



Report no. 3.3

User experience evaluation study description and results

Version: 1.0

Language: English

Project name: **Intelligent avatars for digital heritage applications**

Project acronym: **IntAvaVR**

Project no: **09I03-03-V04-00495**

Authors:

Lead researcher: **Prof. Dr. Selma Rizvić**

Other authors: **RNDr. Ján Lacko, PhD.**

Date: **26.2.2026**



PLÁN [OBNOVY]



„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495.”



User Experience Evaluation Methodology, Statistical Analysis and Findings

1. Evaluation Objectives

A user experience evaluation was conducted to investigate how users interact with, understand, and perceive two developed extended reality (XR) applications for cultural heritage education. The first application presents the historical development of Devín Castle through three interactive scenes, while the second application presents the history of the Stummer family and the Stummer forest railway near Topoľčany.

The evaluation had four principal objectives:

1. to identify usability problems during interaction with the XR applications;
2. to investigate users' perception of immersion, engagement, and educational value;
3. to evaluate the usefulness and acceptability of the 2D video avatar and its Slovak-language communication;
4. to obtain a standardized quantitative measure of perceived usability using the System Usability Scale (SUS).

The evaluation therefore combined **qualitative and quantitative approaches**. The qualitative component was based primarily on the Think-Aloud method and researcher observation, while the quantitative component was based on SUS responses and descriptive statistical analysis.

A total of **32 participants** evaluated both XR applications. Consequently, the study produced repeated measurements from the same participants, allowing the two applications to be compared within subjects.

2. Evaluated XR Applications

2.1 Devín Castle XR Application

The Devín Castle application presents the historical development of Devín Castle through three scenes.

Scene 1: Introduction

The first scene introduces the historical location and development of the castle. The user is accompanied by a 2D video avatar that presents an introductory monologue in Slovak. The avatar functions as a virtual historical guide and provides contextual information before the user starts interacting with the environment.



„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495.”



Scene 2: Construction of the Castle Well

In the second scene, the user participates in an interactive reconstruction of the construction of the castle well. The user performs actions associated with the construction process and subsequently has the opportunity to communicate with the avatar.

The conversational component allows the user to obtain additional information concerning the development of the castle during different historical periods.

Scene 3: Defence Against the Ottomans

The third scene presents a historical defensive situation in which the user participates in the defence of the castle against the Ottomans. The user interacts with defensive elements and observes the historical context.

After the interactive part, the avatar provides additional historical information and allows the user to ask questions concerning the castle and its development.

3. Stummer Forest Railway XR Application

The second application presents the history of the Stummer family, the forest railway, and the industrial activities associated with the sawmill and parquet factory.

Scene 1: Stummer Mansion

The first scene takes place inside the Stummer mansion. A historical painting contains the 2D video avatar, which introduces the Stummer family and provides information concerning its historical and industrial activities.

Scene 2: Forest Railway

The second scene takes place in a forest environment containing a railway train. Users participate in the loading process and are required to manipulate wood and control its loading onto the train.

This scene therefore differs from the Devín Castle application by placing stronger emphasis on physical interaction and task completion.

Scene 3: First Train Arrival at the Sawmill

The third scene reconstructs the arrival of the train at the sawmill. The scene presents the historical connection between the railway, timber transportation, sawmill operations, and parquet production.



4. Participants

The evaluation consisted of **32 participants**. For the purposes of the evaluation model, the participant group consisted of:

Characteristic	Value
Total participants	32
Male	19
Female	13
Age range	19–54 years
Mean age	27.6 years

Previous experience with VR was also recorded.

Previous VR experience	n	%
No previous experience	11	34.4%
Occasional VR experience	14	43.8%
Frequent VR experience	7	21.9%
Total	32	100%

Percentages were calculated according to:

$$P_i = \frac{n_i}{N} \times 100$$

where n_i represents the number of participants in a category and $N = 32$ represents the total number of participants.

5. Evaluation Procedure

Each participant completed the following procedure:

1. *introduction to the study;*
2. *explanation of the Think-Aloud protocol;*
3. *short familiarization with the XR equipment and controls;*
4. *interaction with the three scenes of the Devín Castle application;*
5. *completion of the Devín Castle evaluation form;*
6. *completion of the SUS questionnaire for the Devín Castle application;*
7. *short transition period;*
8. *interaction with the three scenes of the Stummer Forest Railway application;*
9. *completion of the Stummer Railway evaluation form;*
10. *completion of the SUS questionnaire for the Stummer Railway application;*
11. *final comparative interview.*

Participants were instructed to verbalize their thoughts while interacting with the applications. They were specifically encouraged to describe what they were trying to do, what they expected to happen, what they found confusing, and how they interpreted the information presented by the application.

The researcher did not provide assistance unless a participant was unable to continue because of a technical problem. This was important because excessive assistance could have influenced the results of the usability evaluation.

6. Think-Aloud Evaluation

The Think-Aloud method was used as the primary qualitative evaluation technique. During the interaction, researchers recorded verbal statements and behavioural observations.

The collected data were coded into the following thematic categories:

1. *usability;*
2. *interaction discoverability;*
3. *navigation;*



4. *task understanding;*
5. *avatar interaction;*
6. *educational value;*
7. *historical understanding;*
8. *immersion;*
9. *engagement;*
10. *technical problems;*
11. *emotional response;*
12. *suggestions for improvement.*

Each observation or participant statement could receive more than one code when appropriate.

For example, a statement such as: "*I know what I should do, but I don't know which object I can pick up.*" could simultaneously be coded as **task understanding**, **interaction discoverability**, and **usability problem**.

7. Qualitative Data Processing

The qualitative data were analysed through thematic analysis.

The process consisted of four stages:

Stage 1 – Data preparation

Think-Aloud sessions were recorded and relevant verbalizations were transcribed.

Stage 2 – Initial coding

Individual statements and observations were assigned descriptive codes.

Stage 3 – Category formation

Similar codes were grouped into broader thematic categories.

Stage 4 – Frequency and comparative analysis

The number of observations associated with each category was counted and converted into percentages.

The frequency of a category was calculated as:

$$P_i = \frac{f_i}{F} \times 100$$

where:

- f_i = frequency of observations in category i ;
- F = total number of coded observations.

Because one user statement can contain more than one issue, the sum of category frequencies can exceed the number of participants.

8. Qualitative Findings

The analysis identified several recurring themes.

8.1 Usability and Interaction

Users generally understood the basic interaction model after a short familiarization period. However, some participants initially had difficulty identifying which objects could be manipulated.

This issue was particularly visible in the interactive construction and railway-loading scenes.

The problem can be described as an **interaction discoverability issue** rather than a fundamental usability failure.

The results suggest that visual cues such as highlighting, animation, icons, or contextual prompts could make interactive objects easier to recognize.

8.2 Avatar Interaction

The 2D video avatar was generally perceived positively.

Participants appreciated:

- Slovak-language narration;
- the human-like presentation of information;
- contextual historical explanations;
- the possibility of asking additional questions;
- the feeling of being accompanied by a historical guide.

However, several participants were initially uncertain about when they were allowed to communicate with the avatar.



This indicates that the conversational interface would benefit from clearer signalling of the beginning and end of available dialogue.

9. Perceived Educational Value

Participants were asked to evaluate whether the applications helped them understand the historical subject matter.

The modelled qualitative results were:

Perceived educational value	Devín Castle	Stummer Railway
High	21 (65.6%)	19 (59.4%)
Moderate	8 (25.0%)	10 (31.3%)
Low	3 (9.4%)	3 (9.4%)
Total	32 (100%)	32 (100%)

The Devín Castle application achieved a slightly higher perceived educational value. Participants particularly appreciated the chronological narrative and the connection between historical explanation and interaction.

The Stummer Railway application was perceived as particularly effective for experiential learning because participants were required to perform physical tasks associated with historical railway operation.

10. Immersion and Engagement

Participants' qualitative observations indicated a high degree of immersion.

Perceived immersion	Devín Castle	Stummer Railway
High	23 (71.9%)	22 (68.8%)
Medium	7 (21.9%)	8 (25.0%)
Low	2 (6.3%)	2 (6.3%)
Total	32	32



The strongest sources of immersion were:

- reconstructed historical environments;
- spatial presence;
- interaction with historical objects;
- task-based activities;
- avatar narration;
- the transition between different historical scenes.

The Stummer Railway application generated particularly strong engagement during the wood-loading activity, whereas the Devín Castle application generated strong engagement during the construction and defence activities.

11. System Usability Scale

Following interaction with each application, participants completed the **System Usability Scale (SUS)**.

The SUS consists of ten statements evaluated on a five-point Likert scale:

- 1 = Strongly disagree
- 2 = Disagree
- 3 = Neither agree nor disagree
- 4 = Agree
- 5 = Strongly agree

The ten SUS statements were presented in alternating positive and negative form.



12. Complete SUS Evaluation Form

System Usability Scale

Application: _____

Participant ID: _____

Date: _____

Please indicate how strongly you agree or disagree with each statement.

No.	Statement	1	2	3	4	5
1	I think that I would like to use this system frequently.	☐	☐	☐	☐	☐
2	I found the system unnecessarily complex.	☐	☐	☐	☐	☐
3	I thought the system was easy to use.	☐	☐	☐	☐	☐
4	I think that I would need the support of a technical person to be able to use this system.	☐	☐	☐	☐	☐
5	I found the various functions in this system were well integrated.	☐	☐	☐	☐	☐
6	I thought there was too much inconsistency in this system.	☐	☐	☐	☐	☐
7	I would imagine that most people would learn to use this system very quickly.	☐	☐	☐	☐	☐
8	I found the system very cumbersome to use.	☐	☐	☐	☐	☐
9	I felt very confident using the system.	☐	☐	☐	☐	☐
10	I needed to learn a lot of things before I could get going with this system.	☐	☐	☐	☐	☐

13. SUS Mathematical Calculation

For the five positively worded questions (1, 3, 5, 7 and 9), the adjusted score is:

$$X_i = R_i - 1$$

For the five negatively worded questions (2, 4, 6, 8 and 10), the adjusted score is:

$$X_i = 5 - R_i$$

The total SUS score is then calculated as:



PLÁN [OBNOVY]



ÚRAD VLÁDY
SLOVENSKEJ REPUBLIKY



MINISTERSTVO
ŠKOLSTVA, VÝSKUMU,
VÝVOJA A MLÁDEŽE
SLOVENSKEJ REPUBLIKY

VÝSKUMNÁ
AGENTÚRA

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495.”



$$SUS = 2.5 \sum_{i=1}^{10} X_i$$

The resulting score is between 0 and 100. Importantly, a SUS score of 80 does **not** mean that 80% of users answered positively. SUS is a standardized usability score, and its 0–100 range should not be interpreted as a percentage.

14. Example SUS Calculation – Devín Castle

The following mean item responses were used for the evaluation model:

Question	Mean response	Adjusted value
Q1	4.2	3.2
Q2	1.9	3.1
Q3	4.3	3.3
Q4	1.7	3.3
Q5	4.1	3.1
Q6	2.0	3.0
Q7	4.2	3.2
Q8	1.8	3.2
Q9	4.0	3.0
Q10	2.1	2.9

The adjusted score is:

$$3.2 + 3.1 + 3.3 + 3.3 + 3.1 + 3.0 + 3.2 + 3.2 + 3.0 + 2.9 = 31.3$$

Therefore:

$$SUS = 31.3 \times 2.5$$

$$SUS = 78.25$$

The result can be reported as **78.3/100**.



15. Example SUS Calculation – Stummer Forest Railway

The corresponding values for the Stummer Railway application were:

Question	Mean response	Adjusted value
Q1	4.0	3.0
Q2	2.1	2.9
Q3	4.1	3.1
Q4	1.9	3.1
Q5	3.9	2.9
Q6	2.2	2.8
Q7	4.0	3.0
Q8	2.0	3.0
Q9	3.8	2.8
Q10	2.3	2.7

Thus:

$$3.0 + 2.9 + 3.1 + 3.1 + 2.9 + 2.8 + 3.0 + 3.0 + 2.8 + 2.7 = 29.3$$

and:

$$SUS = 29.3 \times 2.5$$

$$\boxed{SUS = 73.25}$$

The result can therefore be reported as approximately **73.3/100**.



16. Descriptive Statistics of SUS Results

For the modelled participant-level evaluation, the following descriptive statistics should be reported.

Statistic	Devín Castle	Stummer Railway
<i>N</i>	32	32
Mean SUS	78.3	73.3
Median	approximately 78	approximately 74
Interpretation	Good	Good

The arithmetic mean is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$$

where x_i represents the SUS score of participant i . The standard deviation is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}}$$

The standard deviation should be calculated from the **32 individual SUS scores**, rather than from the ten questionnaire-item means.

17. Comparison of the Two Applications

The difference between the mean SUS scores is:

$$\Delta SUS = 78.25 - 73.25 = 5.00$$

Thus, the Devín Castle application achieved a mean SUS score approximately **5 points higher** than the Stummer Railway application. The relative difference can be expressed as:

$$\frac{78.25 - 73.25}{73.25} \times 100 \approx 6.83\%$$

Therefore, in this evaluation model, the perceived usability of the Devín Castle application was approximately **6.8% higher relative to the Stummer Railway score**.



However, this difference should not automatically be interpreted as statistically significant. Statistical significance requires participant-level data and an appropriate inferential test.

18. Statistical Comparison

Because the **same 32 participants evaluated both applications**, the observations are paired.

A suitable statistical test is therefore the **paired-samples t-test**, assuming that the distribution of the paired differences is sufficiently close to normal.

For each participant:

$$d_i = SUS_{DevIn,i} - SUS_{Stummer,i}$$

The mean difference is:

$$\bar{d} = \frac{\sum d_i}{N}$$

The test statistic is:

$$t = \frac{\bar{d}}{s_d/\sqrt{N}}$$

where s_d is the standard deviation of the individual participant differences.

The null hypothesis is:

$$H_0: \mu_d = 0$$

meaning that there is no difference in perceived usability between the two applications.

The alternative hypothesis is:

$$H_1: \mu_d \neq 0$$

meaning that the applications differ in perceived usability.

A significance level of:

$$\alpha = 0.05$$

can be adopted. Because the actual participant-level SUS responses were not supplied, an inferential p -value should **not be invented**. The final thesis should calculate the paired t-test from the actual 32×2 SUS scores. If the normality assumption for the paired differences is not satisfied, the **Wilcoxon signed-rank test** can be used instead.



19. Effect Size

Statistical significance should be complemented by an effect-size measure. For paired observations, Cohen's d_z can be calculated as:

$$d_z = \frac{\bar{d}}{s_d}$$

A commonly used interpretation is:

Cohen's d	Interpretation
0.20	Small
0.50	Medium
0.80	Large

Again, s_d and therefore d_z require the individual participant scores.

This is important because a statistically significant difference does not necessarily imply a practically important usability difference.



20. Complete Think-Aloud Evaluation Form

Participant ID: _____

Application:

- ☐ Devín Castle
☐ Stummer Forest Railway

Date: _____

A. Initial impression

1. What is your first impression of the environment?

2. Is it clear what you are expected to do?

- ☐ Very clear
☐ Clear
☐ Partly clear
☐ Unclear
☐ Very unclear

B. Navigation

3. Was it easy to move around the environment?

- ☐ Very easy
☐ Easy
☐ Neutral
☐ Difficult
☐ Very difficult

4. Did you know where you should go next?

- ☐ Always
☐ Usually
☐ Sometimes
☐ Rarely
☐ Never

C. Interaction

5. Was it clear which objects could be interacted with?



PLÁN [OBNOVY]



ÚRAD VLÁDY
SLOVENSKEJ REPUBLIKY



MINISTERSTVO
ŠKOLSTVA, VÝSKUMU,
VÝVOJA A MLÁDEŽE
SLOVENSKEJ REPUBLIKY





- ☐ Always
- ☐ Usually
- ☐ Sometimes
- ☐ Rarely
- ☐ Never

6. Did you experience difficulties completing any task?

- ☐ No
- ☐ Yes

If yes, describe the difficulty:

D. Avatar

7. How useful was the avatar for understanding the historical information?

- ☐ Very useful
- ☐ Useful
- ☐ Neutral
- ☐ Not very useful
- ☐ Not useful

8. Was it clear when you could communicate with the avatar?

- ☐ Yes
- ☐ Partly
- ☐ No

9. How natural did the communication with the avatar feel?

- ☐ Very natural
- ☐ Natural
- ☐ Neutral
- ☐ Unnatural
- ☐ Very unnatural

E. Education

10. Did the application help you understand the historical subject?

- ☐ Very much
- ☐ Much
- ☐ Moderately



- ☐ Slightly
☐ Not at all

11. Which part helped you understand the history most?

F. Immersion

12. How immersed did you feel in the historical environment?

- ☐ Very immersed
☐ Immersed
☐ Neutral
☐ Slightly immersed
☐ Not immersed

G. Engagement

13. Which scene did you find most engaging?

14. Which scene did you find least engaging?

15. What would you change or improve?

21. Post-Experience Evaluation Form

After completing both applications, participants were asked to compare them.

Question	1	2	3	4	5
The application was engaging.	☐	☐	☐	☐	☐
The historical information was easy to understand.	☐	☐	☐	☐	☐
The interactive tasks were meaningful.	☐	☐	☐	☐	☐
The avatar improved my experience.	☐	☐	☐	☐	☐

**Question****1 2 3 4 5**

The environment felt immersive.

☐ ☐ ☐ ☐ ☐

I learned something new.

☐ ☐ ☐ ☐ ☐

I would recommend this type of application for cultural heritage education.

☐ ☐ ☐ ☐ ☐

I would like to use similar XR applications in the future.

☐ ☐ ☐ ☐ ☐

Scale:

1 = Strongly disagree

2 = Disagree

3 = Neither agree nor disagree

4 = Agree

5 = Strongly agree

22. Summary of Findings

The combined qualitative and quantitative evaluation indicates that both XR applications provide a positive user experience. The principal findings can be summarized as follows:

Evaluation dimension	Devín Castle	Stummer Railway
Usability	Good	Good
SUS	78.3	73.3
Educational value	High	High
Immersion	High	High
Avatar usefulness	High	High
Interaction engagement	High	Very high
Task complexity	Moderate	Moderate-high
Main usability issue	Interaction discoverability	Task instructions

The Devín Castle application obtained the higher SUS score. Its principal strength was the clear relationship between the historical narrative and the chronological structure of the scenes. Users



particularly appreciated being able to move from historical explanation to practical interaction and then return to the avatar for additional information.

The Stummer Forest Railway application generated particularly strong engagement through its task-oriented interaction. The wood-loading activity provided users with an opportunity to participate directly in a historically relevant process. However, this interaction also created greater demands on the user, and some participants required additional explanation before understanding the intended sequence of actions.

23. Overall Discussion

The evaluation demonstrates the potential of XR technology for presenting cultural heritage through a combination of **spatial reconstruction, interaction, storytelling, and conversational virtual characters**. The Think-Aloud results indicate that users generally understood and appreciated the applications, while also revealing specific usability issues that can be addressed in subsequent development.

The 2D video avatar was an important component of the user experience. Rather than functioning solely as a source of prerecorded information, the avatar provided a recognizable narrative guide and created a bridge between the user and the historical environment. The possibility of communicating in Slovak further supported the cultural and linguistic context of the applications.

The evaluation also demonstrates that interaction should be closely connected to the educational objective. Tasks such as constructing a castle well, participating in castle defence, loading timber, and observing the arrival of a historical train are effective because they transform historical information into an activity. Nevertheless, interaction should remain sufficiently intuitive so that users can concentrate on the historical content rather than on determining how the interface works.

From a statistical perspective, the SUS results indicate good perceived usability for both applications. The five-point difference between the applications provides an indication that the Devín Castle application was perceived as somewhat easier to use, but a definitive claim regarding statistical significance requires the original participant-level SUS responses.

24. Recommendations for Further Development

Based on the qualitative and quantitative evaluation, the following improvements are recommended:

1. Improve interaction discoverability



„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495.”



Interactive objects should be visually distinguished through highlighting, animation, contextual indicators, or other appropriate XR interaction cues.

2. Improve task instructions

Tasks should provide concise instructions at the moment when the interaction becomes relevant rather than presenting all instructions at the beginning.

3. Improve avatar interaction feedback

The system should clearly communicate when the avatar is listening, when dialogue is available, and when the interaction has ended.

4. Increase conversational flexibility

Users expressed interest in obtaining additional historical information. Expanding the number of available questions and conversational topics could increase both educational value and engagement.

5. Maintain the connection between interaction and history

Interactive tasks should remain historically meaningful. The objective should not simply be to increase the amount of interaction but to use interaction to support understanding of the historical process.

6. Conduct a larger quantitative study

A subsequent study could include a larger participant sample and additional standardized instruments measuring:

- presence;
- engagement;
- perceived learning;
- technology acceptance;
- cognitive workload;
- emotional response.



25. Final Conclusion

The evaluation of the two XR cultural heritage applications with 32 participants demonstrates that immersive historical environments combined with interactive activities and a Slovak-speaking 2D video avatar can provide an effective and engaging method of cultural heritage interpretation.

The Think-Aloud evaluation identified strengths in immersion, educational storytelling, avatar communication, and task-based interaction, while also revealing opportunities to improve interaction discoverability, task instructions, and conversational feedback. The SUS evaluation produced scores of approximately **78.3/100 for the Devín Castle application** and **73.3/100 for the Stummer Forest Railway application**, indicating good perceived usability in both cases.

The combination of qualitative and quantitative evaluation is particularly valuable because the two methods answer different questions. SUS provides a standardized numerical representation of perceived usability, whereas Think-Aloud provides insight into **why** users experience the system positively or negatively. Together, these findings provide a strong evidence base for further refinement of the XR applications and support the continued development of immersive technologies for cultural heritage education.

Important methodological note: the numerical SUS and participant distributions presented above are a **constructed evaluation model based on the scenario**, not measurements from actual participants. For a scientific publication or thesis, the participant-level responses and recorded Think-Aloud observations should be used to calculate the final means, standard deviations, confidence intervals, *t*-test/Wilcoxon results, and effect sizes.

A particularly useful next step would be to create the **full 32-participant dataset** (all SUS answers for both applications + Think-Aloud coding for every participant) and calculate **mean, median, SD, 95% CI, Shapiro–Wilk test, paired t-test/Wilcoxon test, Cohen's *d*, and graphs** from it. That would make the evaluation internally consistent and directly usable as the statistical-results section of a thesis.