IMPACT OF GEOLOCATION DATA ON AUGMENTED REALITY USABILITY: A COMPARATIVE USER TEST

J. Mercier[1,2]*, N. Chabloz[1], G. Dozot[1], C. Audrin[3], O. Ertz[1], E. Bocher[2], D. Rappo[1]

[1] Media Engineering Institute (MEI), School of Engineering and Management Vaud, HES-SO, Yverdon-les-Bains, Switzerland - (julien.mercier, nicolas.chabloz, gregory.dozot, olivier.ertz, daniel.rappo)@heig-vd.ch
[2] Lab-STICC, UMR 6285, CNRS, Université Bretagne Sud, Vannes, France - julien.mercier@univ-ubs.fr, erwan.bocher@cnrs.fr
[3] University of Teacher Education, HES-SO, Lausanne, Switzerland - catherine.audrin@hepl.ch

KEY WORDS: Location-Based Augmented Reality, Usability, Exploration, Focus, Comparative user test, Free Open Source Web AR Application, Cartographic Authoring Tool.

ABSTRACT:

While the use of location-based augmented reality (AR) for education has demonstrated benefits on participants' motivation, engagement, and on their physical activity, geolocation data inaccuracy causes augmented objects to jitter or drift, which is a factor in downgrading user experience. We developed a free and open source web AR application and conducted a comparative user test (n = 54) in order to assess the impact of geolocation data on usability, exploration, and focus. A control group explored biodiversity in nature using the system in combination with embedded GNSS data, and an experimental group used an external module for RTK data. During the test, eye tracking data, geolocated traces, and in-app user-triggered events were recorded. Participants answered usability questionnaires (SUS, UEQ, HARUS). We found that the geolocation data the RTK group was exposed to was less accurate in average than that of the control group. The RTK group reported lower usability scores on all scales, of which 5 out of 9 were significant, indicating that inaccurate data negatively predicts usability. The GNSS group walked more than the RTK group, indicating a partial effect on exploration. We found no significant effect on interaction time with the screen, indicating no specific relation between data accuracy and focus. While RTK data did not allow us to better the usability of location-based AR interfaces, results allow us to assess our system's overall usability as excellent, and to define optimal operating conditions for future use with pupils.

1. INTRODUCTION

This study is part of the ongoing *BiodivAR* project, which attempts to assess the potential benefits of using augmented reality (AR) for outdoor education on biodiversity. In AR interfaces, digital objects can be overlaid on top of users' field of view in real-time, through the screen of a mobile device or a head-mounted display. When used sensibly in an educational setting, it may convey the impression of an enriched environment and make the material more attractive, thus motivating students to learn (Geroimenko, 2020, Alnagrat et al., 2022). The most reported positive effects of AR in education are learning gains and motivation (Bacca et al., 2014). Our research is focused on the use of *location-based* AR in particular, where the position of augmented objects is computed based on their geographic coordinates relative to the user's location as estimated by the mobile device's GNSS. With this technology, augmented objects can be built remotely from any given geodata, as opposed to marker-based AR which requires physical markers to be physically placed on target locations. Location-based AR specially promotes learning in context (Arvola et al., 2021, Chiang et al., 2014), ecological engagement (Bloom et al., 2010), and causes users to experience a positive interdependence with nature (O'Shea et al., 2011), which fosters improved immersion and learning. Last but not least, location-based AR shows positive effects on the physical activity of users across genders, ages, weight status, and prior activity levels (Rauschnabel et al., 2017). However, location-based AR requires steady and continuously accurate data to operate. While GNSS technology has evolved and improved in the past decades, it has been more of an evolution than a revolution. Usability issues have been reported by a number of studies (Chiang et al., 2014, Dunleavy et al., 2009, Ryokai and Agogino, 2013, Admiraal et al., 2011, Lee et al., 2012), most of which blame the inaccuracy of mobile devices' embedded GNSS sensors. Some studies considered that these recurring problems made AR distracting and frustrating and eventually favored marker-based AR, which is more advanced and offers better user experience (Bressler and Bodzin, 2013, Debandi et al., 2018).

2. BACKGROUND

A first proof-of-concept was developed in 2017, featuring a series of geolocated points of interest (POIs) on biodiversity. A test with ten-year-old pupils confirmed the relevance of using AR to support educational field trips (Ingensand et al., 2018) while also revealed usability challenges:

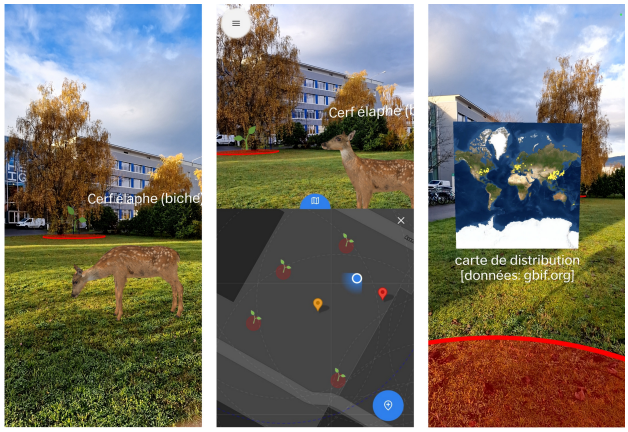
1. The system should allow non-expert users to create AR experiences (Cubillo et al., 2015)
2. Users should be able to publish observations rather than being restricted to a passive viewing role;
3. The instability of augmented objects deteriorates usability. Participants spent 88.5 % of the time looking at the tablet rather than with the surrounding nature. This imbalance could be in part related to inaccurate geolocation data: participants were observed spending considerable time reorienting themselves (Ingensand et al., 2018).

In order to address these identified issues, we developed *BiodivAR*[1], a free and open source (GNU GPLv3.0) web application using a user-centered design process (Mercier et al., 2023).

* Corresponding author

[1] The web application is released under the GNU General Public Li-

It was built using the web framework A-Frame[2], for which we also created a custom library[3] for the creation of WebXR location-based objects in A-Frame. We used the Leaflet[4] library for the interactive maps. *BiodivAR* enables the creation and visualization of geolocated POIs in AR (see Figure 1) as well as a cartographic authoring tool for the collaborative management of AR environments (see Figure 2). They can be shared publicly with or without editing privileges. The application allows anyone without technological know-how to create AR environments by importing/exporting geospatial data and styling POIs by attaching medias to them. Medias can be location-triggered (visible/audible) according to various distance thresholds set by the author.



a) b) c)

Figure 1. *BiodivAR*'s AR interface: a) view of two POIs from a distance; b) the 2D map is opened in split view; c) after entering the radius of a POI, contextual data on the adjacent plant specimen is triggered. https://biodivar.heig-vd.ch/

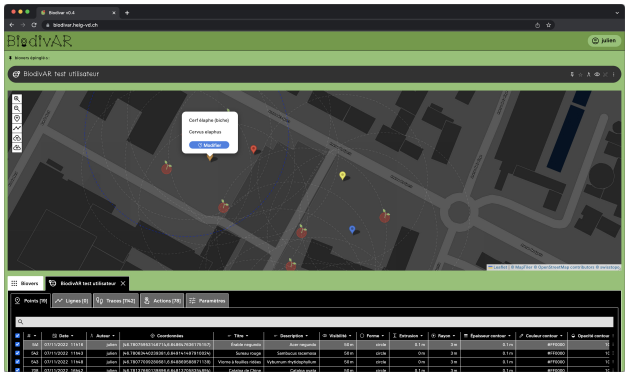


Figure 2. *BiodivAR*'s cartographic authoring tool.

3. RESEARCH GOALS

The purpose of our research *overall* is to assess the potential benefits of using this application in the context of biodiversity education. Before introducing the tool to pupils, it seemed important to ensure its usability. This comparative user test will allow us to define and guarantee the best possible conditions of use for a younger audience. The goals of this study can be synthesized as follows:

1. Assess the overall usability of the AR application.
2. Assess the impact of geolocation data accuracy on usability, exploration, and focus.
3. Gather user feedback for future improvements[5].

The literature review and the observations made during the first iteration led us to propose the following hypothesis: Inaccurate geolocation data negatively affects usability. Additionally, we are looking to investigate the impact that geolocation data accuracy may have on exploration and focus in location-based AR, about which we have not been able to find any literature. The resulting research questions are:

Q1: Does geolocation data accuracy predict usability scores?
Q2: Is geolocation data accuracy related to exploration?[6]
Q3: Is geolocation data accuracy related to focus?[7]

4. MATERIALS AND METHODS

4.1 Experimental design

The present study aims to measure and compare the usability of a location-based AR application used in combination with different geolocation data sources. Using our authoring tool, we created an AR environment with POIs on biodiversity in the surroundings of the School of Engineering and Management Vaud in Yverdon-les-Bains (Switzerland). After a brief introduction to the tool, all participants freely explored the AR environment for 15 minutes using a Samsung Galaxy Tab Active3 tablet with a SIM card for cellular data. As shown in Figure 3, the comparative user test (n = 54) includes in two groups:

GNSS the control group received geolocation data coming from the GNSS sensor embedded in the mobile device
RTK the experimental group received geolocation data coming from an external Ardusimple RTK kit[8].

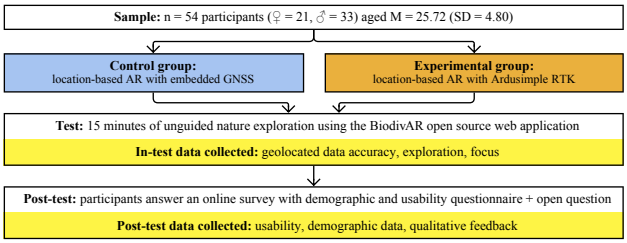


Figure 3. Experimental design of the comparative user test.

4.2 Participants

The sample includes 54 participants (♀ = 21, ♂ = 33), with a mean age of M = 25.72 (SD = 4.80). They are students and collaborators of the School of Engineering and Management Vaud, and they each signed an informed consent form for the use of the data collected. Login credentials (identifier + password)

cense v3.0. It is accessible (no download required) at: https://biodivar.heig-vd.ch. The source code is available at https://github.com/MediaComem/biodivar.
2 https://github.com/aframevr/aframe (MIT License)
3 https://github.com/MediaComem/LBAR.js/ (MIT License)
4 https://github.com/Leaflet/Leaflet (FreeBSD License)
5 The qualitative feedbacks were not included in this paper, as we exclusively focused on the quantitative data and group comparison.
6 Exploration is represented by the distance walked, the number of POIs visited, and the number of times the 2D map was opened.
7 Focus is represented by the ratio of time spent gazing at the screen *versus* the real world.
8 https://www.ardusimple.com/product/rtk-handheld-surveyor-kit/

were created for each participant to record their data separately and facilitate comparison. Among them, 47 agreed to wear eye-tracking glasses, of which 41 successfully recorded data. They were randomly assigned to each group. The control group's (GNSS) mean age is M = 27.5 (SD = 6.09), and it includes 12 ♀ and 15 ♂. The experimental group's (RTK) mean age is M = 24.2 (SD = 2.22) and it includes 9 ♀ and 18 ♂. The first participant eventually had to be excluded from the final results because they experienced numerous crashes due to a bug that was fixed for the subsequent participants. The treatment they received was therefore too different to compare.

4.3 Data collection and processing

The four main concepts our study seeks to connect are "location data accuracy", "usability", "exploration", and "focus". The measurable observations we chose to represent those concepts are listed in Table 1. In our experiment, the two groups (or treatments) operationalize the concept of "geolocation data accuracy". This concept is represented by two variables: *accuracy* and *continuity*. The accuracy attribute is provided by the Geolocation API along with the horizontal location data as latitude and longitude[9]. It denotes the accuracy level of the latitude and longitude coordinates in meters. We use the average accuracy participants were exposed to while in AR mode as the indicator for accuracy. However, in the specific context of location-based AR, sudden changes in data accuracy heavily impact the display of augmented objects in the interface. An indicator for continuity in the data is thus the amount of outliers–i.e. the points that are visibly out of a user's trajectory (as shown in Figure 4). An additional indicator for continuity in the data is the standard deviation of the data accuracy the participants of each group was exposed to. As far as the concept of "usability" goes, it is represented by a series of nine variables whose indicators are the different scales of the three questionnaires (SUS, HARUS, UEQ): *overall usability*, *ease of handling*, *ease of understanding*, *attractability*, *user-friendliness*, *efficiency*, *dependability*, *motivation*, *innovativeness*. The concept of "exploration" is represented by three variables: *quantity*, *diversity*, and *ease*. The distance walked is the indicator of the quantity of exploration. The amount of POIs visited is the indicator of the diversity of exploration. An important use of the 2D map may indicate that participants required assistance in navigating. The amount of times the 2D map was opened is thus the indicator of the ease users had exploring. Finally, the concept of "focus" in our study is represented by a *screen interaction* variable, whose indicator is the amount of time participants spent interacting with the tablet screen *versus* with the real world.

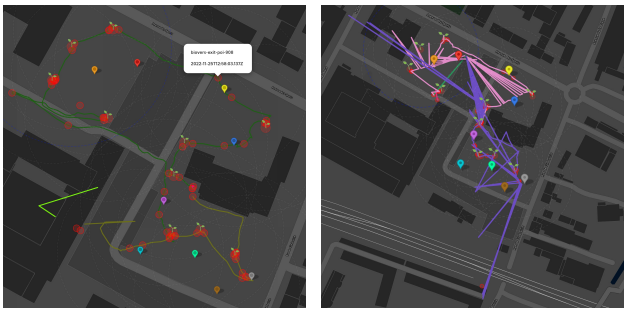
4.3.1 Geolocation data accuracy During the test, participants' geographical coordinates were logged at 1 Hz. Each log also contains an attribute for location accuracy, user ID and a timestamp. The resulting users' trajectories can be visualized in the application (see Figure 4) and downloaded as GeoJSON files for further analysis. The color of the trajectory changes when the AR session is stopped and resumed again. We downloaded the data and calculated the mean location accuracy each participant was exposed to. As shown in Figure 4, the trajectories–in particular that of the RTK group–contained outliers, which were removed manually using the free and open source software QGIS to get a more accurate estimate of the actual distance travelled (as an indicator of our "exploration quantity" variable, see 4.3.3). By calculating the different amount of points before and after this manual processing, the outliers were summed for each participant. Once the data was cleaned, we calculated the total distance walked by each participant. Because there were variations in the duration of each participant's test (min = 9′14, max = 24′11 s), the data was normalized for a duration of 15 minutes. This allowed us to calculate:

1. The average geolocation data accuracy
2. The amount of outliers in the data
3. The standard deviation of the geolocation data accuracy

[9] https://w3c.github.io/geolocation-api

Concept	Variable	Indicator
Geolocation data accuracy	Quality	Average geolocation data accuracy
	Continuity	Amount of outliers
		Standard deviation of data accuracy
Usability	Overall usability	SUS score
	Ease of handling	HARUS (manipulability) score
	Ease of understanding	HARUS (comprehensibility) score
	Attractability	UEQ (attractiveness) score
	User-friendliness	UEQ (perspicuity) score
	Efficiency	UEQ (efficiency) score
	Dependability	UEQ (dependability) score
	Motivation	UEQ (stimulation) score
	Innovativeness	UEQ (novelty) score
Exploration	Quantity	Distance walked
	Diversity	Amount of POIs visited
	Ease	Amount of times 2D map was opened
Focus	Screen interaction	Interaction time with tablet screen

Table 1. Operationalization table.



a) b)

Figure 4. a) A trajectory from the GNSS group. The short light green line is at an impossible location (on top of a tall building), indicating outliers. b) A trajectory from the RTK group. The star-shaped spikes indicate the presence of many outliers.

4.3.2 Usability Immediatly after the test, participants answered an online survey containing demographic questions (age, gender), an open question for qualitative feedback, and three usability questionnaires:

- SUS (System Usability Scale) is a generic, technology-independent 10 item questionnaire on a 5 point Likert scale, frequently used for generic evaluation of a system (Brooke, 1996). The Cronbach's alpha of the SUS questionnaire is 0.79, showing an appropriate internal consistency. In accordance with the instructions of the scale's authors, the SUS score is calculated as follows: 1 point was subtracted from the odd-numbered (phrased positively) items' scores. We subtracted the even-numbered (phrased negatively) items score to 5. The processed scores were added together and then multiplied by 2.5 to get an individual user's score on a scale of 100. While a comparison between two scores is self-explanatory, we used an adjective scale (Bangor, 2009) to qualify the results individually.

- HARUS (Handheld Augmented Reality Usability Scale) is a mobile AR-specific 16 item questionnaire (Santos et al., 2014) on a 7 point Likert scale that focuses on handheld devices and emphasizes perceptual and ergonomic issues. The Cronbach's alpha of the HARUS questionnaire is 0.798, showing appropriate internal consistency. It has two components: *manipulability*—the ease of handling the AR system, and *comprehensibility*—the ease to read the information presented on screen. In accordance with the instructions of the scale's authors, the HARUS scores are calculated as follows: We subtracted the odd-numbered (phrased negatively) items score to 7. 1 point was subtracted from the even-numbered (phrased positively) items' scores. The processed scores for items 1 to 8 were added together, divided by 48, and multiplied by 100 to get the individual "manipulability" score on a scale of 100. Similarly, the processed scores for items 9 to 16 were added together, divided by 48, and multiplied by 100 to get the individual "comprehensibility" score on a scale of 100. HARUS was designed so that its scores are commensurable with SUS scores.
- UEQ (User Experience Questionnaire) is a 26 item questionnaire in the form of semantic differentials: each item is scored on a 7 point scale (from -3 to +3, with 0 as neutral) with two terms with opposite meanings at each extreme (i.e. attractive|unattractive). It provides a comprehensive measure of user experience (Laugwitz et al., 2008). It includes six scales, covering classical usability aspects such as *efficiency* (can users solve their tasks without unnecessary effort?), *perspicuity* (is it easy to learn how to use the application?), and *dependability* (does the user feel in control of the interaction?), as well as broader user experience aspects such as *attractiveness* (do users like the application?), *novelty* (is the application innovative and creative?), and *stimulation* (is it exciting and motivating to use the application?). UEQ is typically routinely used to statistically compare two version of a system to check which one has the better user experience. Thus, the UEQ evaluations of both systems or both versions of a system are compared on the basis of the scale means for Each UEQ scale. *Attractiveness* is calculated by averaging the scores from items 1, 12, 14, 16, 24, and 25. *Perspicuity* is calculated by averaging the scores from items 2, 4, 13, and 21. *Efficiency* is calculated by averaging the scores from items 9, 20, 22, and 23. *Dependability* is calculated by averaging the scores from items 8, 11, 17, and 19. *Stimulation* is calculated by averaging the scores from items 5, 6, 7, and 18. Novelty is calculated by averaging the scores from items 3, 10, 15, and 26. Values range between -3 (horribly bad) and +3 (extremely good), but in general only values in a restricted range will be observed. The calculation of means over a panel of participants make it extremely unlikely to observe values above +2 or below -2, as specified in the UEQ handbook (Schrepp, 2015). As per their interpretation, values between -0.8 and 0.8 correspond to a neutral evaluation of the corresponding scale and values greater than 0,8 represent a positive evaluation.

These questionnaires provided scores for the nine scales reported in Table 1 as indicators of our usability variables.

4.3.3 Exploration During the test, various in-app, user-triggered events were recorded by the application. These included: when the AR session was initiated or exited, when the 2D map was opened or closed, and when the triggering radius of a POI was entered or exited. Each log also contains the coordinates the action took place at, the user ID and a timestamp. The resulting users' action log can be visualized in the application and downloaded as GeoJSON files. Events are represented with red circles on the 2D map (see Figure 4). We downloaded the data and calculated the number of POIs each participant visited as well as how many times they opened the 2D map. These values (POIs visited, 2D map opened) were normalized for a test duration of 15 minutes. This allowed us to calculate:

1. The amount of POIs visited
2. The amount of times the 2D map was opened

The distance walked by each participant was calculated from the geolocation data (see 4.3.1).

4.3.4 Focus The goal of using eye tracking glasses and data in our study is to determine for how long participants were looking in or out of the tablet screen. 47 out of 54 participants were able–and agreed–to wear eye trackers (Tobii Pro Glasses 3), recording their gaze for the duration of the test. The 7 participants that didn't either choose not to or couldn't because they had prescription glasses. Despite rigorous implementation, 6 recordings did not work as expected and no files were saved. The 41 remaining recordings were imported in Tobii's analysis software. Unfortunately, its tools do not support tracking of moving areas of interest (i.e. the surface of the tablet). We exported the videos with the overlaying gaze point and extracted 10 frames per second, resulting in a dataset of 380K images, an instance of which is shown in Figure 5. We attempted to classify the data with openCV pattern recognition, but the variability prevented from obtaining any results. We resolved to train a deep learning multiclass image classifier model by fine-tuning a pretrained vision transformer (ViT) model with our dataset (Dosovitskiy et al., 2020). We first had to manually label a random selection of 10K frames with "in" or "out" labels corresponding to whether the point was in or out of the tablet screen (see Figure 5). After training for only one epoch using Google's colaboratory and obtaining a satisfying validity of 95%, we inferred the whole dataset which provided a label for every frame[10]. They were encoded in order to calculate the ratio of time each user spent looking at the tablet screen *versus* outside of it, at the real world.

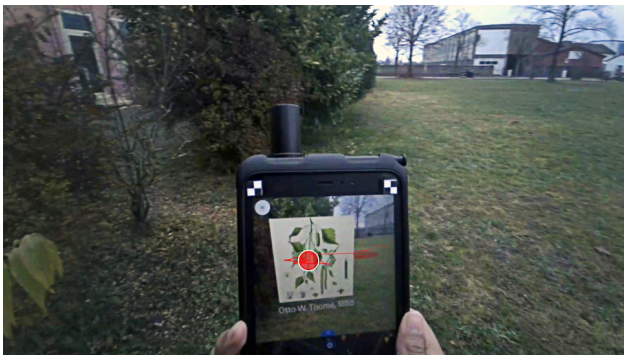


Figure 5. Eye tracking data sample. The user's gaze is located within the tablet screen area.

[10] The code used to fine-tune the ViT model is accessible in the following Jupyter Notebook: https://colab.research.google.com/drive/1sYxbJQ-7FrScr7R87qwmqKZcb837LWhJ#scrollTo=JLseEgvycDGy. The dataset and the trained model are available here: https://huggingface.co/julienmercier.

5. RESULTS

5.1 Data analysis

Statistical analysis were made with the free and open platform Jamovi (The jamovi project, 2022). In the following subsections, we report descriptive statistics (M, SD), and compare our groups (GNSS *versus* RTK) using an independant Student *t*-test to emphasize to which extent both groups differ on our variables of interest. In cases where the homogeneity of variances assumption is not met, we used a Welch *t*-test, which is more robust[11].

5.2 Geolocation data accuracy

5.2.1 Average geolocation data accuracy As shown in Figure 6, the mean accuracy for the GNSS group is M = 11.0 (SD = 15.3), and M = 33.6 (SD = 24.8) for the RTK group. The value is in meters, meaning the data the GNSS group was exposed to was accurate within a 11 meters radius, whereas the RTK group got data accurate within a 33.6 meters radius. A Welch *t*-test was used. The results show a significant difference between the two groups (t(43.5) = -3.99, p = <.001).

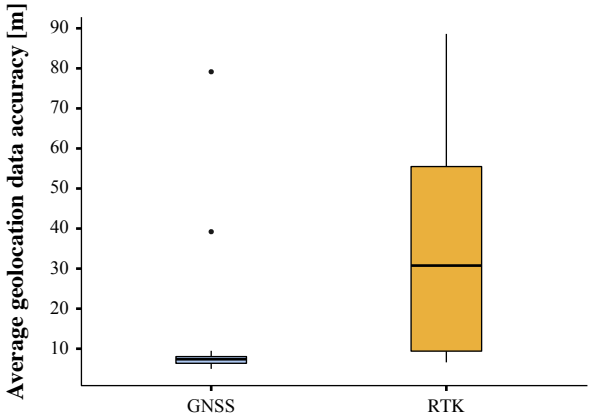


Figure 6. Geolocation data accuracy by group.

5.2.2 Outliers As shown in Figure 7, the GNSS group trajectories contained M = 7.2 (SD = 7.55) outliers, and these of the RTK group M = 46.8 (SD = 40.1). A Welch *t*-test was used. The results show a significant difference between the two groups (t(27.9) = -5.04, p = <.001).

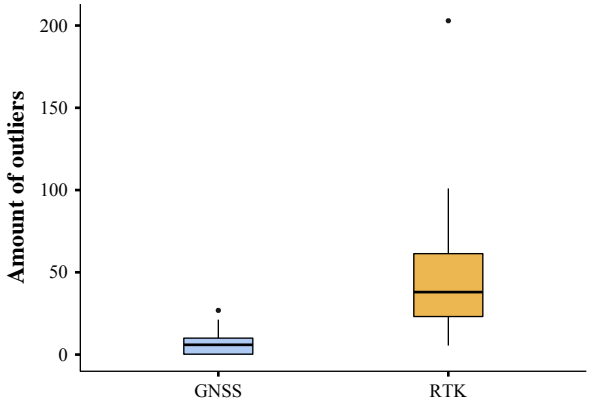


Figure 7. Amount of outliers by group.

[11] The data is available here: https://zenodo.org/record/7845707.

5.2.3 Standard deviation geolocation data accuracy As shown in Figure 8, the data participants from the GNSS group were exposed to had a standard deviation of M = 32.0 (SD = 77.7), and that of the RTK group M = 168.3 (SD = 120.1). A Welch *t*-test was used. The results show a significant difference between the two groups (t(44.7) = -4.93, p = <.001).

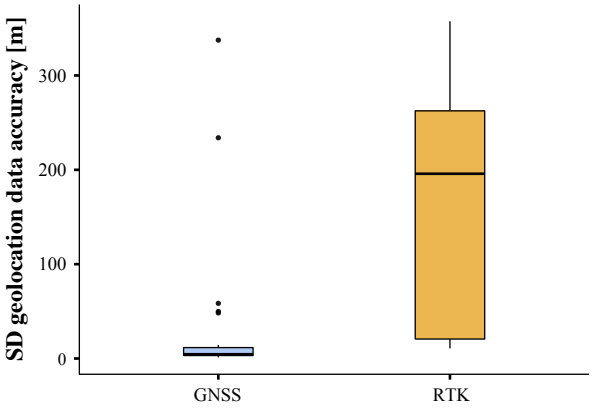


Figure 8. Standard deviation geolocation data accuracy by group.

5.3 Usability

The means of each group for all nine scales from the three usability questionnaires are reported in Table 2 along with *t*-test's p values for significance assessment.

Scale	GNSS	RTK	*t*-test
SUS	M = 81.7 SD = 9.74	M = 74.4 SD = 12.0	t(51) = 2.45, p = 0.018
HARUS (manipulability)	M = 76.7 SD = 13	M = 68.1 SD = 16.1	t(51) = 2.13, p = 0.038
HARUS (comprehensibility)	M = 78.3 SD = 11.3	M = 74.9 SD = 12.9	t(51) = 1.01, p = 0.318
UEQ (attractiveness)	M = 1.72 SD = 0.7	M = 1.1 SD = 0.98	t(51) = 2.65, p = 0.011
UEQ (perspicuity)	M = 2.02 SD = 0.64	M = 1.45 SD = 0.92	t(46.7) = 2.61, p = 0.012
UEQ (efficiency)	M = 1.24 SD = 0.85	M = 0.85 SD = 0.94	t(51) = 1.58, p = 0.121
UEQ (dependability)	M = 1.17 SD = 0.68	M = 1.02 SD = 0.62	t(51) = 0.87, p = 0.39
UEQ (stimulation)	M = 1.84 SD = 0.84	M = 1.31 SD = 1.11	t(51) = 1.93, p = 0.059
UEQ (novelty)	M = 1.8 SD = 0.85	M = 1.21 SD = 0.89	t(51) = 2.45, p = 0.018

Table 2. Usability results by group and *t*-tests.

5.3.1 SUS As shown in Figure 9, the mean SUS score for the GNSS group is M = 81.7 (SD = 9.74). The mean SUS score for the RTK group is M = 74.4 (SD = 12). The results show a significant difference between the two groups (t(51) = 2.45, p = 0.018).

5.3.2 HARUS On the *manipulability* scale (indicating ease of handling the AR system), the mean score for the GNSS group is M = 76.7 (SD = 13) and that of the RTK group is M = 68.1 (SD = 16.1), as shown in Figure 10. The results show a significant difference between the two groups (t(51) = 2.13, p = 0.038). On the *comprehensibility* scale (indicating ease of understanding information presented in the AR interface), the mean score for the GNSS group is M = 78.3 (SD = 11.3) whereas the mean score and that of the RTK group is M = 74.9 (SD = 12.9). The results *do not* show any significant difference between the two groups (t(51) = 1.01, p = 0.318).

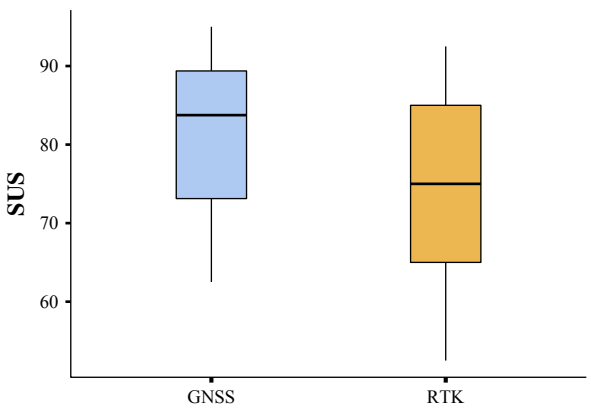


Figure 9. SUS scores by group.

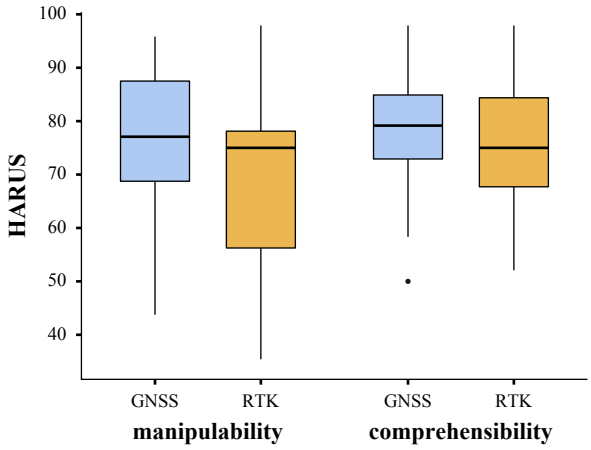


Figure 10. HARUS scores by group.

5.3.3 UEQ As shown in Figure 11, on the *attractiveness* scale, the mean score for the GNSS group is M = 1.72 (SD = 0.7) and that of the RTK group is M = 1.1 (SD = 0.98). The results show a significant difference (t(51) = 2.65, p = 0.011). On the *perspicuity* scale, the mean score for the GNSS group is 2.02 (SD = 0.64) and that of the RTK group is 1.45 (SD = 0.92). A Welch *t*-test was used. The results show a significant difference between the two groups (t(46.7) = 2.61, p = 0.012). On the *efficiency* scale, the mean score for the GNSS group is 1.24 (SD = 0.85) and that of the RTK group is 0.85 (SD = 0.94). The results *do not* show any significant difference (t(51) = 1.58, p = 0.121). On the *dependability* scale, the mean score for the GNSS group is 1.17 (SD = 0.68) and that of the RTK group is 1.02 (SD = 0.62). The results *do not* show any significant difference (t(51) = 0.87, p = 0.39). On the *stimulation* scale, the mean score for the GNSS group is 1.84 (SD = 0.84) and that of the RTK group is 1.31 (SD = 1.11). The results *do not* show any significant difference (t(51) = 1.93, p = 0.059). On the *novelty* scale, the mean score for the GNSS group is 1.8 (SD = 0.85) and that of the RTK group is 1.21 (SD = 0.89). The results show a significant difference (t(51) = 2.45, p = 0.018).

5.4 Exploration

5.4.1 Distance walked As shown in Figure 12, the GNSS group walked an average distance of M = 586.15 (SD = 96.24) meters, whereas the RTK group walked an average distance of

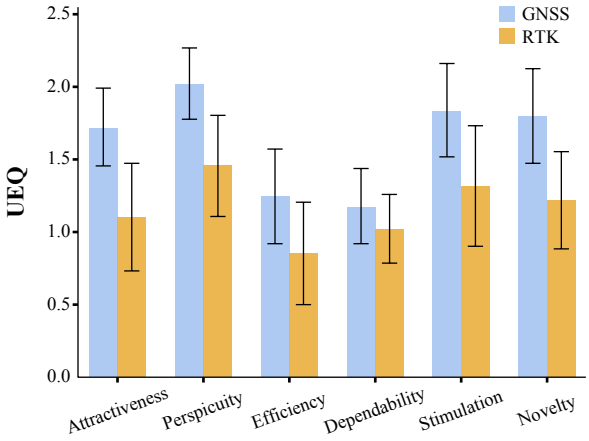


Figure 11. UEQ scores by group.

M = 525.94 (SD = 71.9) meters. The results show a significant difference (t(51) = 2.59, p = 0.013).

5.4.2 POIs visited The GNSS group visited an average of M = 21.09 (SD = 4.02) POIs, whereas the RTK group visited an average of M = 19.29 (SD = 5.87). The results *do not* show any significant difference (t(51) = 1.30, p = 0.199).

5.4.3 Map opened The GNSS group opened the 2D map M = 2.83 (SD = 2.24) times in average, whereas the RTK group opened it M = 1.91 (SD = 2.41) times. The results *do not* show any significant difference (t(51) = 1.44, p = 0.157).

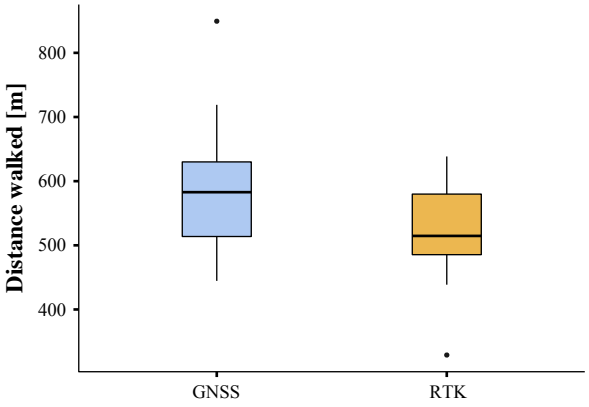


Figure 12. Distance walked by group.

5.5 Focus

The GNSS group spend an average M = 73.3% (SD = 9.81) of the time looking at the tablet screen. The RTK group spend an average M = 69.2% (SD = 12.4) of the time looking at the tablet screen. The results *do not* show any significant difference (t(51) = 1.16, p = 0.251).

6. CONCLUSIONS

The purpose of the study was to assess the impact of geolocation data on the usability of our location-based AR system. To test our hypotheses, we exposed the participants to different geolocation data sources with significantly different accuracies. While we expected RTK data to be more accurate and that it

would enable us to improve usability, analysis highlights that it was significantly less accurate and less continuous than GNSS data. This appears to be due to the fact that the embedded GNSS sensor contains filters that preprocess data and remove most of the outliers. In contrast, RTK data purposefully remains "raw", which is valuable for an advanced user. RTK data accuracy is very efficient when used on an isolated basis (ie. at a 2D map scale), but not particularly suitable for a real-time continuous usage (where location is measured several times per second) on a 1:1, tridimensional scale, at least without any filters applied onto it. Despite this contingency, both the quality and continuity of the geolocation data accuracy the two groups were exposed to was significantly different, which is the essential premise for testing our hypothesis and addressing our research questions. Regarding our main research question, results reveal that the GNSS group, who used the AR application in combination with more accurate and continuous data, reported higher scores in all usability scales, of which five out of nine were statistically significant. This supports our initial hypothesis that poor data accuracy negatively impacts the usability of a location-based AR system. Futures studies should however investigate whether RTK data with proper outlier processing may actually better usability. Our results further highlight that the GNSS group walked more than the RTK group, revealing that the accuracy of geolocation data was partially related to exploration, at least for the quantity indicator. However, due to the manual removal of the outliers–which were significantly more frequent in the RTK group–from the trajectories, the data could be biased. It would be necessary to record a trajectory with both modalities, remove the outliers and observe if there are not significant difference between the measurements to ensure that there are no bias. The comparison on the exploration diversity indicator (amount of POIs visited) was not significantly different. Additionally, although the difference was not significant, the GNSS group opened the 2D map more often than the RTK group in average, suggesting the RTK group could have had more ease exploring. Our results further highlight that there were no significant difference between the ratio of time participants from each group spent interacting with the tablet screen, which would indicate that there is no particular relation between the accuracy of geolocation data and focus.

Although the two experiments cannot be properly compared, because the tests took place 5 years apart under different conditions, we note that participants spent 69.2%–73.3% of the time looking at the tablet screen, which seems to be a meaningful longitudinal progress from the measurement that was made on our 2017 proof-of-concept, where participants interacted with the screen for 88.5 % of the time (Ingensand et al., 2018). While we are not aware of a method to determine the ideal proportion, this measure overall remains an interesting indicator of the importance of the tablet in this type of activity. In a wide review of mobile learning projects, technology was found to dominate the experience in a problematic way in 70% (28/38) of the cases (Goth et al., 2006). While using RTK data did not allow us to positively impact the usability of our system, our study however demonstrated the impact of varying geolocation data accuracy on usability and exploration. The immediate benefit of performing this comparative study is for us to define the most suitable conditions of use before offering our system to a young audience, as well as to ensure an adequate overall level of usability. The overall score reported by the GNSS group allows us to qualify the application's usability as "excellent" according to the SUS adjective scale (Bangor, 2009).

7. ACKNOWLEDGEMENTS

The authors thank Yoann Douillet for his help with the organization of the tests and the eye tracking data collection. Study participation was voluntary, and written informed consent to publish this paper was obtained from all participants involved in the study. Participants were informed that they could withdraw from the study at any point. The data presented in this study is openly available on Zenodo at https://zenodo.org/record/7845707. This research was funded by the Swiss National Science Foundation (SNSF) as part of the NRP 77 "Digital Transformation" (project number 407740_187313) and by the University of Applied Sciences and Arts Western Switzerland (HES-SO): Programme stratégique "Transition numérique et enjeux sociétaux". The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

REFERENCES

Admiraal, W., Huizenga, J., Akkerman, S., Dam, G. t., 2011. The concept of flow in collaborative game-based learning. *Computers in Human Behavior*, 27(3), 1185–1194. doi.org/10.1016/j.chb.2010.12.013.

Alnagrat, A., Ismail, R., Syed Idrus, S. Z., 2022. A Review of Extended Reality (XR) Technologies in the Future of Human Education: Current Trend and Future Opportunity. *Journal of Human Reproductive Sciences*, 1, 81–96. doi.org/10.11113/humentech.v1n2.27.

Arvola, M., Fuchs, I. E., Nyman, I., Szczepanski, A., 2021. Mobile Augmented Reality and Outdoor Education. *Built Environment*, 47(2), 223–242. doi.org/10.2148/benv.47.2.223.

Bacca, J., Baldiris, S., Fabregat, R., Graf, S., Kinshuk, 2014. Augmented Reality Trends in Education: A Systematic Review of Research and Applications. *Journal of Educational Technology & Society*, 17(4), 133–149. jstor.org/stable/jeductechsoci.17.4.133.

Bangor, A., 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale - JUX. *JUX - The Journal of User Experience*. uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/.

Bloom, M. A., Holden, M., Sawey, A. T., Weinburgh, M. H., 2010. Promoting the Use of Outdoor Learning Spaces by K-12 Inservice Science Teachers Through an Outdoor Professional Development Experience. A. M. Bodzin, B. Shiner Klein, S. Weaver (eds), *The Inclusion of Environmental Education in Science Teacher Education*, Springer Netherlands, Dordrecht, 97–110.

Bressler, D., Bodzin, A., 2013. A mixed methods assessment of students' flow experiences during a mobile augmented reality science game. *Journal of Computer Assisted Learning*, 29(6), 505–517. doi.org/10.1111/jcal.12008.

Brooke, j., 1996. SUS: A 'Quick and Dirty' Usability Scale. *Usability Evaluation In Industry*, CRC Press.

Chiang, T. H. C., Yang, S. J. H., Hwang, G.-J., 2014. An Augmented Reality-based Mobile Learning System to Improve Students' Learning Achievements and Motivations in Natural Science Inquiry Activities. *Journal of Educational Technology & Society*, 17(4), 352–365. jstor.org/stable/jeductechsoci.17.4.352.

Cubillo, J., Martin, S., Castro, M., Boticki, I., 2015. Preparing augmented reality learning content should be easy: UNED ARLE-an authoring tool for augmented reality learning environments. *Computer Applications in Engineering Education*, 23(5), 778–789. doi.org/10.1002/cae.21650.

Debandi, F., Iacoviello, R., Messina, A., Montagnuolo, M., Manuri, F., Sanna, A., Zappia, D., 2018. Enhancing cultural tourism by a mixed reality application for outdoor navigation and information browsing using immersive devices. *IOP Conference Series: Materials Science and Engineering*, 364, 012048. doi.org/10.1088/1757-899X/364/1/012048.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. doi.org/10.48550/ARXIV.2010.11929.

Dunleavy, M., Dede, C., Mitchell, R., 2009. Affordances and Limitations of Immersive Participatory Augmented Reality Simulations for Teaching and Learning. *Journal of Science Education and Technology*, 18(1), 7–22. doi.org/10.1007/s10956-008-9119-1.

Geroimenko, V., 2020. *Augmented reality in education: a new technology for teaching and learning*. Springer International Publishing.

Goth, C., Frohberg, D., Schwabe, G., 2006. *The Focus Problem in Mobile Learning*. IEEE, Athens.

Ingensand, J., Lotfian, M., Ertz, O., Piot, D., Composto, S., Oberson, M., Oulevay, S., Da Cunha, M., 2018. *Augmented reality technologies for biodiversity education—a case study*. 12-15 June 2018.

Laugwitz, B., Held, T., Schrepp, M., 2008. *Construction and Evaluation of a User Experience Questionnaire*. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg.

Lee, G., Duenser, A., Kim, S., Billinghurst, M., 2012. *CityViewAR: A mobile outdoor AR application for city visualization*.

Mercier, J., Chabloz, N., Dozot, G., Ertz, O., Bocher, E., Rappo, D., 2023. BiodivAR: A Cartographic Authoring Tool for the Visualization of Geolocated Media in Augmented Reality. *ISPRS International Journal of Geo-Information*, 12(2), 61. doi.org/10.3390/ijgi12020061.

O'Shea, P. M., Dede, C., Cherian, M., 2011. Research Note: The Results of Formatively Evaluating an Augmented Reality Curriculum Based on Modified Design Principles. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, 3(2), 57–66. doi.org/10.4018/jgcms.2011040104.

Rauschnabel, P. A., Rossmann, A., tom Dieck, M. C., 2017. An adoption framework for mobile augmented reality games: The case of Pokémon Go. *Computers in Human Behavior*, 76, 276–286. doi.org/10.1016/j.chb.2017.07.030.

Ryokai, K., Agogino, A., 2013. Off the paved paths: Exploring nature with a mobile augmented reality learning tool. *Journal of Mobile Human Computer Interaction*, 5(2), 21–49. doi.org/10.4018/jmhci.2013040102.

Santos, M. E. C., Taketomi, T., Sandor, C., Polvi, J., Yamamoto, G., Kato, H., 2014. *A usability scale for handheld augmented reality*. VRST '14, Association for Computing Machinery, New York, NY, USA.

Schrepp, M., 2015. *User Experience Questionnaire Handbook*.

The jamovi project, 2022. jamovi Software, Version 2.3. jamovi.org.