CO-DESIGNING SPECIES IDENTIFICATION APPS WITH UNDERGRADUATES: AN AI-ASSISTED, PARTICIPATORY APPROACH BY A NON-PROGRAMMER INSTRUCTOR

SUBJECT/PROBLEM

Repeated, low-stakes retrieval practice with feedback is among the most reliable ways to build durable memory (Roediger & Karpicke, 2006), and few biology skills rely on it as heavily as species identification. Such practice is most powerful when it is formative: when each attempt returns immediate feedback that students can use to find and close their own gaps (Black & Wiliam, 1998). Separating a Semipalmated Sandpiper from a Least Sandpiper, like dozens of other regionally relevant taxa, means recognizing each from photographs that vary in pose, plumage, age, sex, and light. The tool suited to this is demanding to build: a custom trainer holding a deep library of real photographs for every species on the local list, quizzing recognition and giving feedback on each attempt. Until very recently, building one required programming skill or development funding that most biology instructors do not have, leaving educators to rely on generic commercial flashcard platforms that know nothing about their local species list or the field conditions their students will face.

Two developments change this picture. The first is AI-assisted coding, sometimes called "vibe coding," in which an instructor describes a tool in plain language and a large language model writes the working code, putting functional app development within reach of educators who have never written software. The second is the students-as-partners movement in higher education, which reframes learners as co-creators of their own education (Cook-Sather, Bovill, & Felten, 2014). Healey, Flint, and Harrington (2014) describe partnership as a process in which staff and students contribute to learning and teaching on more equal footing, and Bovill (2020) frames such co-creation as relational pedagogy. Together these developments open a new possibility: a non-programmer instructor who can build tools, guided by a partnership in which those tools are designed with students as co-creators. This possibility remains largely undocumented.

This study documents one such effort across a single semester. We ask three questions. First, can a non-programmer instructor, using AI-assisted coding, build and iteratively refine a suite of custom species-identification apps that are genuinely shaped by student feedback? Second, how do students engage with the resulting tools, in terms of how much and how often they use them and with what identification performance? Third, how do students receive both the apps and the experience of helping design them?

DESIGN OR PROCEDURE

Setting and participants

The study took place in one section of an upper-division undergraduate wildlife biology course at a public research university in the northeastern United States during spring 2026 (23 enrolled students). The course requires students to identify roughly 120 regionally relevant, wetland-associated species across four taxonomic units taught in sequence: waterfowl, mammals, herpetofauna, and shorebirds/wading birds. All procedures were reviewed and approved as exempt human-subjects research (protocol IRB2526-175).

The apps, and who built them

For each unit, the instructor built a browser-based identification quiz app. Each app presents hand-curated, research-grade photographs from the iNaturalist platform (iNaturalist, 2026), asks the student to name the species shown, and gives immediate feedback. Students sign in with a de-identified animal-name username, and every response is logged to a lightweight spreadsheet backend. The defining feature of the project is who did the building: the course instructor, whose

only prior computing experience was statistical analysis, wrote and deployed all four apps using AI-assisted, natural-language coding, with no external developer, no grant funding, and free hosting.

The workflow required no software training. The instructor built each app by describing the desired tool in plain language to a conversational AI assistant, testing each version, and reporting bugs and design ideas until it behaved as intended. The same AI assistant supplied step-by-step instructions for the parts a non-coder would otherwise find daunting, such as connecting each app to a Google Sheet so that usage data are captured automatically and publishing the finished app to free web hosting (Netlify). The process grew simpler over the semester. The first app took somewhat more initial setup before the routine settled into ordinary back-and-forth conversation, but at no point did it require prior coding knowledge, which is what makes the approach transferable to other instructors.

The participatory design cycle

After each unit, an anonymous paper survey gathered students' critiques, suggestions, and rank-ordered feature requests for the next app: what confused them, what was missing, and what they wished the tool could do. These responses were the primary design input for the following unit's build. In effect, each app was specified in large part by the students who had just finished using its predecessor, which makes the survey the engine of the participatory cycle and the mechanism that drove each redesign. The specific changes this produced are reported in the findings.

Data sources and analysis

Two data streams inform this proposal. The first is the set of end-of-unit student surveys (roughly 22–23 respondents per unit; response rates near 96%), which capture self-assessed ability before and after each unit, perceived helpfulness, field confidence, usage frequency, and open-ended feedback. The second is the apps' anonymized usage logs, which record every identification attempt with the species shown, the student's answer, whether it was correct, and a timestamp. Analyses here are descriptive: surveys are summarized as means and frequencies; usage logs as counts of logins, study bouts (attempts separated by gaps of more than 30 minutes), attempts, and accuracy, including each login's accuracy early versus late in its own practice.

What the data can and cannot show

These data support a clear picture of how students engaged with the apps and how their accuracy patterned over practice. Because logins are anonymized, use was voluntary, and there was no comparison group, we report the ability and confidence items as self-report and read within-practice accuracy change as a descriptive pattern.

ANALYSES AND FINDINGS

The instructor built all four apps

The non-programmer instructor successfully built and deployed all four apps within the semester. Each shared a common architecture while improving on the one before, and students noticed the progression: "All were very helpful," one wrote at the end of the term, and "they did get better over the semester." The reasons for that improvement are the subject of the next finding.

What students asked for, and what they got

The apps improved because students designed the improvements. The end-of-unit surveys surfaced a steady stream of specific, actionable requests, and the most common ones reappeared in the next build. Five request-to-feature chains illustrate the process. First, answer format: waterfowl and mammal users repeatedly asked for a way to type a species name themselves. With multiple choice, one explained, "you can just narrow it down instead of being confident enough to name it." A write-in mode was accordingly built into the final shorebird app, which a user there judged "super helpful." Second, self-diagnosis: a waterfowl student wished the app would mark "which ones I struggled on so I can study just those," and the mammal app accordingly gained a progress dashboard, a form of formative self-assessment, tracking per-species accuracy, which students used to steer their studying ("it helped guide my quizzes"). Third, learning support: requests for the "key features of the species" became an ID-tips feature in the herpetofauna app, used by one student "to quickly ID species," while a follow-on request to show those tips even after a correct answer points to the next refinement. Fourth, targeted drilling: requests to compare similar species side by side and to "get a quiz specifically on species I struggle with" became the study mode and, in the shorebird app, an adaptive mode. Fifth, photographs: students kept asking for more images and more variety, including birds in flight, in different plumages and lighting, and simply to "keep adding to the Picture Bank." The hand-curated photo library grew with every app, reaching roughly 200 images per species in the later units, so each species could be practiced across many poses, ages, sexes, and conditions. Not every request has been met; several students asked for an audio mode for vocalizations, which remains on the list. The design cycle, in other words, is still running.

Students used the apps heavily, and came back

Engagement was both broad and deep. Adoption was near-universal: between 20 and 23 of the 23 enrolled students used each app. Use was also sustained: across the four apps, students generated 45,105 logged identification attempts in 505 distinct study bouts (Table 1). Each study login returned for an average of roughly four to six separate bouts spread across three to four-and-a-half different calendar days, and the most active logins practiced on ten or more days. Self-reported usage matched the logs, with 86–95% of survey respondents reporting that they used the relevant app three or more times.

Accuracy tracked difficulty, and rose with practice

Overall accuracy in species ID was about 77%, declining predictably from 81% on the relatively distinctive waterfowl to 70% on the more cryptic shorebirds (Table 1). This ordering is what a knowledgeable observer would expect, and it suggests the apps reflect real differences in how hard these taxa are to tell apart. More notable is the pattern within each login's own practice. Comparing the first quarter of a login's attempts to its last quarter, accuracy rose in every app, pooled from roughly 67% early to 80% late, with 84 of the 100 logins that practiced enough to measure improving and only 12 declining. Because the three later apps track each student under a persistent login, we ran a stricter, fully within-student check, comparing a login's accuracy on the very same species early versus late in its own history, which removes any effect of a shifting species mix. The gain held: on these matched species, most logins still improved (16 of 19 in mammals, 18 of 21 in herpetofauna, and 23 of 24 in shorebirds), by roughly 6 to 15 points. We do not read this as proof that the apps caused learning, since practice was self-selected, identities are anonymous, and there was no control group. The pattern is nonetheless consistent across four independently built apps and converges with what students reported about themselves.

Table 1. Engagement and identification performance by app (units in teaching order).

Unit (in order)	Study bouts	ID attempts	Overall accuracy	Accuracy, early→late practice
Waterfowl	158	16,215	81%	74% → 84%
Mammals	93	7,624	80%	70% → 86%
Herpetofauna	133	10,907	75%	65% → 78%
Shorebirds/wading	121	10,359	70%	56% → 71%
Total / pooled	**505**	**45,105**	**~77%**	**~67% → ~80%**

Students shaped their own practice

Many students shaped their own practice. Each app let students choose which species or taxonomic groups to be quizzed on and how long a quiz should run; the herpetofauna and shorebird apps added a study mode for drilling difficult or flagged photographs; and the shorebird app could adaptively concentrate on a student's weaker species. These options saw real use. In three of the four apps, students customized what they practiced in a sizable share of sessions (48% for waterfowl, 36% for herpetofauna, and 39% for shorebirds), isolating subsets such as the swans, the diving ducks, or the wading birds, or invoking study mode to rehearse the species they kept missing; one student wanted to "practice just the shrews/mole/lemming," another to "go back + focus on the species I often got wrong."

Students rated the apps highly, and felt themselves improve

Reception was strong and remarkably stable across all four units: 91–100% of respondents rated each app "Very" or "Extremely" helpful (Table 2). Retrospective self-rated ability rose in every unit, with 65–86% of students reporting an increase and no student in any unit reporting a decline. Self-reported field confidence was moderate-to-high, with 65–73% "Very" or "Extremely" confident. The behavioral accuracy trend and the perceptual self-ratings thus point in the same direction.

Table 2. Survey reception by app.

Unit	Survey n	% rating app Very/Extremely helpful	Self-rated ability (before → after)	% Very/Extremely confident
Waterfowl	23	91%	3.1 → 4.9	65%
Mammals	22	100%	3.5 → 4.9	73%
Herpetofauna	22	95%	3.1 → 4.6	73%
Shorebirds/wading	23	100%	3.0 → 4.9	65%

The open-ended responses were both positive and specific. Students described the apps as their primary or only study method ("it was the only thing I used to study"), preferred them to commercial tools ("would use instead of Quizlet everytime"), and offered superlatives that extended well beyond the course: "best learning tool I've used in an NRS class," "By far the best studying tool I have been given… across my entire academic career," and "I can't imagine learning ID any other way. I wish all my classes had this." Several reported carrying the apps into other courses on their own initiative.

CONTRIBUTION

This work offers biology educators three transferable insights. First, it documents a **replicable, near-zero-cost model** in which an instructor with no programming background uses AI-assisted

coding to build curriculum-specific practice tools, dissolving the skill and budget barriers that long confined custom educational software to well-resourced teams. Second, it demonstrates a **lightweight mechanism for genuine student partnership**: a recurring survey-to-feature loop that turns student feedback into deployed design changes within weeks, with a traceable record connecting request to release. This mechanism is content-agnostic, and the same loop could operate in genetics, anatomy, or ecology wherever students benefit from custom practice. Third, it shows that a free, instructor-built tool can draw **sustained, voluntary practice**, the kind that actually builds memory. The difficulty with retrieval practice has never been whether it works; it is getting students to do enough of it. Here they did, on their own: nearly every enrolled student used every app, together logging about 45,000 identification attempts across 505 study sessions they were never required to complete, with many returning over several days and a few carrying the apps into other courses unprompted. With no comparison group we cannot credit the co-design itself, but a tool built for their exact course pulled the kind of willing, repeated practice that generic study apps rarely manage to.

GENERAL INTEREST

Biology educators have always wanted resources tuned to their own content and context. Building them used to demand programming skill and funding that most did not have, and that barrier has fallen. This proposal speaks to essentially the entire NABT membership, across grade levels and across formal and informal settings, because the method it describes (an ordinary instructor building real, tailored tools through conversation with an AI) is now available to all of them. Its participatory dimension is just as broadly relevant: the approach that produced these species-identification apps can position students as co-designers of their own learning in any discipline, advancing the long-standing call for student-centered, evidence-based biology education (Freeman et al., 2014). NABT members would leave with both a concrete method they could adopt next semester and a candid account of its limits, so that adoption and appropriate claims travel together.

REFERENCES

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Bovill, C. (2020). Co-creating learning and teaching: Towards relational pedagogy in higher education. Critical Publishing.

Cook-Sather, A., Bovill, C., & Felten, P. (2014). Engaging students as partners in learning and teaching: A guide for faculty. Jossey-Bass.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415.

Healey, M., Flint, A., & Harrington, K. (2014). Engagement through partnership: Students as partners in learning and teaching in higher education. Higher Education Academy.

iNaturalist. (2026). iNaturalist [Online biodiversity platform]. California Academy of Sciences & National Geographic Society. https://www.inaturalist.org

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255.