Effect of Image Quality Feedback on Fossil Photo Submissions in Citizen Science

Megan Spielberg
Fontys University of Applied Sciences
Eindhoven, The Netherlands
m.spielberg@student.fontys.nl

Abstract

Citizen science projects often depend on photographs submitted by volunteers. In paleontology, image quality is important for scientific evaluation, yet contributors usually receive no guidance while taking photos. This study examines whether feedback on image quality can improve fossil photographs at the moment of capture. Three prototype systems were compared: no feedback, feedback after the photo was taken, and real-time feedback during image capture. Image quality was measured using objective metrics for lighting, sharpness, and contrast, along with manual ratings for scale presence and viewing angle. Usability was evaluated using the System Usability Scale. The analysis included 183 images and usability responses from 20 participants. Real-time feedback led to a significant improvement in image contrast and higher usability scores compared to the baseline. No significant improvements were found for lighting, sharpness, scale inclusion, or viewing angle. Post-capture feedback showed no significant effects. These results suggest that real-time feedback can support image capture in citizen science, while delayed feedback offers no or limited benefits.

Keywords: Fossil Photography, Image Quality Metrics, Real Time Feedback, Usability Testing, Prototype Evaluation, Citizen Science

1 Introduction

Citizen science has become an increasingly important approach for collecting large volumes of scientific data across diverse domains. These include ecology, biodiversity monitoring, astronomy, and paleontology (Abdul-Rahman, Zwitter, and Haleem, 2025 & Silvertown, 2009). By enabling non-experts to contribute observations, citizen science projects can expand and accelerate data collection in addition to fostering public engagement with scientific research. However, the scientific value of citizen-contributed data is highly dependent on its quality, especially when observations are submitted in the form of photographs (Eijkelboom et al., 2024 & López-Guillén et al., 2024). Image-based submissions are now included in many citizen science platforms, because they allow

Fontys University of Applied Sciences, Master of Applied IT Eindhoven, January 16, 2026.

experts and automated systems to validate observations remotely. In fields such as paleontology, where physical access to specimens is often limited, photographs serve as the primary medium for identification, verification, and archival documentation (Yaqoob et al., 2024 & Yu et al., 2024). Poor image quality can substantially reduce the usefulness of submissions, increasing expert workload, lowering identification accuracy, and limiting the effectiveness of machine learning models trained on such data (Liu et al., 2023 & Sun et al., 2024).

Paleontological citizen science has distinct challenges for image-based data collection. Fossils are often fragmentary, partially embedded in matrix material, and visually similar to their surrounding background. Accurate identification often depends on subtle morphological features that need good lighting, sufficient sharpness, clear contrast between specimen and background, the inclusion of scale references, and multiple, relevant viewing angles (Eijkelboom et al., 2024). As a result, variations in photographic quality can directly affect whether a submission can be meaningfully evaluated by experts or automated systems. Within this context, projects such as LegaSea aim to involve citizen scientists in the documentation and identification of fossil finds. The current submission workflow, implemented through platforms such as *Oervondstchecker* (Naturalis Biodiversity Center, 2025), allows users to upload photographs and metadata for expert review. While this approach is accessible and scalable, it places the responsibility for image quality entirely on the contributor and does not guide the image capture process. Consequently, quality control occurs only after submission, often requiring follow-up communication or resulting in unusable data. Prior research in citizen science and human–computer interaction shows that timely and actionable feedback can play an important role in improving both data quality and participant experience (Sharma et al., 2024 & Wal et al., 2018).

In related domains, real-time feedback has been shown to improve photographic outcomes by guiding users during image capture rather than after the fact (Li, Yang, and Chang, 2020 & Xu et al., 2015). Commercial applications, such as AI-assisted camera guidance in mobile devices, further demonstrate the feasibility of providing immediate, context-aware feedback to non-expert users (Google, 2025). Despite these advances, real-time image quality feedback has received limited attention within paleontological citizen science. Existing research on artificial intelligence in paleontology has mostly focused on post-hoc image analysis, including fossil classification, taxonomic identification, and feature extraction using deep learning techniques (Dodge and Karam, 2016 & Liu et al., 2023 & Yu et al., 2024). While these approaches benefit from large datasets, they are very sensitive to input image quality (iNaturalist, 2023). Fewer studies address the earlier stage of data collection, where image quality could be improved proactively through user guidance (Lotfian, Ingensand, and Brovelli, 2021). This represents a gap, as improving image quality at the point of capture could provide benefits for expert validation, automated analysis, and overall project scalability. The gap this study investigates is visualized in Figure 2. Figure 1 illustrates the difference between a typical citizen science fossil submission and a professionally archived museum specimen. Figure 1a shows an example of a fossil image that was taken by a citizen scientist. Figure 1b shows a fossil image from a paleontological archive at Naturalis Biodiversity Center.

(a) Citizen science submission from Oervondstchecker (b) Archived fossil from Naturalis Biodiversity Center

Figure 1: Comparison between a citizen science fossil submission and an archived museum specimen.

Research in human–computer interaction and cognitive psychology shows that timely, actionable feedback improves task performance, reduces cognitive load, and helps non-experts adhere to complex quality criteria (Sharma et al., 2024 & Wal et al., 2018). In photography, real-time guidance has been shown to improve composition, lighting, and sharpness by informing users of quality issues while the photo is still being taken (Li, Yang, and Chang, 2020 & Xu et al., 2015). Commercial mobile applications demonstrate that AI-assisted feedback can be both accessible and effective for non-expert users (Google, 2025). Despite these advances, paleontological citizen science has not yet shown whether real-time or post-capture feedback can improve fossil image quality. Existing AI research in paleontology focuses primarily on post-hoc image analysis classification, feature extraction, and taxonomic identification (Liu et al., 2023 & Yu et al., 2024). These methods are sensitive to image quality (iNaturalist, 2023), but very few studies have addressed the earlier stage of data collection, where image quality could be improved proactively.

Research gap

This study addresses the gap between post-hoc fossil image analysis and the moment of image capture. Where previous work in paleontological citizen science has concentrated on improving classification models or refining expert review workflows, little attention has been given to how contributors might be supported during the act of taking photographs itself. Improving image quality at the point of capture has the potential to reduce expert workload, increase the scientific usefulness of citizen-generated data, and enhance downstream machine learning pipelines. Yet it remains unclear whether real-time guidance, post-capture feedback, or the absence of feedback produces meaningfully different outcomes in terms of image quality or user experience. This gap is visualized in Figure 2.

Research aim

To investigate this problem, three prototype interfaces were developed that represent different approaches to guiding contributors during fossil photography: one without feedback, one that provides feedback only after the photo has been taken, and one that offers real-time feedback during capture. By comparing these interfaces, the study examines whether guidance delivered at different moments in the workflow influences measurable aspects of photo quality—such as lighting, sharpness, and contrast—as well as contributors' perceptions of usability. In doing so, the research seeks to clarify whether real-time or delayed feedback can effectively support non-expert users in producing higher-quality fossil photographs and how such feedback mechanisms shape their interaction experience.

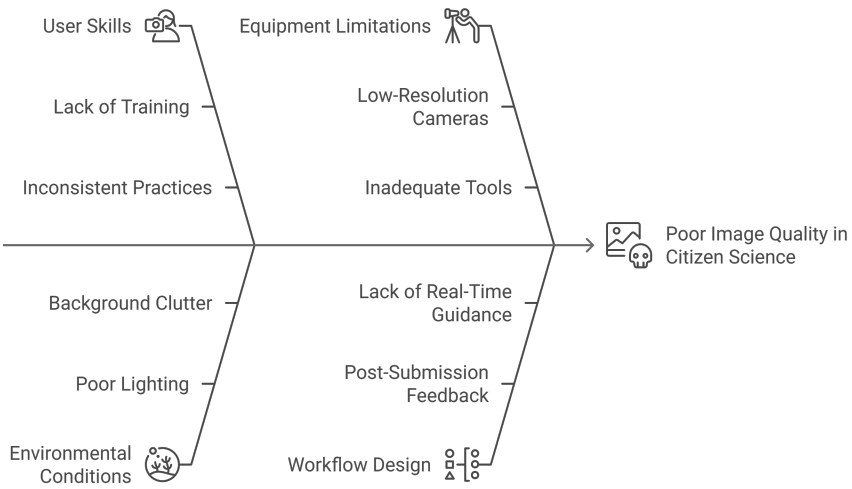


Figure 2: Image quality problems in paleontological citizen science

2 Methodology

This study investigated how different types of feedback on image quality influence the quality of fossil images in a citizen science context. A comparative experimental design was used. This method enabled measurement and evaluation of the effects of various feedback mechanisms on user-submitted images. This experiment is based on three prototypes, each representing a different level of user guidance. This allows for a direct comparison between one control condition and two interventions.

Prototype one serves as the baseline and is the control condition based on existing platforms used for citizen science submissions (Oervondstchecker). In this prototype, users take and upload images without receiving any feedback on their image quality. The second prototype provides feedback on image quality after an image has been taken and encourages users to retake images based on that feedback. The final prototype offers immediate feedback during the image-capturing process, guiding users to take optimal images on their first attempt. The general image capture process is illustrated in Figure 3.

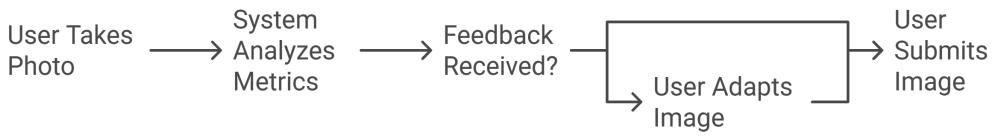


Figure 3: Image capture process

Prototype application

The prototype application is the main tool for data collection. Its design and implementation followed user-centered design principles and an iterative approach. Architectural decisions improve usability and technical performance. The design follows established usability heuristics, emphasizing clarity, consistency, and intuitiveness (Nielsen, 1994). Existing applications, such as Google Lens and Rock Identifier, were analyzed to create an intuitive and familiar user experience. The visual styling is intentionally minimalistic. A natural color palette featuring green and beige tones emphasizes the paleontological context. Using system fonts ensures consistent readability across various devices. These choices are designed to minimize distraction and visual fatigue, allowing users to focus on capturing high-quality images.

Figure 4 presents the interface of each prototype. The left section displays the baseline prototype, which mimics a standard camera view. Users can activate the flashlight and view the images they have taken in the album. The instructions provided for the baseline are the same as those found on the platform Oervondstchecker (Naturalis Biodiversity Center, 2025). In the center of the figure, the post-capture interface is showcased. This interface includes all the functionalities of the baseline, along with additional features that provide users with feedback on lighting, sharpness, and contrast after each image capture. Users are also given suggestions to include a scale and capture images from different angles. The image view features a circular overlay that encourages users to center their object within it. On the right side of the figure, the real-time interface is displayed. This interface shares the same functionalities as the baseline and includes the circular overlay. Additionally, it features three icons that represent lighting, sharpness, and contrast. These icons change color to indicate the quality: red (very poor or poor), yellow (intermediate), and green (good or very good). When any of the icons are pressed, users receive a more detailed explanation of what each one represents, as well as suggestions for adding a scale and taking multiple angles.

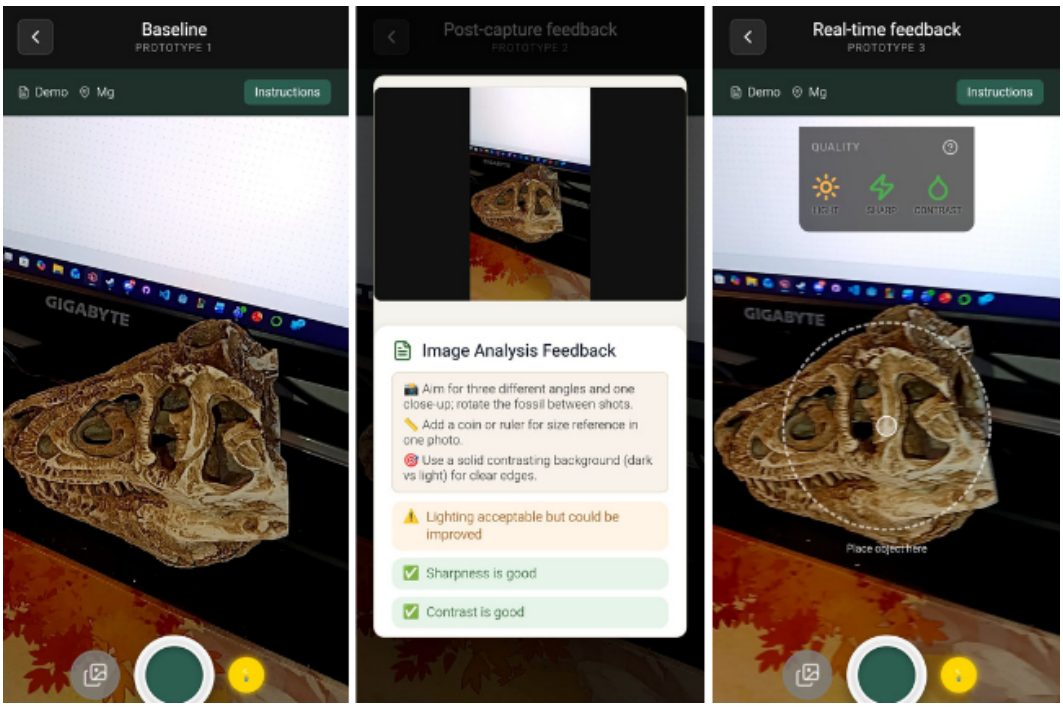


Figure 4: Prototypes
Left: baseline, middle: post-capture feedback, right: real-time feedback

Measuring image quality

The objective was to create a method for measuring image quality that is both objective and reproducible. To achieve this, a quantitative scoring system was developed to calculate image metrics based on an experiment conducted by paleontologist Isaak Eijkelboom. The system incorporates the expert's experience and includes subjective grading. This grading combines both quantitative and qualitative assessments.

The grading system is based on a dataset of 100 fossil images submitted by citizen scientists. These images were evaluated by paleontologist Isaak Eijkelboom using a scale from one (very poor) to five (very good). The criteria for grading included lighting, sharpness, and the contrast between the fossil and its background. Additionally, the evaluation considered the presence of a scale, the type of scale shown in the image, and whether the fossil is depicted from multiple, relevant angles.

To convert the expert's qualitative judgments into actionable decision rules, numerical thresholds were established for each metric. In the first step, the continuous values for lighting, sharpness, and contrast were matched with the expert's 1–5 ratings to create initial threshold candidates for each quality level. In the second step, these thresholds were iteratively refined through in-app testing on actual devices, ensuring that the boundaries between rating levels were both empirically sound and aligned with practical image capture conditions in the field.

For measuring image quality, three metrics can be calculated without a reference image. Lighting is computed as the mean gray value over all pixels in the image. The RGB values are first converted to grayscale using standard luminance weights ($0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$) (International Telecommunication Union, 2011), and then the arithmetic mean across all pixels is calculated. Based on empirical testing, slightly stricter thresholds were defined to more reliably detect under- and overexposed images, with breakpoints at gray values 60, 90, 120, and 180.

Sharpness is measured using the variance of the Laplacian in the central region of the image. A discrete Laplacian operator is applied to the middle 50% of the image area, and the variance of the resulting Laplacian values is computed. Higher variance corresponds to stronger edges and, therefore, better focus. The thresholds 40, 80, 110, and 150 were calibrated using real device data so that slightly shaky or soft images are treated more leniently, while clearly focused images are consistently classified as "good" or "very good."

Contrast between the fossil and its background is captured using a center–edge contrast measure. In the grayscale image, a central circular area (about 60% of the image diameter) is defined as the object region, and a more distant ring is treated as background. The mean brightness of the center region and the edge region is computed, and the absolute difference between them is used as the contrast metric. Higher values indicate stronger separation between the specimen and the background. Thresholds of 15, 30, 40, and 60 impose a relatively strict standard on contrast, so that poorly separated or "blending" fossils are more clearly downgraded.

All three metrics are classified into five qualitative rating levels: "Very Poor," "Poor," "Intermediate," "Good," and "Very Good." A general rating function assigns a continuous metric value to one of these categories by using four thresholds. Values below the first threshold are labeled "Very Poor," those between the first and second thresholds are labeled "Poor," and so on. Values above the highest threshold are labeled "Very Good." The remaining qualitative aspects, such as the presence and type of scale and the relevance of the viewing angle, are manually evaluated and mapped to the same rating levels after data collection concluded.

Data collection

The data collection consists of two main parts. The first part involves collecting images using the prototype, while the second part focuses on usability evaluation through a survey. The collected images are used to assess the impact of the guidance and feedback on image quality, whereas the usability test measures how effectively the feedback conveys actionable steps.

Several preliminary usability tests were conducted with ICT students and project stakeholders from Naturalis Biodiversity Center to evaluate the technical stability and usability of the prototypes. Testers received a link to the live application hosted on a server. For each prototype, testers followed the same process:

1. Taking one or more images.
2. Reviewing the provided feedback mechanism.
3. Being able to delete any unwanted images before submission.
4. Submitting their final selection of images.

Each tester evaluated all three prototypes, and the order in which they were tested was determined by balanced randomization to mitigate any order bias.

In the final test, participants were asked to test the prototype on their phones. Initially, the group included fossil hunters, paleontologists, and citizen scientists. A power analysis indicated that at least 28 participants were needed to detect a medium-sized effect, which was determined using the G*Power application. After receiving only three responses from the fossil hunter group, the criteria were broadened to include anyone who could read/speak English reasonably well and had a phone. The test was advertised on the Fontys ICT campus, as well as in Facebook groups, sub-reddits, and forums related to paleontology. The testing period lasted for 14 days. Unfortunately, the function designed to assign a balanced random testing order failed for this test. As a result, most participants tested the prototype in the same order: p1 (baseline) -> p3 (real-time) -> p2 (post-capture).

Data evaluation

The evaluation is done in two parts, similar to the data collection process. First, the image data is assessed using the rating system described earlier. Next, the calculated metrics are compared among the three prototypes, and their statistical significance is determined. In parallel, the survey responses are evaluated to determine the usability of the prototypes. The survey contains ten questions that allow for the calculation of a System Usability Score (SUS) (Brooke, 1995). These subjective findings help to understand why some feedback prototypes were more accepted than others. Reviewing the responses helps identify features that improved user experience and areas that need work. These findings will guide future iterations of the prototypes, making sure they better meet user needs and improve overall usability.

3 Results

The impact of real-time and post-capture feedback on the quality of fossil photographs submitted by citizen scientists was investigated through analysis of 183 images across three prototype conditions: Baseline (n = 55), Post-Capture Feedback (n = 63), and Real-Time Feedback (n = 65). Image quality was assessed using five metrics: lighting, sharpness, contrast, scale rating, and angle rating. Usability was evaluated using the System Usability Scale (SUS), and user perceptions were collected through a structured survey.

Objective quality metrics

Objective quality metrics were analyzed to determine whether feedback timing influenced the technical quality of submitted images. Descriptive statistics for each metric are presented in Table 1, showing the mean, standard deviation, and median values for lighting, sharpness, contrast, scale rating, and angle rating across each prototype condition. The Real-Time Feedback prototype demonstrated a higher median contrast score (41.88) compared to the Baseline (18.63) and Post-Capture Feedback (31.85) prototypes, suggesting a potential improvement in this dimension. Kruskal-Wallis H-tests were conducted for each metric to assess the statistical significance of differences across prototype conditions. A significant difference was observed only for the contrast metric ($H = 9.03$, $p < \alpha$). Pairwise comparisons using Mann-Whitney U tests with Bonferroni correction ($\alpha = 0.025$) revealed that the difference in contrast between the Baseline and Real-Time Feedback prototypes was statistically significant ($U = 1236.00$, $p < \alpha$). The median contrast score for the Real-Time Feedback prototype was 125% higher than the Baseline, indicating substantial improvement in object-background separation. No significant differences were found for lighting, sharpness, scale rating, or angle rating across the three prototype conditions.

Usability measures

Usability was evaluated using the System Usability Scale (SUS). As shown in Figure 5, the Real-Time Feedback prototype achieved the highest SUS scores, with both the median and mean values clearly surpassing those of the Baseline and Post-Capture Feedback prototypes. Statistical testing confirmed this visual pattern: a significant difference was found between the Baseline and Real-Time Feedback

prototypes, with an observed difference of 14.40 points (W = 4.00, p < α) and a large effect size (r = 0.98). Participants therefore perceived the Real-Time Feedback prototype as substantially more usable than the Baseline. In contrast, no significant difference in SUS scores was found between the Baseline and Post-Capture Feedback prototypes (W = 24.50, p > α). This aligns with the distribution in Figure 5, where the Post-Capture prototype shows greater variability and a lower mean usability score than the Real-Time Feedback interface. Participant comments from the structured survey further supported these findings. Users reported that the Real-Time Feedback prototype was easier to use, more intuitive, and provided clearer instructions compared to both alternative versions. Additionally, 65% of participants strongly agreed that they felt motivated to adjust their photographs based on the guidance provided while testing the real-time prototype. This is a significant increase compared to 40% for the post-capture version and only 10% for the baseline version.

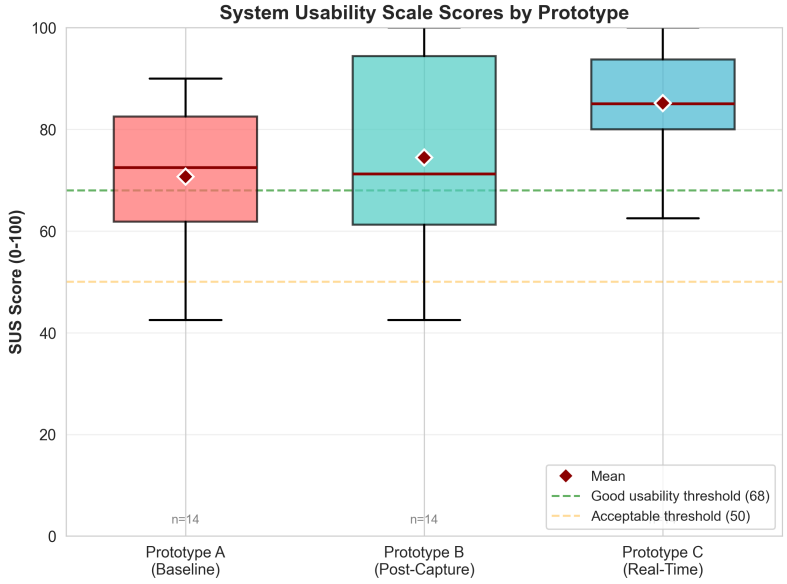


Figure 5: System Usability Scale (SUS) results across prototype conditions.
Each boxplot displays the distribution of SUS scores (0 - 100) for the three prototype interfaces. Red diamonds denote mean values, while dashed lines indicate commonly used thresholds for acceptable (50) and good (68) usability.

Sub-hypotheses

Additional analyses explored the impact of flashlight usage and time spent on image quality. A Mann-Whitney U test revealed no significant difference in lighting quality between images captured with the flashlight enabled (n = 47) and those captured without (n = 136) (U = 3378.00, p > α). Spearman rank correlation was used to assess the relationship between time spent and image quality metrics. Contrary to expectations, no significant positive correlations were found. Instead, significant negative correlations were observed between time spent and lighting quality (active time: $\rho = -0.22$, p < α; total time: $\rho = -0.23$, p < α). Longer sessions were associated with poorer lighting outcomes, potentially indicating that users who struggled with the task spent more time without achieving better results.

4 Discussion

This section interprets the findings of the comparative evaluation of image quality feedback mechanisms for citizen science fossil photography. The results are discussed in relation to the research question, prior work on feedback timing and usability, and the practical constraints of real-world citizen science workflows. In addition, potential sources of bias and methodological limitations are addressed, followed by implications for future research and application development.

Interpretation of image quality outcomes

The primary research question asks whether fossil images captured with real-time feedback have a different image quality from those captured without guidance. The results show that real-time feedback produced a statistically significant improvement for contrast quality, while no significant differences were found for lighting, sharpness, scale presence, or photographed angles. The observed improvement in contrast quality suggests that real-time guidance is effective for image characteristics that depend on immediate spatial adjustments, such as repositioning the specimen or altering the camera angle relative to the background. In contrast to lighting and sharpness, which are often limited by environmental conditions or the device settings, contrast can be directly improved through user actions when prompted during the capture. Post-capture feedback did not result in statistically significant improvements across any of the evaluated metrics. This outcome suggests that delayed feedback may be less effective in guiding users to take better images. When feedback is provided only after an image has been taken, users must interrupt their workflow, reinterpret quality metrics, and decide whether corrective action is needed. Prior research in human–computer interaction shows that delayed feedback increases cognitive load and reduces task efficiency, particularly for less technologically skilled users (Nielsen, 1994 & Norman, 2013). The absence of significant effects for lighting and sharpness across feedback conditions could demonstrate boundary conditions of real-time guidance. Modern smartphone cameras already perform automatic exposure and focus adjustments. This is potentially limiting the small gains achievable through user instruction alone (Hasinoff et al., 2016). Additionally, environmental factors such as ambient lighting and physical stability are not always under the user's control in field-based settings.

Feedback timing and user interaction

The contrast between post-capture and real-time feedback highlights the role of feedback timing in interactive systems. Real-time feedback aligns with established usability principles that emphasize immediate and continuous system responses to user actions (Nielsen, 1994). By providing guidance during the capture process, real-time feedback helps users avoid having to remember earlier instructions or waiting for evaluations to figure out their next steps. This effect is important in citizen science, where participants have different levels of experience and knowledge. Providing immediate feedback can help reduce barriers to participation by guiding users when they make decisions. In contrast, giving feedback later requires careful thought, which may be harder for non-expert contributors to understand. Real-time feedback can still be valuable, even if the changes in image quality are small. While the improvements in clarity and ease of use might seem minor, they can help build confidence and lead to better submissions over time.

Usability and perceived effectiveness

The usability evaluation revealed a statistically significant increase in System Usability Scale (SUS) scores for the real-time feedback prototype compared to the baseline condition. No statistically significant usability difference was observed between the baseline and post-capture feedback prototypes. These findings indicate that participants perceived real-time feedback as more intuitive and supportive, despite its increased functional complexity. This suggests that usability is not determined solely by interface simplicity, but by how well feedback aligns with user expectations and task flow. Continuous guidance appears to reduce uncertainty during image capture, resulting in a smoother in-

teraction experience. The divergence between usability outcomes and objective image quality metrics underscores an important distinction: perceived usability and measurable performance improvements do not necessarily co-vary. In applied systems such as citizen science platforms, perceived usability may play a critical role in sustained engagement, even when immediate performance gains are limited.

Time-on-task and image quality

Contrary to expectations, no positive relationship was observed between time spent on image capture and image quality. Instead, weak but statistically significant negative correlations were found between time-based measures and lighting quality. One interpretation is that increased time spent reflects uncertainty or difficulty rather than deliberate improvement. Users who struggle to achieve good results may spend more time adjusting settings or retaking images without clear guidance. Prior usability research suggests that long task duration can be an indicator of interaction friction rather than thoroughness (Faudzi et al., 2024 & Norman, 2013). These findings further support the role of real-time feedback in reducing inefficient trial-and-error behavior by helping users converge more quickly toward better image quality.

Non-significant findings and boundary conditions

Several hypotheses were not supported, particularly those related to lighting, sharpness, scale inclusion, and viewing angles. These non-significant findings should be interpreted as indicators of the current boundaries of the proposed feedback approach rather than as evidence of ineffectiveness. Lighting and sharpness are influenced by external factors such as device hardware, environmental conditions, and user steadiness, which cannot be fully mitigated through software-based guidance (Kosmala et al., 2016). Similarly, the inclusion of scale and appropriate angles may depend more on user awareness and domain understanding than on automated visual feedback alone. The automated image quality metrics used in this study measure specific and measurable aspects of image quality. However, they might not fully capture expert assessments of whether the images are suitable for research.

Generalizing to other domains

This study focuses on fossil photography, but its findings apply to other citizen-science fields that use photographic data. The real-time feedback system, especially the indicators for lighting, sharpness, and contrast, can be used in biodiversity monitoring, ecological surveys, astronomy, and environmental observation. These areas also face similar issues with visual quality that affect scientific usability. The prototype's feedback system is modular. This means that only elements specific to fossils, like scale inclusion or required viewing angles, need to be modified to fit other domains. For example, observations of plants might require clear visibility of leaves instead of a size reference, while wildlife photography may prioritize unobstructed views or the size of the subject. Overall, the advantages of real-time feedback go beyond paleontology. They align with broader human-computer interaction principles that help contributors during the image capture process. Thus, the approach explored not only addresses issues in fossil photography but also provides a useful framework for improving image quality in various citizen-science applications.

Limitations and potential sources of bias

Several limitations and potential sources of bias should be considered when interpreting the findings of this study. First, the within-subject experimental design required all participants to evaluate all three prototypes. While this design enabled direct comparison between feedback conditions and reduced inter-participant variability, it may also have introduced comparative bias. Participants were aware that they were evaluating multiple versions of the same system, which may have encouraged relative rather than absolute judgments, particularly in usability ratings.

Second, the intended balanced randomization of prototype order failed during the final data collection phase, resulting in the same testing sequence for 82,6% of participants (Baseline → Real-Time Feedback → Post-Capture Feedback). This ordering may have introduced learning effects or contrast effects between conditions. Although the real-time feedback prototype was evaluated before the post-capture prototype, prior exposure to feedback mechanisms may still have influenced subsequent interactions.

Third, the participant sample was not fully representative of the target population of fossil-finding citizen scientists. A large portion of participants consisted of ICT students, lecturers, and project stakeholders, who may possess higher technical proficiency and greater familiarity with digital interfaces than typical users of citizen science platforms. While efforts were made to recruit fossil hunters and citizen scientists, their representation in the final dataset was limited, potentially affecting the generalization of usability findings.

Fourth, participants were aware that they were participating in a research study, which may have influenced their behavior and attentiveness (Hawthorne effect). Finally, the evaluation focused on short-term interactions with prototype systems rather than long-term use, limiting insight into learning effects and long-term behavioral change.

Future work

Several directions for future research follow directly from the limitations of this study. First, the experimental design can be improved by introducing randomized or counterbalanced ordering of the prototypes. This would reduce learning and order effects and allow a clearer separation of the impact of feedback timing from participant familiarity with the task.

Second, future studies should involve a broader and more representative participant group. While this study primarily included ICT students and lecturers, further evaluations should focus on active citizen scientists and fossil hunters exclusively. Testing the system with users who regularly collect fossil data in real-world settings would improve ecological validity and provide insights into practical constraints not observed in controlled environments.

Third, the long-term effects of image quality feedback are still unexplored. Future work should investigate whether repeated exposure to real-time feedback leads to sustained improvements in image quality or learning effects over time. Longitudinal studies could assess whether users internalize quality guidelines and need less guidance as experience increases.

Fourth, the feedback mechanisms themselves can be refined. The current system relies mainly on textual and visual instructions. Future versions could incorporate fossil image examples, other overlays, or adaptive feedback that responds to repeated user errors. More explicit guidance for scale inclusion and viewing angle may be necessary, as these aspects did not improve in the current study.

Finally, the image quality assessment approach can be expanded. Additional features, such as automatic detection of scale objects, fossil orientation, or occlusion, may better reflect scientific usability. Combining these features with AI performance measures could help clarify the connections between capture guidance, image quality, and model accuracy. Together, these improvements would support a more robust evaluation of real-time feedback systems and their role in improving data quality within citizen science projects.

Conclusion

This study examined whether feedback on image quality can improve fossil photographs submitted by citizen scientists. Three prototype systems were compared: no feedback, post-capture feedback, and real-time feedback during image capture. The real-time feedback prototype produced a statistically significant improvement in image contrast and achieved higher usability scores compared to the baseline condition. However, no significant improvements were observed for lighting, sharpness, scale inclusion, or viewing angle across any feedback condition. Post-capture feedback did not

result in significant changes in image quality or usability. Additional analyses revealed that neither time spent on image capture nor flashlight usage correlated positively with image quality outcomes, suggesting that increased time or additional tools alone are insufficient without targeted guidance. These findings indicate that feedback timing influences effectiveness, and that guidance provided during image capture is more effective than feedback delivered afterward. More broadly, the results highlight the challenge of supporting non-expert contributors in producing consistent, high-quality photographic data for scientific use. While real-time feedback demonstrates potential for improving specific aspects such as contrast and perceived usability, its effects remain limited in scope. The study's findings should be interpreted within the context of several methodological constraints, including the controlled experimental setting, the specific set of image quality metrics evaluated, and the short-term nature of user interaction with the prototypes. Future research should investigate more sophisticated feedback mechanisms, assess their impact across a broader range of quality dimensions, and evaluate their effectiveness in authentic field conditions with diverse participant populations. In summary, real-time feedback can support certain aspects of citizen-generated fossil imagery, but substantial opportunities remain for developing more comprehensive and effective guidance systems to improve data quality in paleontological citizen science.

References

Abdul-Rahman, Zwitter, and Haleem (2025). "A systematic literature review on the role of artificial intelligence in citizen science". In: *AI and Society*. DOI: https://doi.org/10.1007/s44163-025-00437-z. URL: https://link.springer.com/article/10.1007/s44163-025-00437-z.

Brooke, John (Nov. 1995). "SUS: A quick and dirty usability scale". In: *Usability Eval. Ind.* 189.

Dodge, Samuel and Lina Karam (2016). "Understanding how image quality affects deep neural networks". In: *2016 eighth international conference on quality of multimedia experience (QoMEX)*. IEEE, pp. 1–6.

Eijkelboom, Isaak et al. (2024). "Making sense of fossils and artefacts: a review of best practices for the design of a successful workflow for machine learning-assisted citizen science projects". In: *PeerJ*. DOI: 10.7717/peerj.18927. URL: https://peerj.com/articles/18927.

Faudzi, Masyura Ahmad et al. (2024). "User interface design in mobile learning applications: Developing and evaluating a questionnaire for measuring learners' extraneous cognitive load". In: *Heliyon* 10.18, e37494. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2024.e37494. URL: https://www.sciencedirect.com/science/article/pii/S2405844024135256.

Google (2025). *How AI-powered Camera Coach on Pixel 10 Helps Take Amazing Photos*. Accessed: 2025-10-06. URL: https://store.google.com/intl/en/ideas/articles/camera-coach.

Hasinoff, Samuel et al. (Nov. 2016). "Burst photography for high dynamic range and low-light imaging on mobile cameras". In: *ACM Transactions on Graphics* 35, pp. 1–12. DOI: 10.1145/2980179.2980254.

iNaturalist (2023). *Creating High-Quality iNaturalist Observations*. Accessed: 2025-10-06. URL: https://www.inaturalist.org/posts/80155-creating-high-quality-inaturalist-observations.

International Telecommunication Union (2011). *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*. Recommendation ITU-R BT.601-7. International Telecommunication Union. Geneva, Switzerland. URL: https://www.itu.int/rec/R-REC-BT.601/.

Kosmala, Margaret et al. (2016). "Assessing data quality in citizen science". In: *Frontiers in Ecology and the Environment* 14.10, pp. 551–560. DOI: https://doi.org/10.1002/fee.1436. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/fee.1436. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/fee.1436.

Li, Yi-Feng, Chuan-Kai Yang, and Yi-Zhen Chang (2020). "Photo Composition with Real-Time Rating". In: *Sensors* 20.3, p. 582. DOI: 10.3390/s20030582. URL: https://www.mdpi.com/1424-8220/20/3/582.

Liu, Xiaokang et al. (2023). "Automatic taxonomic identification based on the Fossil Image Dataset and deep convolutional neural networks". In: *Paleobiology* 49.1, pp. 1–22. DOI: 10.1017/pab.2022.14. URL: https://www.cambridge.org/core/journals/paleobiology/article/automatic-taxonomic-identification-based-on-the-fossil-image-dataset-415000-images-and-deep-convolutional-neural-networks/4863E2FDE20D6115415EE5FE232B9DCD.

López-Guillén, Eduard et al. (Jan. 2024). "Strengths and Challenges of Using iNaturalist in Plant Research with Focus on Data Quality". In: *Diversity* 16, p. 42. DOI: 10.3390/d16010042.

Lotfian, Maryam, Jens Ingensand, and Maria Antonia Brovelli (2021). "The Partnership of Citizen Science and Machine Learning: Benefits, Risks, and Future Challenges for Engagement, Data Collection, and Data Quality". In: *Sustainability* 13.14. ISSN: 2071-1050. DOI: 10.3390/su13148087. URL: https://www.mdpi.com/2071-1050/13/14/8087.

Naturalis Biodiversity Center (2025). *Oervondstchecker*. Accessed: 2025-10-06. URL: https://www.oervondstchecker.nl.

Nielsen, Jakob (1994). *Usability Engineering*. Morgan Kaufmann.

Norman, Don (2013). *The Design of Everyday Things*. Basic Books.

Sharma, Nirwan et al. (Dec. 2024). "Image Recognition as a "Dialogic AI Partner" Within Biodiversity Citizen Science—an empirical investigation". In: *Citizen Science: Theory and Practice*. DOI: 10.5334/cstp.735.

Silvertown, Jonathan (2009). "A new dawn for citizen science". In: *Trends in Ecology Evolution* 24.9, pp. 467–471. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2009.03.017. URL: https://www.sciencedirect.com/science/article/pii/S016953470900175X.

Sun, Jiarui et al. (2024). "Automatic identification and morphological comparison of bivalve and brachiopod fossils based on deep learning". In: *PeerJ*. DOI: https://doi.org/10.7717/peerj.16200. URL: https://peerj.com/articles/16200.

Wal, René van der et al. (2018). "The role of automated feedback in training and retaining biological recorders for citizen science". In: *Conservation Biology*. DOI: https://doi.org/10.1111/cobi.12705. URL: https://conbio.onlinelibrary.wiley.com/doi/full/10.1111/cobi.12705.

Xu, Yan et al. (Apr. 2015). "Real-time Guidance Camera Interface to Enhance Photo Aesthetic Quality". In: pp. 1183–1186. DOI: 10.1145/2702123.2702418.

Yaqoob, Mohammed et al. (2024). "Advancing paleontology: a survey on deep learning methodologies in fossil image analysis". In: *Artificial Intelligence Review*. DOI: https://doi.org/10.1007/s10462-024-11080-y. URL: https://link.springer.com/article/10.1007/s10462-024-11080-y.

Yu, Congyu et al. (2024). "Artificial intelligence in paleontology". In: *Earth-Science Reviews* 252, p. 104765. ISSN: 0012-8252. DOI: https://doi.org/10.1016/j.earscirev.2024.104765. URL: https://www.sciencedirect.com/science/article/pii/S0012825224000928.

A Supplemental Material

A priori power analysis

The following table summarizes the assumptions entered into G Power (left and middle columns) and explains their specific relevance to the study design (right column). Together, these values define the structure of the statistical test and determine the total number of participants required to achieve the desired level of statistical power. Based on these parameters, G Power calculated that a total sample size of 28 participants is needed to achieve 80% statistical power. Figure 6 shows the power analysis within the tool G Power.

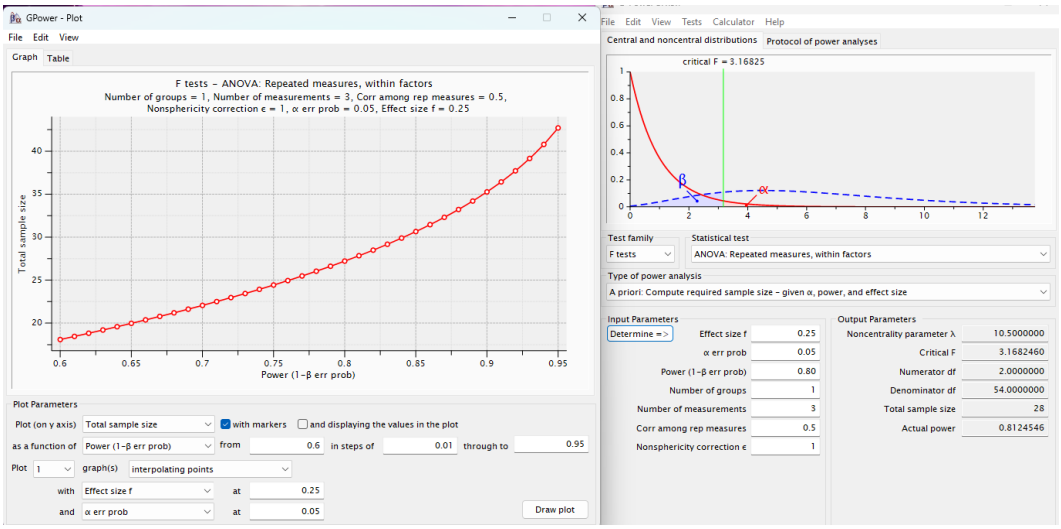


Figure 6: G Power analysis configuration and output

Descriptive Statistics by Prototype

The following table and figure provide a detailed overview of image quality metrics across the three prototypes. Table 1 summarizes the mean, standard deviation, and median values for each metric, while Figure 7 visualizes the distributions using boxplots.

Table 1: Descriptive statistics for image quality metrics across the three prototypes.

Metric	**Prototype**	**Mean (SD)**	**Median**
Lighting	Baseline	141.84 (29.83)	145.84
	Post-Capture	143.71 (26.66)	145.18
	Real-Time	145.34 (24.49)	145.17
Sharpness	Baseline	264.29 (349.06)	147.27
	Post-Capture	240.56 (233.37)	183.34
	Real-Time	225.73 (254.57)	159.58
Contrast	Baseline	27.48 (21.76)	18.63
	Post-Capture	33.42 (23.98)	31.85
	Real-Time	38.43 (20.84)	41.88
Scale Rating	Baseline	4.27 (1.41)	5.00
	Post-Capture	4.19 (1.41)	5.00
	Real-Time	4.38 (1.32)	5.00
Angle Rating	Baseline	4.18 (1.07)	5.00
	Post-Capture	4.21 (0.97)	5.00
	Real-Time	4.23 (0.91)	4.00

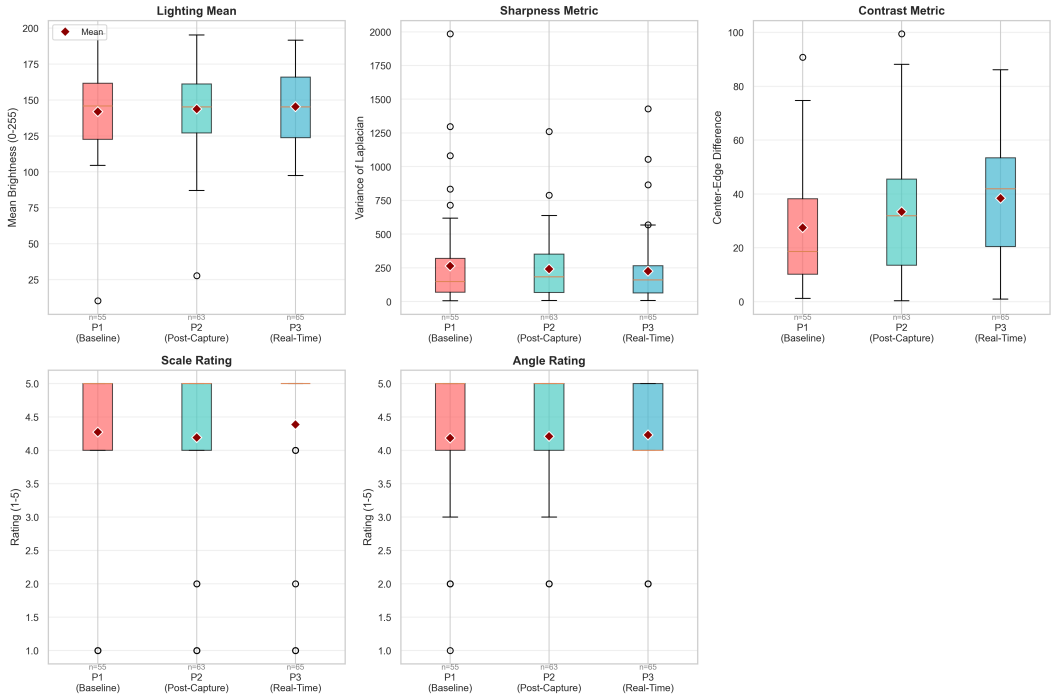


Figure 7: Distribution of image quality metrics across the three prototypes. *Boxplots show median (center line), interquartile range (box), and outliers (whiskers). Red diamonds indicate mean values. P1 (Baseline) had no real-time feedback, P2 (Post-Capture) provided feedback after image capture, and P3 (Real-Time) provided continuous feedback during capture. Sample sizes (n) are shown below each box.*

Image Quality Thresholding Results

To assess automatic grading of image quality, multiple thresholding methods were evaluated. No single method was consistently optimal across all metrics. Performance was quantified using the Mean Absolute Error (MAE) relative to expert ratings. For lighting quality, the K-means method achieved the lowest MAE (approximately 0.96), with strong agreement between predicted and expert grades. Sharpness quality was best captured using the median midpoint method (MAE $\sim$ 0.95). Contrast quality proved the most challenging, with higher error rates across all methods. The grading system showed the greatest discriminatory power at the extremes (grades 1 and 5), while middle grades (2 - 4) exhibited substantial overlap and confusion.

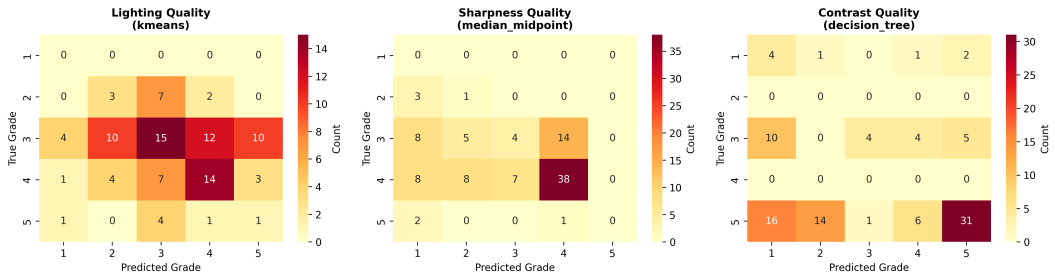


Figure 8: Confusion matrices for the three image quality metrics using their respective best-performing thresholding methods. *Each cell indicates the count of images with a given true grade (y-axis) predicted as a specific grade (x-axis). Diagonal cells represent correct classifications.*

Usability Survey Questions

The usability survey used to assess user experience is included in full below.

Fossil Photography Usability Survey

Thank you for helping us test this prototype.

Your honest input is crucial for developing a smoother, more effective process for capturing high-quality fossil images for citizen science. By 'citizen science' we mean any activity where members of the public help collect or analyse data for scientific research, for example uploading observations or photos to a science platform.)

This survey is expected to take approximately **10 -15 minutes** to complete.

Please be aware that by filling out this survey, you consent to the usage of the information provided for research purposes.

* Indicates required question

Personal Information

We ask a few background questions so we can understand whether different types of users experience the app differently. This helps us improve the design for everyone.

1. What is the username you entered in the Fossil Photography app? *

2. What age group are you in? *

Mark only one oval.

- ◯ below 18
- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ 55-64
- ◯ 65+
- ◯ Prefer not to disclose

3. What is your gender? *

Mark only one oval.

- ◯ Male
- ◯ Female
- ◯ Nonbinary
- ◯ Prefer not to disclose
- ◯ Other: _____

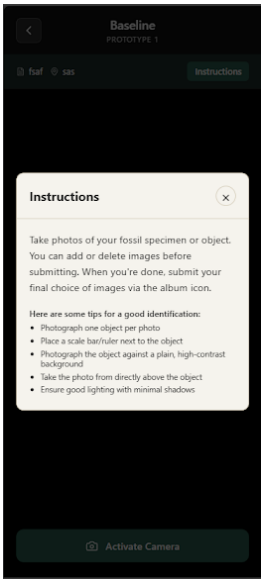
4. How often do you contribute to scientific research or citizen-science projects (e.g., by submitting images or observations of specimens)? *

Mark only one oval.

- ◯ Weekly or more often
- ◯ Monthly
- ◯ A few times per year
- ◯ Less than once per year
- ◯ Never

Evaluating Feedback Mechanisms - Prototype A - Baseline

Prototype A - Baseline



5. For Prototype A, how easy was it to capture an acceptable fossil photograph? *
(An acceptable photo is clear, well-lit, focused on the fossil with a size reference like a coin.)

Mark only one oval.

- ◯ Very difficult
- ◯ Somewhat difficult
- ◯ Neither easy nor difficult
- ◯ Somewhat easy
- ◯ Very easy

6. For Prototype A, how clear were the instructions? *

Mark only one oval.

- ◯ Very unclear
- ◯ Somewhat unclear
- ◯ Neither clear nor unclear
- ◯ Somewhat clear
- ◯ Very clear

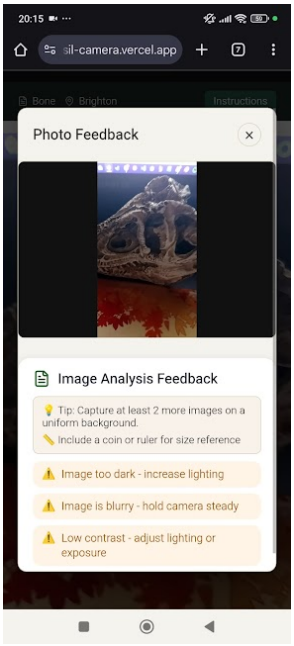
7. I felt motivated to adjust my photograph based on the guidance provided by Prototype A *

Mark only one oval.

- ◯ Strongly disagree
- ◯ Somewhat disagree
- ◯ Neither agree nor disagree
- ◯ Somewhat agree
- ◯ Strongly agree

Evaluating Feedback Mechanisms - Prototype B - Post Capture Feedback

Prototype B - Post Capture Feedback



8. For Prototype B, how easy was it to capture an acceptable fossil photograph? *
(An acceptable photo is clear, well-lit, focused on the fossil with a size reference like a coin.)

Mark only one oval.

- Very difficult
- Somewhat difficult
- Neither easy nor difficult
- Somewhat easy
- Very easy

9. For Prototype B, how clear were the instructions? *

Mark only one oval.

- Very unclear
- Somewhat unclear
- Neither clear nor unclear
- Somewhat clear
- Very clear

10. For Prototype B, how helpful was the feedback in improving your photo? *

Mark only one oval.

- Very unhelpful
- Somewhat unhelpful
- Neither helpful nor unhelpful
- Somewhat helpful
- Very helpful

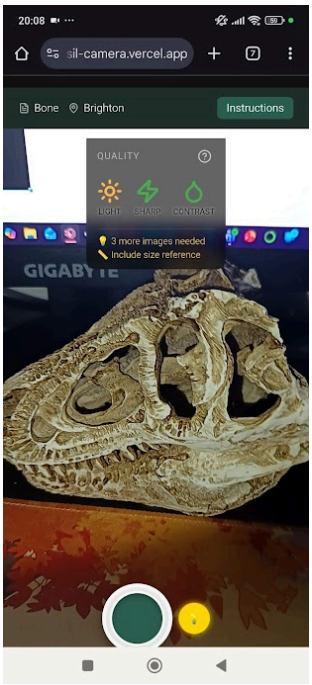
11. I felt motivated to adjust my photograph based on the guidance provided by Prototype B *

Mark only one oval.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Evaluating Feedback Mechanisms - Prototype C - Real Time Feedback

Prototype C - Real Time Feedback



12. For Prototype C, how easy was it to capture an acceptable fossil photograph? *
(An acceptable photo is clear, well-lit, focused on the fossil with a size reference like a coin.)

Mark only one oval.

- Very difficult
- Somewhat difficult
- Neither easy nor difficult
- Somewhat easy
- Very easy

13. For Prototype C, how clear were the instructions? *

Mark only one oval.

- Very unclear
- Somewhat unclear
- Neither clear nor unclear
- Somewhat clear
- Very clear

14. For Prototype C, how helpful was the feedback in improving your photo? *

Mark only one oval.

- Very unhelpful
- Somewhat unhelpful
- Neither helpful nor unhelpful
- Somewhat helpful
- Very helpful

15. I felt motivated to adjust my photograph based on the guidance provided by Prototype C. *

Mark only one oval.

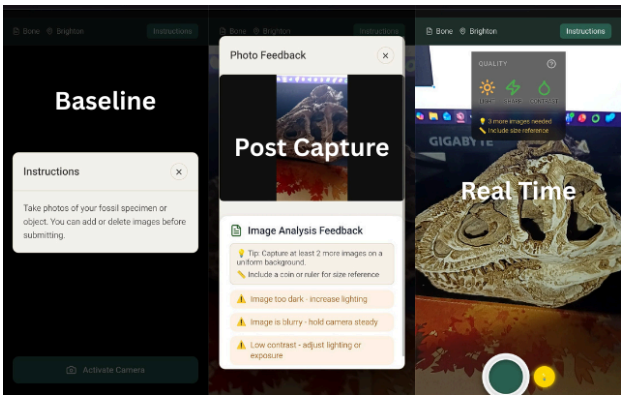
- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

User Experience and Engagement

Please evaluate the following statements for each camera prototype you tested.

1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree

Prototypes
A = Baseline, B = Post Capture, C = Real Time



16. I think that I would like to use this prototype frequently. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A	◯	◯	◯	◯	◯
B	◯	◯	◯	◯	◯
C	◯	◯	◯	◯	◯

17. I found the prototype unnecessarily complex. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A	◯	◯	◯	◯	◯
B	◯	◯	◯	◯	◯
C	◯	◯	◯	◯	◯

18. I thought the prototype was easy to use. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A	◯	◯	◯	◯	◯
B	◯	◯	◯	◯	◯
C	◯	◯	◯	◯	◯

19. I think that I would need the support of a technical person to be able to use this prototype. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A					
B					
C					

20. I found the various functions in this prototype were well integrated. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A					
B					
C					

21. I thought there was too much inconsistency in this prototype. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A					
B					
C					

22. I would imagine that most people would learn to use this prototype very quickly. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A					
B					
C					

23. I found the prototype very difficult to use. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A					
B					
C					

24. I felt very confident using the prototype. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A					
B					
C					

25. I needed to learn a lot of things before I could get going with this prototype. *
1 = Strongly disagree,
2 = Somewhat disagree
3 = Neither agree nor disagree
4 = Somewhat agree
5 = Strongly agree
A = Baseline, B = Post Capture, C = Real Time

Mark only one oval per row.

	1	2	3	4	5
A					
B					
C					

Thank you for testing this fossil photography prototype!

Your feedback is highly appreciated.