



# 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](mailto: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 (Silvertown, 2009; Abdul-Rahman, Zwitter, and Haleem, 2025). 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 (López-Guillén et al., 2024), (Eijkelboom et al., 2024).

Image-based submissions are now included in many citizen science platforms, because they allow 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 (Wal et al., 2018; Sharma et al., 2024). In related domains, real-time feedback has been shown to improve photographic outcomes by guiding users during image capture rather than after the fact (Xu et al., 2015; Li, Yang, and Chang, 2020). 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 (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. 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 1.

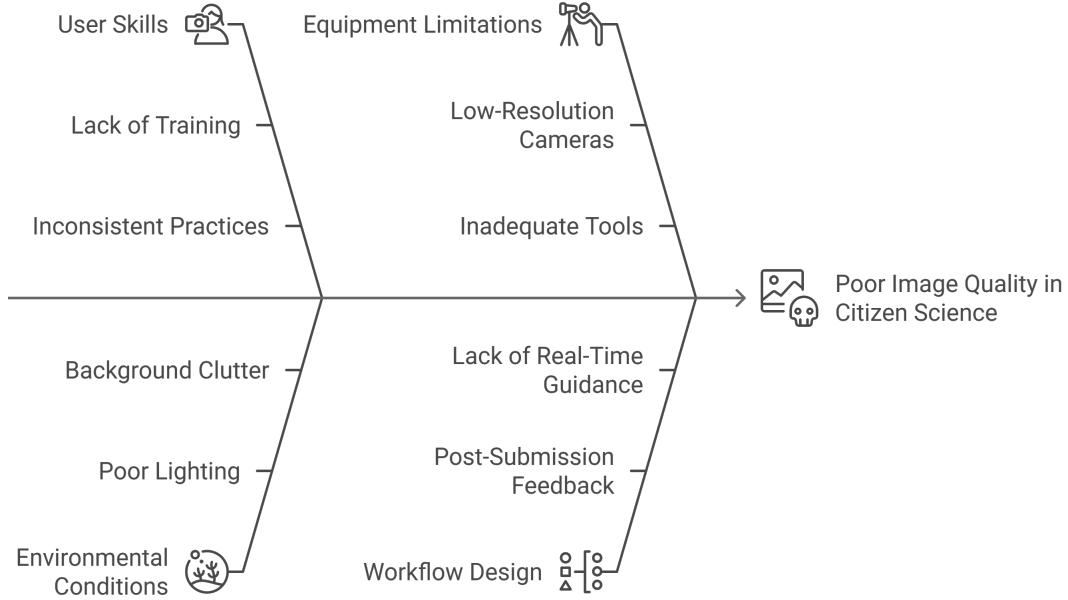


Figure 1: Image quality problems in paleontological citizen science

## 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.

### 2.1 Prototype development

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. Existing applications like Google Lens and Rock Identifier were analyzed to create an intuitive and familiar user experience. The visual styling is intentionally minimalist. 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.

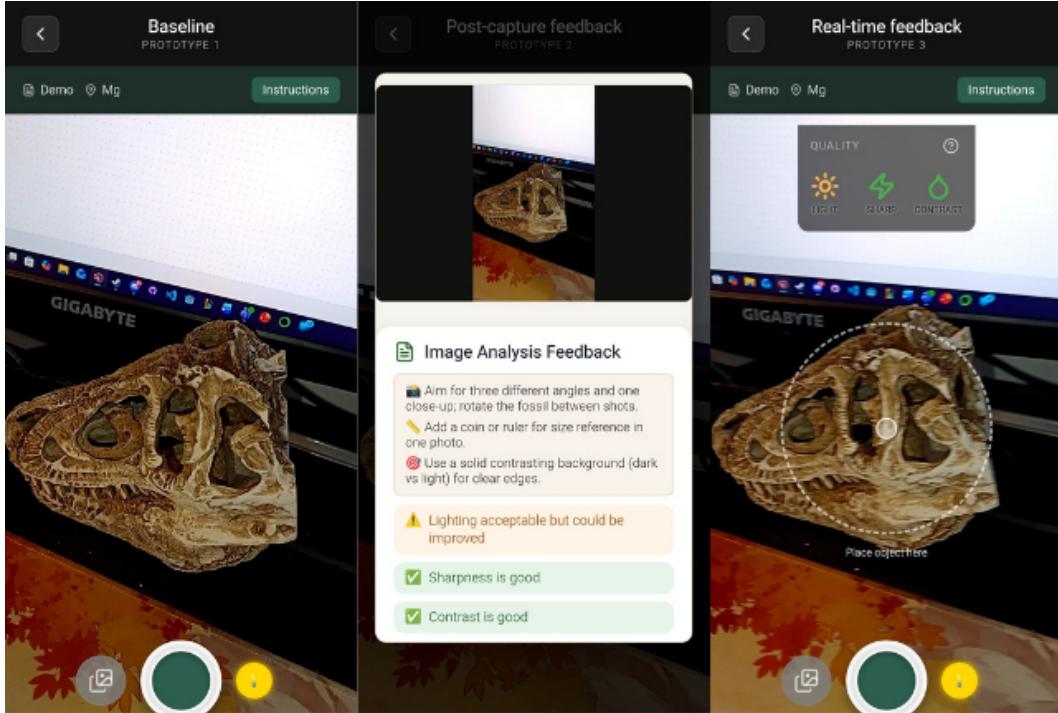


Figure 2: Prototypes

## 2.2 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 a paleontologist. 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 ( $0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$ ), 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. The whole process is visualized in Figure 3.

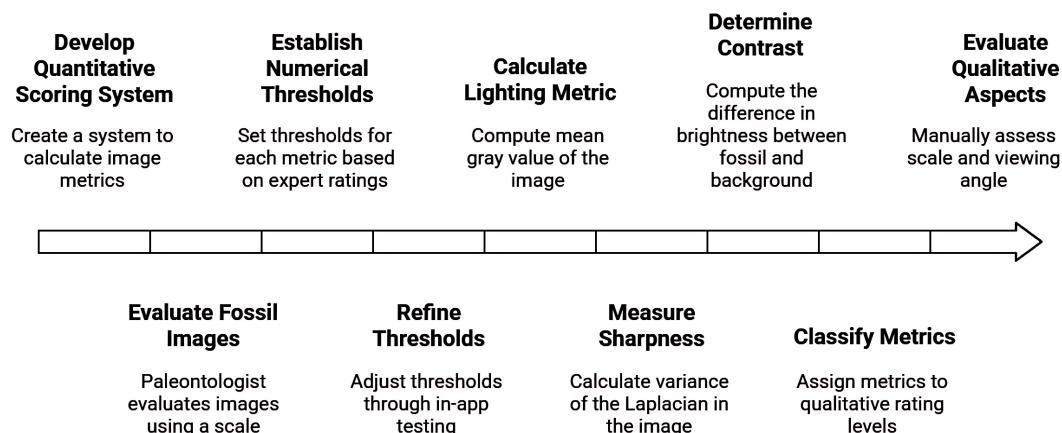


Figure 3: Image quality measurement process

### 2.3 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).

## 2.4 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. The two process evaluation is visualized in Figure 4.

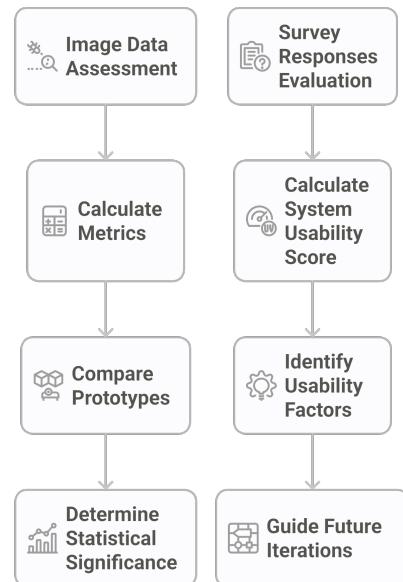


Figure 4: Two-part evaluation process

## 3 Results

A total of 183 images were included in the analysis, distributed across the three prototype conditions: Baseline ( $n = 55$ ), Post-Capture Feedback ( $n = 63$ ), and Real-Time Feedback ( $n = 65$ ).

Five image quality indicators were evaluated for each submission: lighting, sharpness, and contrast (automated metrics), and scale rating and angles rating (manual ratings on a 1–5 scale).

### 3.0.1 Statistical testing

Assumptions of normality and homogeneity of variance were assessed using Shapiro–Wilk and Levene’s tests. Multiple violations of parametric assumptions were observed, and non-parametric statistical tests were therefore applied.

Kruskal–Wallis  $H$  tests were conducted to assess overall differences between prototype conditions for each image quality metric. Where applicable, pairwise Mann–Whitney  $U$  tests were performed comparing the Baseline prototype with each feedback condition. A Bonferroni-corrected significance threshold of  $\alpha = 0.025$  was applied.

### 3.0.2 Hypothesis testing outcomes

Table 1 presents a summary of the hypothesis testing results. For comparisons between the **Baseline** and **Post-Capture Feedback** prototypes ( $H_1$ ), no statistically significant differences were observed

for lighting, sharpness, contrast, scale rating, or angles rating. The contrast metric showed a non-significant trend ( $U = 1470.00, p = 0.079$ ).

For comparisons between the **Baseline and Real-Time Feedback** prototypes ( $H_2$ ), a statistically significant difference was observed for the contrast metric ( $U = 1236.00, p = 0.002$ ). No statistically significant differences were found for lighting, sharpness, scale rating, or angle rating.

Hypothesis	Test Type	Result	Statistic	p-value
$H_1: P1 < P2$ (Lighting)	Mann–Whitney $U$	Not Significant	—	—
$H_1: P1 < P2$ (Sharpness)	Mann–Whitney $U$	Not Significant	—	—
$H_1: P1 < P2$ (Contrast)	Mann–Whitney $U$	Not Significant (trend)	$U = 1470.00$	$= 0.079$
$H_1: P1 < P2$ (Scale Rating)	Mann–Whitney $U$	Not Significant	—	—
$H_1: P1 < P2$ (Angles Rating)	Mann–Whitney $U$	Not Significant	—	—
$H_2: P1 < P3$ (Lighting)	Mann–Whitney $U$	Not Significant	—	—
$H_2: P1 < P3$ (Sharpness)	Mann–Whitney $U$	Not Significant	—	—
$H_2: P1 < P3$ (Contrast)	<b>Mann–Whitney</b> $U$	<b>Significant</b>	$U = 1236.00$	<b><math>0.002^{**}</math></b>
$H_2: P1 < P3$ (Scale Rating)	Mann–Whitney $U$	Not Significant	—	—
$H_2: P1 < P3$ (Angles Rating)	Mann–Whitney $U$	Not Significant	—	—

Table 1: Summary of hypothesis testing results for primary prototype comparisons. P1 = Baseline, P2 = Post-Capture Feedback, P3 = Real-Time Feedback. \*\* indicates  $p < 0.01$ .

### 3.0.3 Behavioral variables

Hypothesis	Test Type	Result	Statistic	p-value
$H_3: \text{Flashlight ON} > \text{OFF}$ (Lighting)	Mann–Whitney $U$	Not Significant	$U = 3378.00$	0.562
$H_4: \text{Active Time} \rightarrow \text{Quality}$	Spearman $\rho$	Not Significant	$\rho = -0.215$	<b><math>0.004^{**}</math></b>
$H_4: \text{Total Time} \rightarrow \text{Quality}$	Spearman $\rho$	Not Significant	$\rho = -0.225$	<b><math>0.002^{**}</math></b>

Table 2: Summary of hypothesis testing results for behavioral and time-related factors. \*\* indicates  $p < 0.01$ .

Flashlight usage was analyzed for its association with lighting quality. No statistically significant difference was found between images captured with the flashlight enabled and disabled ( $U = 3378.00, p = 0.562$ ).

Spearman rank correlation analyses were conducted to assess relationships between time-based behavioral variables and image quality metrics. No positive correlations were observed. Negative correlations were found between time spent and lighting quality ( $\rho = -0.215$  to  $-0.225, p < 0.005$ ).

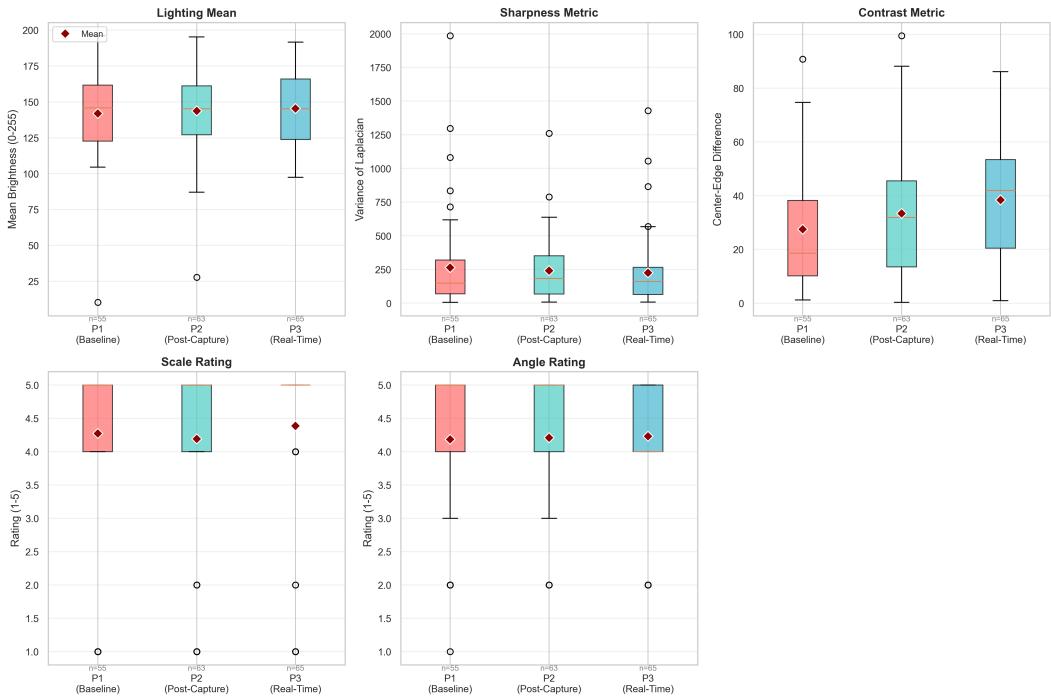


Figure 5: Distribution of image quality metrics across the three prototypes. Boxplots show median (center line), interquartile range (box), and outliers (whiskers). Red diamonds indicate mean values. P1 (Baseline) had no real-time feedback, P2 (Post-Capture) provided feedback after image capture, and P3 (Real-Time) provided continuous feedback during capture. Sample sizes ( $n$ ) are shown below each box.

### 3.1 Prototype usability test results

Prototype usability was evaluated using the System Usability Scale (SUS) based on responses from 20 participants. Each participant evaluated all three prototypes. Due to the repeated-measures design and non-normal score distributions, pairwise Wilcoxon signed-rank tests were conducted to compare the Baseline prototype with each feedback condition. A Bonferroni-corrected significance threshold of  $\alpha = 0.025$  was applied. The comparison between the Baseline and Post-Capture Feedback prototypes yielded a Wilcoxon  $W$  statistic of 24.5 with  $p = 0.0777$ . This result did not meet the corrected significance threshold. The observed difference in SUS scores was 4.1 points. The corresponding rank-biserial effect size was  $r = 0.883$ . The comparison between the Baseline and Real-Time Feedback prototypes yielded a Wilcoxon  $W$  statistic of 4.0 with  $p = 0.0006$ . This result met the corrected significance threshold. The observed difference in SUS scores was 14.4 points. The corresponding rank-biserial effect size was  $r = 0.981$ .

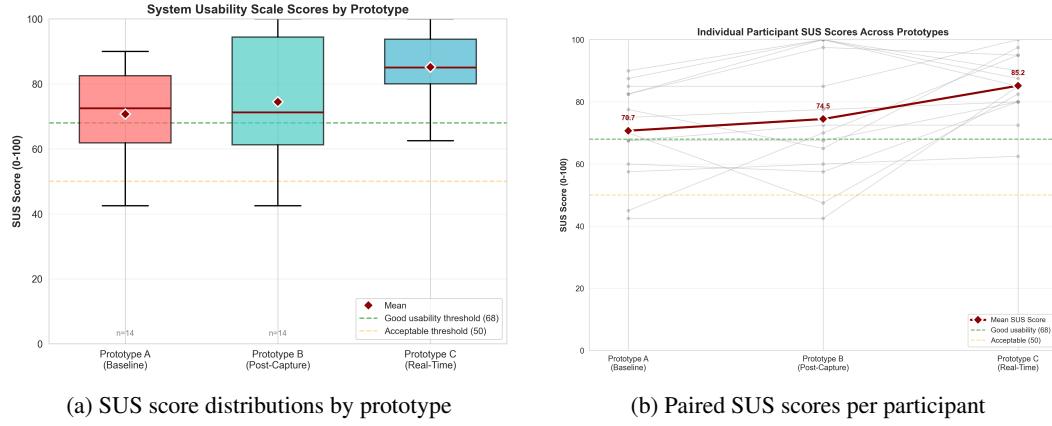


Figure 6: System Usability Scale (SUS) results across prototype conditions

### 3.2 Summary of Results

Statistically significant differences in image quality were observed only for contrast when comparing the Real-Time Feedback prototype with the Baseline condition. No statistically significant differences were observed for lighting, sharpness, scale rating, or angles rating across prototype conditions. Usability scores differed significantly between the Baseline and Real-Time Feedback prototypes, but not between the Baseline and Post-Capture Feedback prototypes.

## 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.

### 4.1 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 (Norman, 2013; Nielsen, 1994).

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. Additionally, environmental factors such as ambient lighting and physical stability are not always under the user's control in field-based settings.

#### **4.2 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.

#### **4.3 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 interaction 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.

#### **4.4 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 (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.

#### **4.5 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. 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.

#### **4.6 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 generalizability 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.

#### **4.7 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.

## 5 Conclusion

This study investigated whether feedback on image quality can improve fossil photographs submitted by citizen scientists. Three prototype systems were compared: a baseline without feedback, post-capture feedback, and real-time feedback during image capture. Image quality was evaluated using objective metrics for lighting, sharpness, and contrast, complemented by manual ratings for scale presence and viewing angle. Usability and user experience were also assessed.

The results show that real-time feedback led to a statistically significant improvement in image contrast compared to the baseline condition. It also achieved higher usability scores. No significant improvements were found for lighting, sharpness, scale inclusion, or viewing angle. Post-capture feedback did not result in significant changes in image quality or usability. These findings suggest that feedback timing might play a role, and that guidance provided during image capture is more effective than feedback delivered afterward.

The study also found no positive relationship between time spent capturing images and image quality. Flashlight usage did not significantly improve results. Together, these outcomes indicate that simply encouraging users to spend more time or add tools is insufficient without targeted, timely guidance.

Overall, this research demonstrates that real-time image quality feedback can support certain aspects of image capture in citizen science, particularly contrast and perceived usability. However, the effects are limited and do not generalize across all quality metrics. The findings highlight the need for better feedback methods and more thorough testing in real-world situations.

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