Using Log Visualizations to Interpret Online Interactions during Self-Quizzing Practice for Taxonomic Identification

Abstract: The lack of practice opportunities for citizen science trainees in taxonomic identification for water quality monitoring adds difficulty to an already complex learning task. We created a self-quizzing feature to complement an online visual teaching and learning platform for citizen scientists in aquatic macroinvertebrate identification. Here we present a preliminary study of fifteen learners whose log interactions and performance on pre- and post-study identification tests provide insights to the design of this quizzing feature. We describe our visualization platform for analyzing data and propose design directions for future work.

Introduction

Citizen science volunteers sample local freshwater bodies to compute water quality scores based on biodiversity counts of benthic macroinvertebrates (e.g. stoneflies, mayflies). Training volunteers to identify these organisms represents a significant challenge (Story 2016), as brief trainings must cover a wide body of knowledge including sampling and equipment use in addition to identification (Authors, 2018; Stepenuck 2018). Moreover, the training materials available to many volunteer organizations are limited: trainers often rely on projected or printed low-resolution images to teach key diagnostic character traits.

To address this visual resource gap, we developed an open source web-based platform of annotated, high-resolution zoomable images of commonly found aquatic macroinvertebrates. Our partner volunteer biomonitoring organizations (VBMO's) in this NSF-funded design-based research project requested we include a feature for practicing identification skills. Therefore, we integrated a self-quizzing feature (Figure 1) that randomly pulls an unlabeled specimen from our collection and asks the user to identify it to the taxonomic level required by most VBMO's, typically order (e.g. "mayfly") or suborder ("net-spinning caddisfly"). This paper describes a preliminary study of this self-quizzing feature as a means of practicing macroinvertebrate identification. The research questions guiding this work are:

RQ1: To what extent will brief daily practice lead to improvements in citizen scientists' identification skills?

RQ2: How do interaction patterns with the practice feature vary between users and across user groups?

RQ3: What internal and external identification resources support improvement in identification?

We present visualizations of log data from practice sessions in conjunction with analysis of pre- and post-tests to explore hypotheses for the factors that contributed to greater gains. Considerations for future work will be discussed.

Background

Macroinvertebrate Identification in Citizen Science

Effective water quality biomonitoring requires reliable taxonomic identification to make assessments of environmental conditions. However, the challenges of training volunteers to accurately identify benthic macroinvertebrates are well documented (Stribling et al., 2008). Nerbonne and Vondracek (2003) looked at volunteers' abilities to correctly sort and ID macroinvertebrates using a family-level identification tool, finding frequent failures to see diagnostic characters as a primary cause of misidentification.

Volunteers are significantly more likely to correctly identify a specimen's order if they have access to a visual depiction of the specimen's family (Nerbonne and Vondracek, 2003). By building a rich visual resource of taxonomic orders and families, we have the potential to improve the accuracy of citizen science data while simultaneously boosting volunteers' confidence and engagement in the task.

Training Citizen Science Volunteers

While training curricula and protocols vary by region and organization, many groups introduce key taxonomic groups during a classroom lecture, followed by hands-on practice identifying specimens either in a lab or at a local waterway (Authors, 2018).

Lectures typically present representative specimens from each of the major taxonomic groups, but variations in size, shape, and color within groups are impossible to fully convey in a brief training (see Figure 2). Fluency in identification comes through exposure to a variety of specimens and verification from a teacher or more expert peer. This apprenticeship often occurs during streamside collection at trainings. Unfortunately, infrequent training and seasonal sampling make it difficult for new learners to have sufficient repeated practice to develop perceptual fluency. Research in human perception has shown that visual expertise develops not only with repeated exposure to the phenomena or object of interest/inquiry, but also tracks with increasingly precise levels of categorical abstraction (Tanaka & Taylor 1991). Our identification practice platform is designed to fill this perceptual training gap by offering opportunities to identify a variety of specimens and, just as trainers verify an identification during live practice, give immediate feedback and point out salient diagnostic features. Through this practice, novices can build robust categorical representations of macroinvertebrate orders required for citizen science.

Data Representations to Assess Learning

Digital technologies provide opportunities to use log data to glean deeper insights into a student's progress. With effective visualizations, educators can understand how learner pathways are built from a different perspective by seeing detailed interactions with the education platform, extracting patterns, and comparing the results among peers (Duval, 2011). Website logs and clickstreams have become useful sources of data in recent years.

One popular approach for analyzing clickstream data is to extract logs from the site and directly feed them into machine learning models to learn general patterns or clusters for visualization (Pardos, & Horodyskyj, 2019; Furr, 2019; Park, 2017). Other work has taken an approach of visualizing user interactions with less stress on grouping but more emphasis on users' change of behavior over time, using color coding and time dimension to indicate the

learning pathways (Cadez, 2000). We employ the latter approach to assess how individual participants interacted with the site through the study period.

Methods

Study Design

The study took place online in the spring of 2020. Participants completed pre- and post-study questionnaires covering background knowledge and experience with identification and conducted self-directed practice sessions with the quiz feature over 10 days. Pre- and post-study identification tests comprised ten benthic macroinvertebrate specimen photos, five of which were repeated between tests. For each specimen, subjects were asked to name the macroinvertebrate, list the physical features that helped them identify it, and report their confidence in that identification (from "Not confident at all" to "Extremely confident"). Subjects were instructed not to use additional identification resources during the pre- and post-tests but were allowed to use resources during practice.

Practice sessions were tracked via a custom URL for each participant. Participants were instructed to spend at least five minutes and identify at least five specimens each day for seven out of ten days. After daily practice, they were required to complete a survey reporting any additional resources they used and optionally give feedback or comments about the day's practice. Upon completion of all activities, subjects received a $25 gift card. The study was overseen by the IRB of the first author.

Recruitment and Participants

Subjects were recruited from past participants of training workshops hosted by project partner VBMO's during this research grant. A total of 22 individuals responded to the study invitation. Nineteen participants consented to participate and completed the pre-study survey. Fifteen subjects completed the full study and are included in this analysis.

Visualization Platform

After subjects finished daily practices, log data were extracted from the server by participant IDs. We used Python to clean the logs and extract JavaScript actions and timestamps into understandable data frames. Data frames were then reformatted into multilayer JSON form in order to build the visualizations using D3.js.

We designed a multi-day timeline visualization which reflects the users interaction patterns across the ten-day period. Each row shows the progress of one subject in one day, including how much time they spent answering a question or inspecting the answer, and whether they got the right answer. We upgraded the chart by creating filters for participant IDs, days, expected versus chosen categories, and user groups to enable more thorough investigation of data. See Figure 3.

Results

Test responses were coded by the study team with consultation from entomology experts. Our coding scheme captured both correctness and level of taxonomic specificity (Table 1). The *lowest correct level* (LCL) was determined for each answer. The fifteen subjects were sorted into groups according to their performance on the tests, as demonstrated by the LCL they reached on each specimen. See Figure 4.

RQ1: Improvements in Identification

This practice quiz feature and the study design were developed for novices. Any gains made by the High and Middle groups during practice were not measurable by our assessments. The six novices (Figure 4, right) showed mixed improvement between pre- and post-test LCLs. All but one novice showed improvement on the paired images, moving from either *no guess* or *incorrect* to a correct answer on one or more of the five paired specimens. The one novice who did not get a paired specimen correct on the posttest (SQ20) did move from *no guess* for three pre-test specimens to plausible (though incorrect) order level guesses on the post-test. With this small preliminary sample we do not aim to make statistical claims; instead we seek to identify practice habits we can support through iterations in the design that show promise to improve performance. RQ2 and RQ3 address these aims.

RQ2: Differences in Interaction Patterns during Practice

We examined daily log files from each group using our timeline visualization tool. At a group level, clear differences were visible in the speed and accuracy overall among the three groups (Figure 5).

We looked at individual novices to investigate patterns that may have contributed to variations in their success. Figure 6 compares the first ten minutes of each practice session for novices SQ18, SQ20, and SQ14. The timeline visualization allows us to disregard outliers (e.g. excessively long IDs that likely indicate the user was off-task) and focus on overall patterns, such as the longer time spent per-specimen by SQ18 and SQ20 compared to SQ14. Larger gaps between specimens indicate more time spent on the answer screen (Figure 1, right). We can see, therefore, that SQ14 seems to have spent little time reflecting on answers whereas SQ18 and SQ20 paused longer between specimens. We also see slight improvements in overall accuracy with SQ18 and SQ20 over the course of the week, in contrast to SQ14 (See Figure 7).

RQ3: Improvements in Identification

More important than whether or not subjects improved is *why* they did or did not make gains. The visible differences in interactions with respect to speed and accuracy during practice can be explained in part by their reported supplemental resources selected for use during practice sessions.

During practice, SQ20 reported referring to a full-collection poster of the specimens in our digital collection, arranged taxonomically and available on our platform. Using this resource allowed the participant to perform a seek-and-find task to match like images. While this method seemed to be somewhat effective in practice sessions, because it does not draw attention to key diagnostic characters, we would not expect the skills to transfer to the new images on the post-test (or to real-life ID work). SQ18, however, reported using a visual ID resource based on diagnostic characteristics during practice. This subject improved most dramatically out of all users over the course of the study.

Discussion

As a preliminary study toward the design of this learning support tool, we approached the analysis with many questions, few of which could be answered statistically with a small sample size. Still even a small number of subjects generated a significant amount of data over the ten

days of the study. Our timeline visualization of log data directed our attention to promising leads of what works for learners and where we can improve the tool. For instance, we thought our answer screen calling out diagnostic characters would guide noticing, yet low performers like SQ14 spent very little time examining it. Making this information more prominent or adding a forcing function requiring them to engage with the information may produce gains. We anticipated more use of the hints feature in the daily quiz, but because no meaningful correlations were found between this feature and performance, re-examination of the functionality is necessary.

Limitations and Future Work

This paper presents initial analysis but leaves many open questions. Further study with more novice participants and with more nuanced pre- and post-tests to study gains in more advanced learners would be warranted. Some users reported occasional discrepancies in the practice quiz answers due to a database update that may have skewed some of the practice results. Our study design included repeated practice over time, but it is not yet clear whether simply extending the study over a longer period might have yielded more dramatic results. It is perhaps worth mentioning that though only five minutes of interaction was required for the study, many subjects opted to actively engage much longer than that, even upwards of 20 minutes per session, indicating that the activity was enjoyable to a variety of users.

Tables and Figures

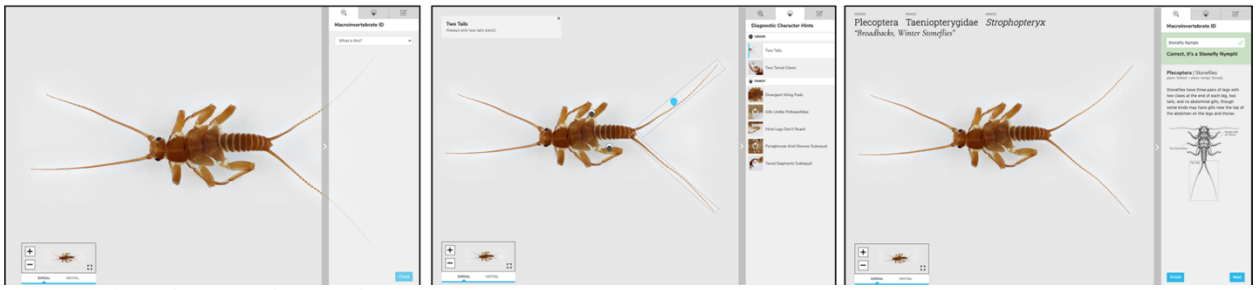

Figure 1. The self-quizzing feature of (URL BLINDED FOR REVIEW).

Figure 2. Screenshot from [URL blinded for review] depicting variation within the Mayfly order. Despite differences in shape, size, and color, these insects have several common physical traits, including three cerci (tails or caudal filaments), a single tarsal claw, and gills on their abdomens. Understanding these diagnostic characters will prevent misidentification as common look-alikes, such as some stoneflies.

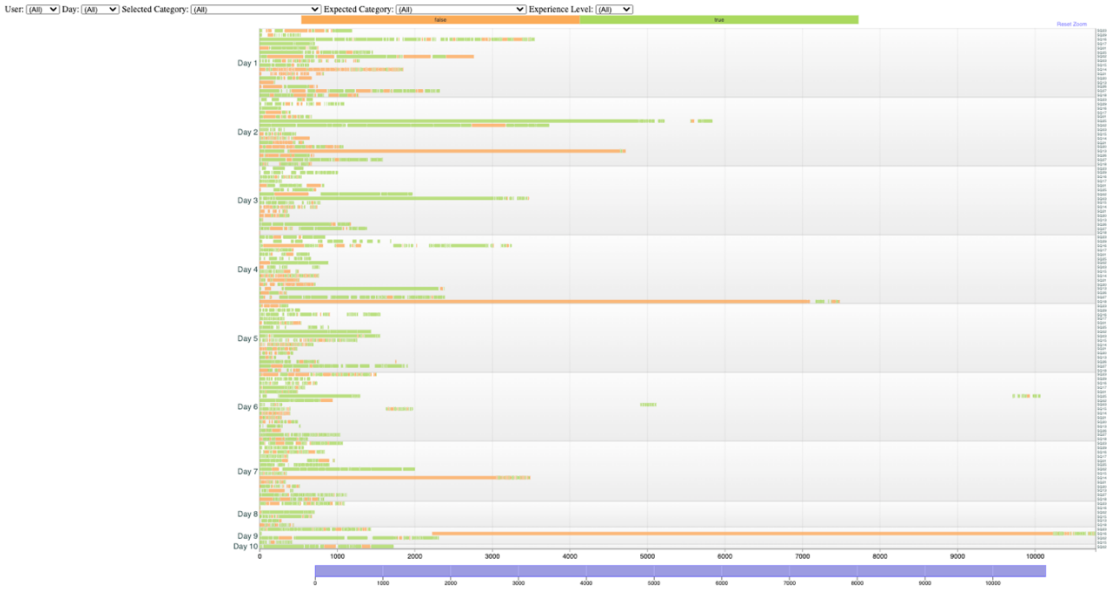

Figure 3. The timeline visualization showing the complete dataset. Bar length indicates amount of time (in seconds) on each specimen. Bars are color-coded to indicate whether the participant's identification was correct (green) or incorrect (orange). Gaps between bars represent time where the answer screen, including annotated line drawings and group descriptions, are visible. Hovering over a bar reveals a popup window listing time spent per identification, the selected category, and the correct category. Menus across the top allow filtering by individual user, practice day, answer, or experience group (High, Middle, or Novice). Adapted from Vasturiano. (n.d.). Vasturiano/timelines-chart. Retrieved August 14, 2020, from https://github.com/vasturiano/timelines-chart

Table 1. Example of coding scheme for an alderfly, Megaloptera Sialidae Sialis. Each answer was coded at the taxonomic Order, Family, and Genus level for whether the answer was correct, incorrect, implied, or no-guess, and whether the subject used the scientific or common name.

Participant	Answer	Order	Family	Genus
SQ21	Caddisfly	incorrect-common	no guess	no guess
SQ23	Mayfly	incorrect-common	no guess	no guess
SQ25	Sialis	correct-implied	correct-implied	correct-scientific
SQ26	Mayfly	incorrect-common	no guess	no guess
SQ27	Hellgrammmite	incorrect-implied	incorrect-implied	incorrect-scientific
SQ29	Sialidae - Alderfly	correct-commonSuborder	correct-scientific	no guess

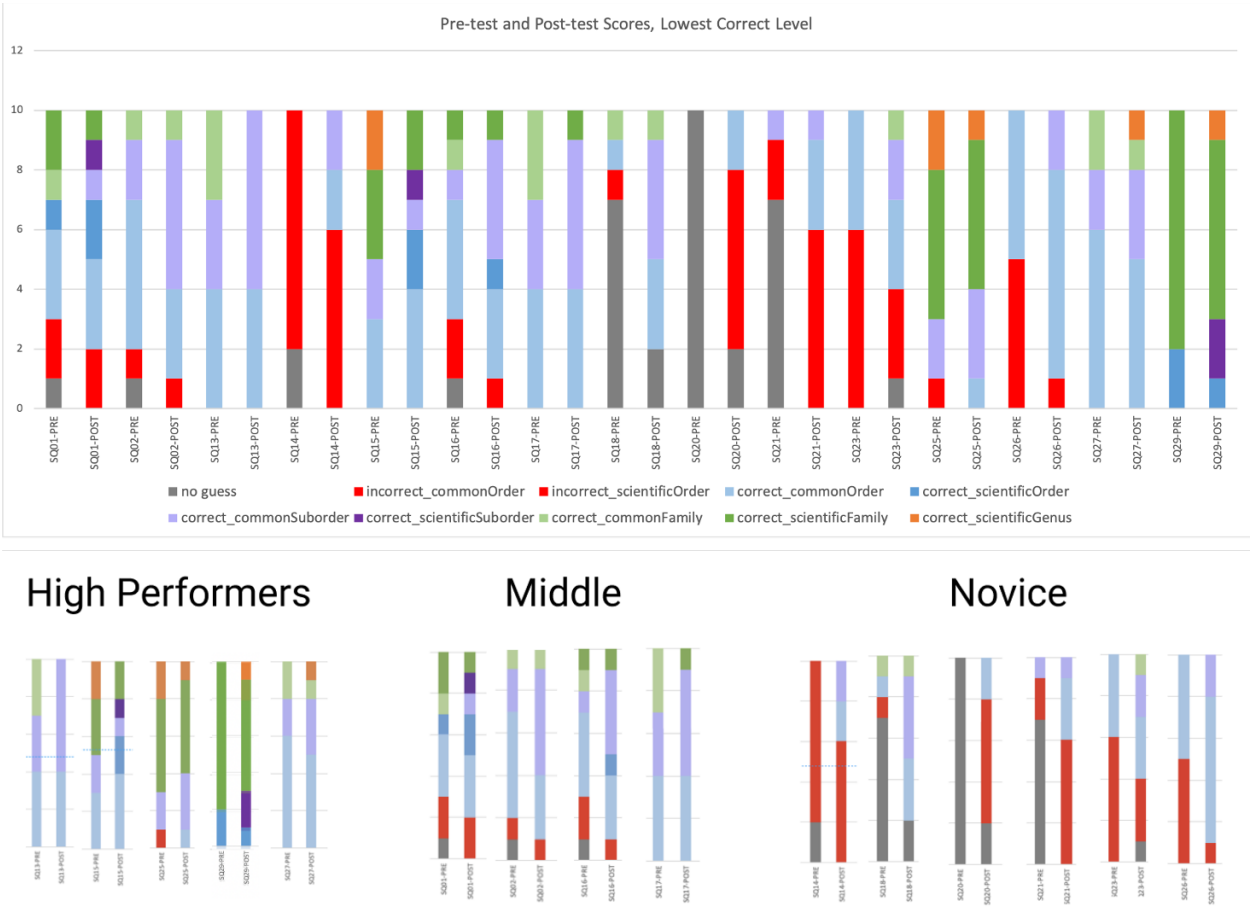

Figure 4. Top: All pre-post Lowest Correct Levels. Bottom: Subjects were grouped according to their test performances. High performers (left) all made at least one genus-level guess (shown in orange) in either the pre- or post-tests. They used scientific names frequently (dark blue and dark green), and had high accuracy even to lower taxonomic levels. Novices, by contrast, (right) had a significant number of no guesses (gray) and/or incorrect answers (red) on the pretest. Subjects in the Middle category (center) gave answers to almost all specimens on the pretest and every specimen on the posttest with high levels of accuracy (>= 80%). They used mostly common names.

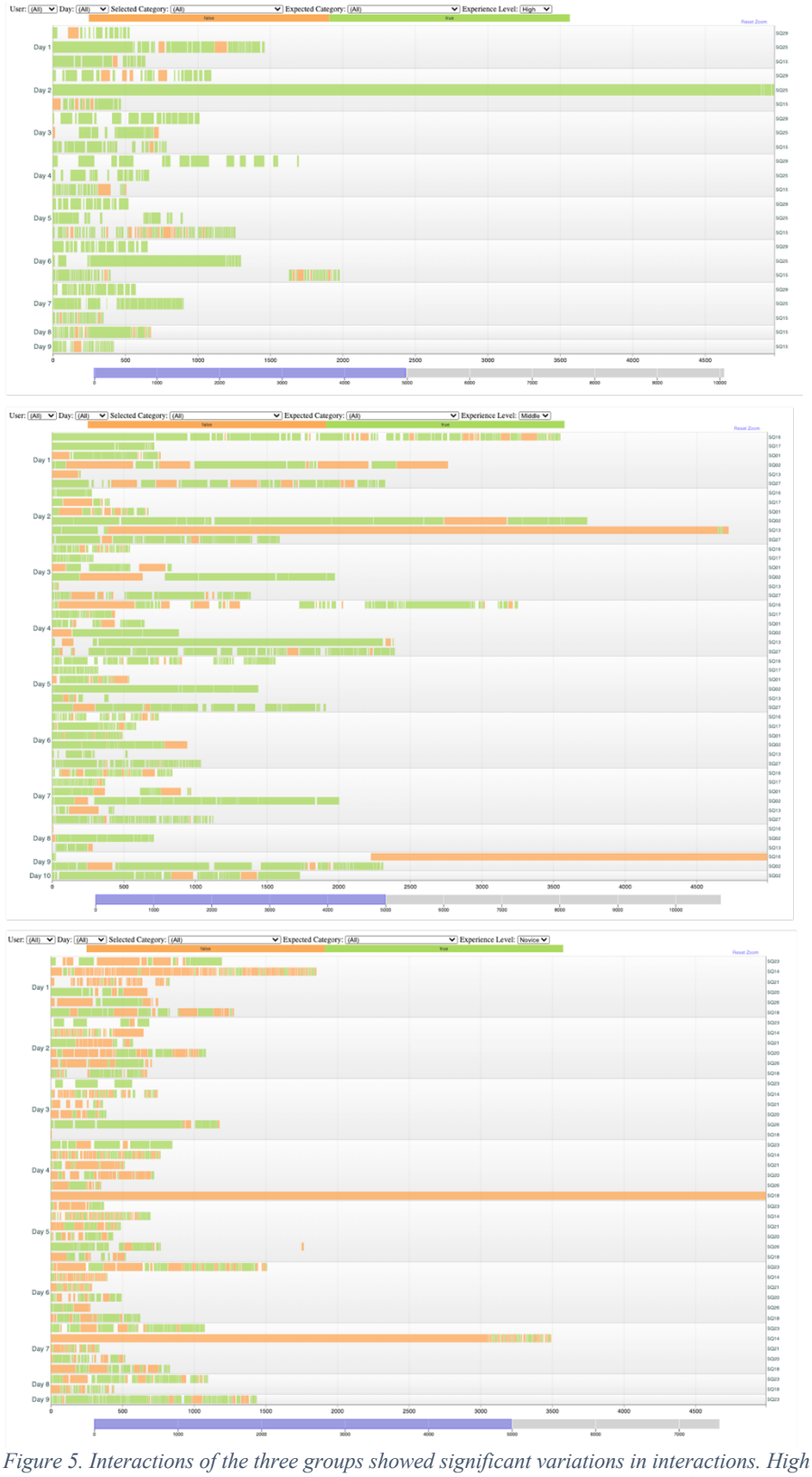

Figure 5. Interactions of the three groups showed significant variations in interactions. High Performers (top) answered most specimens quickly and accurately yet still engaged for a long time. Middle performers (center) had more variety, and Novices (bottom) had higher frequencies of incorrect answers.

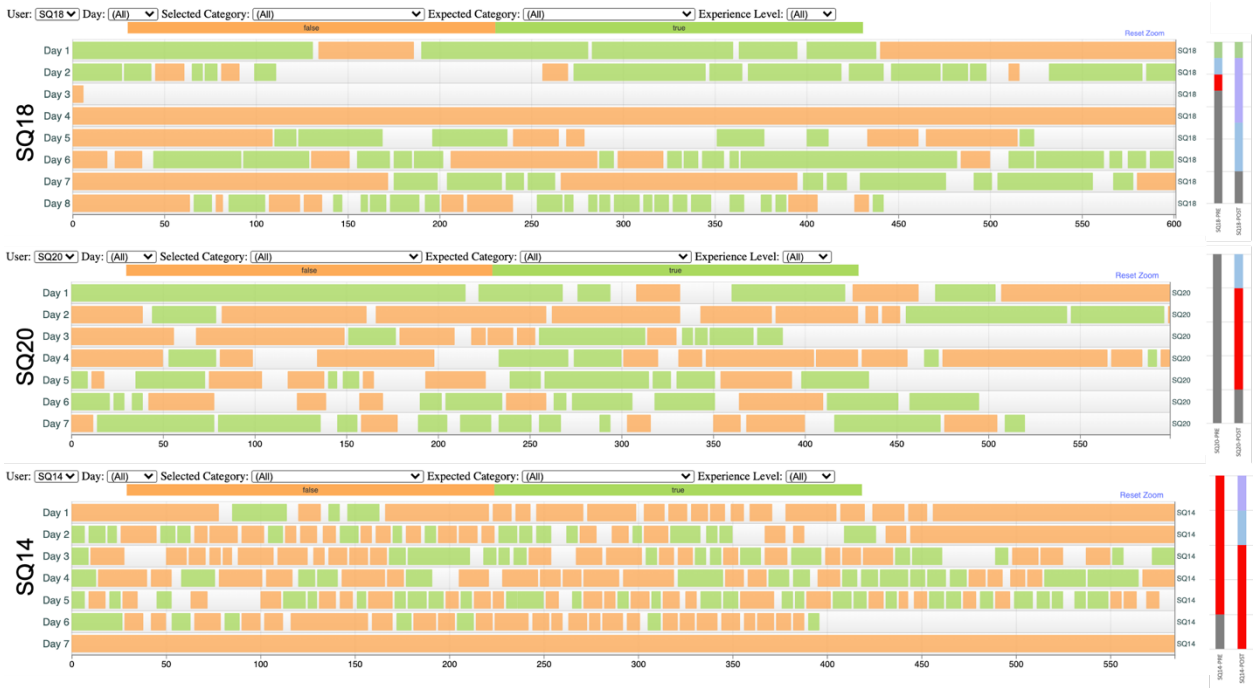

Figure 6. Two novices, SQ18 and SQ20, approached the quiz very differently from SQ14. The latter (bottom) quickly moved through specimens, getting a high percentage incorrect throughout the study with no noticeable improvement over the course of the week. SQ18 and SQ20 saw modest accuracy improvement overall but decreased the overall time spent per specimen without sacrificing accuracy, suggesting steps toward developing fluency in the task.

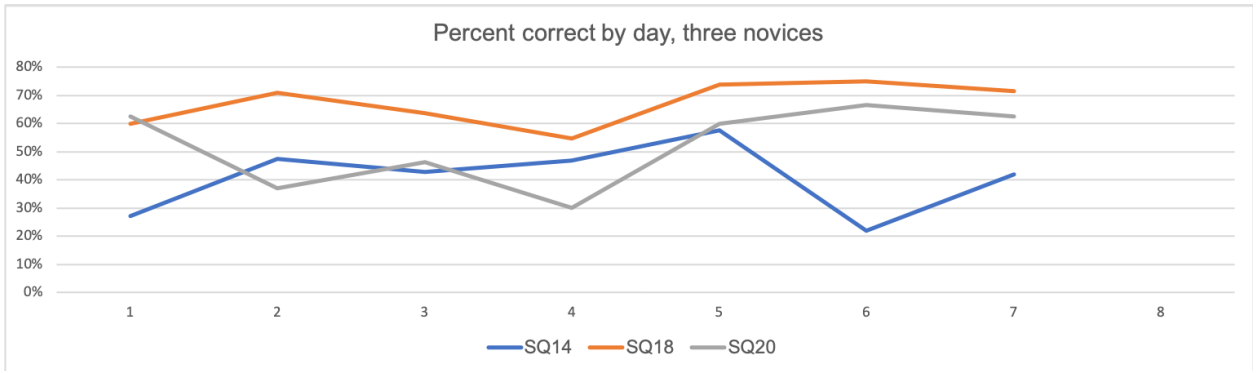

Figure 7. Of these three novices, SQ14 was least consistent in practice performance, with a significant drop in accuracy on Day 6.

References

Authors (2018).

Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2000, August). Visualization of navigation patterns on a web site using model-based clustering. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 280-284).

Duval, E. (2011). Attention Please! Learning Analytics for Visualization and Recommendation. Proceedings of the 1st International Conference on Learning Analytics and Knowledge - LAK '11. doi:10.1145/2090116.2090118

Furr, D. (2019). Visualization and clustering of learner pathways in an interactive online learning environment. In EDM.

Nerbonne, J. F., & Vondracek, B. (2003). Volunteer macroinvertebrate monitoring: assessing training needs through examining error and bias in untrained volunteers. *Journal of the North American Benthological Society*, *22*(1), 152-163.

Park, J., Denaro, K., Rodriguez, F., Smyth, P., & Warschauer, M. (2017, March). Detecting changes in student behavior from clickstream data. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (pp. 21-30).

Stepenuck, K.F., Genskow, K.D. Characterizing the Breadth and Depth of Volunteer Water Monitoring Programs in the United States. Environmental Management 61, 46–57 (2018).

Storey, R. G., Wright-Stow, A., Kin, E., Davies-Colley, R. J., & Stott, R. (2016). Volunteer stream monitoring: Do the data quality and monitoring experience support increased community involvement in freshwater decision making? Ecology and Society, 21(4).

Stribling, J. B., Pavlik, K. L., Holdsworth, S. M., & Leppo, E. W. (2008). Data quality, performance, and uncertainty in taxonomic identification for biological assessments. Journal of the North American Benthological Society, 27(4), 906-919.

Tanaka, J.W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? Cognitive Psychology, 23, 457–482.