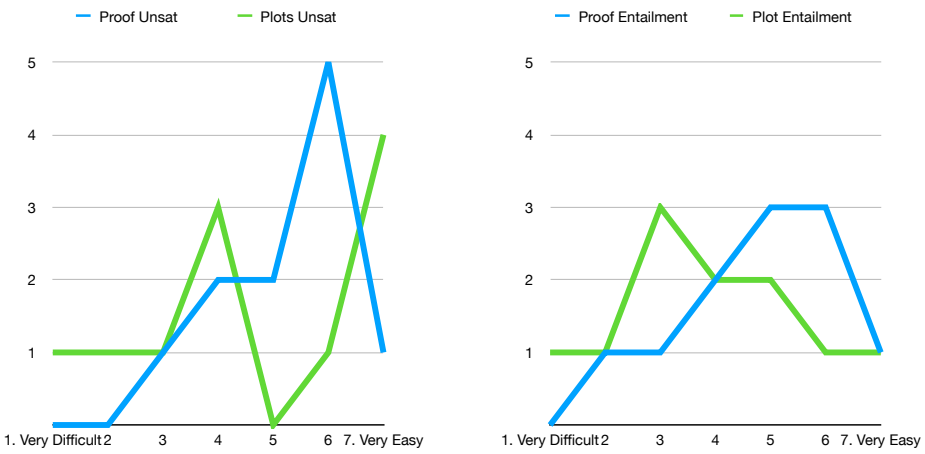


Figure 4: Distribution of Participant Responses to the Single Ease Question in $\mathcal{D}_{\mathbb{Q},lin}$: X-axis represents how difficult was to understand an explanation, and Y-axis indicates the number of participants who selected each class.

4 Results for $\mathcal{D}_{\mathbb{Q},lin}$

This section contains a detailed report, charts and key findings regarding the linear equations domain study.

Regarding SEQ, Figure 4 demonstrates that, across both task types — *unsatisfiability* and consistent *entailment* — *plots* (with average scores of 4.7, SD = 2.2, and 3.9, SD = 1.8, respectively) were perceived as less intuitive than *proofs* (with average scores of 5.3, SD = 1.2, and 4.8, SD = 1.5, respectively). For *plots* participants found it cognitively demanding to combine and interpret multiple variables displayed simultaneously, particularly when dealing with higher dimensions. Distinguishing between single solution or sets of solutions required additional mental effort. Some participants felt uncertain about the system and attempted to verify the correctness of visualizations across projections, which added to the complexity of understanding. An *unsatisfiability* task is perceived little easier than an *entailment* one. Over time, familiarity with the visualization improved, aided by provided instructions, added to the learning curve. For *proof trees* some participants found the step-by-step approach helpful, as it made the task easier to follow and verify compared to *plots*. They appreciated the ability to trace each step and felt confident in the overall result due to its clear mathematical foundation. However, the mental calculations required for verifying were quite challenging. Many participants found the notation less accessible, citing issues such as repetitive left-hand sides of axioms and overly lengthy node labels.

In addition, after each task, we asked participants whether the explanation service – *plots* or *tree proofs* – was useful for understanding. For *tree proofs*, the

response was unanimously positive across both task types. For *plots*, however, the feedback was more varied: 10 participants answered "yes" and 1 was "not sure" for *unsatisfiability*, while for *entailment*, 6 answered "yes", 4 were "not sure", and 1 responded "no".

We asked participants to indicate their subjective preference at two stages: first, based on the theoretical concepts introduced, and second, after observing the implementation. Immediately after the training video and before presenting any tasks, the responses were evenly split, with 5 participants favoring *plots*, 5 favoring *proof trees*, and 1 expressing a preference for both. After seeing Evonne most participants preferred *proofs* (8 vs. 2 for *plots*, 1 for both). However, when linear equations involved only 2-3 variables, preference shifted to *plots* (8 vs. 3 for *proofs*). Three participants noted that the plot examples in the video were easier than those in the tasks, which did not align with their initial expectation.

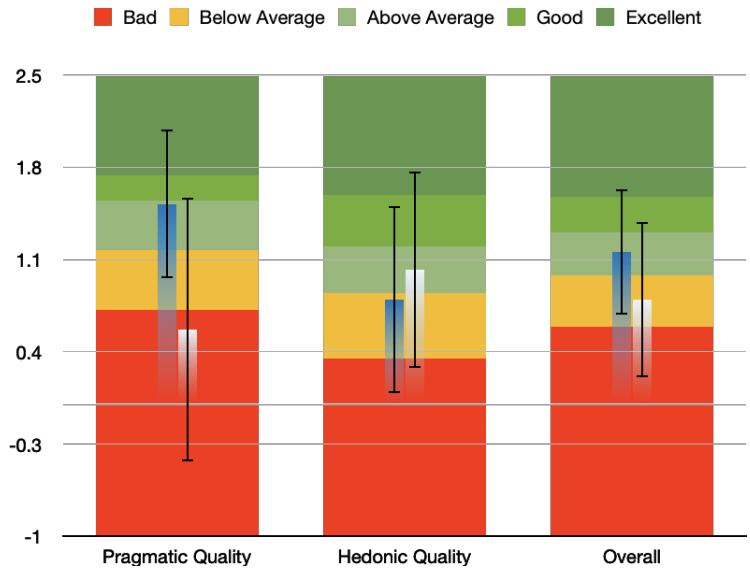


Figure 5: Proof (in blue) and Plot (in white) Quality Means, Confidence Intervals, and the Interpretation of UEQ Scores.

With respect to user experience, Figure 5 illustrates the following findings. *Proofs* received an overall rating of "above average" (1.159). Specifically, they scored "above average" (1.523) for pragmatic qualities (e.g., usability and functionality) but "below average" (0.795) for hedonic qualities (e.g., enjoyment and stimulation). In contrast, *plots* were rated overall as "below average" (0.795). They scored "bad" (0.568) for pragmatic qualities but achieved an "above average" (1.023) rating for hedonic qualities.

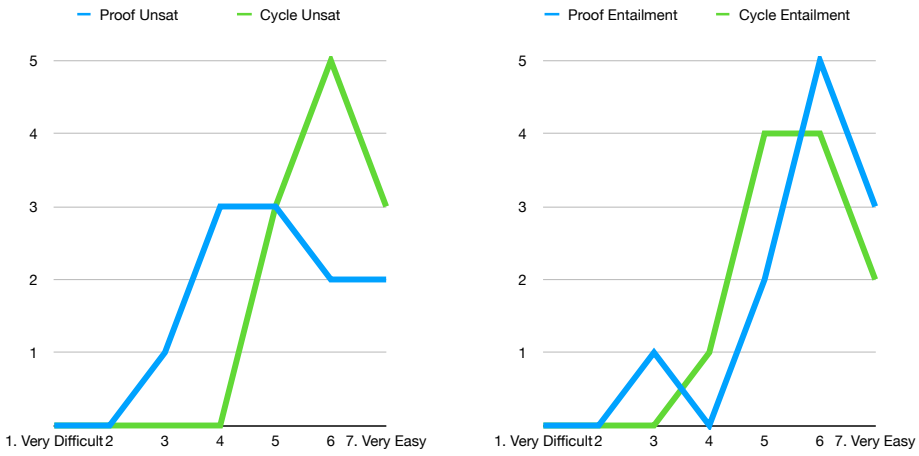


Figure 6: Distribution of Participant Responses to the Single Ease Question in $\mathcal{D}_{\mathbb{Q},diff}$: X-axis represents how difficult was to understand an explanation, and Y-axis indicates the number of participants who selected each class.

All things considered, while *plots* offered valuable insights and they are more enjoyable, they demanded significant cognitive effort and clarity in design to reduce confusion and improve interpretability. Similarly, the *proof tree*'s structured approach aided understanding and verification, but it also highlighted the need for clearer explanations and more intuitive representations.

5 Results for $\mathcal{D}_{\mathbb{Q},diff}$

In this section we provide a detailed analysis of the difference constraint study, featuring charts and highlighting the main results.

With respect to SEQ, Figure 6 demonstrates that for a consistent *entailment* task, *cycles* and *proofs* received similar evaluations (*cycles*: average 5.6, SD = 0.9; *proofs*: average 5.8, SD = 1.2, respectively). However, for *unsatisfiability*, *cycles* (with average score of 6, SD = 0.8) were unanimously perceived as more intuitive than *proofs* (average 5, SD = 1.3). Participants generally found *cycles* and their animations clear and effective for understanding, particularly due to their ability to automate computations, highlight negative cycles, and verify edges. These features made the process intuitive and significantly reduced cognitive effort. However, a few participants noted that understanding the interface and parsing the information required some initial effort, explanations, and practice. In contrast, *proof trees* were perceived slightly less positively. While participants appreciated their ability to track reasoning and clearly indicate which equations to combine, as well as the simplicity of verifying steps due to the absence of coefficients, some challenges were noted. These included sparse information (e.g., repetitive left-hand sides of axioms and overly lengthy node

labels), the need for manual calculations, and difficulties with the semantics of edge operations. Many participants suggested that explanations of rules and clearer naming conventions would greatly improve usability. Opinions varied on the ease of interpreting *proofs* versus *cycles*, with some favoring *proof trees* for inference clarity — especially for understanding consistent *entailments* — while others found the visual representation of *cycles* convincing and well-explained.

After each task, participants were asked whether the explanation service — *plots* or tree *proofs* — was useful for understanding. All but one participant responded "yes" across both task types and both representations. The outlier cited confusion related to the naming of *proof* edges, which specific for $\mathcal{D}_{\mathbb{Q},diff}$, such as "constant too small" and "sum of differences".

Participants were asked to express their **subjective preferenc**e at two stages: first, after learning the theoretical concepts, and second, after using Evonne. Initially, immediately following the training video and before any tasks were presented, all but one participant favoured *cycles*. In the post-test assessment for $\mathcal{D}_{\mathbb{Q},diff}$, this preference became less pronounced: fewer participants preferred *cycles* (8 vs. 2 for *proofs*, with 1 participant favouring both). However, the preference for *cycles* grew slightly stronger as the number of difference constraints increased (9 vs. 2 for *proofs*).

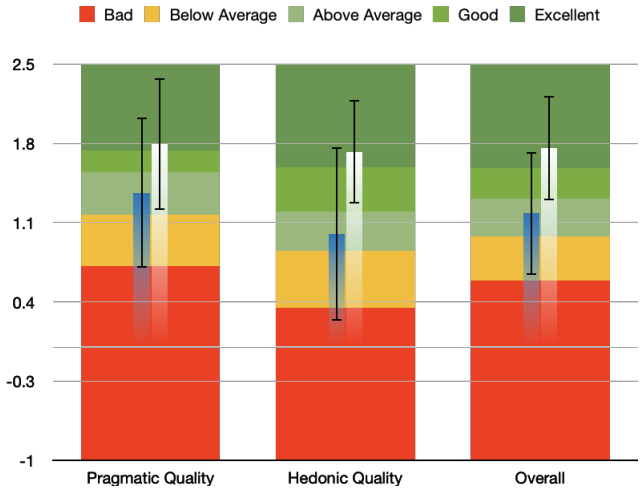


Figure 7: Proof (in blue) and Cycle (in white) Means, Confidence Intervals, and the Interpretation of UEQ Scores.

With respect to user experience, Figure 7 highlights the following results. *Proofs* received an "above average" rating across all qualities, with scores of 1.364 (pragmatic quality, e.g., usability and functionality), 1 (hedonic quality, e.g., enjoyment and stimulation), and 1.182 (overall quality). In contrast, *cycles*

were rated as "excellent" in all three categories, with scores of 1.795 (pragmatic), 1.727 (hedonic), and 1.761 (overall).

Overall, *cycles* was praised for its intuitiveness, clarity, comprehensiveness and enjoyment, though some felt that fully grasping the explanation still required active engagement. The *proof tree* was praised for its clarity in tracking reasoning steps, though the representation could benefit from clearer guidance and labelling.

6 Suggestions for Improvement

This section includes some participant quotes highlighting their main struggles and suggestions for improvement.

6.1 Linear Equations

Participant 1: "The plot service for me personally was difficult to connect to my way of thinking. Generally, I think both services would benefit from presenting the equations below each other, rather than next to each other (this was done in the plot service), but then also by organising them by variable, so that the same variable is always in the same column, more like in a matrix – then it is easier to see what related to what. In the proof view, I would have liked some help in leading my attention to the thing I have to look at – I actually used my fingers for that. In particular: the left hand side of the implication is never important, but takes a lot of space and is difficult to distinguish from the right. A nice idea could be to also highlight the variable that is being eliminated in each step - I think then this would be a super cool way of verifying what happens!"

Participant 2: "So for the plots I really liked that I can look at all the possible configurations. But I did not like that kind of this one confusion where it switched between the highlighted plot. There are too many configurations to look at, so like. I have to freely choose between several parameters like which dimensions I want to see and kind of move the free variables around. So maybe it would if the system could somehow show me some of the necessary ones if that makes sense like the ones that really con that are really convincing, like a few example configurations that are most relevant for the solution. Or like some more guidance, or maybe an animation where it switches through the different configurations, so that I don't have to do all the clicking myself."

Participant 3: "Small tutorial boxes on the plot service would be helpful."

Participant 4: "Equation plots are not intuitive: values for variables that are not used as axis could be changed sometimes for some variables and sometimes for all variables; confusing zooming."

Participant 5: "Slider not super precise, proof tree takes a lot of space. As a suggestion there could be a tunable number of projections."

Participant 6: "For trees, I like that they show each individual step, but I did not figure out the solution myself from trees, instead I trusted the system. For the equation plots, I like that I can see what it means when equations are inconsistent (parallel lines), but I sometimes can't fully grasp the exact solution from the graph. It may be helpful to somehow describe what kind of geometric object the solution is (point, line, plane etc)."

6.2 Difference Constraints

Participant 1: "Proofs take a lot of space, and the edge labelling are unfamiliar. To improve cycles, thicker lines/more contrast colors are needed."

Participant 2: "I don't like that in the proof view, the left-hand side is always very long. It would be much easier if this could somehow be shortened. To improve the negative cycle service: in case of a strong inequality, Evonne can also show the original formula and not only the one with the epsilon."

Participant 3: "Regarding cycles, I have a large screen, it is not my immediate reaction to look to the lower left corner. Maybe the explanation should pop over closer to the node. Animation and cycle values are beneficial. For proof trees: by clicking the labels I was not able to see how the rules work."

Participant 4: "A positive implication is actually very well represented in a proof tree. Cycle highlighting, hovering and animation are super cool. Disadvantages: tree proof rule naming and rule explanation; second negative cycle should be highlighted."

Participant 5: "Cycle animation starts too early, I could not verify the graph first. Proofs are easy to follow but not interactive. They have not intuitive naming; more explanation for edges is needed."

Participant 6: "for unsatisfiability in cycles: I want to see 2 cycles, but it only showed me one cycle. That could be showing both cycles with different colors. With the proof tree, the right hand side of the equations is important; this could be an idea to color it in another way. So it's immediately clear where you have to look at."

References

1. Der Umfragedienst für sächsische Hochschulen und Berufsakademien, `https://bildungsportal.sachsen.de/umfragen/`
2. Kovtunova, A., Alrabbaa, C., Baader, F., Dachselt, R., Méndez, J.: The Concrete Evonne (Feb 2025). `https://doi.org/10.17605/OSF.IO/Y4X5T`

3. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: HCI and Usability for Education and Work, 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings. Lecture Notes in Computer Science, vol. 5298, pp. 63–76. Springer (2008). https://doi.org/10.1007/978-3-540-89350-9_6
4. Sauro, J., Dumas, J.S.: Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009. pp. 1599–1608. ACM (2009). https://doi.org/10.1145/1518701.1518946
5. Sauro, J., Lewis, J.R.: Correlations among prototypical usability metrics: evidence for the construct of usability. In: Jr., D.R.O., Arthur, R.B., Hinckley, K., Morris, M.R., Hudson, S.E., Greenberg, S. (eds.) Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009. pp. 1609–1618. ACM (2009). https://doi.org/10.1145/1518701.1518947, https://doi.org/10.1145/1518701.1518947
6. Schrepp, M.: User Experience Questionnaire Handbook (2015). https://doi.org/10.13140/RG.2.1.2815.0245
7. Schrepp, M.: On the Usage of Cronbachs Alpha to Measure Reliability of UX Scales, vol. 15, p. 247258. Usability Professionals' Association (08 2020)
8. Schrepp, M., Hinderks, A., Thomaschewski, J.: Design and evaluation of a short version of the user experience questionnaire (UEQ-S). Int. J. Interact. Multim. Artif. Intell. **4**(6), 103–108 (2017). https://doi.org/10.9781/IJIMAI.2017.09.001